UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS MATEMÁTICAS Y
DE LA ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA

Stick-breaking processes and related random probability measures

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR (A) EN CIENCIAS

PRESENTA:
María Fernanda Gil Leyva Villa

DIRECTOR DE LA TESIS
Dr. Ramsés Humberto Mena Chávez
IIMAS, UNAM

MIEMBROS DEL COMITÉ TUTOR
Dr. Gerónimo Francisco Uribe Bravo
IMATE, UNAM

Dr. Arno Siri-Jégousse
IIMAS, UNAM

CIUDAD DE MÉXICO, 25/08/2021.

Stick-breaking processes and related random probability measures

Author: María F. Gil–Leyva
IIMAS, Universidad Nacional Autónoma de México
CDMX, México

Advisor: Ramsés H. Mena Chávez
IIMAS, Universidad Nacional Autónoma de México
CDMX, México

August 25, 2021

Abstract

We review important results in probability theory and mathematical statistics, that are essential for a deep understanding of Bayesian non-parametric statistics. In particular we prove the representation theorem for exchangeable sequences, partially exchangeable arrays, exchangeable partitions and symmetric random measures. Latter, we focus on the class of random probability measures with exchangeable increments, better known as species sampling processes. These constitute the building blocks for a wide variety of Bayesian non-parametric models. We study their weak topological support, some convergence results and properties of samples.

The main contribution of the thesis is the introduction of new classes of species sampling processes, characterized by stick-breaking weights with either exchangeable or Markovian length variables. Our models generalize well-known Bayesian non-parametric priors in an unexplored direction. We give conditions to assure the species sampling processes are proper and have full support, which in turn means they are feasible Bayesian non-parametric mixing priors. For a rich sub-class we explain how the stochastic ordering of the weights can be modulated, and Dirichlet and Geometric processes can be recovered. Important quantities related to clustering probabilities are derived, and an MCMC algortithm is proposed for density and clustering estimation.

Acknowledgements

In this randomly chaotic world we live in, being able to conclude this thesis is certainly a privilege. The very least I can do to honour everyone who has supported me, is not to take for granted the opportunity I have been gifted to educate myself and have a positive impact in the lives of those humans and animals around me. This thesis is dedicated to everyone who made this possible.

To my parents, who gave me life, and have dedicated their lives to my happiness and growth. They have taught me that the true language of love and passion is spoken through actions rather than words. Therefore, I am aware that there are no words that can describe how much they have given me and how grateful I am to them, despite this, I am trying to put it into words. To my mom and my best friend, who through her example has taught me to seek courage and kindness at the same time, and that, in those situations where it feels like both can not coexist, there's always a balance. She has tried to taught me to be persistent but not stubborn (which I have not learned so well because both of us are stubborn). She has taught me that success comes in so many different shapes that it turns ambiguous to pursue it, so instead, I should face obstacles and failures with a brave smile, and celebrate small and big achievements with a humble one. My mom has taught me how a real life superhero looks like, and it looks like her. To my dad, who was my first math teacher, and has been the most influential one. He began teaching me about mathematics ever since I was a baby, it never felt like I was studying, yet I was always learning how to solve problems that come in all different kinds of shapes, by simply playing. This has been so essential in my academic career because it got rid of all possible frustration, even before I formally began to study maths. I mean, I truly think that a harder math problem simply means more fun. Due to how diverse were my dad's games, his teachings have extended to all distinct activities in my life. My dad has really tried hard to make me an objective and critical thinker (this has partially worked) and to be responsible and cautious (this hasn't been so effective, but if it weren't for me dad I'd be far worst at this). He always tries to make me think harder and read more by saying: "Piensa Marifer, piensa" and "Te hace falta leer". Both my parents, each in her/his own way, have really done everything in their power for me to reach my full potential, and they have done this with so much love, tolerance, patience (extreme patience), intelligence and strength (to give me a few slaps in the wrist when needed). I am as certain of their unconditional love towards me as I am of my own existence (so let's round it to 100% sure). Mom and dad, thank you for absolutely everything I have.

To the rest of my human family: my two grandmas (titas), my aunts, uncles and cousins. Thank you so much for believing in me even more than I do, and for always cheering for me in everything I attempt to do (except the knives juggling). Thank you so much for the advices, the care, the hugs, the food, the laughter, and overall for your outstanding unconditional support and love. I couldn't have asked for a better family than you. In particular, about finishing this thesis, I just want to say, as my aunt Berthita would say: "Como diría my vecina: -A huevooooo (to be read with a certain chilango accent)-".

To my beautiful girlfriend, Abril, who is the most crazy (in the good way, almost always) and talented girl I know. Thank you for being so fearless and tenacious, and for your unstoppable attempt to outgrow yourself in every way, even if you already are so great. Thank you for teaching me that some risks are worth taking (note that I said

some and not all). Thank you for inspiring me to be better everyday, and for reminding me that it is essential to have fun while doing so. Thank you for being so patient when I have to work hard and for being so respectful of the things I love and the way I am. Thank you for your constant cheering and support at everything I decide to do, and for our complicity in all things we love to do together. Thank you so much for our adventure and the team we make, and overall, thank you for all your love. Oh! I almost forgot I also want to thank you for teaching me how to love cats as much as I love dogs, especially la Loca.

To the queens of my house, Gluck, Leeloo and Gamma. I feel so honoured to give you my food, my bed, my parents bed and the family couch, among others, all in exchange for the usual tail wag. Thank you for trying to stay up and take care of me while I work late at night, although most of the times you don't manage to stay awake, I love so much your fluffy company.

To my best friends, Pame, Alex and Chase, for always being there for me, and taking care of me (even from me). To Pame, with whom I grew up, and is kind of my big sister, thank you for the outstanding woman you are, for your fun craziness and for our multiple adventures together. To Alex, my number one math buddy, for all the philosophical and mathematical talks we continuously have. Thank you for reminding me how much I love mathematics, for all the problems we discuss together and rarely solve, and for the intellectual challenge I have to face every time we talk. To Chase, my favourite traceur, for practising with me our favourite sport, for being my coach, and for teaching me that fails are not only necessary to improve, but are also extremely fun (as long as no bone is broken and everyone is still alive). Overall, I love having the three of you in my life. You have been a cornerstone in my character building, same that helped me accomplished whatever I have so far.

To Elenita, Paula, Carlos and Isra, for your hard work that allows me to focus so much in my career, and for taking such good care of me and my family. In particular, I want to thank Elenita and Paula for spoiling me so much, and for laughing at my terrible jokes. Thank you for your delicious food and for reminding me that I have not had breakfast when I forget to eat it. Thank you for sharing your anecdotes and stories with me, and especially for being my friends.

To my advisor, Ramsés, for guiding me through the academia world, for your assertive words and assignments that correct my flaws and highlight my strengths. Thank you for your tolerance and patience while teaching me how to do research, for all the knowledge you have shared with me, and for encouraging my love for mathematics, in particular probability and statistics. I also want to thank Gerónimo and Arno for being my co-advisors, and for their support throughout my career. I want to thank Antonio, Pierpaolo, Alan and Theo for your careful reading of my work. I am very grateful with Silvia, Lucy, Tere and María Inés, for helping all of us, students, reach our goals. Thanks to the support of CONACyT Phd scholarship program and PAPIIT grant IG100221.

Finally, keeping up with the tradition, I can't help but to thank Darwin and Newton, both, either or neither.

Agradecimientos

En este mundo aleatoriamente caótico en el que vivimos, concluir esta tesis es ciertamente un privilegio. Lo menos que puedo hacer, para agradecer a todos lo que me han apoyado, es no tomar por sentado la oportunidad, que me ha sido regalada, de educarme y tener un impacto positivo en la vida de personas y animales que me rodean. Esta tesis está dedicada a todos los que han hecho esto posible.

A mis papás, quienes me dieron mi vida, y han dedicado la suya a mi felicidad y a mi crecimiento personal. Me han enseñado que el verdadero lenguaje del amor y la pasión se habla con acciones y no palabras. Así que estoy consciente de que no hay palabras que describan lo mucho que me han dado y lo agradecida que estoy con ellos, a pesar de esto, estoy intentando describirlo con palabras. A mi mamá y mejor amiga, quien a través de su ejemplo me ha enseñado a buscar un carácter fuerte y ser bondadosa al mismo tiempo, y que en aquellas situaciones en las que parece que ambas no pueden coexistir, siempre hay un balance. Me ha intentado enseñar a ser persistente pero no necia (que no he aprendido tan bien porque ambas somos necias). Me ha enseñado que el éxito toma tantas formas diferentes que se vuelve ambiguo perseguirlo, en lugar, simplemente debo enfretar los fracasos y obstáculos con una sonrisa valiente, y celebrar los pequeños y grades logros con una sonrisa humilde. Mi mamá me ha enseñado como se ve un superhéroe de la vida real, y se ve como ella. A mi papá quien fue mi primer maestro de matemáticas y el más influyente. Comenzó a enseñarme matemáticas desde que yo era una bebé, nunca sentí que estuviera estudiando, sin embargo siempre estaba aprendiendo a resolver problemas que toman distintas formas, simplemente jugando. Esto ha sido esencial en mi carrera académica porque eliminó las frustaciones desde antes que empezara a estudiar matemáticas formalmente. Es decir, verdaderamente pienso que un problema difícil en matemáticas se traduce a más diversión. Gracias a la diversidad de los juegos que me ponía mi papá, sus enseñanzas se han extendido a las diferentes actividades en mi vida. Mi papá ha procurado que yo sea una pensadora objetiva y crítica (esto ha funcionado parcialmente) y se ha esforzado mucho en enseñarme a ser responsable y precavida (esto no ha sido tan efectivo, pero sería mucho peor si no fuera por él). Mi papá siempre intenta que yo piense y lea más, diciéndome: “Piensa Marifer, piensa” y “Te hace falta leer”. Mis dos papás, cada quien a su manera, han hecho todo lo que esta a su alcance para que yo logre alcanzar todo mi potencial, y lo han hecho con mucho amor, tolerancia, paciencia (paciencia extrema), inteligencia y valentía (para darme sapes cuando es necesario). Estoy tan segura de su amor incondicional como lo estoy de mi propia existencia (así que vamos a redondearlo a 100% segura). Mamá y papá gracias por absolutamente todo lo que tengo.

Al resto de mi familia humana: mis dos abuelas (titas), mis tías, tíos y primos. Muchas gracias por creer en mí más de lo que yo lo hago, y por siempre echarme porras y apoyarme en todo lo que intento hacer (excepto malabarear cuchillos). Muchas gracias por sus consejos, su cuidado, sus abrazos, su comida, las risas, y en general por su apoyo y amor incondicional. No hubiera podido pedir una familia mejor que ustedes. En particular, acerca de terminar esta tesis, solo quiero decir, como diría mi tía Berthita: “Como diría my vecina: -A huevooooo (léase con un cierto acento chilango)-”.

A mi hermosa novia, Abril, quien es la niña más loca (en el buen sentido, casi siempre) y talentosa que conozco. Gracias por ser tan intrépida y tenaz, y por tu constante esfuerzo por superarte a ti misma en todos los aspectos, a pesar de que ya eres genial.

Gracias por enseñarme que algunos riegos valen la pena tomarse (nota que dije algunos y no todos). Gracias por inspirarme a ser mejor cada día y por recordarme que es esencial divertirme mientras lo hago. Gracias por ser tan paciente cuando tengo mucho trabajo y por ser tan respetuosa de las cosas que amo y de mi forma de ser. Gracias por apoyarme en todas las cosas que decido hacer, y por nuestra complicidad en todas las cosas que hacemos juntas. Muchas gracias por nuestra aventura y por el equipo que formamos, y sobre todo muchas gracias por todo tu amor. ¡Ah! casi lo olvido, también te quiero agradecer el enseñarme a amar a los gatitos tanto como amo a los perritos, especialmente a la Loca.

A las reinas de mi casa, Gluck, Leeloo y Gamma. Me siento muy honrada de darles mi comida, mi cama, la cama de mis papás, el sillón de la familia, entre otras cosas, todo a cambio de que muevan su rabito feliz. Gracias por intentar matenerse despiertas y cuidar de mí cuando me quedo trabajando en las noches, aunque casi nunca lo logran y se quedan dormidas, amo demasido su compañía peluda.

A mis mejores amigos, Pame, Alex y Chase, por siempre estar ahí para mí y protegerme (incluso de mí misma). A Pame, con quien crecí, y es como mi hermana mayor, gracias por la excepcional mujer que eres, por tu locura tan divertida y por nuestras múltiples aventuras juntas. A Alex, mi compañero de matemáticas número uno, gracias por las pláticas filosóficas y matemáticas que frecuentemente tenemos. Gracias por recordarme lo mucho que amo las matemáticas, por todos los problemas que discutimos y rara vez resolvemos, y por el reto intelectual que representa platicar contigo. A Chase, mi traceur favorito, gracias por practicar conmigo nuestro deporte favorito, por ser mi entrenador, y por enseñarme que los golpes y fallas no solo son necesarios para mejorar, sino que también son extremadamente divertidos (siempre y cuando no haya huesos rotos y todos sigan vivos). Amo que ustedes tres estén en mi vida, han sido parte esencial en la construcción de mi carácter, mismo que me ha ayudado a lograr lo que sea que haya logrado hasta ahora.

A Elenita, Paula, Carlos e Isra, por todo su trabajo que me permite enfocarme tanto en mi carrera, y por cuidarnos tan bien a mi familia y a mí. Particularmente, quiero agradecer a Elenita y a Paula por consentirme tanto, y por reírse de mis pésimos chistes. Muchas gracias por la deliciosa comida y por recordarme que no he desayunado cuando se me olvida hacerlo. Gracias por compartir conmigo sus anécdotas e historias y sobre todo gracias por nuestra amistad.

A mi tutor, Ramsés, por guiarme a través del mundo de la academia, por tus palabras y tareas asertivas que corrigen mis defectos y resaltan mis cualidades. Muchas gracias por tu tolerancia y paciencia mientras me enseñas a ser investigadora, por todos los conocimientos que has compartido conmigo, y por alentar mi amor por las matemáticas, en partícular probabilidad y estadística. También quiero agradecer a Gerónimo y a Arno por ser parte de mi cómite tutor y por todo su apoyo a lo largo de mi carrera. Quiero agradecer a Antonio, Pierpaolo, Alan y Theo por su cuidadosa lectura de mi trabajo. Estoy muy agradecida con Silvia, Lucy, Tere y María Inés por apoyarnos, a todos los estudiantes, a alcanzar nuestras metas. Gracias al apoyo del programa de becas doctorales de CONACyT y a PAPIIT por la beca IG100221.

Finalmente, como ya es costumbre, no puedo evitar agradecer a Darwin y a Newton, a ambos, a alguno o a ninguno.

Contents

Introduction

The concept of symmetry is intrinsic in the way we understand our universe and relate to it. The human brain naturally recognizes patterns we would characterize as symmetric almost everywhere around us, moreover we have a tendency to search for symmetry to understand and explain the world we live in. Indeed, in many areas of knowledge the concept of symmetry is fundamental. For example in Biology, the notion of symmetry is vastly used to describe body shapes: bilateral animals, highly influenced by the purpose of movement, are roughly symmetric with respect to the sagittal plane, which divides the body into left and right; many plants such as sea anemones often have radial symmetry, which suits them for nutrition and self defence reasons; other animals, like a starfish, have fivefold symmetry. In physics, the concept has become one of the most powerful tools, as many laws of nature originate in symmetry, Anderson (1972) even wrote - it is only overstating the case to say that physics is the study of symmetry.- Outside natural sciences, for instance in arts, we find exceptional uses of symmetric patterns to create marvellous pieces. To name a couple of notable examples, famous paintings of M.C. Escher consist in repeating a figure that perfectly bonds with itself[1], and in music we find the work of Bach, who widely exploited permutations and invariance in his compositions[2]. Of course, these are only a few examples that lead us to the conclusion that symmetry and the way we understand our universe are inseparable. In mathematics, the story is no different, most sub-areas strongly rely on symmetric objects to specialize or expand their theories. Broadly, we refer to an object as symmetric, if it remains somehow invariant under the action of a mathematical transformation. The object to be studied, the transformation and in which way it remains invariant, vary from subject to subject. In this thesis one of the central topics, and the main incentive behind our models, are a very special kind of symmetric random objects, referred to as exchangeable.

The motivations to study exchangeable elements are extensive and diverse. From a theoretical perspective the study of exchangeability leads to representation theorems of great significance, in particular, these explain deep connections between distinct types of random objects and different kinds of distributional symmetries. From an applied viewpoint, exchangeability and generalizations of it, lead to extremely flexible models that can be adjusted to a wide variety of practical problems. Indeed, exchangeable random objects, being intrinsically symmetric, are mathematically tractable enough so we can work with them, while leading to ductile classes of statistical models. In fact, in Bayesian statistics, a prevalent assumption we make on data points, is that they can be model by an exchangeable sequence. These symmetric infinite collections embody one of the most basic forms of exchangeability. Namely, a sequence of random variables, $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$, is called exchangeable if its distribution is invariant under the action of permutations, that is, $\mathbf{X}$ is equal in distribution to $(\mathbf{x}_{\sigma(i)})_{i\geq 1}$ for every bijection $\sigma : \mathbb{N} \to \mathbb{N}$. The first representation theorem for exchangeable sequences was proven by de Finetti (1931), latter, authors such as Hewitt and Savage (1955) and Ryll-Nardzewski (1957) generalized the result to cover richer spaces. This celebrated theorem named in honour to Bruno de Finetti, states that a sequence, $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$, taking values in a space that is measurably isomorph to $\mathbb{R}$, is exchangeable if and only if there exist an almost surely unique

[1] https://mcescher.com/gallery/symmetry/

[2] https://www.youtube.com/watch?v=xUHQ2ybTejU

random probability measure, $\boldsymbol{\mu}$, such that given $\boldsymbol{\mu}$, the elements of $\mathbf{X}$ are independent and identically distributed according to $\boldsymbol{\mu}$. It is mainly due to de Finetti's theorem that Bayesian statistics take exchangeability as such a pliable assumption over certain type of data. To spell this out, assuming that data points come from independent and identically distributed random variables, is clearly more restrictive than to assume that the order in which data points were sampled is irrelevant. Seemingly, working under the latter assumption is much harder than to assume the former one, however, de Finetti's theorem explains that in terms of mathematical tractability, there is not much loss of assuming exchangeability over independence and identical distribution. Of course this is true provided that we are able to understand and work with random probability measures, which is not such a minor detail. To overpass this obstacle many Bayesian models assume that the random probability measure, $\boldsymbol{\mu}$, degenerates on a parametric family of distributions. With the aim of avoiding this restrictive parametric assumption, Bayesian non-parametric statistics undertakes the problem of studying random probability measures. Naturally, the analysis of these objects, in all their generality, can be extremely difficult, which brings us back to symmetry, but this time through exchangeable random probability measures, better known as species sampling processes. Just as it occurs with exchangeable sequences, for exchangeable random probability measures there exist a representation theorem that allows us to decompose them into components we are able to understand and work with. For this reason species sampling processes have become the building blocks for the vast majority of Bayesian non-parametric models and are the main topic of this thesis.

The thesis is organized as follows. In Section 1, mainly following the work of Kallenberg (2002, 2017), we will introduce basic notions of Borel spaces and random probability measures. As mentioned above, exchangeability, even in one of its most basic forms, can not be properly dealt with if we do not introduce the concept of random probability measures. Section 2 studies three broad classes of closely related exchangeable random objects: Exchangeable sequences (de Finetti; 1931; Hewitt and Savage; 1955; Ryll-Nardzewski; 1957; Aldous; 1985), exchangeable partitions of the set of natural numbers (Kingman; 1982; Aldous; 1985; Pitman; 2006), and exchangeable probability measures (Kallenberg; 2005, 2017). The main objective of this section, being to derive the corresponding representation theorems. Latter, Section 3 undertakes a deeper study of species sampling processes. In particular, we review the concept of full support, which an essential requirement for species sampling processes in Bayesian non-parametric modelling (Datta; 1991; Ghosal et al.; 1999; Wu and Ghosal; 2008; Bissiri and Ongaro; 2014). In this section we also specialize de Finetti's theorem to exchangeable sequences driven by species sampling process. Towards the end of Section 3 we will go through some of the most famous constructions of the random probability measures in question, such as using urn schemes (Blackwell and MacQueen; 1973; Pitman; 1996b), by means of normalization of completely random measures (Regazzini et al.; 2003; Prünster; 2003; James et al.; 2009; Hjort et al.; 2010), through the stick-breaking decomposition (Sethuraman; 1994; Ishwaran and James; 2001; Pitman; 2006), and the most recently introduced method, exploiting random subsets of $\mathbb{N}$, (Walker; 2007; Fuentes-García et al.; 2010; De Blasi et al.; 2020; Gil-Leyva; 2021). The last part of Section 3, analyses as an example, the canonical and most popular model, the Dirichlet process (Ferguson; 1973; Blackwell and MacQueen; 1973; Sethuraman; 1994) as well as the Geometric process introduced by Fuentes-García et al. (2010). Up to this point, no major novel contribution has been explored, though some existent connections have been clarified. In Section 4 we arrive to the main contribution of this work, which consist in the introduction of new classes of Bayesian non-parametric priors, exploiting the stick-breaking construction of species sampling processes and connections previously established. To be precise, most efforts have concentrated in exploring stick-breaking processes based on independent variables. There are only a handful of examples (Fuentes-García et al.; 2010; Favaro et al.; 2012, 2016) of stick-breaking processes based on explicitly dependent variables. The dependent case has remained somehow elusive due to some mathematical hurdles to overcome. Our proposal here, represents to some extent, the first general treatment of the dependent case. Explicitly, we study stick-breaking processes based on exchangeable and Markovian variables, and derive sufficient and necessary conditions for these processes to lead to appropriate Bayesian non-parametric models (Gil-Leyva et al.; 2020; Gil-Leyva and Mena; 2021). We also explain that Dirichlet and Geometric processes belong to the novel classes, and are, in some way, the extreme points of stick-breaking processes based on stationary random variables, giving these famous Bayesian non-parametric priors a new interpretation. Section 5 illustrates to how to put all this theory into practice with the aid of Markov chain Monte Carlo methods. Of special relevance, we propose a Gibbs sampling algorithm for the novel classes of Bayesian non-parametric models, introduced in Section 4, by modifying the slice samplers proposed by Walker (2007) and Kalli et al. (2011). The main part of the thesis ends with a small section where we present final comments and possible future work related to the thesis. We shall also mention that all the proofs of the results are deferred to the Appendix. Explicitly, Appendix A includes the proofs of the results in Section 1, Appendix B contains the proofs of the results corresponding to Section 2, and so on.

1 Preliminaries

In this preliminary section we will introduce essential notions of random measures over Borel spaces. The main motivation of working in Borel spaces is that these provide a wide enough framework that covers most relevant cases, yet not so general to entice in technical difficulties. For example $\mathbb{R}$, $\mathbb{R}^n$ and $\mathbb{R}^\infty$ are well-known examples of Borel spaces. More generally, any Polish space, that is a complete and separable metric space, together with its Borel σ-algebra, is a Borel space, whenever working with a distance is required we will further assume the spaces we are working on are Polish. A Borel space $(S, \mathscr{B}_S)$ is in particular a measurable space, so we might consider the set of all σ-finite measures over $(S, \mathscr{B}_S)$, the set of all finite measures over the same space, and the one containing all probability measures, we denote these spaces by $\mathcal{M}(S)$, $\mathcal{F}(S)$ and $\mathcal{P}(S)$, respectively. Roughly speaking, a random measure is simply a random object that takes values in $\mathcal{M}(S)$, in order to define it properly as a measurable mapping, first of all we must endow measure spaces with a suitable σ-algebra, we do this in Section 1.1. In Section 1.2 we introduce the notion of kernels, which is a concept slightly more general than random measures, in fact if a kernel is defined over a probability space we attain a random measure. Inhere we also study important kernel operators that will eventually allow us to measurably transform random measures. Section 1.3 is dedicated to structural properties of random measures, we explain how to characterize the law of these random objects and derive their atomic decomposition. Latter in Section 1.3 we introduce the simplest kinds of random measures, termed basic point processes, and some elementary transformations of them. As we will illustrate in Section 1.3.3, basic point process constitute the building blocks of more complex kinds of random measures such as completely random measures. Sections 1.1 - 1.3 are mainly based on the work of Kingman (1993) Kallenberg (2002), Daley and Vere-Jones (2008) and Kallenberg (2017).

Most of the focus of subsequent sections is in random probability measures, in times we will be required to study convergence properties of them and analyse the support of their distributions. To this aim, in Section 1.4 we will endow $\mathcal{P}(S)$ with a metric that generates the same σ-algebra introduced earlier in Section 1.1. This will then allow us to define some modes of convergence of random probability measures as well as the concept of weak support. The work of Parthasarathy (1967), Billingsley (1968) and Kallenberg (2017) are main bibliography in which Section 1.4 is based on.

1.1 Borel and measure spaces

Two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$ are said to be Borel-isomorphic if and only if there exist a bijection $f : S \to T$ such that f and its inverse function f^{-1} are measurable. A measurable space $(S, \mathcal{S})$ is a Borel space if it is Borel-isomorphic to some Borel subset of $\mathbb{R}$, and in this case we call $\mathcal{S}$ the Borel σ-algebra of S, and denote it by $\mathscr{B}_S$. The following result shows that $\mathbb{R}^n$, $\mathbb{R}^\infty$, $\mathbb{R}_+$, are examples of Borel spaces.

Theorem 1.1. *Any Polish space, together with its Borel σ-algebra, that is the σ-algebra generated by the topology induced by the metric, is a Borel space.*

For a proof of Theorem 1.1, we refer the reader to Theorem 1.1 of Kallenberg (2017). Hereinafter every space is assumed to be Borel, and specifically Polish, if working with a metric is required. Additionally, every Borel space is understood to be localized, that is we can find a sub-collection $\hat{\mathcal{S}}$ of $\mathscr{B}_S$, such that

a) $\hat{\mathcal{S}}$ is a ring, i.e. it is closed under finite unions, finite intersections and proper differences.

b) For every $B \in \hat{\mathcal{S}}$ and $A \in \mathscr{B}_S$ we have that $A \cap B \in \hat{\mathcal{S}}$.

c) There exist $(S_n)_{n=1}^{\infty} \in \hat{\mathcal{S}}$ such that $S_n \nearrow S$ and $\hat{\mathcal{S}} = \bigcup_{n=1}^{\infty}(\mathscr{B}_S \cap S_n)$,

where $\mathscr{B}_S \cap B = \{B \cap A : A \in \mathscr{B}_S\}$. The collection $(S_n)_{n=1}^{\infty}$ in (c) is called a localizing sequence, and we refer to the elements of $\hat{\mathcal{S}}$ as bounded sets. The triplet $\left(S, \mathscr{B}_S, \hat{\mathcal{S}}\right)$ is called localized Borel space. Given any sequence $(S_n)_{n=1}^{\infty} \subseteq \mathscr{B}_S$, such that $S_n \nearrow S$, the class $\bigcup_{n=1}^{\infty}(\mathscr{B}_S \cap S_n)$ forms a localizing ring. It can be easily shown that localizing rings on Borel spaces generate the Borel σ-algebra, that is $\sigma(\hat{\mathcal{S}}) = \mathscr{B}_S$. To see this note that $\sigma(\hat{\mathcal{S}}) \subseteq \mathscr{B}_S$, conversely if we fix a localizing sequence, $(S_n)_{n=1}^{\infty}$, for every $A \in \mathscr{B}_S$, $A \cap S_n \nearrow A$ hence $\mathscr{B}_S \subseteq \sigma(\hat{\mathcal{S}})$. When working with Polish spaces the typical localizing ring is the set of all metrically bounded sets. Localizing rings are very important classes of sets since they allow us to prove certain important properties locally, that is for bounded sets, and latter extend them to the Borel σ-algebra by a monotone class argument.

For a localized Borel space, $\left(S, \mathscr{B}_S, \hat{\mathcal{S}}\right)$, we say a measure, μ over $(S, \mathscr{B}_S)$, is locally finite whenever $\mu(B) < \infty$ for every $B \in \hat{\mathcal{S}}$. We denote by $\mathcal{M}(S)$ to the set of all locally finite measures over $(S, \mathscr{B}_S)$, and endow it with the smallest σ-algebra, $\mathscr{B}_{\mathcal{M}(S)}$, that makes all the projection maps, $\left\{\pi_B : \mu \mapsto \mu(B) \,\middle|\, B \in \hat{\mathcal{S}}\right\}$, measurable. This way $\left(\mathcal{M}(S), \mathscr{B}_{\mathcal{M}(S)}\right)$ becomes a measurable space in its own right. Equivalently, $\mathscr{B}_{\mathcal{M}(S)}$ can be defined as the smallest σ-algebra generated by all the integration maps, $\left\{\pi_f : \mu \mapsto \mu(f) = \int f d\mu \,\middle|\, f : S \to \mathbb{R}_+, \text{ measurable}\right\}$. The following lemma shows that these couple of ways of defining $\mathscr{B}_{\mathcal{M}(S)}$ are in fact equivalent.

Lemma 1.2 (Borel σ-algebra of $\mathcal{M}(S)$)**.** *For any localized Borel space, $\left(S, \mathscr{B}_S, \hat{\mathcal{S}}\right)$, the following generate the same σ-algebra of $\mathcal{M}(S)$*

I. *The projection maps $\left\{\pi_B : \mu \mapsto \mu(B) \mid B \in \hat{\mathcal{S}}\right\}$.*

II. *The projection maps $\{\pi_B : \mu \mapsto \mu(B) \mid B \in \mathscr{B}_S\}$.*

III. *The integration maps $\{\pi_f : \mu \mapsto \mu(f) \mid f : S \to \mathbb{R}_+,$ measurable$\}$.*

The proof of Lemma 1.2 can be found in Appendix A.1. Notice that the space of all finite measures over $(S, \mathscr{B}_S)$,

$$\mathcal{F}(S) = \{\mu \in \mathcal{M}(S) : \mu(S) < \infty\} = \pi_S^{-1}[\mathbb{R}_+],$$

and the space of all probability measures over $(S, \mathscr{B}_S)$,

$$\mathcal{P}(S) = \{\mu \in \mathcal{M}(S) : \mu(S) = 1\} = \pi_S^{-1}[\{1\}].$$

are both measurable subsets of $\mathscr{B}_{\mathcal{M}(S)}$. The next result, whose proof can be found in Kallenberg (2017), justifies the notation $\mathscr{B}_{\mathcal{M}(S)}$.

Theorem 1.3. *For a localized Borel space, $\left(S, \mathscr{B}_S, \hat{\mathcal{S}}\right)$, the measure spaces $(\mathcal{M}(S), \mathscr{B}_{\mathcal{M}(S)})$, $(\mathcal{F}(S), \mathscr{B}_{\mathcal{F}(S)})$ and $(\mathcal{P}(S), \mathscr{B}_{\mathcal{P}(S)})$ are Borel spaces, where $\mathscr{B}_{\mathcal{F}(S)} = \mathscr{B}_{\mathcal{M}(S)} \cap \mathcal{F}(S)$, and $\mathscr{B}_{\mathcal{P}(S)} = \mathscr{B}_{\mathcal{M}(S)} \cap \mathcal{P}(S)$.*

1.2 Kernels and operators

Definition 1.1 (Kernel)**.** *Let $(S, \mathcal{S})$ be a measurable space and $(T, \mathscr{B}_T)$ a Borel space. A (locally finite) kernel $\boldsymbol{\mu}$, from S into T, denoted by $\boldsymbol{\mu} : S \to T$, is a mapping $\boldsymbol{\mu} : S \times \mathscr{B}_T \to \overline{\mathbb{R}}_+$ such that*

i) For every $s \in S$ fixed, $\boldsymbol{\mu}_s = \boldsymbol{\mu}(s, \cdot)$ is a locally finite measure.

ii) For every $B \in \mathscr{B}_T$ fixed, $\boldsymbol{\mu}(B) = \boldsymbol{\mu}(\cdot, B)$ is a measurable function from S into $\mathscr{B}_{\overline{\mathbb{R}}_+}$

Naturally if $\boldsymbol{\mu}_s$ is finite for every $s \in S$, we say $\boldsymbol{\mu}$ is finite, and we call $\boldsymbol{\mu}$ a probability kernel when $\boldsymbol{\mu}_s(T) = 1$, for every $s \in S$.

Perhaps, the simplest way of regarding a kernel is simply as measurable function from S into $\mathcal{M}(T)$. Indeed, by definition $\boldsymbol{\mu}_s$ is a locally finite measure for every $s \in S$, and by Lemma 1.2 together with the fact that $\boldsymbol{\mu}(B)$ is a measurable function for every $B \in \mathscr{B}_T$, we get that $\boldsymbol{\mu}_{(\cdot)} : S \to \mathcal{M}(T)$ is measurable with respect to $\mathcal{S}$ and $\mathscr{B}_{\mathcal{M}(T)}$.

Now, if we denote by S_+ to the set of all measurable functions, $f : S \to \overline{\mathbb{R}}_+$, (and analogously for T), any kernel $\boldsymbol{\mu} : S \to T$ can be identified with an operator $\mathcal{A}^{\boldsymbol{\mu}} : T_+ \to S_+$, given by $\mathcal{A}^{\boldsymbol{\mu}}(f) = g$, where

$$g(s) = \boldsymbol{\mu}(s, f) = \boldsymbol{\mu}_s(f) = \int f d\boldsymbol{\mu}_s = \int f(t)\boldsymbol{\mu}_s(dt),$$

for every $f \in T_+$. Namely, using the standard machinery (first proving the result for a simple function f, and then approximating non-negative functions through simple functions) we obtain that $\mathcal{A}^{\boldsymbol{\mu}}(f) \in S_+$ for every $f \in T_+$. This small discussion is formalized by Theorem 1.4, further details of its proof of are given in Appendix A.2.

Theorem 1.4 (kernel definitions)**.** *The following are equivalent:*

I. *$\boldsymbol{\mu}$ is a kernel from S into T.*

II. *$\mathcal{A}^{\boldsymbol{\mu}} : T_+ \to S_+$.*

III. *$\boldsymbol{\mu}_{(\cdot)} : S \to \mathcal{M}(T)$ is measurable.*

Remark 1.1 (Notation: Kernels and operators)**.** *For simplicity the operator, $\mathcal{A}^{\boldsymbol{\mu}}$, corresponding to the kernel $\boldsymbol{\mu} : S \to T$, will be denoted by the same greek letter. That is, hereinafter $\boldsymbol{\mu}(f)$ stands for $\mathcal{A}^{\boldsymbol{\mu}}(f)$ for every $f \in T_+$*

For kernels $\boldsymbol{\mu}, \boldsymbol{\nu} : S \to T$ we define the sum of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ as the kernel $\boldsymbol{\mu} + \boldsymbol{\nu} : S \to T$ given by

$$\boldsymbol{\mu} + \boldsymbol{\nu}(s, B) = \boldsymbol{\mu}(s, B) + \boldsymbol{\nu}(s, B),$$

for every $s \in S$ and $B \in \mathscr{B}_T$. And, for fixed $A \in \mathscr{B}_T$, we define the restriction of $\boldsymbol{\mu}$ to A as the kernel $\mathbf{1}_A\boldsymbol{\mu} : S \to T$ satisfying

$$\mathbf{1}_A\boldsymbol{\mu}(s, B) = \boldsymbol{\mu}(s, A \cap B),$$

for every $s \in S$ and $B \in \mathscr{B}_T$. Note that for a measurable set B, if $B \subseteq A$, $\mathbf{1}_A\boldsymbol{\mu}(B) = \boldsymbol{\mu}(B)$ and if $B \cap A = \emptyset$, $\mathbf{1}_A\boldsymbol{\mu}(B) = \boldsymbol{\mu}(\emptyset) = 0$. Trivially, the restriction and the sum of kernels results in a kernel. Taking into account Remark 1.1, other transformations of kernels that lead to a kernel are defined below.

Definition 1.2 (Product and composition of kernels)**.** *Consider two Borel spaces* $(T, \mathscr{B}_T)$ *and* $(U, \mathscr{B}_U)$*, and a measurable space* $(S, \mathcal{S})$*. Let* $\boldsymbol{\mu} : S \to T$ *and* $\boldsymbol{\nu} : T \to U$ *be* σ*-finite kernels.*

i) We define the product of $\boldsymbol{\mu}$ *and* $\boldsymbol{\nu}$ *as the kernel,* $\boldsymbol{\mu\nu} : S \to U$*, given by* $\boldsymbol{\mu\nu}(f) = \boldsymbol{\mu}(\boldsymbol{\nu}(f))$ *for every* $f \in U_+$*. Explicitly,*

$$(\boldsymbol{\mu\nu})_s(f) = \int \int f(u) \boldsymbol{\nu}_t(du) \boldsymbol{\mu}_s(dt)$$

for every $s \in S$ *and* $f \in U_+$*.*

ii) We define the composition of $\boldsymbol{\mu}$ *and* $\boldsymbol{\nu}$ *as the kernel,* $\boldsymbol{\mu} \circ \boldsymbol{\nu} : S \to T \times U$*, given by* $\boldsymbol{\mu} \circ \boldsymbol{\nu}(f) = \boldsymbol{\mu}(\boldsymbol{\nu}(f))$*, for every* $f \in (T \times U)_+$*. Explicitly,*

$$(\boldsymbol{\mu} \circ \boldsymbol{\nu})_s(f) = \int \int f(t, u) \boldsymbol{\nu}_t(du) \boldsymbol{\mu}_s(dt)$$

for every $s \in S$ *and* $f \in (T \times U)_+$*.*

Definition 1.3 (Product kernel)**.** *Consider two Borel spaces* $(T, \mathscr{B}_T)$ *and* $(U, \mathscr{B}_U)$*, and a measurable space* $(S, \mathcal{S})$*. Let* $\boldsymbol{\mu} : S \to T$ *and* $\boldsymbol{\nu} : S \to U$ *be* σ*-finite kernels we define the product kernel* $\boldsymbol{\mu} \otimes \boldsymbol{\nu} : S \to T \times U$ *by*

$$(\boldsymbol{\mu} \otimes \boldsymbol{\nu})_s(f) = \int \int f(t, u) \boldsymbol{\mu}_s(dt) \boldsymbol{\nu}_s(du) = \int \int f(t, u) \boldsymbol{\nu}_s(du) \boldsymbol{\mu}_s(dt)$$

for every measurable $f : T \times U \to \mathbb{R}_+$*.*

In the context of the above definition, note that if $f(t, u) = \mathbf{1}_A(t)\mathbf{1}_B(u)$ for some $A \in \mathscr{B}_T$ and $B \in \mathscr{B}_U$ we get $\boldsymbol{\mu} \otimes \boldsymbol{\nu}(A \times B) = \boldsymbol{\mu}(A)\boldsymbol{\nu}(B)$. Clearly Definition 1.3 can be extended to define the product of finitely many kernels. For the case of probability kernels, this definition can be further extended to the product of countably many kernels as follows. Let $(S, \mathcal{S})$ be a measurable space and consider the Borel spaces $\{(T_n, \mathscr{B}_{T_n})\}_{n \geq 1}$. For each $n \geq 1$ let $\boldsymbol{\mu}^{(n)} : S \to T_n$ be a probability kernel. Set $\mathcal{N} = \{(n_1, \ldots, n_k) : k \in \mathbb{N}, n_i \neq n_j \in \mathbb{N}\}$, and for $(n_1, \ldots, n_k)$ let us denote

$$\boldsymbol{\mu}^{(n_1, \ldots, n_k)} = \bigotimes_{i=1}^{k} \boldsymbol{\mu}^{(n_i)} = \boldsymbol{\mu}^{(n_1)} \otimes \cdots \otimes \boldsymbol{\mu}^{(n_k)}.$$

For each $s \in S$ fixed, the family $\left\{\boldsymbol{\mu}_s^{(n_1, \ldots, n_k)}\right\}_{(n_1, \ldots, n_k) \in \mathcal{N}}$ satisfies the hypothesis of Kolmogorov's consistency theorem, hence there exist a unique probability measure $\boldsymbol{\mu}_s$ over $\left(\prod_{n \geq 1} T_n, \bigotimes_{n \geq 1} \mathscr{B}_{T_n}\right)$, such that for every $(n_1, \ldots, n_k) \in \mathcal{N}$, its projection to $\left(\prod_{i=1}^{k} T_{n_i}, \bigotimes_{i=1}^{k} \mathscr{B}_{T_{n_i}}\right)$ is precisely $\boldsymbol{\mu}_s^{(n_1, \ldots, n_k)}$. To see that $\boldsymbol{\mu} : S \to \prod_{n \geq 1} T_n$ is a kernel it suffices to see that for each $B \in \bigotimes_{n \geq 1} \mathscr{B}_{T_n}$ fixed, $\boldsymbol{\mu}(\cdot, B)$ is a measurable function. This can be easily be done by a monotone class argument with the π-system

$$\mathcal{C} = \left\{ \prod_{n \geq 1} B_n \in \bigotimes_{n \geq 1} \mathscr{B}_{T_n} : B_n \in \mathscr{B}_{T_n} \text{ and } B_n \neq T_n \text{ for finitely many indexes } n \right\}$$

and the λ-system $\mathcal{D} = \{B \in \bigotimes_{n \geq 1} \mathscr{B}_{T_n} : \boldsymbol{\mu}(\cdot, B) \text{ is measurable}\}$. In general, such probability kernel, $\boldsymbol{\mu}$, over $\left(\prod_{n \geq 1} T_n, \bigotimes_{n \geq 1} \mathscr{B}_{T_n}\right)$ will be denoted by $\bigotimes_{n \geq 1} \boldsymbol{\mu}^{(n)}$. Particularly, if $T_n = T$ and $\boldsymbol{\mu}^{(n)} = \boldsymbol{\nu}$ for every $n \geq 1$ we simply write $\boldsymbol{\nu}^{\infty}$ instead of $\bigotimes_{n \geq 1} \boldsymbol{\mu}^{(n)}$.

Remark 1.2. *Although we will not be using it, we shall mention that the definition of product kernel can be further extend to uncountably many probability kernels.*

1.3 Random measures

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and a Borel space $(S, \mathscr{B}_S)$, a (locally finite) random measure over $(S, \mathscr{B}_S)$ is defined as kernel $\boldsymbol{\mu} : \Omega \to S$, or equivalently, it is a random element taking values in $\big(\mathcal{M}(S), \mathscr{B}_{\mathcal{M}(S)}\big)$. As usual, the law of a random measure, $\boldsymbol{\mu}$, is defined as the push-forward probability measure Q over $\big(\mathcal{M}(S), \mathscr{B}_{\mathcal{M}(S)}\big)$ given by

$$\mathsf{Q}(B) = \mathbb{P}\left[\boldsymbol{\mu} \in B\right] = \mathbb{P}\left[\boldsymbol{\mu}^{-1}[B]\right]$$

for every $B \in \mathscr{B}_{\mathcal{M}(S)}$ and where $\boldsymbol{\mu}^{-1}[B] = \{\omega \in \Omega : \boldsymbol{\mu}_\omega = \boldsymbol{\mu}(\omega, \cdot) \in B\}$. Two random measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ over a Borel space $(S, \mathscr{B}_S)$, are said to be equal in distribution denoted by $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$, whenever their laws coincide. The following theorem allows us to characterize in different ways the law of a random measure.

Theorem 1.5. *Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be locally finite random measures over the same Borel space $(S, \mathscr{B}_S)$. Then the following are equivalent*

I. $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$

II. $(\boldsymbol{\mu}(B_1), \dots, \boldsymbol{\mu}(B_n)) \overset{d}{=} (\boldsymbol{\nu}(B_1), \dots, \boldsymbol{\nu}(B_n))$ *for every $B_1, \dots, B_n \in \mathscr{B}_S$.*

III. $(\boldsymbol{\mu}(A_1), \dots, \boldsymbol{\mu}(A_n)) \overset{d}{=} (\boldsymbol{\nu}(A_1), \dots, \boldsymbol{\nu}(A_n))$ *for every mutually disjoint measurable sets $A_1, \dots, A_n \in \mathscr{B}_S$.*

IV. $\boldsymbol{\mu}(f) \overset{d}{=} \boldsymbol{\nu}(f)$ *for every $f \in S_+$.*

V. $\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right] = \mathbb{E}\left[e^{-\boldsymbol{\nu}(f)}\right]$ *for every $f \in S_+$.*

See Appendix A.3 for a proof. For a random measure $\boldsymbol{\mu}$ over $(S, \mathscr{B}_S)$, the mapping $f \mapsto \mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right]$, is called Laplace transform of $\boldsymbol{\mu}$. This tool is specially useful to characterize the law of a random measure, as it leads to mathematically analytical expressions.

Now, we turn to characterize almost sure equalities of random measures. Two random measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are said to be equal almost surely, denoted by $\boldsymbol{\mu} \overset{a.s.}{=} \boldsymbol{\nu}$ if and only if there exist $A \in \mathcal{F}$ with $\mathbb{P}[A] = 1$, and such that for every $\omega \in A$, $\boldsymbol{\mu}_\omega = \boldsymbol{\nu}_\omega$. Clearly this implies $\boldsymbol{\mu}(B) \overset{a.s.}{=} \boldsymbol{\nu}(B)$ for every $B \in \mathscr{B}_S$. As the following proposition states, the converse is also true.

Proposition 1.6. *Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be random measures over $(S, \mathscr{B}_S)$ and defined on the same probability space. Then $\boldsymbol{\mu} \overset{a.s.}{=} \boldsymbol{\nu}$ if and only if for every $B \in \mathscr{B}_S$, $\boldsymbol{\mu}(B) \overset{a.s.}{=} \boldsymbol{\nu}(B)$.*

The proof of Proposition 1.6 appears in Appendix A.4. Before stating the following result, recall that for $s \in S$, Dirac's delta at s, denoted by δ_s, is the measure over $(S, \mathscr{B}_S)$ that assigns a mass of 1 to any measurable set that contains s, and assigns a mass of 0 to any measurable set which does not contain s, that is

$$\delta_s(B) = \mathbf{1}_B(s) = \begin{cases} 1 & \text{if } s \in B \\ 0 & \text{if } s \notin B, \end{cases}$$

for every $B \in \mathscr{B}_S$.

Theorem 1.7 (Atomic decomposition)**.** *Let $\boldsymbol{\mu}$ be a (locally finite) random measure over a Borel space $(S, \mathscr{B}_S)$. We can measurably decompose $\boldsymbol{\mu}$ as*

$$\boldsymbol{\mu} \overset{a.s.}{=} \sum_{j \leq \boldsymbol{\kappa}} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j} + \boldsymbol{\nu},$$

for some random elements: $\boldsymbol{\kappa}$ taking values in $\mathbb{Z}_+ \cup \{\infty\}$, $\{(\boldsymbol{\alpha}_j, \boldsymbol{\xi}_j)\}_{j \leq \boldsymbol{\kappa}}$ with state space $\mathbb{R}_+ \times S$, and a random measure, $\boldsymbol{\nu}$, satisfying $\boldsymbol{\nu}(\omega, \{s\}) = 0$ for every $s \in S$, and $\omega \in \Omega$. Moreover, the decomposition is almost surely unique up to the order of the elements in $\{(\boldsymbol{\alpha}_j, \boldsymbol{\xi}_j)\}_{j \leq \boldsymbol{\kappa}}$.

The proof of Theorem 1.7 can be found in Appendix A.5. In such context, to the elements in $(\boldsymbol{\xi}_j)_{j \leq \boldsymbol{\kappa}}$ we call atoms or locations of $\boldsymbol{\mu}$, to $\boldsymbol{\alpha}_j$ we call the size of the atom $\boldsymbol{\xi}_j$, and to $\boldsymbol{\nu}$ we call the diffuse part of $\boldsymbol{\mu}$. If $\boldsymbol{\mu} \overset{a.s.}{=} \boldsymbol{\nu}$ we say $\boldsymbol{\mu}$ is diffuse almost surely. In the opposite case where $\boldsymbol{\mu} \overset{a.s.}{=} \sum_{j \leq \boldsymbol{\kappa}} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$, we call $\boldsymbol{\mu}$ discrete almost surely, in particular if $\boldsymbol{\alpha}_j \in \mathbb{N}$ for all j, we say $\boldsymbol{\mu}$ is a point process. Furthermore if $\boldsymbol{\alpha}_j = 1$ for all j, so that $\boldsymbol{\mu} \overset{a.s.}{=} \sum_{j \leq \boldsymbol{\kappa}} \delta_{\boldsymbol{\xi}_j}$, and $\boldsymbol{\xi}_i \neq \boldsymbol{\xi}_j$ almost surely for all $i \neq j$, we say $\boldsymbol{\mu}$ is simple. For simple point processes, their law can be further characterized by the avoidance probability $\mathbb{P}[\boldsymbol{\mu}(B) = 0]$ for $B \in \hat{\mathcal{S}}$.

Lemma 1.8. *Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be two simple point processes over $(S, \mathscr{B}_S)$. Then $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$ if and only if $\mathbb{P}[\boldsymbol{\mu}(B) = 0] = \mathbb{P}[\boldsymbol{\nu}(B) = 0]$, for every $B \in \hat{\mathcal{S}}$.*

See Appendix A.6 for a proof.

1.3.1 Basic point processes

One of the simplest random measures one can possibly think of, is $\boldsymbol{\mu} = \delta_{\boldsymbol{\xi}}$, for some random element $\boldsymbol{\xi}$ taking values in the Borel space $(S, \mathscr{B}_S)$. Clearly $\boldsymbol{\mu}(B) \in \{0, 1\}$ for every $B \in \mathscr{B}_S$ and $\boldsymbol{\mu}(B) = 1$ if and only if $\boldsymbol{\xi} \in B$. If μ_0 is the distribution of $\boldsymbol{\xi}$ (denoted by $\boldsymbol{\xi} \sim \mu_0$) we get

$$\mathbb{P}[\boldsymbol{\mu}(B) = x] = \begin{cases} \mu_0(B) & \text{if } x = 1 \\ 1 - \mu_0(B) & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

That is $\boldsymbol{\mu}(B) \sim \mathsf{Ber}(\mu_0(B))$. A natural generalization of the above, is to consider a finite collection $(\boldsymbol{\xi}_j)_{j=1}^n$ of independent and identically distributed (i.i.d.) random elements with distribution μ_0 (here denoted as $(\boldsymbol{\xi}_j)_{j=1}^n \overset{\text{iid}}{\sim} \mu_0$) and define $\boldsymbol{\mu} = \sum_{j=1}^n \delta_{\boldsymbol{\xi}_j}$. This random measure assigns a mass of 1 to each element in $(\boldsymbol{\xi}_j)_{j=1}^n$, thus $\boldsymbol{\mu}(B)$ counts the number of random locations that fall into B, and it has a binomial distribution with parameters n and $\mu_0(B)$, namely $\boldsymbol{\mu}(B) \sim \mathsf{Bin}(n, \mu_0(B))$, for this reason $\boldsymbol{\mu}$ is called a binomial process based on (n, μ_0). If instead we consider a sequence $(\boldsymbol{\xi}_j)_{j \geq 1} \overset{\text{iid}}{\sim} \mu_0$, and a random variable, $\boldsymbol{\kappa}$, taking values in $\mathbb{N}$, independent of the locations, then the random measure $\boldsymbol{\mu} = \sum_{j=1}^{\boldsymbol{\kappa}} \delta_{\boldsymbol{\xi}_j}$ is called a mixed binomial process based on $(\boldsymbol{\kappa}, \mu_0)$. Of course, conditionally given $\boldsymbol{\kappa}$, $\boldsymbol{\mu}(B)$ follows a binomial distribution with parameters $\boldsymbol{\kappa}$ and μ_0. One interesting subfamily of this type of random measures, arises when $\boldsymbol{\kappa}$ follows a $\mathsf{Poi}(\theta)$ distribution, in this scenario $\boldsymbol{\mu}(B)$ has a $\mathsf{Poi}(\theta\mu_0(B))$ distribution and for disjoint measurable sets $\{B_i\}_{i \geq 1} \subseteq \mathscr{B}_S$, $(\boldsymbol{\mu}(B_i))_{i \geq 1}$ forms an independent collection of random variables. This type of mixed Binomial processes is a very important one and will be properly defined below.

Definition 1.4 (Poisson random measure, mixed poisson random measure, Cox process)**.** *Let $\mu \in \mathcal{M}(S)$ be a measure over $(S, \mathscr{B}_S)$. A random measure, $\boldsymbol{\mu}$, over $(S, \mathscr{B}_S)$ is called a Poisson process or Poisson random measure directed by μ if*

i) $(\boldsymbol{\mu}(B_i))_{i\geq 1}$ are independent random variables whenever the measurable sets $(B_i)_{i\geq 1}$ are disjoint.

ii) $\boldsymbol{\mu}(B) \sim \mathsf{Poi}(\mu(B))$ for every $B \in \mathscr{B}_S$.

Further if $\boldsymbol{\kappa}$ is random variable taking values in $\mathbb{N}$, and given $\boldsymbol{\kappa}$, $\boldsymbol{\mu}$ is a Poisson process directed by $\boldsymbol{\kappa}\mu$, we say that $\boldsymbol{\mu}$ is a mixed Poisson process directed by $(\boldsymbol{\kappa}, \mu)$. More generally if $\boldsymbol{\nu}$ is a random measure taking values in $\mathcal{M}(S)$ and conditionally given $\boldsymbol{\nu}$, $\boldsymbol{\mu}$ is a Poisson random measure directed by $\boldsymbol{\nu}$, we say that $\boldsymbol{\mu}$ is a Cox process directed by $\boldsymbol{\nu}$.

In the context of the above definition, the directing random measures μ and $\boldsymbol{\nu}$ are also called the intensities of the point process $\boldsymbol{\mu}$. If $\boldsymbol{\mu}$ is a Poisson process

$$\mathbb{E}[\boldsymbol{\mu}(B)] = \mu(B), \quad B \in \mathscr{B}_S,$$

and in the case of Cox processes we have a similar situation conditionally given $\boldsymbol{\nu}$, that is

$$\mathbb{E}[\boldsymbol{\mu}(B)|\boldsymbol{\nu}] = \boldsymbol{\nu}(B), \quad B \in \mathscr{B}_S.$$

Also, if $\Xi = (\boldsymbol{\xi}_j)_{j\geq 1}$ is a collection of random elements such that the counting measure $\boldsymbol{\mu} = \sum_j \delta_{\boldsymbol{\xi}_j}$ is Binomial, Poisson or Cox process, we say Ξ defines a Binomial, Poisson or Cox process, respectively.

We will refer to the point processes defined above as basic point process, the term comes from the fact that these constitute the building blocks for wider and more general classes of random measures, such as random measures with independent increments and random measures with exchangeable increments.

Proposition 1.9 (Laplace transforms of basic point process)**.** *Let μ_0 be a probability measure over $(S, \mathscr{B}_S)$, let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be random measures taking values in $\mathcal{M}(S)$, and consider a random element $\boldsymbol{\kappa}$ taking values in $\mathbb{N}$.*

i) *If $\boldsymbol{\mu}$ is a mixed Binomial process based in $(\boldsymbol{\kappa}, \mu_0)$. Then*

$$\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)} \,\middle|\, \boldsymbol{\kappa}\right] = \left(\mu_0\left(e^{-f}\right)\right)^{\boldsymbol{\kappa}}, \quad f \in S_+$$

ii) *If $\boldsymbol{\mu}$ is a Cox process directed by $\boldsymbol{\nu}$. Then*

$$\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)} \,\middle|\, \boldsymbol{\nu}\right] = \exp\left\{-\boldsymbol{\nu}\left(1 - e^{-f}\right)\right\}, \quad f \in S_+$$

Remark 1.3. *In the statement of the above proposition we are using the short notation of the integrals, and we will continue to do so. In the conventional notation, the above equations read as*

$$\mathbb{E}\left[\exp\left\{-\int f(s)\boldsymbol{\mu}(ds)\right\} \,\middle|\, \boldsymbol{\kappa}\right] = \left(\int e^{-f(s)}\mu_0(ds)\right)^{\boldsymbol{\kappa}},$$

and

$$\mathbb{E}\left[\exp\left\{-\int f(s)\boldsymbol{\mu}(ds)\right\} \,\middle|\, \boldsymbol{\nu}\right] = \exp\left\{-\int \left(1 - e^{-f(s)}\right)\boldsymbol{\nu}(ds)\right\},$$

for every $f \in S_+$.

The proof of Proposition 1.9 can be found in Appendix A.7. As aforementioned, if $\boldsymbol{\mu}$ is a mixed Binomial process based on $(\boldsymbol{\kappa}, \mu_0)$, where $\boldsymbol{\kappa}$ is Poisson distributed, then $\boldsymbol{\mu}$ is a Poisson process. This can be easily corroborated by taking expectations in (i) of Proposition 1.9, and comparing it to (ii) whenever $\boldsymbol{\nu}$ is non-random. With the aid of these Laplace transforms, it can be further proved that $\boldsymbol{\mu}$ is a mixed Poisson or binomial process if and only if the restriction, $\mathbf{1}_B\boldsymbol{\mu}$, is a mixed Binomial process for every bounded set $B \in \hat{\mathcal{S}}$. Latter, this result can be used to prove the existence of Cox processes directed by a locally finite random measure (for further details see for instance Kallenberg; 2017). Another corollary of Proposition 1.9 is that the law of a Cox process, $\boldsymbol{\mu}$, and the one of its intensity measure, $\boldsymbol{\nu}$, characterize each other. Indeed, by taking expectations in (ii), we get that for every $f \in S_+$

$$\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right] = \mathbb{E}\left[\exp\left\{-\boldsymbol{\nu}\left(1-e^{-f}\right)\right\}\right],$$

by Theorem 1.5, the assertion follows. To explain in more detail the relationship between a Cox process and its intensity measure, we have the following result, whose proof appears in Appendix A.8

Proposition 1.10. *Let $\boldsymbol{\mu}$ be a Cox process directed by $\boldsymbol{\nu}$ over the Borel space $(S, \mathscr{B}_S)$. Then*

a) *$\boldsymbol{\mu}$ is simple almost surely if and only if $\boldsymbol{\nu}$ is diffuse, almost surely.*

b) $\mathbf{1}_{\{\boldsymbol{\mu}(f)<\infty\}} = \mathbf{1}_{\{\boldsymbol{\nu}(f\wedge 1)<\infty\}}$ *almost surely, for every $f \in S_+$.*

As we will see eventually, the above Proposition is extremely useful when constructing Poisson (or Cox) random measures.

1.3.2 Transforms of basic point processes

Definition 1.5 ($\boldsymbol{\nu}$-transform of $\boldsymbol{\mu}$). *Let $(S, \mathscr{B}_S)$, $(T, \mathscr{B}_T)$ be Borel spaces, let $\boldsymbol{\mu} = \sum_{j\le\kappa} \delta_{\boldsymbol{\xi}_j}$ be a simple random measure over (S, B_S), also let $\boldsymbol{\nu} : S \to T$ be a probability kernel. Let $(\boldsymbol{\tau}_j)_{j\ge1}$ be random elements taking values in T and such that they are conditionally independent given $(\boldsymbol{\xi}_j)_{j\ge1}$, and $\boldsymbol{\tau}_j$ has distribution $\boldsymbol{\nu}_{\boldsymbol{\xi}_j}$. Define the random measure over $(T, \mathscr{B}_T)$, $\boldsymbol{\eta} = \sum_{j\le\kappa} \delta_{\boldsymbol{\tau}_j}$, whenever $\boldsymbol{\eta}$ is locally finite we say that $\boldsymbol{\eta}$ is a $\boldsymbol{\nu}$-transform of $\boldsymbol{\mu}$.*

Remark 1.4. *The above definition extends naturally for point processes. Simply note that any realization μ of point process, $\boldsymbol{\mu}$, can be expressed as $\mu = \sum_{j\le\kappa} \delta_{s_j}$, allowing $s_i = s_j$. Thus, in Definition 1.5, $\boldsymbol{\mu}$ needs not to be simple as long as it is a point process.*

Let us highlight a few important cases. Let $\boldsymbol{\eta} = \sum_{j\le\boldsymbol{\kappa}} \delta_{\boldsymbol{\tau}_j}$ be a random measure over $(T, \mathscr{B}_T)$ which is a $\boldsymbol{\nu}$-transform of the (possibly random) measure $\boldsymbol{\mu} = \sum_{j\le\boldsymbol{\kappa}} \delta_{\boldsymbol{\xi}_j}$ over $(S, \mathscr{B}_S)$. If $T = S \times U$ where U is Borel, $\boldsymbol{\rho} : S \to U$ is a probability kernel, and $\boldsymbol{\nu}_s = \delta_s \otimes \boldsymbol{\rho}_s$ then $\boldsymbol{\eta}$ is called a $\boldsymbol{\rho}$-randomization of $\boldsymbol{\mu}$. Here, $\boldsymbol{\tau}_j = (\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)$ for some conditionally independent random elements $(\boldsymbol{\alpha}_j)_{j\ge1}$ such that, given $\boldsymbol{\xi}_j$, $\boldsymbol{\alpha}_j$ has distribution $\boldsymbol{\rho}_{\boldsymbol{\xi}_j}$, that is

$$\boldsymbol{\eta} = \sum_{j\le\boldsymbol{\kappa}} \delta_{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)} = \sum_{j\le\boldsymbol{\kappa}} \delta_{\boldsymbol{\xi}_j} \otimes \delta_{\boldsymbol{\alpha}_j}.$$

If the projection $\boldsymbol{\eta}(\cdot \times U)$ is simple almost surely on S, or equivalently $\boldsymbol{\eta}(\{s\} \times U) \leq 1$ for every $s \in S$, then $\boldsymbol{\eta}$ is also called a U-marking of $\boldsymbol{\mu}$ or U-marked point process. Figure 1 illustrates two $\boldsymbol{\rho}$-randomizations of point processes, the one on the right is not a U-marked point process, since $\boldsymbol{\xi}_2 = \boldsymbol{\xi}_5$. It is easy to see that $\boldsymbol{\eta}$ is a U-marking of $\boldsymbol{\mu}$ if and only if $\boldsymbol{\mu}$ is simple. Despite this, if $U = \mathbb{R}_+$, to the point process $\boldsymbol{\eta} = \sum_{j \leq \boldsymbol{\kappa}} \delta_{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)}$ over $S \times \mathbb{R}_+$ we can assign the purely atomic measure over $(S, \mathscr{B}_S)$, $\tilde{\boldsymbol{\eta}} = \sum_{j \leq \kappa} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$. If $\boldsymbol{\eta}$ is a marked point process this identification is unique, and $\tilde{\boldsymbol{\eta}}$ is also called a marked point processes or marking of $\boldsymbol{\mu}$.

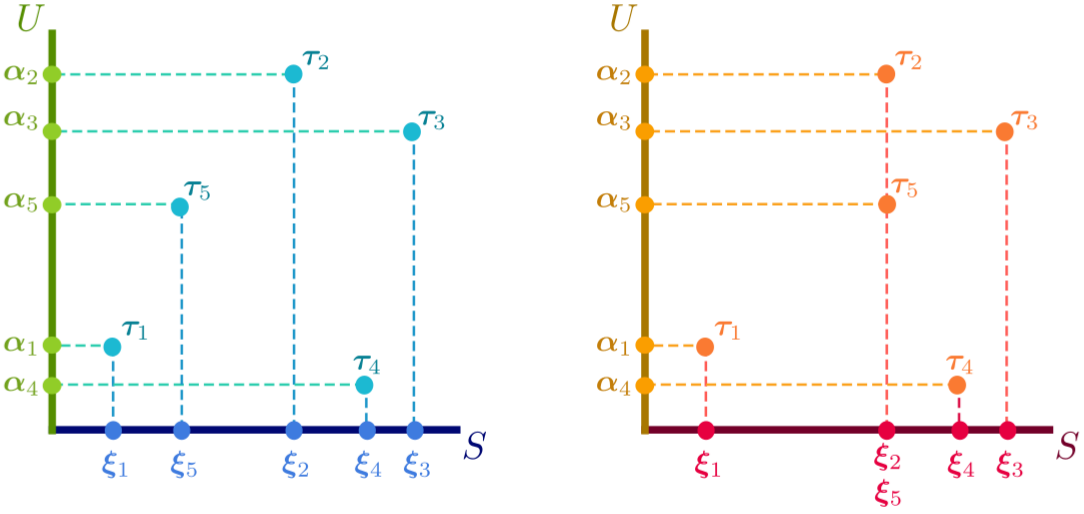


Figure 1: Realizations of two $\boldsymbol{\rho}$-randomizations of point processes.

Whether $\boldsymbol{\eta}$ is a marking or not. If $U = [0, 1]$ and $\boldsymbol{\rho}_s$ is the Lebesgue measure for each $s \in S$, then $\boldsymbol{\eta}$ is called a uniform randomization of $\boldsymbol{\mu}$. Finally, if $U = \{0, 1\}$ then $\boldsymbol{\rho}$ necessarily takes the form

$$\boldsymbol{\rho}_s = p(s)\mathbf{1}_{\{1\}} + (1 - p(s))\mathbf{1}_{\{0\}}$$

for some measurable function $p : S \to [0, 1]$, that is, $\boldsymbol{\rho}_s$ is a $\mathsf{Ber}(p(s))$ distribution. In this instance $\tilde{\boldsymbol{\eta}}$ is an integer valued random measure such that each atom of size 1, $\boldsymbol{\xi}_j$, of $\boldsymbol{\mu}$ is an atom of $\tilde{\boldsymbol{\eta}}$ with probability $p(\boldsymbol{\xi}_j)$ or is dismissed with probability $1 - p(\boldsymbol{\xi}_j)$, and each atom of size $\mathbf{n}$, $\boldsymbol{\xi}_j$, is an atom of size $0 \leq \mathbf{k} \leq \mathbf{n}$ with probability $\binom{\mathbf{n}}{\mathbf{k}} p(\boldsymbol{\xi}_j)^k (1 - p(\boldsymbol{\xi}_j))^{\mathbf{n}-\mathbf{k}}$. An alternative way to represent this process is:

$$\tilde{\boldsymbol{\eta}} = \sum_{j \geq 1} \mathbf{1}_{\{\mathbf{u}_j \leq p(\boldsymbol{\xi}_j)\}} \delta_{\boldsymbol{\xi}_j}$$

where $(\mathbf{u}_j)_{j \geq 1}$ are independent $\mathsf{Unif}(0, 1)$ random variables, in this instance $\tilde{\boldsymbol{\eta}}$ is called a p-thinning. An equivalent way to describe a p-thinning is stated in the following definition.

Definition 1.6 (p-thinning)**.** *Consider a (possibly random) point process $\boldsymbol{\mu}$. Let $\boldsymbol{\eta}$ be a uniform randomization of $\boldsymbol{\mu}$, and let $p : S \to [0, 1]$ be a measurable function. We define the p-thinning of $\boldsymbol{\mu}$ as the random measure over $(S, \mathscr{B}_S)$, $\tilde{\boldsymbol{\eta}}$, that satisfies*

$$\tilde{\boldsymbol{\eta}}(f) = \boldsymbol{\eta}(f_p)$$

for every measurable function $f : S \to \mathbb{R}_+$ and where $f_p(s,u) = f(s)\mathbf{1}_{\{[0,p(s)]\}}(u)$. If $p(s) = c \in [0,1]$ for every $s \in S$ we simply say $\tilde{\boldsymbol{\eta}}$ is a c-thinning.

Lemma 1.11 (Laplace functionals)**.** *Set $(S, \mathscr{B}_S)$ and $(T, \mathscr{B}_T)$ two Borel spaces. Let $\boldsymbol{\mu}$ be a locally finite point process over $(S, \mathscr{B}_S)$, let $\boldsymbol{\nu} : S \to T$ be a probability kernel and consider a measurable function $p : S \to [0,1]$.*

i) *If $\boldsymbol{\eta}$ is a $\boldsymbol{\nu}$-transform of $\boldsymbol{\mu}$ then for any $f : T \to \mathbb{R}_+$,*

$$\mathbb{E}\left[e^{-\boldsymbol{\eta}(f)} \mid \boldsymbol{\mu}\right] = \exp\left\{\boldsymbol{\mu}\left(\log\left\{\boldsymbol{\nu}\left(e^{-f}\right)\right\}\right)\right\}.$$

ii) *If $\tilde{\boldsymbol{\eta}}$ is a p-thinning of $\boldsymbol{\mu}$, then*

$$\mathbb{E}\left[e^{-\tilde{\boldsymbol{\eta}}(f)} \mid \boldsymbol{\mu}\right] = \exp\left\{\boldsymbol{\mu}\left(\log\left\{1 - p\left(1 - e^{-f}\right)\right\}\right)\right\}.$$

See Appendix A.9 for a proof.

1.3.3 Completely random measures

Using basic point process and their transform we can now characterize the class of random measures with independent increments.

Definition 1.7 (Completely random measure)**.** *A random measure $\boldsymbol{\mu}$ over $(S, \mathscr{B}_S)$ is called a completely random measure whenever its has pairwise independent increments, that is for every disjoint $B_1, \ldots, B_n \in \mathscr{B}_S$, the random variables $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$ are independent.*

The following theorem, due to Kingman and Itô, specializes the atomic decomposition (See Theorem 1.7) for locally finite completely random measures.

Theorem 1.12. *Let $\boldsymbol{\mu}$ be a locally finite random measure over the Borel space $(S, \mathscr{B}_S)$. Then, $\boldsymbol{\mu}$ is completely random if and only if it can be almost surely uniquely decomposed as*

$$\boldsymbol{\mu} = \beta + \sum_{j \geq 1} \boldsymbol{\gamma}_j \delta_{s_j} + \sum_{j \geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j},$$

where

a) *$(s_j)_{j \geq 1}$ are fixed elements of S and $(\boldsymbol{\gamma}_j)_{j \geq 1}$ are independent $\mathbb{R}_+$-valued random variables.*

b) *$\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j \geq 1}$ are i.i.d. $(S \times \mathbb{R}_+)$-valued random elements. Moreover, $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j \geq 1}$ defines a Poisson process over $S \times \mathbb{R}_+$, directed by some diffuse measure ν that satisfies*

$$\int_{\mathbb{R}_+} (x \wedge 1)\, \nu(B, dx) < \infty, \tag{1.1}$$

for every $B \in \hat{\mathcal{S}}$.

c) *β is a non-random, locally finite and diffuse measure over $(S, \mathscr{B}_S)$.*

The proof of Theorem 1.12 can be found in the Appendix A.10. To the countable subset $(s_j)_{j\geq 1} \subseteq S$, we call the set of fixed atoms of $\boldsymbol{\mu}$. Inhere, whenever we work with a completely random measure we will assume it has no fixed atoms and that the intensity ν can be decomposed as $\nu(ds, dx) = \mu(ds)\varrho(dx)$, so that equation (1.1) becomes

$$\mu(B) \int_{\mathbb{R}_+} (x \wedge 1)\, \varrho(dx) < \infty, \tag{1.2}$$

for every $B \in \hat{\mathcal{S}}$. In other words, μ is locally finite and ϱ satisfies $\int_{\mathbb{R}_+} (x \wedge 1)\, \varrho(dx) < \infty$. This assumption implies that the random atoms $(\boldsymbol{\xi}_j)_{j\geq 1}$ and the random jumps $(\boldsymbol{\alpha}_j)_{j\geq 1}$ are independent. Additionally, if we also consider the diffuse component, $\beta = c\,\mu$, for some constant $c \geq 0$, we get that for every mutually disjoint measurable sets $B_1, \ldots, B_n \in \mathscr{B}_S$, such that $\mu(B_i) = \mu(B_j)$, the random variables $\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n)$ are not only independent, but also they are i.i.d., in which case we say $\boldsymbol{\mu}$ is a μ-homogeneous completely random measure. Now, one of our main topics of interest are random probability measures, and we will be using completely random measures to build them through normalization. To this aim, there are further requirements to make. Indeed, in order for $\boldsymbol{\mu}/\boldsymbol{\mu}(S)$ to be well defined, we must have $0 < \boldsymbol{\mu}(S) < \infty$. Evidently, if $\boldsymbol{\mu}$ is a μ-homogeneous completely random, to attain $\boldsymbol{\mu}(S) < \infty$, is it enough to ask $\mu(S) < \infty$ and $\int_{\mathbb{R}_+} (x \wedge 1)\, \varrho(dx) < \infty$. On the other side, $\boldsymbol{\mu}(S) > 0$ can be assured by either assuming the diffuse component $\beta(S) = c\mu(S) > 0$, that is $c > 0$, or by requiring $\boldsymbol{\mu}$ to jump infinitely often, that is in every set B with $\mu(B) > 0$, there will be infinitely many very small jumps, formally, when this occurs we say that $\boldsymbol{\mu}$ has infinite activity, and this property can be corroborated through $\int_{\mathbb{R}_+} \varrho(dx) = \infty$, which, from (1.2) is easily seen to be equivalent to $\int_{[0,1]} \varrho(dx) = \infty$.

Example 1.1 (Subordinators)**.** *Set $S = \mathbb{R}_+$ and let λ denote the Lebesgue measure over $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$. Consider a locally finite, λ-homogeneous, completely random measure, $\boldsymbol{\mu}$, and define the stochastic process $\boldsymbol{\phi} = \{\boldsymbol{\phi}(t)\}_{t\geq 1}$ given by $\boldsymbol{\phi}(t) = \boldsymbol{\mu}([0,t])$. Clearly $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$ determine each other completely, and the following hold.*

i) *$\boldsymbol{\phi}(0) = 0$ almost surely.*

ii) *The paths of $\boldsymbol{\phi}$ are right-continuous and their left limits exist.*

iii) *$\boldsymbol{\phi}$ has independent increments. That is, for every $0 = t_0 < t_1 \cdots < t_n$, the random variables $(\boldsymbol{\phi}(t_i) - \boldsymbol{\phi}(t_{i-1}))_{i=1}^n$ are independents.*

iv) *$\boldsymbol{\phi}$ has independent stationary increments, i.e. for every $0 \leq r < t$, $\boldsymbol{\phi}(t-r) \overset{d}{=} \boldsymbol{\phi}(t) - \boldsymbol{\phi}(r)$.*

We call a subordinator to any stochastic process that satisfies (i) – (iv)*. As a consequence of Theorem 1.12 we get that any subordinator, $\boldsymbol{\phi}$, can be decomposed as*

$$\boldsymbol{\phi}(t) = ct + \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}([0,t]),$$

for every $t \geq 1$, where $c \geq 0$ and $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ defines a Poisson process over $\mathbb{R}_+ \times \mathbb{R}_+$, with intensity, ν, that decomposes as $\nu(ds, dx) = \lambda(ds)\varrho(dx) = ds\, \varrho(dx)$.

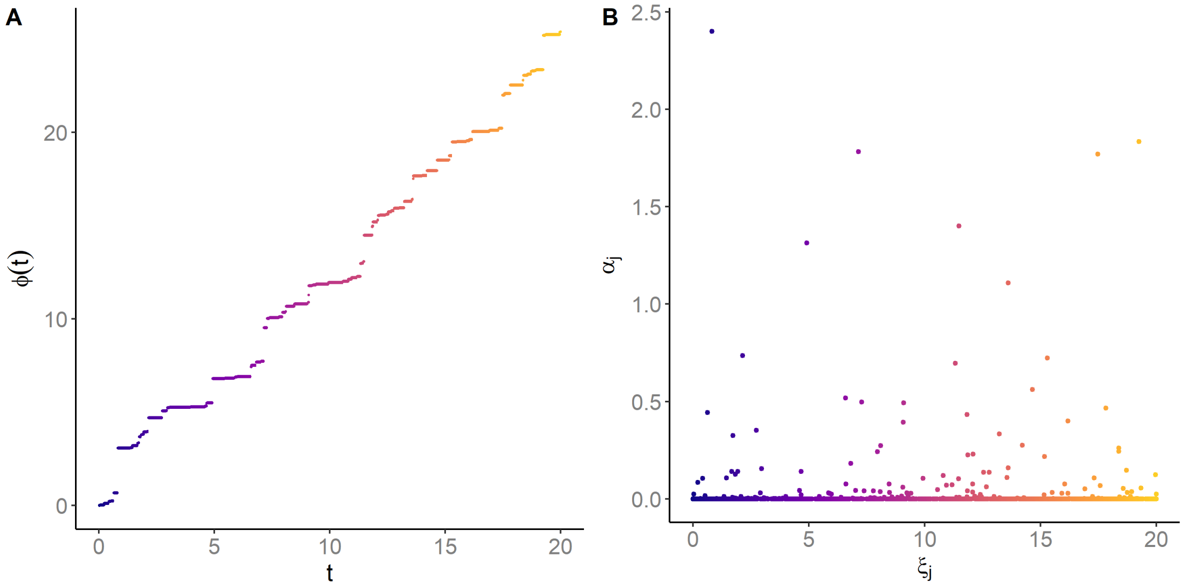


Figure 2: Simulation of a subordinator, $\boldsymbol{\phi}(t) = \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}([0,t])$ with infinite activity and with no drift, i.e. $c = 0$ (A). The graph in the right side (B) shows the corresponding point process $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ over $\mathbb{R}_+ \times \mathbb{R}_+$, that encodes the jumps of the subordinator and the locations of such jumps.

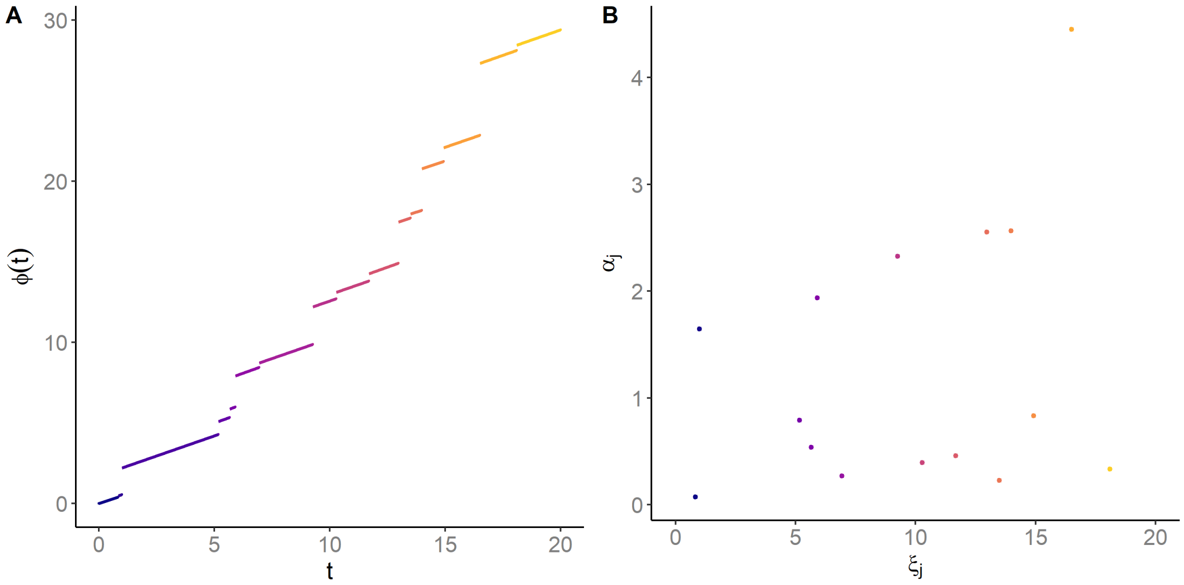


Figure 3: Simulation of a subordinator, $\boldsymbol{\phi}(t) = ct + \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}([0,t])$ with finite activity and with positive drift, i.e. $c > 0$ (A). The graph in the right side (B) shows the corresponding point process $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ over $\mathbb{R}_+ \times \mathbb{R}_+$, that encodes the jumps of the subordinator and the locations of such jumps.

Figures 2 and 3 show some simulations of subordinators such that $\int_{\mathbb{R}_+}(x \wedge 1)\varrho(dx) < \infty$. *The one in Figure 2 corresponds to a subordinator whose diffuse component is identically zero and with infinite activity, i.e.* $\int_{[0,1]} \varrho(dx) = \infty$, *so that infinitely many tiny jumps occur in a finite interval. On other side, the subordinator represented by*

Figure 3 has a non-zero diffuse component and finite activity, this is, only a finite number of jumps occur in any finite interval, this holds whenever $\int_{[0,1]} \varrho(dx) < \infty$. *Note that in either case we have that for every* $T > 0$, $0 < \boldsymbol{\phi}(T) < \infty$ *almost surely, hence through the normalization* $\mathbf{F} = \boldsymbol{\phi}/\boldsymbol{\phi}(T)$ *we can construct a random distribution function over* $\big([0, T], \mathscr{B}_{[0,T]}\big)$.

1.4 Random probability measures

Consistently with Section 1.1, we denote by $\mathcal{P}(S)$, to the space of all probability measures over the Borel space $(S, \mathscr{B}_S)$. A random probability measure over $(S, \mathscr{B}_S)$, $\boldsymbol{\mu}$, is simply a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into $\big(\mathcal{P}(S), \mathscr{B}_{\mathcal{P}(S)}\big)$, equivalently it is a probability kernel from Ω into S, $\boldsymbol{\mu} : \Omega \to S$. The main objective of this section is to endow $\mathcal{P}(S)$ with a metric that induces $\mathscr{B}_{\mathcal{P}(S)}$. To this aim we will consider that (S, d) is a metric space and that $\mathscr{B}_S$ is precisely the σ-algebra generated by the topology induced by the metric d. Providing a metric on $\mathcal{P}(S)$ will then allow us to define weak convergence of random probability measures as well as the concept of weak support.

1.4.1 Topology of weak convergence

For $\mu, \mu_1, \mu_2, \ldots \in \mathcal{P}(S)$ we say that $(\mu_n)_{n\geq 1}$ converges weakly to μ, denoted by $\mu_n \overset{w}{\to} \mu$, whenever

$$\mu_n(f) = \int f d\mu_n \to \int f d\mu = \mu(f),$$

for every continuous and bounded function $f : S \to \mathbb{R}$. The Portmanteau theorem gives alternative definitions of this mode of convergence, in terms of closed, open and μ-continuity sets, this last one refers to Borel sets, whose boundary has μ-measure equal to 0.

Theorem 1.13 (Portmanteau)**.** *For* $\mu, \mu_1, \mu_2, \ldots \in \mathcal{P}(S)$, *the following statements are equivalent:*

I. $\mu_n \overset{w}{\to} \mu$.

II. $\limsup_n \mu_n(A) \leq \mu(A)$, *for every closed set* A.

III. $\liminf_n \mu_n(A) \geq \mu(A)$, *for every open set* A.

IV. $\mu_n(A) \to \mu(A)$, *for every Borel set with* $\mu(\partial A) = 0$, *where* ∂A *denotes the boundary of* A.

We provide the proof of Theorem 1.13 in Appendix A.11. For $\mu, \nu \in \mathcal{P}(S)$, we define the Lévy-Prokhorov metric, by

$$d_{\mathcal{P}}(\mu, \nu) = \inf\{\varepsilon > 0 : \mu(A) \leq \nu\left(A^{\varepsilon}\right) + \varepsilon, \nu(A) \leq \mu\left(A^{\varepsilon}\right) + \varepsilon, \forall A \in \mathscr{B}_S\}, \tag{1.3}$$

where $A^{\varepsilon} = \{s \in S : d(s, A) < \varepsilon\}$ and $d(s, A) = \inf\{d(a, s) : a \in A\}$. The following theorem justifies the adopted terminology.

Theorem 1.14 (Prokhorov)**.** $d_{\mathcal{P}}$ *as in* (1.3) *is a metric on* $\mathcal{P}(S)$ *and, if* S *is separable, then* $d_{\mathcal{P}}(\mu_n, \mu) \to 0$ *and* $\mu_n \overset{w}{\to} \mu$ *are equivalent.*

As can be seen in the proof of Theorem 1.14 (see Appendix A.12), in general $d_{\mathcal{P}}(\mu_n,\mu)\to 0$ implies $\mu_n \overset{w}{\to} \mu$, it is for the converse result that we require S to be separable. Under this assumption we even have that $(\mathcal{P}(S),d_{\mathcal{P}})$ is separable, and in this case, for a countable dense set, D, in S,

$$\left\{\sum_{j=1}^{k} w_k\delta_{s_k} : s_1,s_2,\ldots,\in D, w_1,w_2,\ldots\in[0,1]\cap\mathbb{Q}, \sum_{j=1}^{k} w_k = 1\right\},$$

is countable and dense in $\mathcal{P}(S)$. The proof of this affirmations as well the proof of the following theorem can be found in Kallenberg (2017) or Parthasarathy (1967).

Theorem 1.15. *If (S,d) is a Polish space, then $(\mathcal{P}(S),d_{\mathcal{P}})$ is also Polish.*

With the Lévy-Prokhorov metric, $d_{\mathcal{P}}$, in place we can define the topology of weak convergence, denoted by $\tau_{\mathcal{P}}$, as the topology induced by $d_{\mathcal{P}}$, and consider the coarsest σ-algebra, $\mathscr{B}_{\mathcal{P}(S)}$, containing $\tau_{\mathcal{P}}$. The Borel σ-algebra, $\mathscr{B}_{\mathcal{P}(S)}$, makes all the integration maps, $\{\pi_f:\mu\mapsto\mu(f)\mid f:S\to\mathbb{R}\text{ is continuous and bounded}\}$, measurable. Further, if S is Polish we can even write

$$\mathscr{B}_{\mathcal{P}(S)}=\sigma(\{\pi_f:\mu\mapsto\mu(f)\mid f:S\to\mathbb{R}\text{ is continuous and bounded}\}).$$

The following theorem shows that, for a Polish space S, $\mathscr{B}_{\mathcal{P}(S)}$ as defined in Section 1.1, and as defined here are identical.

Theorem 1.16 (Borel σ-algebra of $\mathcal{P}(S)$)**.** *For a Polish space (S,d) with Borel σ-algebra $\mathscr{B}_S$, the following generate the same σ-algebra of $\mathcal{P}(S)$.*

I. *The weakly open sets of $\mathcal{P}(S)$.*

II. *The integration maps $\{\pi_f:\mu\mapsto\mu(f)\mid f:S\to\mathbb{R}$ is continuous and bounded$\}$.*

III. *The integration maps $\{\pi_f:\mu\mapsto\mu(f)\mid f:S\to\mathbb{R}_+$ is measurable$\}$.*

IV. *The projection maps $\{\pi_B:\mu\mapsto\mu(B)\mid B\in\mathscr{B}_S\}$.*

The proof of the Theorem 1.16 appears in Appendix A.13. As a consequence of this result, a random probability measure $\boldsymbol{\mu}$ can be regarded as an operator that transforms continuous and bounded functions, $f:S\to\mathbb{R}$, into random variables, $\boldsymbol{\mu}(f)$. In terms of the law of random probability measures, Theorem 1.5 can be extended as follows (see Appendix A.14, for a proof).

Lemma 1.17. *Consider a Polish space (S,d) with Borel σ-algebra $\mathscr{B}_S$. Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be random probability measures over $(S,\mathscr{B}_S)$. Then the following are equivalent.*

I. $\boldsymbol{\mu}\overset{d}{=}\boldsymbol{\nu}$

II. $\boldsymbol{\mu}(f)\overset{d}{=}\boldsymbol{\nu}(f)$ *for every continuous and bounded function $f:S\to\mathbb{R}$.*

III. $\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right]=\mathbb{E}\left[e^{-\boldsymbol{\nu}(f)}\right]$ *for every continuous and bounded function $f:S\to\mathbb{R}$.*

1.4.2 Convergence of random probability measures

For random probability measures $\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \ldots$ over $(S, \mathscr{B}_S)$, defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we say that $\left(\boldsymbol{\mu}^{(n)}\right)_{n\geq 1}$ converges weakly almost surely to $\boldsymbol{\mu}$, denoted by $\boldsymbol{\mu}^{(n)} \overset{a.s.w}{\to} \boldsymbol{\mu}$, whenever $\boldsymbol{\mu}^{(n)}(f) \overset{a.s.}{\to} \boldsymbol{\mu}(f)$ for every continuous and bounded function $f : S \to \mathbb{R}$. Analogously we say $\left(\boldsymbol{\mu}^{(n)}\right)_{n\geq 1}$ converges weakly in $\mathcal{L}_p$, in probability, or in distribution, to $\boldsymbol{\mu}$, denoted by $\boldsymbol{\mu}^{(n)} \overset{\mathcal{L}_p w}{\to} \boldsymbol{\mu}$, $\boldsymbol{\mu}^{(n)} \overset{\mathbb{P}w}{\to} \boldsymbol{\mu}$ and $\boldsymbol{\mu}^{(n)} \overset{dw}{\to} \boldsymbol{\mu}$, respectively, whenever $\boldsymbol{\mu}^{(n)}(f) \to \boldsymbol{\mu}(f)$ in the corresponding mode of convergence, for every continuous and bounded function $f : S \to \mathbb{R}$. Evidently $\boldsymbol{\mu}^{(n)} \overset{a.s.w}{\to} \boldsymbol{\mu}$ and $\boldsymbol{\mu}^{(n)} \overset{\mathcal{L}_p w}{\to} \boldsymbol{\mu}$ are both sufficient conditions for $\boldsymbol{\mu}^{(n)} \overset{\mathbb{P}w}{\to} \boldsymbol{\mu}$, which in turn implies $\boldsymbol{\mu}^{(n)} \overset{dw}{\to} \boldsymbol{\mu}$.

The following lemma, whose proof appears in Appendix A.15, allows us to characterize convergence of random probability measure in terms of its components as determined by Theorem 1.7.

Lemma 1.18 (Infinite mixtures). *Let* $\Delta_\infty = \{(w_1, w_2, \ldots) : w_j \geq 0, \sum_{j\geq 1} w_j = 1\}$ *denote the infinite dimensional simplex. Consider two Borel spaces* $(S, \mathscr{B}_S)$ *and* $(T, \mathscr{B}_T)$, *such that* S *and* T *together with suitable metrics are Polish, and let* $\nu : S \to T$ *be a probability kernel such that* $\nu_{s_n} \overset{w}{\to} \nu_s$ *as* $s_n \to s$ *in* S. *Then the following mappings are continuous with respect to the product and weak topologies.*

i) $[(w_1, w_2, \ldots), (\mu_1, \mu_2, \ldots)] \mapsto \sum_{j\geq 1} w_j \mu_j$, *from* $\Delta_\infty \times \mathcal{P}(S)^\infty$ *into* $\mathcal{P}(S)$.

ii) $[(w_1, w_2, \ldots), (s_1, s_2, \ldots)] \mapsto \sum_{j\geq 1} w_j \delta_{s_j}$, *from* $\Delta_\infty \times S^\infty$ *into* $\mathcal{P}(S)$.

iii) $\mu = \sum_{j\geq 1} w_j \delta_{s_j} \mapsto \int \nu_s \, \mu(ds) = \sum_{j\geq 1} w_j \nu_{s_j}$, *from* $\mathcal{P}(S)$ *into* $\mathcal{P}(T)$.

iv) $[(w_1, w_2, \ldots), (s_1, s_2, \ldots)] \mapsto \sum_{j\geq 1} w_j \nu_{s_j}$, *from* $\Delta_\infty \times S^\infty$ *into* $\mathcal{P}(T)$.

Indeed, for random probability measures, its atomic decomposition (Theorem 1.7) reduces to

$$\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \boldsymbol{\nu},$$

where $(\boldsymbol{\xi}_j)_{j\geq 1}$ are random elements of taking values in S, $(\mathbf{w}_j)_{j\geq 1}$ are non-negative random variables with $\sum_{j\geq 1} \mathbf{w}_j \leq 1$, and $\boldsymbol{\nu}$ is a diffuse random probability measure over $(S, \mathscr{B}_S)$. So, for example, Lemma 1.18 assures that if $\left(\mathbf{w}_j^{(n)}, \boldsymbol{\xi}_j^{(n)}\right)_{j\geq 1}$ converges is distribution to $(\mathbf{w}_j, \boldsymbol{\xi}_j)_{j\geq 1}$, where $\sum_{j\geq 1} \mathbf{w}_j^{(n)} = 1 = \sum_{j\geq 1} \mathbf{w}_j$. Then,

$$\boldsymbol{\mu}^{(n)} = \sum_{j\geq 1} \mathbf{w}_j^{(n)} \delta_{\boldsymbol{\xi}_j^{(n)}} \overset{dw}{\to} \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} = \boldsymbol{\mu}.$$

A similar argument follows for the almost sure convergence.

1.4.3 Weak support of random probability measures

Recall that for a probability measure μ on a second countable topological space, the support of μ, denoted by $\mathbb{S}(\mu)$ is the defined as the intersection of all closed sets C such that $\mu(C^c) = 0$. Further, for a random variable $\mathbf{x} \sim \mu$, the support of $\mathbf{x}$, $\mathbb{S}(\mathbf{x})$, is defined as the support of μ. The last notion we will analyse in this preliminary section is the concept of weak support of a random probability measure, that is we are interested in defining the support of, $\boldsymbol{\mu} \sim \mathsf{Q}$, where $\boldsymbol{\mu}$ is a random probability measure.

With this in mind, we find that for every $\mu \in \mathcal{P}(S)$, an open base of neighbourhoods of μ for the topology for weak convergence is the class of sets

$$\mathcal{U}_{\varepsilon_1,\dots,\varepsilon_k}(\mu; B_1, \dots, B_k) = \bigcap_{i=1}^{k} \left\{\nu \in \mathcal{P}(S) : |\mu(B_i) - \nu(B_i)| < \varepsilon_i\right\},$$

where k is a positive integer, $B_1, \dots, B_k$ are μ-continuity sets and $\varepsilon_1, \dots, \varepsilon_k$ are positive real numbers. So, for a random probability measure $\boldsymbol{\mu}$ with distribution Q, we define its (weak) support, $\mathbb{WS}(\boldsymbol{\mu}) = \mathbb{WS}(\mathsf{Q})$, as the set of all probability measures, φ, such that

$$\mathsf{Q}\left(\mathcal{U}_{\varepsilon_1,\dots,\varepsilon_k}(\varphi; B_1, \dots, B_k)\right) = \mathbb{P}\left[\boldsymbol{\mu} \in \mathcal{U}_{\varepsilon_1,\dots,\varepsilon_k}(\varphi; B_1, \dots, B_k)\right] > 0$$

for all $\varepsilon_1, \dots, \varepsilon_k > 0$, every k-tuple of φ-continuity sets and every positive integer k.

Proposition 1.19. *Let (S, d) be a Polish space with Borel σ-algebra $\mathscr{B}_S$, let $\boldsymbol{\mu} \sim \mathsf{Q}$ be a random probability measure over $(S, \mathscr{B}_S)$, with $\mathbb{E}[\boldsymbol{\mu}] = \mu_0$. Then*

$$\mathbb{WS}(\mathsf{Q}) = \mathbb{WS}(\boldsymbol{\mu}) \subseteq \{\varphi \in \mathcal{P}(S) : \mathbb{S}(\varphi) \subseteq \mathbb{S}(\mu_0)\}.$$

The proof of Proposition 1.19 is provided in Appendix A.16. This result establishes an upper bound for the weak support of a random measure. As we will explain in Section 3, from a Bayesian perspective we are interested in those random probability measures $\boldsymbol{\mu}$, such that its support is as wide as can be, that is $\mathbb{WS}(\boldsymbol{\mu}) = \{\varphi \in \mathcal{P}(S) : \mathbb{S}(\varphi) \subseteq \mathbb{S}(\mathbb{E}[\boldsymbol{\mu}])\}$.

2 Exchangeable random elements

The study of symmetrically distributed random objects is crucial in both theoretic and applied probability, particularly in Bayesian non-parametric statistics the concept of exchangeability plays a fundamental role. Roughly speaking a random element is exchangeable if its distribution remains invariant under the action of permutations, and generally for infinite dimensional exchangeable random objects there exist a representation theorem. The representation theorem for exchangeable sequences taking values in $\{0,1\}$ was first proven by de Finetti (1931), latter, authors such as Hewitt and Savage (1955) and Ryll-Nardzewski (1957) generalized this results to richer spaces. The celebrated de Finetti's theorem has been extended in various directions, for example Diaconis and Freedman (1980) derived a representation theorem for exchangeable Markov chains, and Aldous (1981) and Kallenberg (1989) analysed exchangeable and partially exchangeable arrays of random variables. Apart from these, Kingman (1978a,b), motivated by applications in genetics (Ewens; 1972), introduced the concept of partition structures, and eventually Kingman (1982) and Aldous (1985), among others, studied exchangeability for random partitions. Other random structures for which a representation theorem has been derived are stochastic process and random measures with exchangeable increments (Kallenberg; 2005, 2017). Extraordinary compilations on the subject are the monographs by Aldous (1985), Pitman (2006) and Kallenberg (2005, 2017).

This section is dedicated to the study of three closely related classes of exchangeable random objects and their corresponding representation theorems: Exchangeable random sequences taking values in some Borel space, exchangeable random partitions of $\mathbb{N}$, and locally finite exchangeable random measures over some Borel space. The representation theorem for exchangeable sequences, better known as de Finetti's representation theorem, states that a sequence $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ is exchangeable if and only if there exist an almost surely unique, $\mathbf{X}$-measurable, random probability measure, $\boldsymbol{\mu}$, known as the directing random measure of $\mathbf{X}$, such that conditionally given $\boldsymbol{\mu}$, $\mathbf{X}$ becomes a sequence of independent and identically distributed (i.i.d.) random elements with common distribution $\boldsymbol{\mu}$. This theorem makes evident that the law of an exchangeable sequence can be expressed in terms of the law of some random probability measure, explicitly if we denote by Q to the distribution of $\boldsymbol{\mu}$ then we have

$$\mathbb{P}[\mathbf{X} \in \cdot\,] = \int \mu^{\infty} \mathsf{Q}(d\mu).$$

Now, given an exchangeable sequence $\mathbf{X}$, the sequential sampling from it together with the equivalence relation over $\mathbb{N}$,

$$i \sim j \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j,$$

defines a random partition of $\mathbb{N}$, say $\boldsymbol{\Pi}$, whose distribution remains invariant under permutations of the elements of its blocks. The representation theorem for exchangeable random partitions, attributed to Kingman, states that every exchangeable partition of $\mathbb{N}$ can be generated this way. This makes clear the relation between exchangeable sequences and exchangeable partitions of $\mathbb{N}$. As to the relation between exchangeable random partitions and random probability measures, putting together both mentioned representation theorems should clarify it. Explicitly, if $\boldsymbol{\Pi}$ is constructed through the sequential sampling from an exchangeable sequence $\mathbf{X}$, the relative sizes of the blocks of $\mathbf{\Pi}$ that contain an infinite number of elements are precisely the sizes of the atoms of the directing random measure of the sequence, and the proportion of elements that contribute as a singleton to $\mathbf{\Pi}$, is the size of the diffuse part of the directing random measure. Notice that $\mathbf{\Pi}$ does not contain any information about a the values of the atoms, nor the shape of the diffuse part.

The rest of this section is organized as follows. In Section 2.1, following Aldous (1985) and Kallenberg (2005) we analyse exchangeable sequences and separately exchangeable arrays and derive the corresponding representation theorems. As an example, we illustrate with de Finetti's theorem for $\{0,1\}$-valued random variables. Section 2.2 is primarily based on the work of Pitman (2006), and it is concerned with exchangeable partitions of the set of positive integers. Here we introduce basic concepts on the subject and derive Kingman's representation theorem, latter we study some constructions of these random objects, and towards the end of this part, following Pitman (1995, 1996a), we present an overview of partially exchangeable partitions. Finally, Section 2.3, which is based in the work by Kallenberg (2017), studies random measures with exchangeable increments.

2.1 Random sequences and arrays

2.1.1 Exchangeable sequences

In this section every sequence, $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$, is assumed to take values in the Borel space $(S, \mathscr{B}_S)$, unless explicitly stated otherwise.

Definition 2.1 (Exchangeable sequences)**.**

i) We say that $(\mathbf{x}_i)_{i=1}^n$ is exchangeable if for every permutation, σ, of $\{1,\ldots,n\}$ we have that

$$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \overset{d}{=} \left(\mathbf{x}_{\sigma(1)}, \mathbf{x}_{\sigma(2)}, \ldots, \mathbf{x}_{\sigma(n)}\right).$$

ii) A sequence $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ is said to be exchangeable if for every $n \in \mathbb{N}$, the subcollection $(\mathbf{x}_i)_{i=1}^n$ is exchangeable. Equivalently, for every finite permutation of $\mathbb{N}$, σ, we have that $\sigma(\mathbf{X}) \overset{d}{=} \mathbf{X}$, where $\sigma(\mathbf{X}) = \left(\mathbf{x}_{\sigma(i)}\right)_{i\geq 1}$.

In the above definition, by a finite permutation of $\mathbb{N}$ we mean a permutation, $\sigma : \mathbb{N} \to \mathbb{N}$, that only shuffles finitely many numbers, so that there exist $n \in \mathbb{N}$, such that for every $m > n$, $\sigma(m) = m$. Note that if $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ is exchangeable, then $\sigma(\mathbf{X}) \overset{d}{=} \mathbf{X}$ also holds for arbitrary permutations of $\mathbb{N}$, this due to the fact that the law of a sequence is characterized by its finite dimensional distributions. Thus, in the second part of Definition 2.1, we could have not restricted to class of finite permutations. Moreover we could have restricted such class even more. For instance, consider the finite permutation of $\mathbb{N}$, σ_m, such that $\sigma_m(1) = m$, $\sigma_m(m) = 1$ and $\sigma_m(k) = k$ for any natural number k other than 1 and m, it is easy to see that any finite permutation of $\mathbb{N}$ can be written as finite composition of elements in $(\sigma_m)_{m\in\mathbb{N}}$, which implies that $\mathbf{X}$ is exchangeable if and only if for every $m \in \mathbb{N}$, $\sigma_m(\mathbf{X}) \overset{d}{=} \mathbf{X}$.

An important fact about exchangeable sequences is that their elements are equally distributed, this can be easily corroborated by fixing $m \in \mathbb{N}$ and $B \in \mathscr{B}_S$ and noting that

$$\mathbb{P}[\mathbf{x}_1 \in B] = \mathbb{P}\left[(\mathbf{x}_{\sigma_m(1)} \in B) \cap \bigcap_{i=1}^{m} (\mathbf{x}_{\sigma_m(i)} \in S)\right] = \mathbb{P}[\mathbf{x}_m \in B].$$

A first example of an exchangeable sequence, is a collection of i.i.d. random elements, $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$. Recall that for such a sequence there exist a (deterministic) probability measure μ, over $(S, \mathscr{B}_S)$ such that for every $n \in \mathbb{N}$ and each $\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$,

$$\mathbb{P}\left[\bigcap_{i=1}^{n} (\mathbf{x}_i \in B_i)\right] = \prod_{i=1}^{n} \mu(B_i),$$

or equivalently $\mathbb{P}\left[\mathbf{X} \in \cdot\,\right] = \mu^\infty$, denoted here by $(\mathbf{x}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \mu$, or $\mathbf{X} \sim \mu^\infty$. From the above equation, the exchangeability of $\mathbf{X}$ follows trivially. For i.i.d. sequences, as a consequence of the strong law of large numbers, it is true that their law is given by the almost sure limit

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}(B) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{x}_i \in B\}} \to \mathbb{P}[\mathbf{x}_i \in B] = \mu(B),$$

as $n \to \infty$.

In terms of the dependence between the elements, at the other side of the spectrum, we find that a second example of an exchangeable sequence is a collection of almost surely equal random elements. That is $(\mathbf{x}_i)_{i\geq 1}$, where $\mathbf{x}_i = \mathbf{x}$, almost surely, for some random element $\mathbf{x}$. For this sequence it is clear that

$$\mathbb{P}\left[\bigcap_{i=1}^{n} (\mathbf{x}_i \in B_i) \,\middle|\, \mathbf{x}\right] = \delta_{\mathbf{x}}\left(\bigcap_{i=1}^{n} B_i\right) = \prod_{i=1}^{n} \delta_{\mathbf{x}}(B_i),$$

for every $n \in \mathbb{N}$ and $B_1, \ldots, B_n \in \mathscr{B}_S$. If μ denotes the law of $\mathbf{x}$, by taking expectations, we obtain

$$\mathbb{P}\left[\bigcap_{i=1}^{n} (\mathbf{x}_i \in B_i)\right] = \mu\left(\bigcap_{i=1}^{n} B_i\right).$$

Note that $\delta_{\mathbf{x}} \overset{a.s.}{=} n^{-1}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$, so trivially $n^{-1}\sum_{i=1}^{n} \delta_{\mathbf{x}_i} \to \delta_{\mathbf{x}}$, almost surely, as $n \to \infty$.

This said, the main objective of this section is to generalize the above observations to arbitrary exchangeable sequences. Namely, we are interested in showing that $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ is exchangeable if and only if there exist an almost surely unique random probability measure $\boldsymbol{\mu}$, such that $(\mathbf{x}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, conditionally given $\boldsymbol{\mu}$. Furthermore, in this instance, $\boldsymbol{\mu}$ is given by the almost sure limit of the empirical distributions, $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}$, and the joint law of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is given by $\mathbb{E}[\boldsymbol{\mu}^n]$ for every $n \geq 1$. To this aim, let us consider the following notions that will be latter proven equivalent to exchangeability.

Definition 2.2 (Conditionally i.i.d).

i) We say $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ is conditionally i.i.d. given a σ-algebra $\mathcal{G}$, if for every $n \geq 1$ and $\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$

$$\mathbb{P}\left[\bigcap_{i=1}^{n} (\mathbf{x}_i \in B_i) \,\middle|\, \mathcal{G}\right] = \prod_{i=1}^{n} \mathbb{P}\left[\mathbf{x}_i \in B_i \,|\, \mathcal{G}\right] = \prod_{i=1}^{n} \boldsymbol{\mu}(B_i),$$

almost surely, for some random probability measure $\boldsymbol{\mu}$, or equivalently

$$\mathbb{P}[\mathbf{X} \in \cdot \mid \mathcal{G}] = \boldsymbol{\mu}^\infty \quad a.s.$$

ii) A random sequence $\mathbf{X} = (\mathbf{x}_i)_{i \geq 1}$ *is called conditionally i.i.d. if there exist a sub-*σ*-algebra* $\mathcal{G}$ *such* $\mathbf{X}$ *is conditionally i.i.d. given* $\mathcal{G}$*.*

Definition 2.3 (Mixture of i.i.d)**.** *A random sequence* $\mathbf{X} = (\mathbf{x}_i)_{i \geq 1}$ *is said to be a mixture of i.i.d. if for every* $n \geq 1$ *and* $\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$

$$\mathbb{P}\left[\bigcap_{i=1}^{n}(\mathbf{x}_i \in B_i)\right] = \int_{\mathcal{P}(S)} \prod_{i=1}^{n} \mu(B_i)\mathsf{Q}(d\mu)$$

or equivalently

$$\mathbb{P}[\mathbf{X} \in \cdot\,] = \int_{\mathcal{P}(S)} \mu^{\infty}\mathsf{Q}(d\mu)$$

for some probability measure Q *over* $(\mathcal{P}(S), \mathscr{B}_{\mathcal{P}(S)})$*.*

Definition 2.4 (Contractable sequence)**.** *A random sequence* $\mathbf{X} = (\mathbf{x}_i)_{i \geq 1}$ *is said to be contractable if for every* $n \geq 1$ *and* $0 < k_1 < k_2 < \ldots < k_n$*,*

$$(\mathbf{x}_1, \ldots, \mathbf{x}_n) \overset{d}{=} (\mathbf{x}_{k_1}, \ldots, \mathbf{x}_{k_n})$$

Equivalently, every (infinite) subsequence $\tilde{\mathbf{X}} = (\mathbf{x}_{k_j})_{j \geq 1} \subseteq (\mathbf{x}_i)_{i \geq 1}$ *where* $k_1 < k_2 < \cdots$*, satisfies* $\mathbf{X} \overset{d}{=} \tilde{\mathbf{X}}$*.*

It is a direct consequence of the above definitions that if $\mathbf{X}$ is conditionally i.i.d. then it is a mixture of i.i.d which implies $\mathbf{X}$ is exchangeable, this in turn, means that $\mathbf{X}$ is contractable. So to prove that the aforementioned definitions are all equivalent it suffices to show that if $\mathbf{X}$ is contractable, then it is a conditionally i.i.d. A somewhat rough argument, but that provides certain intuition of why this happens, can be given as follows: If $\mathbf{X} = (\mathbf{x}_i)_{i \geq 1}$ is contractable, then we might think it is a sub-sequence of larger sequence (with the cardinality of the natural numbers) which is equally distributed to $\mathbf{X}$. Insomuch as the index set, $\mathbb{N}$, was choosen arbitrarily, and as $\mathbb{Z}$ and $\mathbb{N}$ have the same cardinalities, by a simple duplication trick (after possibly enlarging the original probability space) we might even regard $\mathbf{X}$ as a subsequence of $(\mathbf{x}_i)_{i \in \mathbb{Z}}$. Then, the contractability of $(\mathbf{x}_i)_{i \in \mathbb{Z}} \overset{d}{=} (\mathbf{x}_i)_{i \geq 1}$ means that $((\mathbf{x}_i)_{i \leq 0}, \mathbf{x}_1) \overset{d}{=} ((\mathbf{x}_i)_{i \leq 0}, \mathbf{x}_k)$ for every $k \geq 2$, so that the information that $(\mathbf{x}_i)_{i \leq 0}$ provides about $\mathbf{x}_1$ is the same information that it provides about $\mathbf{x}_k$, this is

$$\mathbb{P}[\mathbf{x}_1 \in \cdot \mid (\mathbf{x}_i)_{i \leq 0}] = \mathbb{P}[\mathbf{x}_k \in \cdot \mid (\mathbf{x}_i)_{i \leq 0}],$$

a.s. The above equation states that given $(\mathbf{x}_i)_{i \leq 0}$, $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ are equally distributed. Also, the assumed contratability implies $((\mathbf{x}_i)_{i \leq k}, \mathbf{x}_{k+1}) \overset{d}{=} ((\mathbf{x}_i)_{i \leq 0}, \mathbf{x}_{k+1})$, so that the finitely many random elements $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ provide no more information about $\mathbf{x}_{k+1}$ than $(\mathbf{x}_i)_{i \leq 0}$ does, that is

$$\mathbb{P}[\mathbf{x}_{k+1} \in \cdot \mid (\mathbf{x}_i)_{i \leq 0}] = \mathbb{P}[\mathbf{x}_{k+1} \in \cdot \mid (\mathbf{x}_i)_{i \leq 0}, \mathbf{x}_1, \ldots, \mathbf{x}_k],$$

a.s. Meaning that for every $k \geq 1$, $\mathbf{x}_{k+1}$ is conditionally independent of $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ given $(\mathbf{x}_i)_{i \leq 0}$. Putting everything together, this must mean that $\mathbf{X}$ is conditionally i.i.d. given $(\mathbf{x}_i)_{i \leq 0}$.

Theorem 2.1 (Ryll-Nardzewski)**.** *Let $\mathbf{X} = (\mathbf{x}_i)_{i \geq 1}$ be a sequence taking values in a Borel space $(S, \mathscr{B}_S)$, then the following statements are equivalent*

I. $\mathbf{X}$ *is conditionally i.i.d.*

II. $\mathbf{X}$ *is a mixture of i.i.d.*

III. $\mathbf{X}$ *is exchangeable.*

IV. $\mathbf{X}$ *is contractable.*

The complete and formal proof of Theorem 2.1 appears in Appendix B.1. Notice that if there exist a random probability measure $\boldsymbol{\mu}$ such that $(\mathbf{x}_i)_{i \geq 1} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, given $\boldsymbol{\mu}$, so that $\mathbb{P}[\mathbf{X} \in \cdot \mid \boldsymbol{\mu}] = \boldsymbol{\mu}^\infty$, then by Theorem 2.1 we know that $\mathbf{X}$ is exchangeable. Conversely, if $\mathbf{X}$ is exchangeable then by the same Theorem we know that there exist a sub-σ-algebra, $\mathcal{G}$, such that $\mathbf{X}$ is conditionally i.i.d. given $\mathcal{G}$. If we let $\boldsymbol{\mu}$ be a regular version of $\mathbb{P}[\mathbf{x}_1 \in \cdot \mid \mathcal{G}]$, then, since $\boldsymbol{\mu}$ is $\mathcal{G}$-measurable, by the tower property of conditional expectation we also have that $\mathbf{X}$ is conditionally i.i.d. given $\boldsymbol{\mu}$ with $\mathbb{P}[\mathbf{x}_1 \in \cdot \mid \boldsymbol{\mu}] = \boldsymbol{\mu}$. Thus, Theorem 2.1 partially proves the result we were interested in, it still remains to show the almost sure uniqueness of $\boldsymbol{\mu}$ and that $\lim_{n\to\infty} n^{-1} \sum_{i=1}^n \delta_{\mathbf{x}_i} = \boldsymbol{\mu}$ almost surely. The following Theorem, whose proof can be found Appendix B.2, shows the remaining part.

Theorem 2.2. *Let $\mathbf{X} = (\mathbf{x}_i)_{i \geq 1}$ be an exchangeable sequence such that $\mathbb{P}[\mathbf{X} \in \cdot \mid \mathcal{G}] \overset{a.s.}{=} \boldsymbol{\mu}^\infty$ for some sub-σ-algebra $\mathcal{G}$ and some random probability measure $\boldsymbol{\mu}$ over $(S, \mathscr{B}_S)$. Then,*

a) $\mathbf{X}$ *is independent of $\mathcal{G}$ given $\boldsymbol{\mu}$.*

b) $\mathbf{X}$ *is conditionally i.i.d. given $\boldsymbol{\mu}$.*

c) $\boldsymbol{\mu}$ *is unique almost surely, $\mathbf{X}$-measurable and given by the almost sure limit*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}(B) \overset{a.s.}{=} \boldsymbol{\mu}(B) \quad B \in \mathscr{B}_S$$

d) $\mathbb{P}[\mathbf{X} \in \cdot] = \int_{\mathcal{P}(S)} \mu^\infty \mathsf{Q}(d\mu)$ *if and only if Q is the law of $\boldsymbol{\mu}$.*

In the context of the above theorem, $\boldsymbol{\mu}$, which is almost surely unique and $\mathbf{X}$-measurable is called the directing random measure of $\mathbf{X}$, we equivalently say that $\mathbf{X}$ is directed by $\boldsymbol{\mu}$ and denote this by $\{\mathbf{X} \mid \boldsymbol{\mu}\} \sim \boldsymbol{\mu}^\infty$ or $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$. Another important highlight here, is that the laws of $\mathbf{X}$ and Q characterize each other, and Q is known as the de Finetti measure of $\mathbf{X}$. The name of Q comes from the fact that the most widely known version of the Theorem 2.1, known as de Finneti's representation theorem (in honour to de Finetti (1931)), states, without involving the concept of contractability, that every sequence, $\mathbf{X} = (\mathbf{x}_i)_{i \geq 1}$, is exchangeable if and only if it is conditionally i.i.d. together with Theorem 2.2. The proof of de Finetti's theorem, instead of using the tail σ-algebra of the sequence (see Appendix B.1), uses the so-called exchangeable σ-algebra defined as $\mathcal{H} = \bigcap_{n \in \mathbb{N}} \mathcal{H}_n$, where

$$\mathcal{H}_n = \sigma(\theta_n(\mathbf{X}), f(\mathbf{x}_1, \ldots, \mathbf{x}_n) : f \text{ is a symmetric function}),$$

and $\theta_n(\mathbf{X}) = (\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots)$. Roughly speaking, $\mathcal{H}_n$ contains the complete information of $\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots$ and partial information of $\mathbf{x}_1, \ldots \mathbf{x}_n$, more precisely, the information that $\mathcal{H}_n$ contains about the latter group of random elements is through symmetric functions, so for instance $\sum_{i=1}^n \mathbf{x}_i$ and $\prod_{i=1}^n \mathbf{x}_i$ are both $\mathcal{H}_n$-measurable functions. Also, for $1 \leq k \leq n$, and a measurable function $f : S \to \mathbb{R}_+$, $\mathbb{E}[f(\mathbf{x}_1)|\mathcal{H}_n] \overset{a.s.}{=} \mathbb{E}[f(\mathbf{x}_k)|\mathcal{H}_n]$. That said, this alternative proof consist in noting that $\{\mathcal{H}_n\}_{n\in\mathbb{N}}$ is a reverse filtration, and that for every $1 \leq k \leq n$ and bounded measurable function $f : S^k \to \mathbb{R}$ we have that

$$\mathbb{E}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_k)|\mathcal{H}_n\right] = \frac{1}{|N_{n,k}|} \sum_{(j_1,\ldots,j_k)\in N_{n,k}} f(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k}), \tag{2.1}$$

a.s. where $N_{n,k} = \{(j_1, \ldots, j_k) \in \{1, \ldots, n\}^k : j_i \neq j_l, \text{ for all } i \neq j\}$, this can be easily seen by noting that the right hand side is a symmetric function of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Now, the left hand side of equation (2.1) is a reverse martingale, hence converges a.s. to $\mathbb{E}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_k)|\mathcal{H}\right]$ as $n \to \infty$, meanwhile the right hand side is asymptotically equivalent to

$$\frac{1}{n^k} \sum_{j_1=1}^n \cdots \sum_{j_k=1}^n f(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k})$$

thus,

$$\frac{1}{n^k} \sum_{j_1=1}^n \cdots \sum_{j_k=1}^n f(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_k}) \overset{a.s.}{\to} \mathbb{E}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_k)|\mathcal{H}\right].$$

The choice $k = 1$ and $f = \mathbf{1}_B$ for some $B \in \mathscr{B}_S$ proves

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_B(\mathbf{x}_i) \overset{a.s.}{=} \mathbb{P}[\mathbf{x}_1 \in B|\mathcal{H}],$$

and the choice $f = \prod_{i=1}^k \mathbf{1}_{B_i}$ for some $\{B_i\}_{i=1}^k \in \mathscr{B}_S$ proves the desired conditional independence.

The next example discusses the above characterizations for some of the simplest sequences of exchangeable random variables.

Example 2.1. *If $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ is an exchangeable sequence of $\{0, 1\}$-valued random variables, the results we have developed so far simplify to:*

i) The long-run proportion of random variables that take the value $x \in \{0, 1\}$ must be

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{x\}}(\mathbf{x}_i) \overset{a.s.}{=} \mathbf{p}^x(1 - \mathbf{p})^{1-x}, \quad x \in \{0, 1\}$$

for some $[0, 1]$-valued random variable $\mathbf{p}$.

ii) Conditionally given $\mathbf{p}$, $\mathbf{X}$ is a collection of i.i.d. random variables such that $\mathbf{x}_i \sim \mathsf{Ber}(\mathbf{p})$, and

iii) For every $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \{0, 1\}^n$

$$\mathbb{P}[\mathbf{x}_1 = x_1, \ldots, \mathbf{x}_n = x_n] = \int_{[0,1]} p^x(1 - p)^{n-x} \mathsf{Q}(dp),$$

for some probability measure Q over $([0, 1]; \mathscr{B}_{[0,1]})$, and where $x = \sum_{i=1}^n x_i$.

In order words, i), ii) and iii) explain that the most general sequence of exchangeable $\{0,1\}$-valued random variables can be constructed by choosing a $[0,1]$-valued random variable $\mathbf{p}$ and conditionally given $\mathbf{p}$, sampling Bernoulli random variables with such parameter. In particular, if $\mathbf{p} \sim \mathsf{Be}(\alpha,\beta)$

$$\begin{aligned}\mathbb{P}[\mathbf{x}_1 = x_1, \ldots, \mathbf{x}_n = x_n] &= \int_{[0,1]} p^x(1-p)^{n-x}\mathsf{Q}(dp) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{\alpha+x-1}(1-p)^{\beta+n-x-1}dp \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+n-x)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)} \\ &= \frac{(\alpha)_x(\beta)_{n-x}}{(\alpha+\beta)_n}\end{aligned} \tag{2.2}$$

where $(z)_m = \prod_{k=0}^{m-1}(z+k)$ with the convention that the empty product equals one.

The next example is a particular and fun case of Example 2.1.

Example 2.2 (Pólya Eggenberger urn)**.** *Imagine an urn contains β white balls and α black balls. At each step of the experiment, a ball is drawn and returned together with another one of the same color. Define the event*

$$A_i = \text{ the ith ball drawn is black}$$

and let $\mathbf{x}_i = \mathbf{1}_{A_i}$. Then it is easy to see that $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ is a sequence of exchangeable $\{0,1\}$-valued random variables, as for $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \{0,1\}$ and σ permutation of $\{1,\ldots,n\}$,

$$\mathbb{P}\left(\bigcap_{i=1}^n(\mathbf{x}_i = x_i)\right) = \frac{(\alpha)_x(\beta)_{n-x}}{(\alpha+\beta)_n} = \mathbb{P}\left(\bigcap_{i=1}^n(\mathbf{x}_{\sigma(i)} = x_i)\right). \tag{2.3}$$

Comparing equations (2.2) and (2.3), we see that $\mathbf{X}$ could have been equivalently constructed by independently letting $\mathbf{x}_i \sim \mathsf{Ber}(\mathbf{p})$ (conditionally given $\mathbf{p}$) for some $\mathbf{p} \sim \mathsf{Be}(\alpha,\beta)$. Moreover by i) in Example 2.1 we see that the long-run proportion of black balls in the urn is $\mathsf{Be}(\alpha,\beta)$ distributed whilst the long-run proportion of white balls in the urn is $\mathsf{Be}(\beta,\alpha)$ distributed.

Before moving on let us highlight one important fact. A key piece, to characterize exchangeable (contractable) sequences, is the fact that they are infinite sequences, this way we can define the tail σ-algebra, the exchangeable σ-algebra or the limit in Theorem 2.2 c), this limiting information is precisely what allows any of the above characterizations. For a finite collection of exchangeable random variables, $(\mathbf{x}_i)_{i=1}^n$, we still have that if it is conditionally i.i.d. then it is a mixture of i.i.d., which in turn implies it is exchangeable. Despite the equivalence holds only if the finite collection can be regarded as subset of some infinite exchangeable sequence. The next example shows a finite exchangeable sequence that can not be thought as a sub-collection of a larger one.

Example 2.3. *Let $\mathbf{X} = (\mathbf{x}_i)_{i\in\{1,2\}}$ such that*

$$\mathbb{P}[\mathbf{x}_1 = 0, \mathbf{x}_2 = 1] = \mathbb{P}[\mathbf{x}_1 = 1, \mathbf{x}_2 = 0] = \frac{1}{2}$$

and

$$\mathbb{P}[\mathbf{x}_1 = 1, \mathbf{x}_2 = 1] = \mathbb{P}[\mathbf{x}_1 = 0, \mathbf{x}_2 = 0] = 0.$$

$\mathbf{X}$ *is clearly a finite collection of exchangeable random variables, now if it could be regarded as a subset of an infinite exchangeable sequence, then the entire sequence would be* $\{0,1\}$*-valued and iii) in Example 2.1 must have held, thus, particularly*

$$0 = \int_{[0,1]} p^2 \mathsf{Q}(dp) = \int_{[0,1]} (1-p)^2 \mathsf{Q}(dp),$$

that is, Q *assigns a mass of one to* 0 *and* 1*, which is impossible.*

2.1.2 Partially exchangeable arrays

In this section, we will discuss a generalization of exchangeability to a higher dimension, so instead of working with sequences we will be working with arrays, here denoted by $\mathbf{X} = \left(\mathbf{x}_i^{(j)}\right)_{i\in\mathbb{N}, j\in\mathrm{J}}$, where the index set, J, is at most countable. The sequence $\mathbf{X}^{(j)} = \left(\mathbf{x}_i^{(j)}\right)_{i\in\mathbb{N}}$ is going to be thought as jth column of the array, whilst $\mathbf{X}_i = \left(\mathbf{x}_i^{(j)}\right)_{j\in\mathrm{J}}$ denotes the ith row. We will be particularly interested in deriving a representation theorem for the arrays whose distribution is preserved under permutations of entries that leave each element in the same column (see Figure 4). Before doing so, we define the notion of exchangeability over some random element and highlight some results concerning such concept, these will make our objective become trivial. The results in this part are corollaries of Theorems 2.1 and 2.2, that is why we will call them such way.

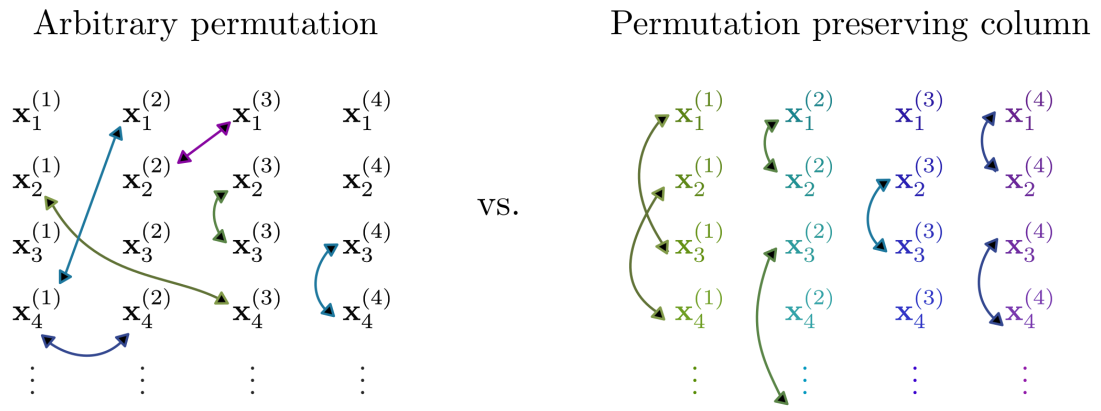


Figure 4: Illustration of two permutations of an array, the one in the left represents an arbitrary shuffle, whilst the one in the right leaves each element in its original column.

Definition 2.5. *We say* $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ *(taking values in the Borel space* $(S, \mathscr{B}_S)$*) is exchangeable over a random element* $\boldsymbol{\zeta}$ *(with range the Borel space* $(T, \mathscr{B}_T)$*) if* $\boldsymbol{\eta} = ((\mathbf{x}_i, \boldsymbol{\zeta}))_{i\geq 1}$ *forms and exchangeable sequence.*

Note that if $\mathbf{X}$ is exchangeable over $\boldsymbol{\zeta}$ then for $\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$

$$\mathbb{P}\left[\bigcap_{i=1}^n [(\mathbf{x}_i, \boldsymbol{\zeta}) \in (B_i \times T)]\right] = \mathbb{P}\left[\bigcap_{i=1}^n (\mathbf{x}_i \in B_i)\right]$$

which implies $\mathbf{X}$ is exchangeable itself. By the third part of Theorem 2.2, for $B \in \mathscr{B}_S$ and $A \in \mathscr{B}_T$

$$\delta_{\boldsymbol{\zeta}}(A)\left(\frac{1}{n}\sum_{i\leq n}\mathbf{1}_B(\mathbf{x}_i)\right) \to \boldsymbol{\mu}(B)\delta_{\boldsymbol{\zeta}}(A),$$

almost surely, as $n \to \infty$, where $\boldsymbol{\mu}$ is the directing random measure of $\mathbf{X}$. Simultaneously, we have

$$\delta_{\boldsymbol{\zeta}}(A)\left(\frac{1}{n}\sum_{i\leq n}\mathbf{1}_B(\mathbf{x}_i)\right) = \frac{1}{n}\sum_{i\leq n}\mathbf{1}_{B\times A}(\mathbf{x}_i,\boldsymbol{\zeta}) \to \boldsymbol{\nu}(B\times A),$$

almost surely, where $\boldsymbol{\nu}$ is the directing random measure of $\boldsymbol{\eta}$. This means $\boldsymbol{\mu}\otimes\delta_{\boldsymbol{\zeta}}$ coincides (almost surely) with the directing random measure of $\boldsymbol{\eta} = ((\mathbf{x}_i,\boldsymbol{\zeta}))_{i\geq 1}$. In particular, we have that the σ-algebra generated by $\boldsymbol{\nu}$ and the one generated by $(\boldsymbol{\zeta},\boldsymbol{\mu})$ are equal, except for null sets, and

$$\mathbb{P}[\mathbf{X}\in\cdot \mid \boldsymbol{\zeta},\boldsymbol{\mu}] = \mathbb{P}[\mathbf{X}\in\cdot|\boldsymbol{\nu}] = \boldsymbol{\mu}^\infty,$$

almost surely, so we have proven the following result.

Corollary 2.3. *Let $(S,\mathscr{B}_S)$ and $(T,\mathscr{B}_T)$ some Borel spaces, Let $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$ taking values in S be exchangeable over a random element $\boldsymbol{\zeta}$ with range space T. Then*

a) *The directing random measure of $\boldsymbol{\eta} = ((\mathbf{x}_i,\boldsymbol{\zeta}))_{i\geq 1}$ is $\boldsymbol{\nu} = \boldsymbol{\mu}\otimes\delta_{\boldsymbol{\zeta}}$, where $\boldsymbol{\mu}$ is the directing random measure of $\mathbf{X}$, and*

b) *$\mathbf{X}$ is conditionally independent of $\boldsymbol{\zeta}$ and $\boldsymbol{\nu}$ given $\boldsymbol{\mu}$.*

Definition 2.6. *Let $\mathrm{J} \subseteq \mathbb{N}$. For $j \in \mathrm{J}$, let $\mathbf{X}^{(j)} = \left(\mathbf{x}_i^{(j)}\right)_{i\geq 1}$ be a collection of random elements taking values in some Borel space $(S,\mathscr{B}_S)$.*

i) We say $\mathbf{X} = \left(\mathbf{x}_i^{(j)}\right)_{i\in\mathbb{N},j\in\mathrm{J}}$ is a completely exchangeable or contractable array if its distribution is invariant under arbitrary permutations of its elements, or equivalently, any sub-array of the same size has the same distribution as the original.

ii) We say $\mathbf{X} = \left(\mathbf{x}_i^{(j)}\right)_{i\in\mathbb{N},j\in\mathrm{J}}$ is separately exchangeable or contractable if for any family of permutations of $\mathbb{N}$, or contractions $\sigma = (\sigma_j)_{j\in\mathrm{J}}$

$$\sigma(\mathbf{X}) = \left(\sigma_j\left(\mathbf{X}^{(j)}\right)\right)_{j\in\mathrm{J}} \overset{d}{=} \mathbf{X}$$

where $\sigma_j\left(\mathbf{X}^{(j)}\right) = \left(\mathbf{x}_{\sigma_j(i)}^{(j)}\right)_{i\in\mathbb{N}}$.

There are not many new things to say about completely exchangeable arrays, after all, they are countable sets of exchangeable random elements, hence we can reorder such elements and treat them as exchangeable sequences. So let us focus in separately exchangeable arrays. For the separately exchangeable or contractable array, $\mathbf{X}$, let us denote

$$\mathbf{X}\setminus\mathbf{X}^{(k)} = \left(\mathbf{X}^{(j)}\right)_{j\neq k},\ \mathbf{X}^{(k)}\setminus\mathbf{x}_m^{(k)} = \left(\mathbf{x}_i^{(k)}\right)_{i\neq m},\ \text{ and } \mathbf{X}\setminus\mathbf{x}_m^{(k)} = \left(\mathbf{x}_i^{(j)}\right)_{(i,j)\neq(m,k)}.$$

In the context of the Definition 2.6, by letting σ_j be the identity function for each $j \neq k$, we obtain that $\mathbf{X}^{(k)}$ is exchangeable (or contractable) over $\mathbf{X} \setminus \mathbf{X}^{(k)}$ hence it is exchangeable itself and has a directing random measure, say $\boldsymbol{\mu}^{(k)}$. By the second part of Corollary 2.3 we have that $\mathbf{X}^{(k)}$ is conditionally independent of $\mathbf{X} \setminus \mathbf{X}^{(k)}$ given $\boldsymbol{\mu}^{(k)}$ and by the representation theorem for exchangeable sequences, we know $\mathbf{x}_m^{(k)}$ is conditionally independent of $\mathbf{X}^{(k)} \setminus \mathbf{x}_m^{(k)}$ given $\boldsymbol{\mu}^{(k)}$, these together imply that $\mathbf{x}_m^{(k)}$ must be conditionally independent of $\mathbf{X} \setminus \mathbf{x}_m^{(k)}$ given $\boldsymbol{\mu}^{(k)}$. Moreover, the second part of Corollary 2.3 also states that $\mathbf{X}^{(k)}$ is conditionally independent of $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$, given $\boldsymbol{\mu}^{(k)}$ that is

$$\mathbb{P}\left[\mathbf{X}^{(k)} \in \cdot \,\middle|\, \left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}\right] = \mathbb{P}\left[\mathbf{X}^{(k)} \in \cdot \mid \boldsymbol{\mu}^{(k)}\right] = \left(\boldsymbol{\mu}^{(k)}\right)^{\infty}.$$

These results are summarized as follows.

Corollary 2.4. *Let $(S, \mathscr{B}_S)$ be a Borel space, $\mathrm{J} \subseteq \mathbb{N}$ and $\mathbf{X} = \left(\mathbf{X}^{(j)}\right)_{j\in\mathrm{J}}$ be separately exchangeable with range space S, such that $\mathbf{X}^{(j)} = \left(\mathbf{X}_i^{(j)}\right)_{i\in\mathbb{N}}$ has directing random measure $\boldsymbol{\mu}^{(j)}$. Then, conditionally given $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$, the elements of $\left(\mathbf{x}_i^{(j)}\right)_{i\in\mathbb{N}, j\in\mathrm{J}}$, are independent and $\mathbf{x}_i^{(j)} \sim \boldsymbol{\mu}^{(j)}$.*

Conversely if $\mathbf{X}$ is an array, whose columns are conditionally independent given $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$, and $\left\{\mathbf{X}^{(k)} \,\middle|\, \left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}\right\} \sim \left(\boldsymbol{\mu}^{(k)}\right)^{\infty}$. Then

$$\mathbb{P}\left[\mathbf{X} \in \cdot\right] = \int_{\mathcal{P}(S)^{|\mathrm{J}|}} \prod_{j\in\mathrm{J}} \mu_j^{\infty} \mathsf{Q}(d\mu), \tag{2.4}$$

where $\mu = (\mu_j)_{j\in\mathrm{J}}$, and Q denotes the law of $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$. Equivalently, for every finite subset K of J, $n_k \in \mathbb{N}$ and $B_{k,1}, \ldots, B_{k,n_k} \in \mathscr{B}_S$,

$$\mathbb{P}\left[\bigcap_{k\in\mathrm{K}} \left(\bigcap_{i=1}^{n_k} \left\{\mathbf{x}_i^{(k)} \in B_{k,i}\right\}\right)\right] = \int_{\mathcal{P}(S)^{|\mathrm{K}|}} \prod_{k\in\mathrm{K}} \left(\prod_{i=1}^{n_k} \mu_k(B_{k,i})\right) \mathsf{Q}_{\mathrm{K}}(d\mu_1, \ldots, d\mu_k),$$

where Q_{K} denotes the law of $\left(\boldsymbol{\mu}^{(k)}\right)_{k\in\mathrm{K}}$. In particular if $\mathrm{K} = \{1, 2\}$, for every $n_1, n_2 \in \mathbb{N}$ and $A_1, \ldots, A_{n_1}, B_1, \ldots, B_{n_2} \in \mathscr{B}_S$, we obtain

$$\begin{aligned}\mathbb{P}&\left[\mathbf{x}_1^{(1)} \in A_1, \ldots, \mathbf{x}_{n_1}^{(1)} \in A_{n_1}, \mathbf{x}_1^{(2)} \in B_1, \ldots, \mathbf{x}_{n_2}^{(2)} \in B_{n_2}\right] \\ &= \int_{\mathcal{P}(S)^2} \left(\prod_{i=1}^{n_k} \mu_1(A_i)\right) \left(\prod_{i=1}^{n_2} \mu_2(B_i)\right) \mathsf{Q}_{\{1,2\}}(d\mu_1, d\mu_2).\end{aligned}$$

where $\mathsf{Q}_{\{1,2\}}$ denotes the law of $\left(\boldsymbol{\mu}^{(k)}\right)_{k=1}^{2}$. From equation (2.4) it is clear that $\mathbf{X}$ is separately exchangeable. This way, we have proven the representation theorem for separately exchangeable arrays.

Theorem 2.5. *Let $\mathbf{X} = \left(\mathbf{x}_i^{(j)}\right)_{i\in\mathbb{N}, j\in\mathrm{J}}$ be an array whose elements take values in the Borel space $(S, \mathscr{B}_S)$, and where $\mathrm{J} \subseteq \mathbb{N}$ is at most countable. Then the following are equivalent:*

I. $\mathbf{X}$ *is separately exchangeable.*

II. *There exist an almost surely unique collection of random measures* $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$*, such that the elements of* $\mathbf{X}$ *are conditionally independent given* $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$ *and its columns satisfy* $\left\{\mathbf{X}^{(k)}\,\middle|\,\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}\right\}\sim\left(\boldsymbol{\mu}^{(k)}\right)^{\infty}$.

III. *For every finite subset* K *of* J*,* $n_k\in\mathbb{N}$ *and* $B_{k,1},\dots,B_{k,n_k}\in\mathscr{B}_S$*,*

$$\mathbb{P}\left[\bigcap_{k\in\mathrm{K}}\left(\bigcap_{i=1}^{n_k}\left\{\mathbf{x}_i^{(k)}\in B_{k,i}\right\}\right)\right]=\int_{\mathcal{P}(S)^{|\mathrm{K}|}}\prod_{k\in\mathrm{K}}\left(\prod_{i=1}^{n_k}\mu_k(B_{k,i})\right)\mathsf{Q}(d\mu),$$

for some probability measure Q *over* $\left(\mathcal{P}(S)^{|J|},\mathscr{B}_{\mathcal{P}(S)^{|J|}}\right)$*, and where* μ *denotes* $(\mu_j)_{j\in\mathrm{J}}$.

Theorem 2.5 shows that the most general separately exchangeable array, $\mathbf{X}=\left(\mathbf{x}_i^{(j)}\right)_{j\in\mathrm{J},i\in\mathbb{N}}$, can be constructed by first considering a collection of random measures $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$, and then sampling $\mathbf{x}_i^{(j)}\sim\boldsymbol{\mu}^{(j)}$ (conditionally given $\boldsymbol{\mu}^{(j)}$) for every $i\geq 1$. The dependence scheme between the columns of the array and the directing random measures is illustrated below, recalling that $\mathbf{X}^{(j)}=\left(\mathbf{x}_i^{(j)}\right)_{i\geq 1}$ denotes the jth column of the array.

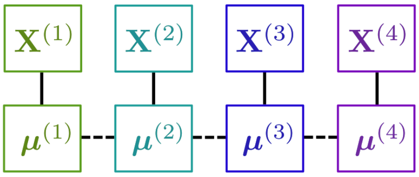


Figure 5: Dependence scheme between the first four columns of a separately exchangeable array and their directing random measures.

Notice that in Theorem 2.5, no further requirements are made about the mutual dependence of the random probability measures in $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$. This said, we can recognize two opposite scenarios. The first one, where the elements in $\left(\boldsymbol{\mu}^{(j)}\right)_{j\in\mathrm{J}}$ are mutually independent (see Figure 6) so that $\mathbf{X}^{(1)},\mathbf{X}^{(2)},\dots$ are simply, mutually independent, exchangeable sequences. And the second one, where $\boldsymbol{\mu}^{(j)}=\boldsymbol{\mu}$, for every $j\in\mathrm{J}$ and some random probability measure, $\boldsymbol{\mu}$, in which case $\{\mathbf{X}\mid\boldsymbol{\mu}\}\sim\boldsymbol{\mu}^{\infty}$, and we obtain that $\mathbf{X}$ is even exchangeable (see Figure 7).

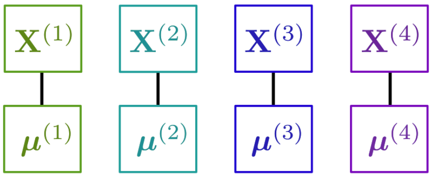


Figure 6: Dependence scheme between the first four columns of a separately exchangeable array and their mutually independent directing random measures.

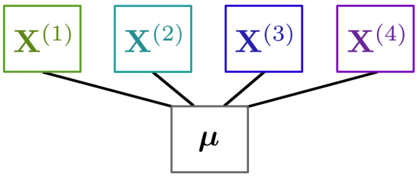


Figure 7: Dependence scheme between the first four columns of an exchangeable array and their unique directing random measure.

Form a practical point of view, if one encounters grouped data, say,

$$\left\{X^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, \ldots\right), X^{(2)} = \left(x_1^{(2)}, x_2^{(2)}, \ldots\right), \ldots, X^{(n)} = \left(x_1^{(n)}, x_2^{(n)}, \ldots\right)\right\},$$

modelling it as if sampled from a separately exchangeable array $\mathbf{X} = \left(\mathbf{x}_i^{(j)}\right)_{i \geq 1, j \in [n]}$, is a great idea. This, due to the fact that the dependence between the groups can range between complete independence, meaning that we might as well model each group separately, and the case where the groups behave as though they were one unique larger group. Furthermore, under this assumption, within each group the data is exchangeable, roughly speaking this means that the order in which the data points, $x_1^{(j)}, x_2^{(j)}, \ldots$, we sampled, is irrelevant. Depending on the context this consideration can be as flexible as possible without losing mathematical tractability of the model. Examples of Bayesian non-parametric models that exploit this type of symmetric arrays can be found in the work of Camerlenghi et al. (2019).

A final remark on this topic is that separately exchangeable arrays are often called partially exchangeable instead. This section was mainly based in the work of Kallenberg (2005), so we adopted the terminology used there.

2.2 Random partitions

2.2.1 Preliminary definitions

Given an arbitrary non-empty set, S, a partition of S is an unordered collection of subsets of S, $A = \{A_j\}_{j \geq 1}$, called blocks, such that $A_j \neq \emptyset$ for every $j \geq 1$, $A_i \cap A_j = \emptyset$ for every $i \neq j$ and $\bigcup_{j \geq 1} A_j = S$. We will mainly be working with partitions of $\mathbb{N}$ or $[n] = \{1, \ldots, n\}$, we denote by $\mathcal{P}_{[n]}^k$ to set of all partitions of $[n]$ into exactly k blocks, by $\mathcal{P}_{[n]}$ to set of all partitions of $[n]$, and by $\mathcal{P}_{\mathbb{N}}$ to the set of all partitions of $\mathbb{N}$. For a positive integer n, a composition of n into $1 \leq k \leq n$ parts is a sequence $(n_1, \ldots, n_k)$ of positive integers such that $\sum_{i=1}^k n_i = n$. We will denote by $\mathcal{C}_n^k$ to set of all compositions of n into exactly k parts, and by $\mathcal{C}_n$ to set of all compositions of n. Note that in a partition the ordering of the blocks is irrelevant, for instance, $\{\{1,2\},\{3\}\}$ is exactly the same partition as $\{\{3\},\{1,2\}\}$, but in a composition the order do matters so $(1,2)$ and $(2,1)$ are two distinct compositions of 3. We say a partition is ordered when we assign an ordering to the blocks, it is straight forward that for each partition with k blocks there are $k!$ ordered partitions that correspond to it. Note that given an ordered partition, $(A_1, \ldots, A_k)$, of $[n]$, the vector $(|A_1|, \ldots, |A_k|)$ defines a composition of n into k parts. We can also define an unordered composition of n into k parts as any set $\{n_j\}_{j=1}^k$ of positive integers such that $\sum_{j=1}^k n_j = n$. Any unordered composition of n into k parts, $\{n_j\}_{j=1}^k$, can be uniquely identified with and element in $\mathcal{M}_n^k = \{(m_1, \dots, m_n) \in \mathbb{Z}_+^n : \sum_{i=1}^n im_i = n, \sum_{i=1}^n m_i = k\}$, through $m_i = \sum_{j=1}^k \mathbf{1}_{\{n_j=i\}}$, that is m_i counts the number of parts of the composition that equal i. Clearly, any partition, $A = \{A_1, \dots, A_k\}$, of $[n]$ defines an element in $\mathcal{M}_n^k$ through $m_i = |\{A_j \in A : |A_j| = i\}|$, for every $i \in [n]$.

Definition 2.7 (Internal permutation of a partition)**.** *Let $A \in \mathcal{P}_{[n]}$ and σ be a permutation of $[n]$ (or let $A \in \mathcal{P}_{\mathbb{N}}$ and σ be a permutation of $\mathbb{N}$). We define the σ-internal permutation of A by*

$$\sigma(A) = \{\sigma(A_j) : A_j \in A\}$$

where $\sigma(A_j) = \{\sigma(i) : i \in A_j\}$ for every block, A_j, of A.

Example 2.4. *Consider $A = \{\{1, 6\}, \{2, 5\}, \{3\}, \{4\}\} \in \mathcal{P}_{[6]}$, and the permutation of $[6]$, σ, given by $\sigma(1) = 3$, $\sigma(2) = 5$, $\sigma(3) = 6$, $\sigma(4) = 4$, $\sigma(5) = 1$ and $\sigma(6) = 2$, or in compact cycle notation $\sigma = (13625)(4)$, then*

$$\sigma(A) = \{\{3, 2\}, \{5, 1\}, \{6\}, \{4\}\}$$

Definition 2.8 (Restriction of partitions)**.** *Let $m \leq n$. Let $A \in \mathcal{P}_{[n]}$, or $A \in \mathcal{P}_{\mathbb{N}}$, we define the restriction of A to $[m]$ by*

$$A\big|_{[m]} = \{A_j \cap [m] : A_j \in A, A_j \cap [m] \neq \emptyset\}.$$

Example 2.5. *Let A and σ be as in Example 2.4, then*

a) $A\big|_{[5]} = \{\{1\}, \{2, 5\}, \{3\}, \{4\}\}$

b) $\sigma(A)\big|_{[4]} = \{\{3, 2\}, \{1\}, \{4\}\}$

Let $m \leq n$ and $A \in \mathcal{P}_{[m]}$. Let us denote by $\mathcal{P}_{[n]}(A)$ to the set of all partitions of $[n]$ such that its restriction to $[m]$ is A, that is

$$\mathcal{P}_{[n]}(A) = \{B \in \mathcal{P}_{[n]} : B\big|_{[m]} = A\}$$

and analogously

$$\mathcal{P}_{\mathbb{N}}(A) = \{B \in \mathcal{P}_{\mathbb{N}} : B\big|_{[m]} = A\}$$

Example 2.6. *Consider $A = \{\{1, 3\}, \{2\}\}$. Then*

$$\mathcal{P}_{[4]}(A) = \Big\{\{\{1, 3, 4\}, \{2\}\}, \{\{1, 3\}, \{2, 4\}\}, \{\{1, 3\}, \{2\}, \{4\}\}\Big\}.$$

In general, for every $n \in \mathbb{N}$, and $A = \{A_1, \dots, A_k\} \in \mathcal{P}_{[n]}$

$$\mathcal{P}_{[n+1]}(A) = \{A^{(j)} : j \in \{1, \dots, k + 1\}\}$$

where $A^{(j)} = \{A_i\}_{i\neq j} \cup \{A_j \cup \{n + 1\}\}$ and $A^{(k+1)} = \{A_1, \dots, A_k, \{n + 1\}\}$.

Let us denote by $\mathcal{P}_\infty$ to the set of all infinite families of partitions $(\Pi_n)_{n\geq1}$ such that $\Pi_n \in \mathcal{P}_{[n]}$ and for every $1 \leq m \leq n$, $\Pi_m = \Pi_n\big|_{[m]}$. Notice that there is a correspondence one to one between elements of $\mathcal{P}_\infty$ and $\mathcal{P}_{\mathbb{N}}$. Namely, for a partition of $\mathbb{N}$, Π, the collection

$\left(\Pi\big|_{[n]}\right)_{n\geq 1} \in \mathcal{P}_\infty$, conversely for $(\Pi_n)_{n\geq 1} \in \mathcal{P}_\infty$ there exist a unique element $\Pi \in \mathcal{P}_\mathbb{N}$ such that $\Pi_n = \Pi\big|_{[n]}$, for every $n \geq 1$. For this reason, we simply call the elements of $\mathcal{P}_\infty$ as partitions of $\mathbb{N}$.

Due to its finite nature, the partition space $\mathcal{P}_{[n]}$ is Borel and its Borel σ-algebra is $\mathscr{B}_{\mathcal{P}_{[n]}} = 2^{\mathcal{P}_{[n]}}$, where 2^S denotes the set of all subsets of S. This said, a random partition $\mathbf{\Pi}_n$ of $[n]$ is a measurable mapping, $\mathbf{\Pi}_n : \Omega \to \mathcal{P}_{[n]}$. Note that together with a random partition of $[n]$, say $\mathbf{\Pi}_n$, there other random variables implicitly defined, such as the number of blocks of $\mathbf{\Pi}_n$, which we will denote by $\mathbf{K}_n$, and the unordered composition of n corresponding to the frequencies of the blocks which we will denote by $\mathbf{N}_n = \{\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}_n}\}$. Alternatively, we can encrypt this unordered composition through the random vector, $\mathbf{M}_n = (\mathbf{m}_1, \ldots, \mathbf{m}_n)$, given by $\mathbf{m}_i = \sum_{j=1}^{\mathbf{K}_n} \mathbf{1}_{\{\mathbf{n}_j = i\}}$. Note that

$$\mathbb{P}[\mathbf{K}_n = k] = \sum_{A \in \mathcal{P}^k_{[n]}} \mathbb{P}[\mathbf{\Pi}_n = A],$$

so the distribution of $\mathbf{K}_n$ is known (perhaps not analytically) given the distribution of $\mathbf{\Pi}_n$, this is not different for the distribution of $\mathbf{N}_n$ or $\mathbf{M}_n$.

2.2.2 Consistent families of random partitions

Definition 2.9 (Strongly consistent random partitions). *For $n \geq 1$ let $\mathbf{\Pi}_n$ be a random partition of $[n]$. We say that $(\mathbf{\Pi}_n)_{n\geq 1}$ is (strongly) consistent or a random partition of $\mathbb{N}$, if for every $m \leq n$, $\mathbf{\Pi}_n \in \mathcal{P}_{[n]}(\mathbf{\Pi}_m)$, or equivalently $\mathbf{\Pi}_n\big|_{[m]} = \mathbf{\Pi}_m$ almost surely.*

Definition 2.10 (Projective distributions). *Let π_n be a probability measure on $(\mathcal{P}_{[n]}, \mathscr{B}_{\mathcal{P}_{[n]}})$ for every $n \geq 1$. We say that $(\pi_n)_{n\geq 1}$ is a projective family of distributions if for every $m \leq n$ and every $A \in \mathcal{P}_{[m]}$*

$$\pi_m(\{A\}) = \sum_{B \in \mathcal{P}_{[n]}(A)} \pi_n(\{B\})$$

Definition 2.11 (Consistent in distribution random partitions). *Let $\mathbf{\Pi}_n$ be a random partition of $[n]$ defined on the probability space $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$. We say that $(\mathbf{\Pi}_n)_{n\geq 1}$ is consistent in distribution, if for every $m \leq n$ and $A \in \mathcal{P}_{[m]}$*

$$\mathbb{P}_m[\mathbf{\Pi}_m = A] = \sum_{B \in \mathcal{P}_{[n]}(A)} \mathbb{P}_n[\mathbf{\Pi}_n = B]. \tag{2.5}$$

Or equivalently, $(\pi_n)_{n\geq 1}$ defines a projective family, where π_n denotes the distribution of $\mathbf{\Pi}_n$.

Remark 2.1. *It is easy to see that equation 2.5 is equivalent to*

$$\mathbb{P}_n[\mathbf{\Pi}_n = A] = \sum_{B \in \mathcal{P}_{[n+1]}(A)} \mathbb{P}_{n+1}[\mathbf{\Pi}_{n+1} = B]$$

for every $n \in \mathbb{N}$ and $A \in \mathcal{P}_{[n]}$. In particular, if $(\Omega_n, \mathbf{F}_n, \mathbb{P}_n) = (\Omega, \mathbf{F}, \mathbb{P})$ for every $n \in \mathbb{N}$, the above reduces to

$$\mathbb{P}[\mathbf{\Pi}_n = A] = \sum_{B \in \mathcal{P}_{[n+1]}(A)} \mathbb{P}[\mathbf{\Pi}_{n+1} = B].$$

Another trivial remark concerning the aforementioned definitions is that if $(\mathbf{\Pi}_n)_{n\geq 1}$ is a (strongly) consistent family of random partitions then it is also consistent in distribution. In fact, for every random partition $\mathbf{\Pi}_n$ of $[n]$, and each $1 \leq m \leq n$, we have that the distribution of $\mathbf{\Pi}_m = \mathbf{\Pi}_n\big|_{[m]}$ is completely characterized by that of $\mathbf{\Pi}_n$ through

$$\mathbb{P}[\mathbf{\Pi}_m = A] = \sum_{B\in\mathcal{P}_{[n]}(A)} \mathbb{P}[\mathbf{\Pi}_n = B].$$

for every partition, A, of $[m]$. The converse is clearly false.

2.2.3 Exchangeable partitions of $[n]$

Definition 2.12 (Exchangeable random partition). *Let $\mathbf{\Pi}_n$ be a random partition of $[n]$, we say that it is exchangeable if, for every permutation, σ, of $[n]$, and every $A \in \mathcal{P}_{[n]}$,*

$$\mathbb{P}[\mathbf{\Pi}_n = A] = \mathbb{P}[\mathbf{\Pi}_n = \sigma(A)]$$

where $\sigma(A)$ denotes the σ-internal permutation of A, equivalently $\mathbf{\Pi}_n \overset{d}{=} \sigma(\mathbf{\Pi}_n)$.

Proposition 2.6. *Let $\mathbf{\Pi}_n$ be an exchangeable random partition of $[n]$, for some $n \geq 2$. Then, for every $m < n$, $\mathbf{\Pi}_m = \mathbf{\Pi}_n\big|_{[m]}$ is an exchangeable partition of $[m]$.*

See Appendix B.3 for a proof. Roughly speaking, $\mathbf{\Pi}_n$ being exchangeable means that the probability of the event $\{\mathbf{\Pi}_n = A\}$ does not depends of which elements of $[n]$ are grouped together, at most it depends on the number of blocks of A and the frequency of each block. This leads to the following equivalent definition.

Definition 2.13 (Exchangeable random partition/ EPPF). *Let $\mathbf{\Pi}_n$ be a random partition of $[n]$, we say that it is exchangeable if for every $A = \{A_1, \ldots, A_k\} \in \mathcal{P}_{[n]}$,*

$$\mathbb{P}[\mathbf{\Pi}_n = A] = \pi_n(|A_1|, \ldots, |A_k|)$$

for some symmetric function of its arguments, $\pi_n : \mathcal{C}_n \to [0, 1]$. In this case π_n is called exchangeable partition probability function or EPPF for short.

Note that for a fixed composition of n into k parts, $(n_1, \ldots, n_k) \in \mathcal{C}_n^k$, there are exactly $n!/(k!\prod_{i=1}^k n_i!)$ partitions of $[n]$ with k blocks and with block sizes given by $\{n_1, \ldots, n_k\}$. Insomuch as for every random partition of $[n]$ we have that

$$\sum_{A\in\mathcal{P}_{[n]}} \mathbb{P}[\mathbf{\Pi}_n = A] = 1,$$

and since the any EPPF π_n, is symmetric, we get that

$$\sum_{(n_1,\ldots,n_k)} \frac{n!}{k!\prod_{i=1}^k n_i!}\pi_n(n_1, \ldots, n_k) = 1 \tag{2.6}$$

(where the sum ranges over the set of compositions of n, $\mathcal{C}_n$). In fact the most general EPPF of $[n]$ is any positive integer-valued symmetric function such that (2.6) holds.

As one can intuit from the definition of an exchangeable partition, the block sizes and the number of blocks are of primary interest. Next we describe some ways to encode the block frequencies as a random composition of n.

Definition 2.14 (Orderings of the blocks)**.** *Let $\mathbf{\Pi}_n = \{\mathbf{\Pi}_{n,1}, \ldots, \mathbf{\Pi}_{n,\mathbf{K}_n}\}$ be a random partition of $[n]$ having $\mathbf{K}_n$ blocks with corresponding frequencies $\{\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}_n}\}$, so that $|\mathbf{\Pi}_{n,j}| = \mathbf{n}_j$.*

a) We say that $\tilde{\mathbf{\Pi}}_n = \left(\tilde{\mathbf{\Pi}}_{n,1}, \ldots, \tilde{\mathbf{\Pi}}_{n,\mathbf{K}_n}\right)$ is the ordering of the blocks of $\mathbf{\Pi}_n$ according to the least element, or in order of appearance, if $\tilde{\mathbf{\Pi}}_{n,1}$ is the block containing 1, under the event $\mathcal{A}_1 = [n] \setminus \tilde{\mathbf{\Pi}}_{n,1} \neq \emptyset$, $\tilde{\mathbf{\Pi}}_{n,2}$ is the block of $\mathbf{\Pi}_n$ that contains $\min \mathcal{A}_1$, if $\mathcal{A}_2 = \mathcal{A}_1 \setminus \tilde{\mathbf{\Pi}}_{n,2} \neq \emptyset$, $\tilde{\mathbf{\Pi}}_{n,3}$ is the block of $\mathbf{\Pi}_n$ that contains $\min \mathcal{A}_2$, and so on. To the frequencies of the blocks in order of appearance, we denote by $\tilde{\mathbf{n}}_j = |\tilde{\mathbf{\Pi}}_{n,j}|$, and we write $\tilde{\mathbf{N}} = (\tilde{\mathbf{n}}_1, \ldots, \tilde{\mathbf{n}}_{\mathbf{K}_n})$ for the corresponding composition of n.

b) We say that $\mathbf{\Pi}^{\downarrow}_n = \left(\mathbf{\Pi}^{\downarrow}_{n,1}, \ldots, \mathbf{\Pi}^{\downarrow}_{n,\mathbf{K}_n}\right)$ is a decreasing ordering of $\mathbf{\Pi}_n$, if $\mathbf{n}^{\downarrow}_j = |\mathbf{\Pi}^{\downarrow}_{n,j}| \geq |\mathbf{\Pi}^{\downarrow}_{n,j+1}| = \mathbf{n}^{\downarrow}_{j+1}$, for every $j \geq 1$ such that $\mathbf{\Pi}^{\downarrow}_{n,j+1} \neq \emptyset$. To the entries of the composition of n, $\mathbf{N}^{\downarrow} = \left(\mathbf{n}^{\downarrow}_1, \ldots, \mathbf{n}^{\downarrow}_{\mathbf{K}_n}\right)$, we call ranked frequencies of $\mathbf{\Pi}_n$.

c) We say that $\mathbf{\Pi}^{\mathrm{ex}}_n = \left(\mathbf{\Pi}^{\mathrm{ex}}_{n,1}, \ldots, \mathbf{\Pi}^{\mathrm{ex}}_{n,\mathbf{K}_n}\right)$ is in exchangeable random order, if the order blocks, $\mathbf{\Pi}^{\mathrm{ex}}_{n,1}, \ldots, \mathbf{\Pi}^{\mathrm{ex}}_{n,\mathbf{K}_n}$, were obtained by uniformly permuting the blocks of $\mathbf{\Pi}_n$. To the corresponding frequencies of $\mathbf{\Pi}^{\mathrm{ex}}_n$, we denote by $\mathbf{N}^{\mathrm{ex}} = \left(\mathbf{n}^{\mathrm{ex}}_1, \ldots, \mathbf{n}^{\mathrm{ex}}_{\mathbf{K}_n}\right)$.

Proposition 2.7. *Let $\mathbf{\Pi}_n$ be an exchangeable partition of $[n]$ with EPPF π_n. Let $\mathbf{K}_n$ be the random number of blocks of $\mathbf{\Pi}_n$, $\mathbf{N}_n = \{\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}_n}\}$ be the unordered composition of n induced by $\mathbf{\Pi}_n$ and let $\mathbf{M}_n = (\mathbf{m}_1, \ldots, \mathbf{m}_n)$ be given by $\mathbf{m}_i = \sum_{j=1}^{\mathbf{K}_n} \mathbf{1}_{\{\mathbf{n}_j = i\}}$, that is $\mathbf{m}_j$ counts the number of blocks of $\mathbf{\Pi}_n$ with exactly i elements. Also let $\tilde{\mathbf{N}} = (\tilde{\mathbf{n}}_1, \ldots, \tilde{\mathbf{n}}_{\mathbf{K}_n})$, $\mathbf{N}^{\downarrow} = \left(\mathbf{n}^{\downarrow}_1, \ldots, \mathbf{n}^{\downarrow}_{\mathbf{K}_n}\right)$ and $\mathbf{N}^{\mathrm{ex}} = \left(\mathbf{n}^{\mathrm{ex}}_1, \ldots, \mathbf{n}^{\mathrm{ex}}_{\mathbf{K}_n}\right)$ be as in Definition 2.14. Then, for any positive integers $n_1, \ldots, n_k$ such that $\sum_{j=1}^k n_j = n$, and where $m_i = \sum_{j=1}^k \mathbf{1}_{\{n_i = j\}}$ we have that*

$$\mathbb{P}[\mathbf{M}_n = (m_1, \ldots, m_n)] = \mathbb{P}[\mathbf{N}_n = \{n_1, \ldots, n_k\}] = \frac{n!}{\prod_{i=1}^n (i!)^{m_i}(m_i!)} \pi_n(n_1, \ldots, n_k),$$

$$\mathbb{P}\left[\tilde{\mathbf{N}}_n = (n_1, \ldots, n_k)\right] = \frac{n!}{\prod_{j=1}^k \left(\sum_{i \geq j} n_i\right)(n_j - 1)!} \pi_n(n_1, \ldots, n_k),$$

$$\mathbb{P}[\mathbf{N}^{\downarrow}_n = (n_1, \ldots, n_k)] = \frac{n!}{\prod_{i=1}^n (i!)^{m_i}(m_i!)} \pi_n(n_1, \ldots, n_k) \mathbf{1}_{\{n_1 \geq n_2 \geq \cdots \geq n_k\}},$$

and

$$\mathbb{P}[\mathbf{N}^{\mathrm{ex}}_n = (n_1, \ldots, n_k)] = \frac{n!}{k! \prod_{j=1}^k n_j!} \pi_n(n_1, \ldots, n_k).$$

See Appendix B.4 for a proof of Proposition 2.7. This result shows how to compute the mass probability function of certain compositions of n induced by an exchangeable partition of $[n]$ in terms of the corresponding EPPF. The following Proposition, conversely, explains how to derive the EPPF if the distribution of a composition of n is available.

Proposition 2.8. *Let $\mathbf{\Pi}_n$ be an exchangeable partition of $[n]$, say that $(\mathbf{\Pi}_{n,1},\ldots,\mathbf{\Pi}_{n,\mathbf{K}_n})$ is an arbitrary ordering of the blocks of $\mathbf{\Pi}_n$, with corresponding frequencies $\mathbf{N}_n = (\mathbf{n}_1,\ldots,\mathbf{n}_{K_n})$, and that*

$$\mathbb{P}[\mathbf{N}_n = (n_1,\ldots,n_k)] = \pi_n^*(n_1,\ldots,n_k).$$

Then, the EPPF of $\mathbf{\Pi}_n$ is given by

$$\pi_n(n_1,\ldots,n_k) = \frac{\prod_{j=1}^k n_j!}{n!}\sum_{\sigma}\pi_n^*(n_{\sigma(1)},\ldots,n_{\sigma(k)}) \tag{2.7}$$

where the sum ranges over all $k!$ permutations of $[k]$.

See Appendix B.5 for a proof.

Corollary 2.9. *Let $\mathbf{\Pi}_n$ be an exchangeable partition of $[n]$ and $\mathbf{N}^{\downarrow} = (\mathbf{n}_1^{\downarrow},\ldots,\mathbf{n}_{\mathbf{K}_n}^{\downarrow})$ its ranked frequencies of the blocks. Then, the conditional law of $\mathbf{\Pi}_n$ given $\mathbf{N}^{\downarrow}$, is that of, $\mathbf{\Pi}(\mathbf{x}_{1:n}) = \mathbf{\Pi}(\mathbf{x}_1,\ldots,\mathbf{x}_n)$, the partition of $[n]$ generated by the random equivalence relation $i \sim j$ if and only if $\mathbf{x}_i = \mathbf{x}_j$, where conditionally given $\mathbf{N}^{\downarrow}$, $\mathbf{x}_1,\ldots,\mathbf{x}_n$ were sampled without replacement from a set $\{x_1,x_2,\ldots\}$ with $\mathbf{n}_j^{\downarrow}$ values equal to j, i.e. $|\{i : x_i = j\}| = \mathbf{n}_j^{\downarrow}$.*

The proof of Corollary 2.9 appears in Appendix B.6. Our next aim is to define and study exchangeable partitions of $\mathbb{N}$, in particular we are interested in deriving an analogous result (Kingman's representation theorem) to that of Corollary 2.9 , for the infinite case.

2.2.4 Exchangeable partitions of $\mathbb{N}$

Definition 2.15 (Exchangeable partition of $\mathbb{N}$ / infinite EPPF)**.**

i) By an exchangeable partition of $\mathbb{N}$ we mean a consistent family, $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\in\mathbb{N}}$, of exchangeable partitions. Equivalently, we say that $\mathbf{\Pi}$ is an exchangeable partition of $\mathbb{N}$ is $\mathbf{\Pi} \overset{d}{=} \sigma(\mathbf{\Pi})$, for every permutation, σ, of $\mathbb{N}$, and where $\sigma(\mathbf{\Pi})$ denotes the σ-internal permutation of $\mathbf{\Pi}$.

ii) We say $\pi : \bigcup_{k\in\mathbb{N}}\mathbb{N}^k \to [0,1]$, is an infinite exchangeable partition probability function (EPPF) if there exist an exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\in\mathbb{N}}$, such that

$$\mathbb{P}[\mathbf{\Pi}_n = A] = \pi(|A_1|,\ldots,|A_k|)$$

for every $n \in \mathbb{N}$ and any partition A of $[n]$.

If π is an infinite EPPF, for every sequence of positive integers $(n_1,\ldots,n_k)$ and any permutation, σ, of $[k]$, $\pi(n_1,\ldots,n_k) = \pi\left(n_{\sigma(1)},\ldots,n_{\sigma(k)}\right)$, that is, π is symmetric, and for every $n \in \mathbb{N}$

$$\sum_{(n_1,\ldots,n_k)\in\mathcal{C}_n}\frac{n!}{k!\prod_{i=1}^k n_i!}\pi(n_1,\ldots,n_k) = 1. \tag{2.8}$$

In other words, the restriction of π to the set $\mathcal{C}_n = \left\{(n_1,\ldots,n_k) : \sum_{j=1}^k n_k = n\right\}$ is a (finite) EPPF of $[n]$.

The fact that the laws of a partition of $\mathbb{N}$ forms a projective family, translates to the infinite EPPF as the the so-called addition rule:

$$\pi(n_1,\ldots,n_k) = \pi(n_1,\ldots,n_k,1) + \sum_{j=1}^{k} \pi(n_1,\ldots,n_{j-1},n_j+1,n_{j+1},\ldots,n_k). \tag{2.9}$$

Note that if $\pi : \bigcup_{k\in\mathbb{N}} \mathbb{N}^k \to [0,1]$, is an arbitrary symmetric function of its arguments, satisfying (2.9), then using the symmetry of π and rearranging the terms of the sum we get

$$\sum_{(n_1,\ldots,n_k)\in\mathcal{C}_n} \frac{n!}{k!\prod_{i=1}^{k} n_i!}\pi(n_1,\ldots,n_k) = \sum_{(n_1,\ldots,n_k)\in\mathcal{C}_{n+1}} \frac{(n+1)!}{k!\prod_{i=1}^{k} n_i!}\pi(n_1,\ldots,n_k)$$

which means that if $\pi(1) = 1$, then (2.8) also holds for π. This leads to the following equivalent definition of an infinite EPPF.

Definition 2.16 (Infinite EPPF)**.** *We say $\pi : \bigcup_{k\in\mathbb{N}} \mathbb{N}^k \to [0,1]$, is an infinite exchangeable partition probability function (EPPF) if it is a symmetric function, $\pi(1) = 1$, and for every sequence of positive integers $(n_1,\ldots,n_k)$*

$$\pi(n_1,...n_k) = \pi(n_1,...,n_k,1) + \sum_{j=1}^{k} \pi(n_1,...,n_{j-1},n_j+1,n_{j+1},...,n_k).$$

Evidently, for every exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\in\mathbb{N}}$, its law is completely characterized by an infinite EPPF, π. And conversely, by Kolmogorov's consistency theorem, for each infinite EPPF, π, there exist an exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\in\mathbb{N}}$, whose law is described by π.

In terms of the EPPF, it is also easy to derive a prediction rule for the consistent family of exchangeable random partitions, $(\mathbf{\Pi}_n)_{n\geq 1}$, as follows. Let $A = \{A_1,\ldots,A_k\} \in \mathcal{P}_{[n]}$, with $n_j = |A_j|$, and consider $A^{(j)} = \{A_i\}_{i\neq j} \cup \{A_j \cup \{n+1\}\}$, for $j \in [k]$ and $A^{(k+1)} = \{A_1,\ldots,A_k,\{n+1\}\}$. Then, as A is the only partition on $\mathcal{P}_{[n]}$ such that $A^{(j)}\big|_{[n]} = A$, for each $j \in [k+1]$, we get

$$\begin{aligned} \mathbb{P}[\mathbf{\Pi}_{n+1} = A^{(k+1)}|\mathbf{\Pi}_n = A] &= \frac{\mathbb{P}[\mathbf{\Pi}_{n+1} = A^{(k+1)}]}{\mathbb{P}[\mathbf{\Pi}_n = A]} = \frac{\pi(n_1,\ldots,n_k,1)}{\pi(n_1,\ldots,n_k)}, \text{ and} \\ \mathbb{P}[\mathbf{\Pi}_{n+1} = A^{(j)}|\mathbf{\Pi}_n = A] &= \frac{\pi(n_1,\ldots,n_{j-1},n_j+1,n_{j+1},\ldots,n_k)}{\pi(n_1,\ldots,n_k)}, \end{aligned} \tag{2.10}$$

for $j = [k]$. To the collection of numbers $\{\pi(j \mid n_1,\ldots,n_k)\}_{(n_1,\ldots,n_k)\in\mathcal{C}_n, j\in[k+1]}$, given by

$$\begin{aligned} \pi(j \mid n_1,\ldots,n_k) &= \frac{\pi(n_1,\ldots,n_{j-1},n_j+1,n_{j+1},\ldots,n_k)}{\pi(n_1,\ldots,n_k)}, \\ \pi(k+1 \mid n_1,\ldots,n_k) &= \frac{\pi(n_1,\ldots,n_k,1)}{\pi(n_1,\ldots,n_k)}, \end{aligned} \tag{2.11}$$

for $j \in [k]$, we call the prediction rule for the infinite EPPF π. Clearly an infinite EPPF and its prediction rule characterize each other completely.

2.2.5 Random partitions of $\mathbb{N}$ generated by sampling without replacement

Let $\mathbf{X} = (\mathbf{x}_i)_{i\in\mathbb{N}}$ be a sequence of random elements taking values in some Borel space $(S, \mathscr{B}_S)$. And let $\mathbf{\Pi}(\mathbf{X}) = \mathbf{\Pi}(\mathbf{x}_{1:\infty}) = (\mathbf{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$, be the random partition of $\mathbb{N}$ generated by the random equivalence relation $i \sim j$ if and only if $\mathbf{x}_i = \mathbf{x}_j$. So for example, say that for some realization $\mathbf{x}_1 = a, \mathbf{x}_2, = a, \mathbf{x}_3 = b, \mathbf{x}_4 = c$ and $\mathbf{x}_5 = b$, then the under such event $\mathbf{\Pi}(\mathbf{x}_{1:5}) = \{\{1,2\},\{3,5\},\{4\}\}$. Note that this transformation is measurable since the diagonal part $\{(s_1, s_2) \in S^2 : s_1 = s_2\}$, of S^2 is. By definition it is obvious that for $1 \leq m \leq n \leq \infty$, $\mathbf{\Pi}(\mathbf{x}_{1:n})\big|_{[m]} = \mathbf{\Pi}(\mathbf{x}_{1:m})$. Indeed, for $n \geq 1$, $\mathbf{x}_{n+1}$ necessarily takes a value already observed in $\mathbf{x}_1, \ldots, \mathbf{x}_n$, in which case $n+1$ will be added to an existing block of $\mathbf{\Pi}(\mathbf{x}_{1:n})$, or $\mathbf{x}_{n+1}$ is distinct from $\mathbf{x}_1, \ldots, \mathbf{x}_n$, in which case $\{n+1\}$ will be added as a new, additional block to $\mathbf{\Pi}(\mathbf{x}_{1:n})$, either way $\mathbf{\Pi}(\mathbf{x}_{1:(n+1)}) \in \mathcal{P}_{[n+1]}(\mathbf{\Pi}(\mathbf{x}_{1:n}))$, almost surely. Moreover, if $\mathbf{X}$ is exchangeable then for every permutation σ of $\mathbb{N}$, we have that $\mathbf{X} \overset{d}{=} \sigma(\mathbf{X})$, hence $\mathbf{\Pi}(\mathbf{X}) \overset{d}{=} \mathbf{\Pi}(\sigma(\mathbf{X}))$, which means that for every $n \geq 1$, the distribution of $\mathbf{\Pi}(\mathbf{x}_{1:n})$ is invariant under internal permutations. That is, $\mathbf{\Pi}(\mathbf{x}_{1:n})$ is exchangeable. Thus, by sequentially sampling from exchangeable sequences, we can generate exchangeable partitions of $\mathbb{N}$. Theorem 2.10 shows that every exchangeable partition of $\mathbb{N}$ can be generated this way.

Theorem 2.10 (Kingman's representation theorem)**.** *Let $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\geq 1}$ be an exchangeable random partition of $\mathbb{N}$. For each $n \in \mathbb{N}$, let $\mathbf{K}_n$ be the number of blocks of $\mathbf{\Pi}_n$, let $\left(\mathbf{n}^{\downarrow}_{n,1}, \ldots, \mathbf{n}^{\downarrow}_{n,\mathbf{K}_n}\right)$ be ranked frequencies of $\mathbf{\Pi}_n$ and set $\mathbf{n}^{\downarrow}_j = 0$ for $j \geq \mathbf{K}_n$. Then for each $j \geq 1$, $\mathbf{n}^{\downarrow}_{n,j}/n$ has an almost sure limit $\mathbf{w}^{\downarrow}_j$, as $n \to \infty$. Moreover the conditional distribution of $\mathbf{\Pi}$ given $(\mathbf{w}^{\downarrow}_1, \mathbf{w}^{\downarrow}_2, \ldots)$ is that of $\mathbf{\Pi}(\mathbf{x}_{1:\infty})$ where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{iid}{\sim} \boldsymbol{\mu}$ for some random probability measure, $\boldsymbol{\mu}$, with ranked sizes of atoms $(\mathbf{w}^{\downarrow}_1, \mathbf{w}^{\downarrow}_2, \ldots)$.*

Last theorem, whose proof can be found in Appendix B.7, sets up a bijection between probability distributions of infinite exchangeable random partitions, as specified by an infinite EPPF, and probability distributions of $(\mathbf{w}^{\downarrow}_1, \mathbf{w}^{\downarrow}_2, \ldots)$ on the set

$$\overline{\Delta}^{\downarrow}_{\infty} := \left\{ (w_1, w_2, \ldots) : w_1 \geq w_2 \geq \ldots \geq 0, \sum_{i=1}^{\infty} w_i \leq 1 \right\} \tag{2.12}$$

This bijection is known in literature as Kingman's correspondence or Kingman's bijection, explicitly:

a) Given a probability distribution P over $\left(\overline{\Delta}^{\downarrow}_{\infty}, \mathscr{B}_{\overline{\Delta}^{\downarrow}_{\infty}}\right)$ we can choose a random element $(\mathbf{w}^{\downarrow}_1, \mathbf{w}^{\downarrow}_2, \ldots)$ with distribution P, and let $\boldsymbol{\mu}$ be any random probability measure over any Borel space $(S, \mathscr{B}_S)$ taking the form

$$\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}^{\downarrow}_j \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}^{\downarrow}_j\right) \mu_0$$

where $(\boldsymbol{\xi}_j)_{j\geq 1}$ are deterministic or random distinct elements of S, and μ_0 is a diffuse probability measure. Now, let $\mathbf{X} = (\mathbf{x}_k)_{k\geq 1}$ be a sequence of exchangeable random variables with directing random measure $\boldsymbol{\mu}$, and define the exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi} = \mathbf{\Pi}(\mathbf{x}_{1:\infty})$. As $\mathbf{\Pi}$ is exchangeable, it has an (infinite) EPPF, π, which corresponds to P.

b) Conversely, consider an (infinite) EPPF π and let $\mathbf{\Pi}$ be some exchangeable partition of $\mathbb{N}$, whose distribution is characterized by such EPPF. Kingman's representation theorem establishes (following the notation of the statement of this theorem) that the almost sure limits,

$$\lim_{n\to\infty} \frac{\mathbf{n}^{\downarrow}_{n,j}}{n} = \mathbf{w}^{\downarrow}_{j},$$

exist and that the limiting random variables $(\mathbf{w}^{\downarrow}_1, \mathbf{w}^{\downarrow}_2, \ldots)$ are such that $(\mathbf{w}^{\downarrow}_j)_{j\geq 1} \in \overline{\Delta}^{\downarrow}_{\infty}$ almost surely, hence its distribution is some probability measure over $\left(\overline{\Delta}^{\downarrow}_{\infty}, \mathscr{B}_{\overline{\Delta}^{\downarrow}_{\infty}}\right)$, P corresponding to π.

Definition 2.17 (Orderings of weights)**.** *Let $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$ be a sequence of non-negative random variables such that $0 < \sum_{j\geq 1} \mathbf{w}_j \leq 1$ almost surely. To the elements of $\mathbf{W}$ we call weights.*

i) We say $\mathbf{W}^{\downarrow} = (\mathbf{w}^{\downarrow}_1, \mathbf{w}^{\downarrow}_2, \ldots)$ is the decreasing permutation of $\mathbf{W}$ if $\mathbf{W}^{\downarrow}$ is a permutation of $\mathbf{W}$ and $\mathbf{w}^{\downarrow}_1 \geq \mathbf{w}^{\downarrow}_2, \geq \cdots$.

ii) We call $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots)$, a size-biased permutation of $\mathbf{W}$, or say that it is invariant under size-biased permutations, if $\tilde{\mathbf{W}}$ is obtained by sampling without replacement from $\mathbf{W}$, with probabilities proportional to $\mathbf{W}$. That is

$$\mathbb{P}\left[\tilde{\mathbf{w}}_1 \in \cdot \mid \mathbf{W}\right] = \sum_{j\geq 1} \frac{\mathbf{w}_j}{\sum_{k\geq 1} \mathbf{w}_k} \delta_{\mathbf{w}_j},$$

and for $i \geq 1$,

$$\mathbb{P}\left[\tilde{\mathbf{w}}_{i+1} \in \cdot \mid \mathbf{W}, \tilde{\mathbf{w}}_1, \ldots \tilde{\mathbf{w}}_i\right] = \frac{\sum_{j\geq 1} \mathbf{w}_j \delta_{\mathbf{w}_j} - \sum_{j=1}^{i} \tilde{\mathbf{w}}_j \delta_{\tilde{\mathbf{w}}_j}}{\left(\sum_{k\geq 1} \mathbf{w}_k - \sum_{l=1}^{i} \tilde{\mathbf{w}}_l\right)},$$

if $\left(\sum_{k\geq 1} \mathbf{w}_k - \sum_{l=1}^{i} \tilde{\mathbf{w}}_l\right) > 0$ (so that there exist $\mathbf{w}_j > 0$ such that $\mathbf{w}_j \neq \tilde{\mathbf{w}}_l$ for every $l \in [i]$), and $\mathbb{P}\left[\tilde{\mathbf{w}}_{i+1} \in \cdot \mid \mathbf{W}, \tilde{\mathbf{w}}_1, \ldots \tilde{\mathbf{w}}_i\right] = \delta_0$, otherwise (so that for every $\mathbf{w}_j$ satisfying $\mathbf{w}_j \neq \tilde{\mathbf{w}}_l$ for every $l \in [i]$, we get $\mathbf{w}_j = 0$).

iii) We call $\tilde{\mathbf{W}}' = (\tilde{\mathbf{w}}'_1, \tilde{\mathbf{w}}'_2, \ldots)$, a size-biased pseudo-permutation of $\mathbf{W}$ if

$$\mathbb{P}\left[\tilde{\mathbf{w}}'_1 \in \cdot \mid \mathbf{W}\right] = \sum_{j\geq 1} \mathbf{w}_j \delta_{\mathbf{w}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \delta_0,$$

and for every $i \geq 1$,

$$\mathbb{P}\left[\mathbf{w}_{i+1} \in \cdot \mid \mathbf{W}, \tilde{\mathbf{w}}'_1, \ldots, \tilde{\mathbf{w}}'_i\right] = \frac{\sum_{j\geq 1} \mathbf{w}_j \delta_{\mathbf{w}_j} - \sum_{j=1}^{i} \tilde{\mathbf{w}}'_j \delta_{\tilde{\mathbf{w}}'_j} + \left(1 - \sum_{k\geq 1} \mathbf{w}_k\right) \delta_0}{\left(1 - \sum_{l=1}^{i} \tilde{\mathbf{w}}'_l\right)}$$

if $\left(1 - \sum_{l=1}^{i} \tilde{\mathbf{w}}'_l\right) > 0$, and $\mathbb{P}\left[\tilde{\mathbf{w}}_{i+1} \in \cdot \mid \mathbf{W}, \tilde{\mathbf{w}}_1, \ldots \tilde{\mathbf{w}}_i\right] = \delta_0$, otherwise.

Remark 2.2. *To understand the difference between size-biased permutations and size-biased pseudo-permutations note that a size-biased permutation, $\tilde{\mathbf{W}}$, of $\mathbf{W}$ simply permutes the elements of $\mathbf{W}$ in such way that larger weights tend to appear before smaller weights. For a size-biased pseudo-permutation, $\tilde{\mathbf{W}}'$, of $\mathbf{W}$ it is still true that larger weights tend to appear before smaller weights, however, this second re-ordering allows the possibility of zeros appearing in the sequence even if $\mathbf{w}_j > 0$ for every $j \geq 1$. For example, consider the deterministic weights sequence $\mathbf{W} = (1/4, 1/8, 1/16, \ldots)$, so that $\mathbf{w}_j = (1/2)^{j+1}$, and $\sum_{j\geq 1} \mathbf{w}_j = 1/2$. For this weights sequence, $(0, 0, 1/8, 1/64, 0, 1/4, 1/128, \ldots)$ could be a realization of $\tilde{\mathbf{W}}'$ but not of $\tilde{\mathbf{W}}$, removing the zeros from $\tilde{\mathbf{W}}'$, say $(1/8, 1/64, 1/4, 1/128, \ldots)$, is a realization of $\tilde{\mathbf{W}}$. In general $\tilde{\mathbf{W}}$ has the same number of zeros than $\mathbf{W}$ does and they always appear at the end of the sequence. In contrast, if $\sum_{j\geq 1} \mathbf{w}_j < 1$, $\tilde{\mathbf{W}}'$, has infinitely many zeros intercalated with the elements of $\mathbf{W}$. For a second example say that $\mathbf{W} = (1/4, 1/4, 0, 0, \ldots)$, so that $\sum_{j\geq 1} \mathbf{w}_j = 1/2$. For this weights sequence the only possible realization of a size-biased permutation is $\tilde{\mathbf{W}} = \mathbf{W} = (1/4, 1/4, 0, 0, \ldots)$, alternatively, there are infinitely many possible realizations of $\tilde{\mathbf{W}}'$, for example $(0, 1/4, 0, 1/4, 0, \ldots)$ and $(1/4, 0, 0, 0, 1/4, 0, \ldots)$ are two distinct possible realizations of $\tilde{\mathbf{W}}'$. In fact,*

$$\mathbb{P}\left[\tilde{\mathbf{w}}_1' = 0\right] = \frac{1}{2} = \mathbb{P}\left[\tilde{\mathbf{w}}_1' = 1/4\right].$$

Now, if the sequence $\mathbf{W}$ satisfies $\sum_{j\geq 1} \mathbf{w}_j = 1$ almost surely, then it is straight forward from Definition 2.17 that a size-biased pseudo-permutation and a size-biased permutation are exactly the same.

Remark 2.3. *Although this case is absolutely not interesting, for the sequence $\mathbf{W} = (0, 0, \ldots)$ we say that $\mathbf{W}$ is the decreasing reordering of itself, and a size-biased permutation and pseudo-permutation of itself. This vacuous remark is to cover all possible cases in the following proposition.*

Proposition 2.11. *Let $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\geq 1}$ be an exchangeable random partition of $\mathbb{N}$. For each $n \in \mathbb{N}$, let $\left(\tilde{\mathbf{\Pi}}_{n,1}, \ldots, \tilde{\mathbf{\Pi}}_{n,\mathbf{K}_n}\right)$ be the ordering of the blocks of $\mathbf{\Pi}_n$ according to the least element, with corresponding block sizes $(\tilde{\mathbf{n}}_{n,1}, \ldots, \tilde{\mathbf{n}}_{n,\mathbf{K}_n})$ and consider the ranked frequencies $\left(\mathbf{n}^{\downarrow}_{n,1}, \ldots, \mathbf{n}^{\downarrow}_{n,\mathbf{K}_n}\right)$, set $\mathbf{n}^{\downarrow}_{n,j} = \tilde{\mathbf{n}}_{n,j} = 0$, for all $j > \mathbf{K}_n$. Define $\mathbf{w}^{\downarrow}_j = \lim_{n\to\infty} \mathbf{n}^{\downarrow}_{n,j}/n$. Then*

$$\lim_{n\to\infty} \frac{\tilde{\mathbf{n}}_{n,j}}{n} \to \tilde{\mathbf{w}}_j,$$

almost surely, for every $j \geq 1$, and where $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots)$ is a size-biased pseudo-permutation of $\mathbf{W}^{\downarrow} = (\mathbf{w}^{\downarrow}_1, \mathbf{w}^{\downarrow}_2, \ldots)$. In particular, if $\sum_{j\geq 1} \mathbf{w}^{\downarrow}_j = 1$ almost surely, $\tilde{\mathbf{W}}$ is invariant under size-biased permutations.

The proof of Proposition 2.11 appears in Appendix B.8. For a distribution P over $(\overline{\Delta}^{\downarrow}_{\infty}, \mathscr{B}_{\overline{\Delta}^{\downarrow}_{\infty}})$ we can recognize two important mutually exclusive cases.

- *Proper case:* This scenario happens when $\mathbf{W}^{\downarrow} = (\mathbf{w}^{\downarrow}_j)_{j\geq 1} \sim \mathsf{P}$, satisfies $\sum_{j\geq 1} \mathbf{w}^{\downarrow}_j = 1$ almost surely. In this case the corresponding exchangeable partition of $\mathbb{N}$ and its EPPF are also termed proper. By Kingman's representation theorem it is clear that for a proper exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi}$, every $i \in \mathbb{N}$ will belong to a block containing infinitely many elements almost surely. For proper EPPF's Kingman's correspondence can be made more explicit by noting that for any partition $A = \{A_1, \ldots, A_k\}$ of $[n]$,

$$\mathbb{P}[\mathbf{\Pi}_n = A \mid \mathbf{W}^{\downarrow}] = \sum_{(j_1,\ldots,j_k)} \prod_{i=1}^{k} \left(\mathbf{w}_{j_i}^{\downarrow}\right)^{|A_i|},$$

where the sum ranges over all k-tuples of distinct positive integers. Thus by the tower property and monotone convergence theorem we get

$$\pi(n_1, \ldots, n_k) = \sum_{(j_1,\ldots,j_k)} \mathbb{E}\left[\prod_{i=1}^{k} \left(\mathbf{w}_{j_i}^{\downarrow}\right)^{n_i}\right].$$

for any positive integers $n_1, \ldots, n_k$.

- *Improper case:* This case takes place when $\mathbf{W}^{\downarrow} = (\mathbf{w}_j^{\downarrow})_{j\geq 1} \sim \mathsf{P}$, satisfies

$$\mathbb{P}\left[\sum_{j\geq 1} \mathbf{w}_j^{\downarrow} < 1\right] > 0$$

In this instance we also say that the corresponding exchangeable partition of $\mathbb{N}$ and its EPPF are improper. For such a partition $\mathbf{\Pi}$, from Kingman's representation theorem, we get that for every $i \in \mathbb{N}$, i will either belong to a block of $\mathbf{\Pi}$ containing infinitely many elements, or will contribute to the partition as a singleton, almost surely.

2.2.6 The ordered paintbox

Here we provide a construction that might help us understand the connection between exchangeable partitions, sequences of exchangeable random variables and completely random measures, particularly subordinators. This construction is just an alternative way of looking at Kingsman's correspondence.

Let $\mathcal{R}$ be a random closed subset of $[0, 1]$ such that the open complement $\mathcal{R}^c := [0, 1] \setminus \mathcal{R}$ has a canonical representation as a disjoint union of countably many open interval components, which we are going to call gaps of $\mathcal{R}$. That is $\mathcal{R}^c = \bigcup_{j=1}^{\infty} \mathcal{R}_j$ where $\mathcal{R}_j$ is an open interval of $[0, 1]$, and $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$ for every $i \neq j$. If this representation turns out to be the union of m open intervals, just set $\mathcal{R}_j = \emptyset$ for every $j > m$.

Imagine we have a countable number of uncoloured balls, and we decide to color them according to the next procedure. Let $\mathbf{u}_1, \mathbf{u}_2, ...$ be $\mathsf{Unif}(0, 1)$ independent random variables which are also independent of $\mathcal{R}$ and assume that to each gap of $\mathcal{R}$ we assign a different colour. Now if $\mathbf{u}_i$ falls into $\mathcal{R}_j$ we are going to paint the ith ball with the color j (previously assigned to $\mathcal{R}_j$), if on the other hand $\mathbf{u}_i$ falls into $\mathcal{R}$ we are going to paint the ith ball with a unique color different from the colors assigned to the gaps and also different from any other previously used color. Note that the colors of the balls generate an exchangeable partition $\mathbf{\Pi}$ of $\mathbb{N}$ by the equivalence relation $i \sim k$ if and only if the ith ball and the kth ball to be painted have the same colour. It is clear that if $\mathbf{u}_i$ falls into

$\mathcal{R}$ the ith ball will contribute to a singleton in $\mathbf{\Pi}$. Usually $\mathcal{R}$ is referred as the paintbox and $\mathbf{\Pi}$ as above is called the partition generated by the paintbox $\mathcal{R}$.

Let $\mathbf{w}_j$ be the length of $\mathcal{R}_j$ so that $\mathbb{P}[\mathbf{u}_i \in \mathcal{R}_j] = \mathbf{w}_j$, note that the $\mathbf{w}_j$'s can be thought as the sizes of the atoms of some random distribution function $\mathbf{F}$. So that if we sample $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mid \mathbf{F}\} \overset{\text{iid}}{\sim} \mathbf{F}$ and consider $\mathbf{\Pi}(\mathbf{x}_{1:\infty})$, then this one has the same distribution as $\mathbf{\Pi}$ generated by the coloured balls. Or conversely, given a random distribution function $\mathbf{F}$, the closure of the set $\{\mathbf{F}(x) : \mathbf{F} \text{ is continuous at } x\}$ defines a closed random subset of $[0,1]$ whose lengths of the gaps coincide with the sizes of the atoms of $\mathbf{F}$. See Figure 8 for an illustration of this paragraph.

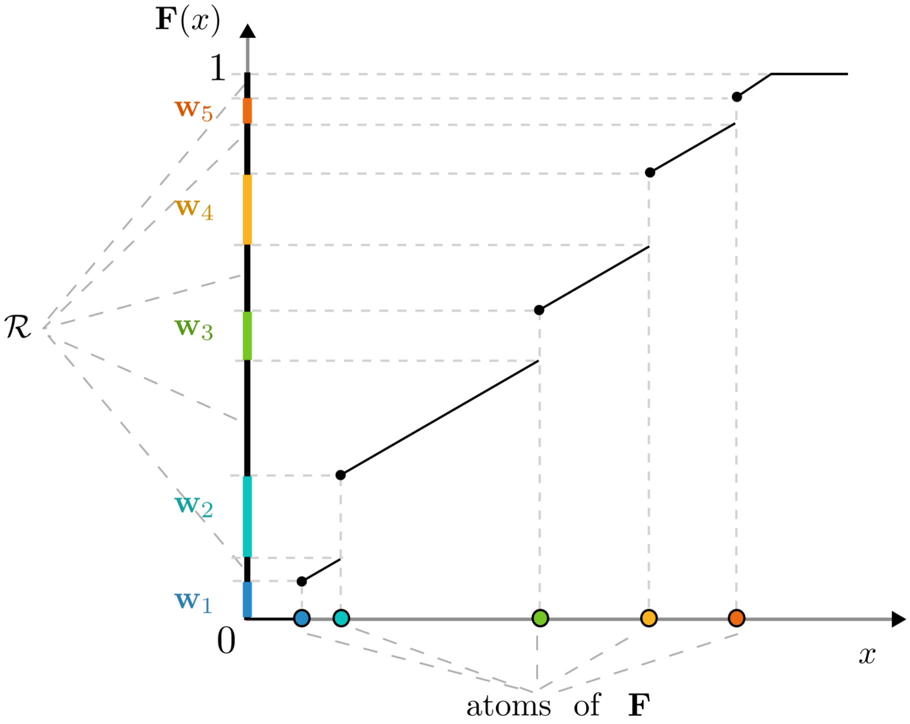


Figure 8: Construction of paintbox through random distribution function.

As mentioned in Example 1.1 a subordinator $\{\boldsymbol{\phi}(t)\}_{t\geq 0}$ is a stochastic process such that $\boldsymbol{\phi}(0) = 0$ a.s., has independent and stationary increments, and its trajectories are non-decreasing, right-continuous and their left limit exists. Note that any subordinator is very similar to a random distribution function except for the fact that $\lim_{t\to\infty} \boldsymbol{\phi}(t) \neq 1$, nevertheless by suitably restraining and normalizing, we can transform it into a random distribution function and generate partitions by sampling from it. Explicitly, let $\boldsymbol{\varsigma} = \inf\{t : \boldsymbol{\phi}(t) = \infty\}$, with the convention $\inf(\emptyset) = \infty$ and let $\mathscr{R}$ be the closure of $\{\boldsymbol{\phi}(t) : 0 \leq t < \boldsymbol{\varsigma}\}$. $\mathscr{R}$ is called the range of $\boldsymbol{\phi}$, it is clear by definition that $\mathscr{R}$ is a closed interval of $[0,\infty]$. Let $\boldsymbol{\tau}$ be a stopping time and consider the closed interval of $[0, \boldsymbol{\phi}(\boldsymbol{\tau})]$, $\mathscr{R}_{\boldsymbol{\tau}} = [0, \boldsymbol{\phi}(\boldsymbol{\tau})] \cap \mathscr{R}$. Let $\mathscr{R}_{\tau}^{c} = [0, \boldsymbol{\phi}(\boldsymbol{\tau})] \setminus \mathscr{R}_{\boldsymbol{\tau}}$ be the open component of $\mathscr{R}_{\tau}$ with canonical representation $\mathscr{R}_{\tau}^{c} = \bigcup_{j=1}^{\infty} \mathscr{R}_j$ (where $\mathscr{R}_j$ is open an interval and $\mathscr{R}_j \cap \mathscr{R}_i = \emptyset$ whenever $i \neq j$). Let $\boldsymbol{\alpha}_i$ be the Lebesgue measure of $\mathscr{R}_i$, so that $(\boldsymbol{\alpha}_i)_{i=1}^{\infty}$ are the jumps of the subordinator up to time $\boldsymbol{\tau}$. Now, consider the transformation $G : [0, \boldsymbol{\phi}(\boldsymbol{\tau})] \to [0,1]$ given by $G(t) = t/\boldsymbol{\phi}(\boldsymbol{\tau})$, that is, we normalize the set $[0, \boldsymbol{\phi}(\boldsymbol{\tau})]$, then $\mathcal{R} = G(\mathscr{R}_{\boldsymbol{\tau}})$ is a random closed subset of $[0,1]$, hence $\mathcal{R}$ could be thought as a paintbox. Define, once again,

$\mathcal{R}^c = [0,1] \setminus \mathcal{R}$ and let $\bigcup_{j=1}^{\infty} \mathcal{R}_j$ be its canonical representation. It is then clear that (by possibly renaming the open intervals) $\mathcal{R}_j = G(\mathscr{R}_j)$, hence the Lebesgue measure of $\mathcal{R}_j$ will be given by $\mathbf{w}_j = \boldsymbol{\alpha}_j / \boldsymbol{\phi}(\boldsymbol{\tau})$. Now, given the random closed subset $\mathcal{R}$ of $[0,1]$ we can generate a random exchangeable partition, by the procedure described above. Figure 9 illustrates this construction.

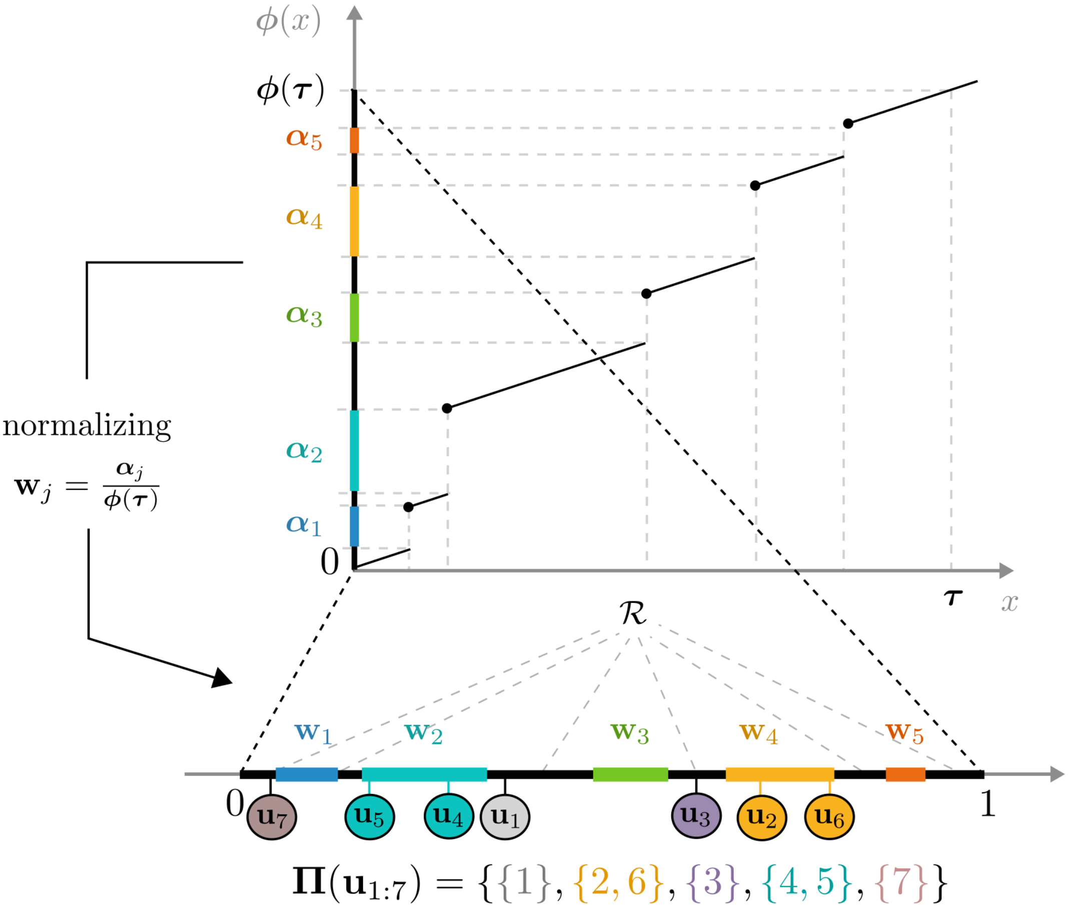


Figure 9: Construction of paintbox through subordinator.

In order for this construction to be well defined we require $0 < \boldsymbol{\phi}(\boldsymbol{\tau}) < \infty$ almost surely. In particular, if $\boldsymbol{\phi}(t) < \infty$ for all $t \in \mathbb{R}$, as explained in Section 1.3.3, we can write

$$\boldsymbol{\phi}(t) = ct + \sum_{j \geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}([0,t]), \quad t \geq 0$$

where $c \geq 0$ and $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j \geq 1}$ defines a Poisson process over $\mathbb{R}_+ \times \mathbb{R}_+$, with intensity ν that can be decomposed as $\nu(ds, dx) = ds\varrho(dx)$, where $\int_{\mathbb{R}_+} (x \wedge 1)\varrho(dx) < \infty$. It is easily seen that if $c = 0$, for every stopping time $\boldsymbol{\tau}$, $\mathscr{R}_{\boldsymbol{\tau}}$ will have a Lebesgue measure zero almost surely thus the derived paintbox $\mathcal{R}$ will also have Lebesgue measure of zero a.s. Therefore we stand in the *proper* scenario, i.e. $\sum_{i=1}^{\infty} \mathbf{w}_i = 1$ a.s. In this case, we must also require that $\int_{[0,1]} \varrho(dx) = \infty$, so that infinitely many very small jumps occur in a finite interval, and we can assure that $\boldsymbol{\phi}(\boldsymbol{\tau}) > 0$ almost surely. If a paintbox $\mathcal{R}$ is constructed through the normalization of a subordinator, then the partition generated by $\mathcal{R}$ is also called the partition generated by the subordinator $\boldsymbol{\phi}$.

2.2.7 Chinese restaurant construction and partially exchangeable partitions

Imagine there exist a Chinese restaurant with numbered tables, each one allowing infinitely many customers seating at once. Let π be an infinite EPPF, with prediction rule $\pi(\cdot \mid \dots)$, as in (2.11). When the first customer arrives he/she will be seated at table number 1. After n customers have arrived and are currently occupying k tables, with n_j customers seating at table j. The customer $(n+1)$th will seat at table j with probability $\pi(j \mid n_1, \dots, n_k)$, for $j \in [k+1]$, as illustrated in Figure 10.

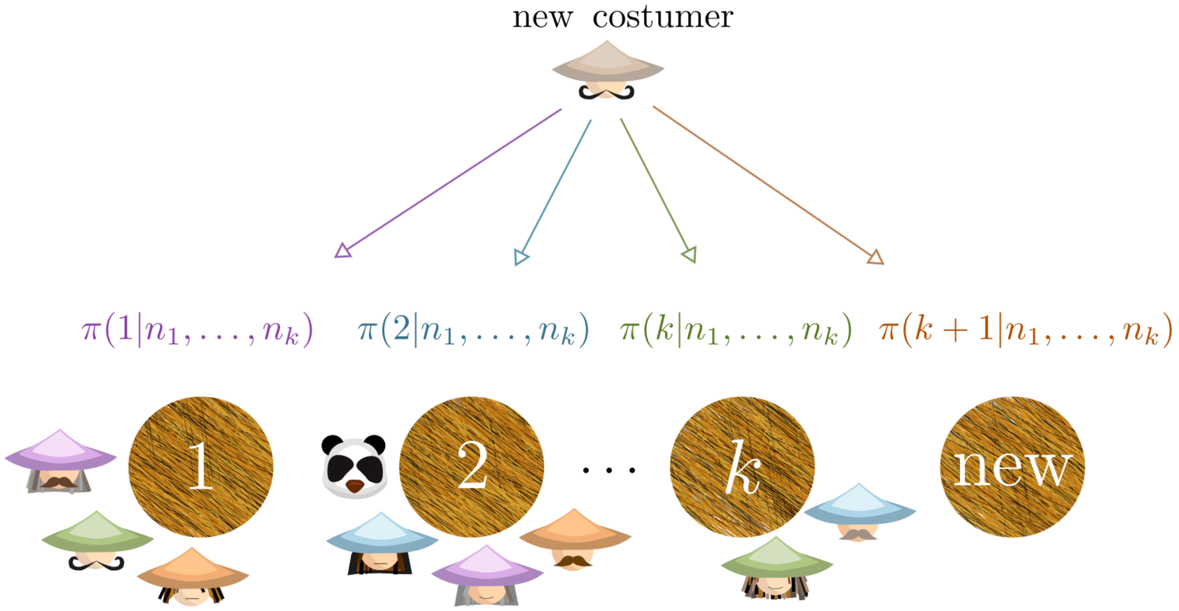


Figure 10: Step of the chinese restaurant process with a given prediction rule, conditioning on event that there are currently k occupied tables with n_j costumers seated at table j, for all $j \in [k]$.

Let $\mathbf{\Pi}_n$ be the partition of $[n]$ generated by the equivalence relation $i \sim j$ if and only if the ith and the jth customers to arrive at sitting at the same table. Then by construction $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\geq 1}$ is an exchangeable partition of $[n]$ and its infinite EPPF is precisely π. It is also obvious that every exchangeable partition of $\mathbb{N}$ can be constructed this way.

For example, fix two real numbers $0 \leq \sigma < 1$ and $\theta > -\sigma$, and say that as above when the first customer arrives he/she will seat at table 1, and after n customers have arrived and are occupying k tables, with n_j customers seating at table j. The customer $(n+1)$th will seat at table j with probability $(n_j - \sigma)/(n+\theta)$, for $j \in [k]$, or will seat at a new table with probability $(\theta + k\sigma)/(n+\theta)$, as illustrated in Figure 11. Note that this is well defined as

$$\frac{\theta + k\sigma}{n+\theta} + \sum_{j=1}^{k} \frac{n_j - \sigma}{n+\theta} = 1.$$

Moreover, if $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\geq 1}$ is the partition of $\mathbb{N}$ generated through the two-parameter scheme, a simple counting argument gives

$$\pi(n_1,\dots,n_k) = \mathbb{P}[\mathbf{\Pi}_n = A] = \frac{(\theta+\sigma)_{k-1\uparrow\sigma}\prod_{j=1}^{k}(1-\sigma)_{n_j-1}}{(\theta+1)_n} \tag{2.13}$$

for every partition $A = \{A_1,\dots,A_k\}$ of $[n]$ with $|A_i| = n_i$, and where $(x)_{m\uparrow\alpha} = \prod_{i=0}^{m-1}(x+i\alpha)$, and $(x)_m = (x)_{m\uparrow 1}$. That is $\mathbf{\Pi}$ is exchangeable, π as in (2.13) describes an infinite EPPF (since n was arbitrary) and

$$\pi(j \mid n_1,\dots,n_k) = \frac{n_j-\sigma}{n+\theta}, \quad \pi(k+1 \mid n_1,\dots,n_k) = \frac{\theta+k\sigma}{n+\theta}, \tag{2.14}$$

for $j \in [k]$, is its prediction rule. To $\mathbf{\Pi}$ we call a two parameter partition

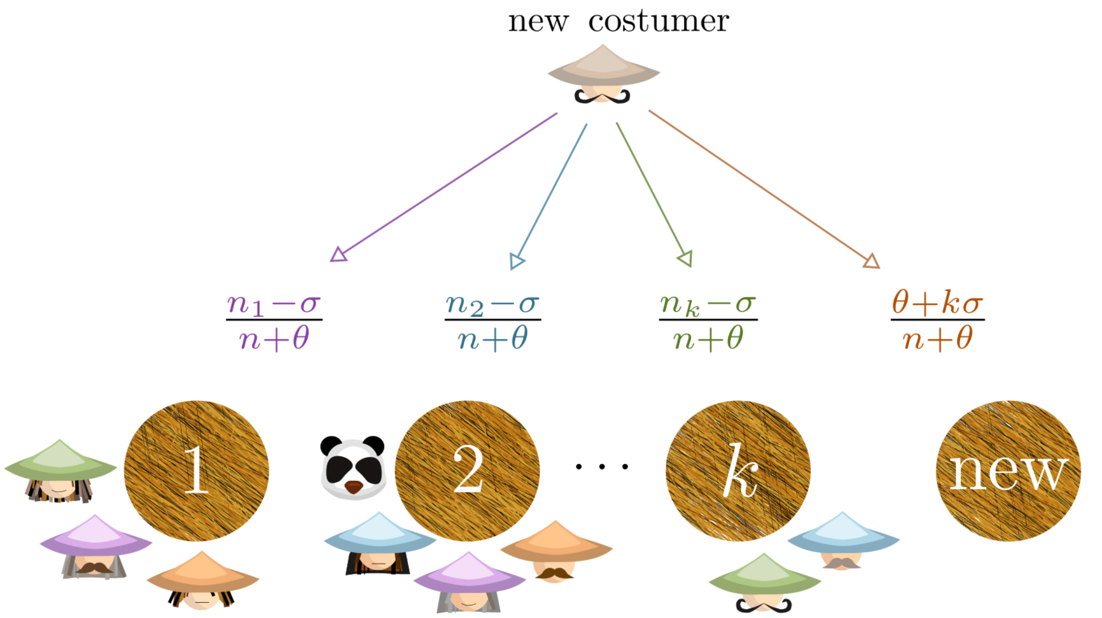

new costumer
$\frac{n_1-\sigma}{n+\theta}$
$\frac{n_2-\sigma}{n+\theta}$
$\frac{n_k-\sigma}{n+\theta}$
$\frac{\theta+k\sigma}{n+\theta}$
1
2
$\cdots$
k
new

Figure 11: Step of the chinese restaurant process for the two parameter model, conditioning on event that there are currently k occupied tables with n_j costumers seated at table j, for all $j \in [k]$.

In particular if $\sigma = 0$ and $\theta > 0$, the new to customer to arrive will chose where to sit with probabilities proportional to $(n_1,\dots,n_k,\theta)$. That is, the prediction rule simplifies as

$$\pi(j \mid n_1,\dots,n_k) = \frac{n_j}{n+\theta}, \quad \pi(k+1 \mid n_1,\dots,n_k) = \frac{\theta}{n+\theta} \tag{2.15}$$

for $j \in [k]$, and the correponding EPPF to

$$\pi(n_1,\dots,n_k) = \mathbb{P}[\mathbf{\Pi}_n = A] = \frac{\theta^k\prod_{j=1}^{k}(n_j-1)!}{(\theta)_n}. \tag{2.16}$$

In this case we call $\mathbf{\Pi}$ a Dirichlet partition, (2.16) is also known as Ewens sampling formulae and has a very interesting relation to stochastic models in genetic populations (Ewens; 1972).

Proposition 2.12. *Fix $0 \leq \sigma < 1$ and $\theta > -\sigma$, and consider the two parameter chinese restaurant model. Let $\mathbf{w}_j$ be the long-run proportion of customers that will end up sitting at table j. Then*

i) *$\mathbf{w}_1 = \mathbf{v}_1$ and for $j \geq 2$, $\mathbf{w}_j = \mathbf{v}_j \prod_{i=1}^{j-1}(1-\mathbf{v}_i)$ where $(\mathbf{v}_i)_{i\geq 1}$ are independent random variables with $\mathbf{v}_i \sim \mathsf{Be}(1-\sigma, \theta + i\sigma)$.*

ii) *The generated exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi}$, is proper, that is $\sum_{j\geq 1} \mathbf{w}_j = 1$ almost surely.*

iii) *$(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ are in size-biased random order, in other words, they are invariant under size-biased permutations.*

The proof of Proposition 2.12 can be found in Appendix B.9. In general if $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\geq 1}$ is the exchangeable partition of $\mathbb{N}$ generated by the chinese restaurant scheme, and for each $n \in \mathbb{N}$, $\left(\tilde{\mathbf{\Pi}}_{n,1}, \ldots, \tilde{\mathbf{\Pi}}_{n,\mathbf{K}_n}\right)$ is the ordering of the blocks of $\mathbf{\Pi}_n$ according to the least element, then $\tilde{\mathbf{\Pi}}_{n,j}$ describes precisely which customers are sitting at table j after n customers arrived. So it is a straight-forward consequence of Proposition 2.11, that if $\tilde{\mathbf{w}}_j$ is the long run-proportion of customers that will end up sitting at table j, then $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots)$ is a size-biased pseudo-permutation of some weights sequence, and if $\mathbf{\Pi}$ is proper $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots)$ is also invariant under size-biased permutations.

Now, let $(\mathbf{w}_j)_{j\geq 1}$ be an arbitrary sequence of weights (not necessarily proper, nor invariant under size-biased permutations) and consider the following random seating plan. The first customer to arrive will always sit at table 1, after n customers have arrived and are occupying k tables, the next customer to arrive will seat at table j with probability $\mathbf{w}_j$, for $j \in [k]$, or will seat at table $k+1$ with probability $1 - \sum_{j=1}^{k} \mathbf{w}_j$, as illustrated in Figure 12.

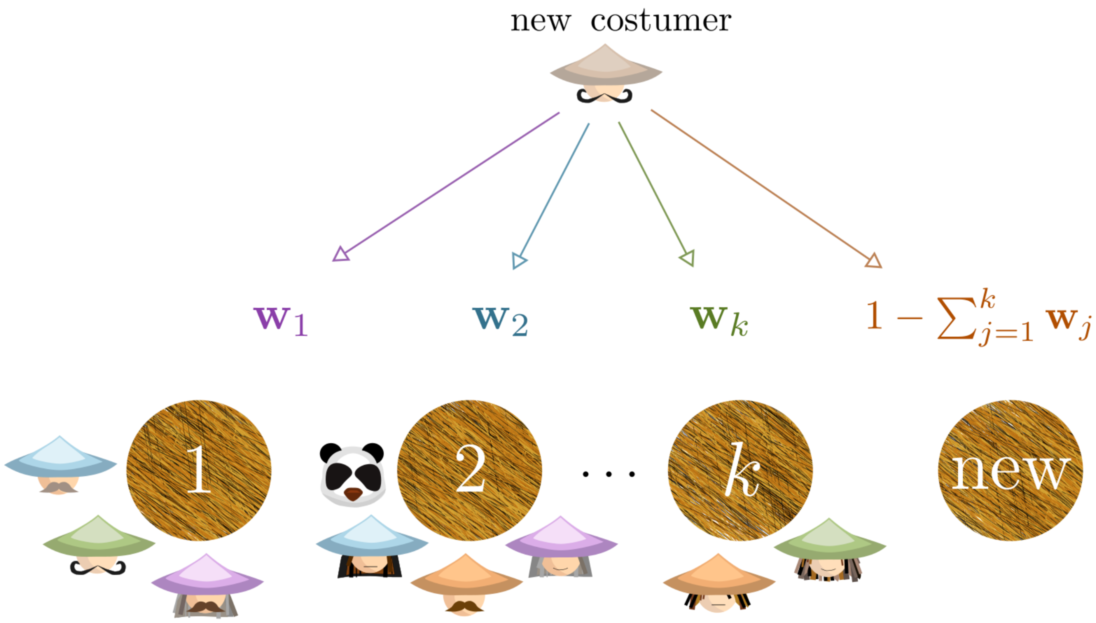


Figure 12: Step of the chinese restaurant process with a random seating plan, conditioning on event that there are currently k occupied tables.

Given $\mathbf{w}_1, \ldots, \mathbf{w}_k$ it is quite easy to compute the conditional probability that after $n = \sum_{j=1}^k n_j$ customers have arrived, there are $n_1, \ldots, n_k$ of them sitting at tables $1, \ldots, k$, respectively, and it is

$$\prod_{j=1}^{k} \mathbf{w}_j^{n_j-1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^{j} \mathbf{w}_j\right)$$

Hence,

$$\pi'(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{w}_j^{n_j-1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^{j} \mathbf{w}_i\right)\right] \tag{2.17}$$

is the (unconditional) probability that for every $j \in [k]$ there are exactly n_j customers seated at table j after $\sum_{j=1}^k n_j$ already arrived. Evidently the random seating plan generates partition of $\mathbb{N}$, $\mathbf{\Pi}' = (\mathbf{\Pi}'_n)_{n \geq 1}$, by means of the equivalence relation $i \sim j$ if and only if the ith and the jth customer to arrive sit at the same table. The first thing one might wonder is if $\mathbf{\Pi}'$ is exchangeable, or in other words, if π' as in (2.17) defines an EPPF. Although in some cases it might, the general answer is no. Note that through the random seating plan we are forcing the long-run proportion of elements in the block containing 1 to be $\mathbf{w}_1$, the long-run proportion of elements in the block containing the smallest element that is not in the block that contains 1 to be $\mathbf{w}_2$, and so on. So if $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ is not a size-biased pseudo-permutation, by Proposition 2.11, it is not possible that $\mathbf{\Pi}'$ is exchangeable. Consider the following example.

Example 2.7. *Let $\mathbf{w}_1 = 1/4$, $\mathbf{w}_2 = 1/2$, $\mathbf{w}_3 = 1/4$, and $\mathbf{w}_j = 0$ for $j > 3$, almost surely. Let $\mathbf{\Pi}' = (\mathbf{\Pi}'_n)_{n \geq 1}$ be the partition of $\mathbb{N}$ generated by the chinese restaurant process with random seating plan and weights as above. Let us focus in $\mathbf{\Pi}_3$. Then from equation (2.17),*

$$\mathbb{P}[\mathbf{\Pi}'_3 = \{\{1\}, \{2\}, \{3\}\}] = \pi'(1,1,1) = \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) = \frac{3}{16},$$

$$\mathbb{P}[\mathbf{\Pi}'_3 = \{\{1,2,3\}\}] = \pi'(3) = \left(\frac{1}{4}\right)^2 = \frac{1}{16},$$

$$\mathbb{P}[\mathbf{\Pi}'_3 = \{\{1\}, \{2,3\}\}] = \pi'(1,2) = \left(\frac{3}{4}\right)\left(\frac{1}{2}\right) = \frac{6}{16},$$

and

$$\mathbb{P}[\mathbf{\Pi}'_3 = \{\{1,2\}, \{3\}\}] = \mathbb{P}[\mathbf{\Pi}'_3 = \{\{1,3\}, \{2\}\}] = \pi'(2,1) = \left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{3}{16},$$

Note that $\mathbf{\Pi}$ is not exchangeable as $\{\{1,2\},\{3\}\}$ is an internal permutation of $\{\{1\},\{2,3\}\}$ but $\mathbb{P}[\mathbf{\Pi}_3 = \{\{1\},\{2,3\}\}] \neq \mathbb{P}[\mathbf{\Pi} = \{\{1,2\},\{3\}\}]$. Now, let $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots)$ be a size-biased permutation of $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$. So that $\tilde{\mathbf{w}}_j = 0$ almost surely for every $j > 3$, and

$$\mathbb{P}[(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \tilde{\mathbf{w}}_3) = (1/2, 1/4, 1/4)] = \frac{1}{2}, \quad \mathbb{P}[(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \tilde{\mathbf{w}}_3) = (1/4, 1/2, 1/4)] = \frac{1}{3},$$

and

$$\mathbb{P}[(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \tilde{\mathbf{w}}_3) = (1/4, 1/4, 1/2)] = \frac{1}{6}.$$

Let $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\geq 1}$ be the partition of $\mathbb{N}$ generated by the chinese restaurant process with random seating plan and weights $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \tilde{\mathbf{w}}_3, \ldots)$. Again, let us focus in $\mathbf{\Pi}_3$. Then, from (2.17)

$$\begin{aligned}\mathbb{P}[\mathbf{\Pi}_3 = \{\{1,2\},\{3\}\}] = \mathbb{P}[\mathbf{\Pi}_3 = \{\{1,3\},\{2\}\}]\\ &= \pi'(2,1) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{3}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\\ &= \frac{1}{8} + \frac{1}{16} + \frac{1}{32} = \frac{7}{32},\end{aligned}$$

$$\begin{aligned}\mathbb{P}[\mathbf{\Pi}_3 = \{\{1\},\{2,3\}\}] &= \pi'(1,2)\\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{3}\right)\left(\frac{3}{4}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{6}\right)\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)\\ &= \frac{1}{16} + \frac{1}{8} + \frac{1}{32} = \frac{7}{32},\end{aligned}$$

$$\begin{aligned}\mathbb{P}[\mathbf{\Pi}_3 = \{\{1,2,3\}\}] &= \pi'(3)\\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^2 + \left(\frac{1}{3}\right)\left(\frac{1}{4}\right)^2 + \left(\frac{1}{6}\right)\left(\frac{1}{4}\right)^2\\ &= \frac{1}{8} + \frac{1}{32} = \frac{5}{32},\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}[\mathbf{\Pi}_3 = \{\{1\},\{2\},\{3\}\}] &= \pi'(1,1,1)\\ &= \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{3}\right)\left(\frac{3}{4}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{6}\right)\left(\frac{3}{4}\right)\left(\frac{1}{2}\right)\\ &= \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{6}{32}.\end{aligned}$$

The above example together with the chinese restaurant construction with random seating plan motivate the following definition.

Definition 2.18. *Let $\mathbf{\Pi}'_n$ be a random partition of $[n]$. We say that $\mathbf{\Pi}'_n$ is partially exchangeable if for any partition $A = \{A_1, \ldots, A_k\}$ of $[n]$, where $A_1, \ldots, A_k$ are in order of appearance*

$$\mathbb{P}[\mathbf{\Pi}'_n = A] = \pi'_n(|A_1|, \ldots, |A_n|)$$

for some function $\pi' : \bigcup_{k=1}^n \mathbb{N}^k \to [0,1]$ in such case π'_n is called a (finite) partially exchangeable partition probability function (pEPPF) of $[n]$.

Proposition 2.13. *For some $n \geq 1$, let $\mathbf{\Pi}'_{n+1}$ be a partially exchangeable random partition of $[n+1]$ with pEPPF π'_{n+1}. Then, $\mathbf{\Pi}'_n = \mathbf{\Pi}'_n\big|_{[n+1]}$ is a partially exchangeable partition of $[n]$, and the pEPPF of $\mathbf{\Pi}'_n$ is given by*

$$\pi'_n(n_1, \ldots, n_k) = \pi'_{n+1}(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} \pi'_{n+1}(n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_k).$$

From Proposition 2.13, whose proof can be found in Apendix B.10, and a simple inductive argument we get that if $\mathbf{\Pi}'_n$ is a partially exchangeable partition of $[n]$ for some $n \geq 2$, and for $m < n$, $\mathbf{\Pi}'_m = \mathbf{\Pi}'_n\big|_{[m]}$, then $\mathbf{\Pi}'_m$ is a partially exchangeable partition of $[m]$. This justifies the following definition

Definition 2.19. *Let $\mathbf{\Pi}' = (\mathbf{\Pi}'_n)_{n\geq 1}$ be a partition of $\mathbb{N}$. We say that $\mathbf{\Pi}'$ is partially exchangeable if for every $n \geq 1$ and any partition $A = \{A_1, \ldots, A_k\}$ of $[n]$, where $A_1, \ldots, A_k$ are in order of appearance*

$$\mathbb{P}[\mathbf{\Pi}'_n = A] = \pi'(|A_1|, \ldots, |A_n|)$$

for some function $\pi' : \bigcup_{k\in\mathbb{N}} \mathbb{N}^k \to [0, 1]$ in such case π' is called (infinite) partially exchangeable partition probability function (pEPPF).

It is straight forward from Proposition 2.13 that every infinite pEPPF π' satisfies the addition rule

$$\pi'(n_1, \ldots, n_k) = \pi'(n_1, \ldots, n_k, 1) + \sum_{j=1}^{k} \pi'(n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_k).$$

This is simply a consequence from the fact that the laws of any random partition of $\mathbb{N}$ form a projective family. As one can intuit, many of the basic definitions and results extend naturally from exchangeable partitions to partially exchangeable partitions. For instance if one substitutes the EPPF, π by the pEPPF π' in equations (2.10) and (2.11), one obtains the prediction rule of a partially exchangeable partition of $\mathbb{N}$. For now, we get back to the Chinese restaurant process with random seating plan.

Proposition 2.14. *The partition of $\mathbb{N}$ generated by the chinese restaurant scheme with random seating plan is always partially exchangeable and its pEPPF is given by* (2.17).

Proposition 2.14 is straight-forward from equation (2.17) and the definition of partially exchangeable partition. Another fact that is obvious from the corresponding definitions is that if $\mathbf{\Pi}'$ is partially exchangeable, then its pEPPF is symmetric if and only if $\mathbf{\Pi}'$ is further exchangeable. This said, the following Theorem, together with Proposition 2.14 and equation (2.17), formally proves that the chinese restaurant process with random seating plan determined by the weights, $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$, generates an exchangeable partition of $\mathbb{N}$ if and only if $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ is a size-biased pseudo-permutation.

Theorem 2.15. *Let $(\mathbf{w}_j)_{j\geq 1}$ be a weights sequence. Then the following are equivalent*

i) $(\mathbf{w}_j)_{j\geq 1}$ *is a size-biased pseudo-permutation.*

ii) $\pi'(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{w}_j^{\mathbf{n}_j-1} \prod_{j=1}^{k-1}\left(1 - \sum_{i=1}^{j} \mathbf{w}_j\right)\right]$ *is a symmetric function of its arguments.*

Corollary 2.16. *Let $(\mathbf{w}_j)_{j\geq 1}$ be a weights sequence. Then the following are equivalent*

i) $\sum_{j\geq 1} \mathbf{w}_j = 1$ *almost surely and $(\mathbf{w}_j)_{j\geq 1}$ is a size-biased permutation.*

ii) $\pi'(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{w}_j^{n_j-1} \prod_{j=1}^{k-1}\left(1 - \sum_{i=1}^{j} \mathbf{w}_j\right)\right]$ *is a symmetric function of its arguments, and $\mathbf{w}_1 > 0$ almost surely.*

Corollary 2.17. *Let $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$ be an arbitrary weights sequence. Let $\mathbf{\Pi}' = (\mathbf{\Pi}'_n)_{n\geq 1}$ be the partially exchangeable partition of $\mathbb{N}$ constructed through the chinese restaurant with random seating plan driven by a size-biased pseudo-permutation, $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_j)_{j\geq 1}$, of $\mathbf{W}$. Also consider the exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi} = (\mathbf{\Pi}_n = \mathbf{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$, generated by sequentially sampling, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{iid}{\sim} \boldsymbol{\mu}$, from a random probability measure $\boldsymbol{\mu}$ with sizes of the atoms given by $(\mathbf{w}_j)_{j\geq 1}$. Then $\mathbf{\Pi}'$ is equal in distribution to $\mathbf{\Pi}$, and the EPPF of $\mathbf{\Pi}$ is given by*

$$\pi(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} (\tilde{\mathbf{w}}_j)^{n_j - 1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^{j} \tilde{\mathbf{w}}_j\right)\right]. \tag{2.18}$$

Proofs of the above results can be found in Appendix B.11, B.12 and B.13, respectively . For an arbitrary weights sequence, $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$, we have shown that we can construct a partially exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi}' = (\mathbf{\Pi}'_n)_{n\geq 1}$ through the chinese restaurant with random seating plan driven by $\mathbf{W}$, and that the pEPPF, π', of $\mathbf{\Pi}'$ is given by (2.17). But also we can construct an exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi} = (\mathbf{\Pi}_n = \mathbf{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$ by sequentially sampling, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\}$, from a random probability measure $\boldsymbol{\mu}$ with sizes of the atoms given by $\mathbf{W}$. The way $\mathbf{\Pi}'$ and $\mathbf{\Pi}$ relate to each other is through the size-biased pseudo-permutation of the weights, as Corollary 2.17 shows. Although it would we great to be able to write the EPPF, π, of $\mathbf{\Pi}$ in terms of π', this is very hard to do. The reason this is, is because in general π is not a symmetrization of π', if it were, in particular we would have $\pi(n) = \pi'(n)$ for every $n \geq 1$, and as Example 2.7 illustrates, this is not case more often than not. To the best of our knowledge, so far there are only two explicit formulae that express the EPPF in terms of the weights sequence. One of them is equation (2.18), and if the weights are proper, the other one is

$$\pi(n_1, \ldots, n_k) = \sum_{(i_1, \ldots, i_k)} \mathbb{E}\left[\prod_{j=1}^{k} (\mathbf{w}_{i_j})^{n_j - 1}\right], \tag{2.19}$$

where the sum ranges over all k-tuples of distinct positive integers. The derivation of (2.19) is explained immediately after Proposition 2.11, for the case where the weights are decreasing, but for arbitrary weights sequence it is completely analogous. The clear advantage of (2.18) over (2.19) is that seemingly no infinite sum is involved, and it is true even for the improper case, its disadvantage is that we would require to know how the size-biased pseudo-permutation distributes.

2.2.8 Final remarks

For an arbitrary weights sequence $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$ let us denote $\mathbf{W}^{(\sigma)}$ to any sequence such that the removal of zeros out of it results in a sequence that is a permutation of the sequence obtained by removing the zeros of $\mathbf{W}$. For simplicity we will simply say $\mathbf{W}^{(\sigma)}$ is a permutation of $\mathbf{W}$. Also, for a distribution P over $\overline{\Delta}_\infty = \left\{(w_1, w_2, \ldots) : w_j \geq 0, \sum_{j\geq 1} w_j \leq 1\right\}$, let us denote

$$\sigma(\mathsf{P}) = \left\{\mathsf{P}^{(\sigma)} : \text{ if } \mathbf{W} \sim \mathsf{P} \text{ then some } \mathbf{W}^{(\sigma)} \sim \mathsf{P}^{(\sigma)}\right\}.$$

In other words, $\sigma(\mathsf{P})$ denotes the equivalence class of P, generated by the equivalence relation P is related to P_0 if and only if P and P_0 are the distributions of permutations of the same weights sequence. So far, give or take, we have explained two methods of constructing a random partition of $\mathbb{N}$ given P over $\overline{\Delta}_\infty$

I. The first construction consists in taking $\mathbf{W} \sim \mathsf{P}$, letting $\boldsymbol{\mu}$ be any random probability measure with atom's sizes given by $\mathbf{W}$, sampling sequentially from $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$ and considering $\boldsymbol{\Pi} = (\boldsymbol{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$. For this construction we know

  a) $\boldsymbol{\Pi}$ is always exchangeable (hence has an infinite EPPF) and the long run proportion of elements in the jth block in order of appearance is $\tilde{\mathbf{w}}_j$, where $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \dots)$ is a size-biased paseudo-permutation of $\mathbf{W}$.

  b) By Kingman's representation theorem, every exchangeable partition, $\boldsymbol{\Pi}$, of $\mathbb{N}$ can be constructed this way. Moreover, this weights sequence is precisely the long-run proportions of elements in the blocks of $\boldsymbol{\Pi}$, according to some ordering.

  c) For P and $\mathsf{P}^{(\sigma)} \in \sigma(\mathsf{P})$ the partitions generated are identically distributed, meaning that through this construction they generate the same infinite EPPF. Thus we have a one to one correspondence between infinite EPPF's and equivalence clases $\sigma(\mathsf{P})$'s. To spell this out, Kingman chose a representative of the class, which is the distribution of the decreasing rearrangement of the weights.

II. The second construction consists in taking $\mathbf{W} \sim \mathsf{P}$, and consider the chinese restaurant process with random seating plan driven by $\mathbf{W}$, then let $\boldsymbol{\Pi}'$ be the random partition of $\mathbb{N}$ generated by the equivalence relation $i \boldsymbol{\sim} j$ if and only if the ith and the jth costumers to arrive end up sitting at the same table.

  a) $\boldsymbol{\Pi}'$ is always partially exchangeable (hence has an infinite pEPPF, given by (2.17)), and the long run proportion of elements in the jth block in order of appearance is $\mathbf{w}_j$. In particular, if $\mathbf{W}$ is a size-biased pseudo permutation, then $\boldsymbol{\Pi}$ is even exchangeable and its EPPF coincides with that corresponding to $\sigma(\mathsf{P})$ in the first construction.

  b) There is a representation theorem for partially exchangeable partitions of $\mathbb{N}$ (Pitman; 1995), which states that every partially exchangeable partition, $\boldsymbol{\Pi}'$ of $\mathbb{N}$ can be constructed through the chinese restaurant with random seating plan, driven by some weights sequence. Furthermore, this weights sequence is given by the long-run proportions of elements in the blocks of $\boldsymbol{\Pi}$, according the least element.

  c) For two distinct probability measures, P and P_0, over $\overline{\Delta}_\infty$, the pEPPF corresponding to them is different, even if $\mathsf{P}_0 \in \sigma(\mathsf{P})$. So this this construction sets up a one to one correspondence between distributions over $\overline{\Delta}_\infty$ and infinite pEPPF's.

This is illustrated in Figure 13.

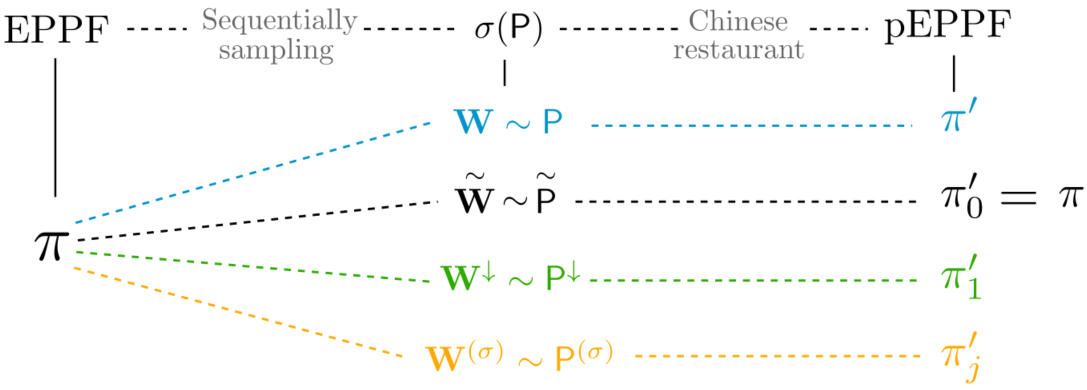

EPPF
Sequentially sampling
$\sigma(\mathsf{P})$
Chinese restaurant
pEPPF
$\mathbf{W} \sim \mathsf{P}$
π'
$\tilde{\mathbf{W}} \sim \tilde{\mathsf{P}}$
$\pi'_0 = \pi$
π
$\mathbf{W}^{\downarrow} \sim \mathsf{P}^{\downarrow}$
π'_1
$\mathbf{W}^{(\sigma)} \sim \mathsf{P}^{(\sigma)}$
π'_j

Figure 13: Correspondence between a weights sequence, permutations of it, and the respective EPPF and pEPPF's. $\mathbf{W}$ denotes an arbitrary weights sequence, $\tilde{\mathbf{W}}$ a size-biased pseudo permutation of $\mathbf{W}$, $\mathbf{W}^{\downarrow}$ its decreasing rearrangement, and $\mathbf{W}^{(\sigma)}$ another arbitrary permutation of $\mathbf{W}$.

Further reading about exchangeable and partially exchangeable partitions can be found in Pitman (2006), Pitman (1995) and Pitman and Yakubovich (2017).

2.3 Random measures

The last big class of exchangeable elements we will be analysing are random measures. In the preliminaries section we formally defined what is a random measure, we also characterized a sub-class of exchangeable random measures which we termed homogeneous completely random measures. Consistently with Section 1, Borel spaces, $(S, \mathscr{B}_S)$, are asummed to be localized and denote the localizing ring by $\hat{\mathcal{S}}$. Recall that $\mathcal{M}(S)$ denotes the space of all locally finite measures over $(S, \mathscr{B}_S)$, so a measure, μ, over $(S, \mathscr{B}_S)$ belongs to $\mathcal{M}(S)$ if and only of $\mu(B) < \infty$ for all $B \in \hat{\mathcal{S}}$.

Definition 2.20 (Symmetric random measure)**.** *Let λ be a diffuse measure in $\mathcal{M}(S)$, and $\boldsymbol{\mu}$ be a locally finite random measure over $(S, \mathscr{B}_S)$.*

i) We say $\boldsymbol{\mu}$ is λ-symmetric if for every measurable function $f : S \to S$, such that $\lambda = \lambda(f^{-1}[\cdot])$ we have that $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\mu}(f^{-1}[\cdot])$.

ii) We say that $\boldsymbol{\mu}$ has λ-exchangeable increments if for every disjoint sets $B_1, \ldots, B_n \in \mathscr{B}_S$, with $\lambda(B_i) = \lambda(B_j)$, we get $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$ is exchangeable.

iii) If $S = \mathbb{R}_+$ or $S = [0,1]$, and λ stands for the Lebesgue measure, for every $B \in \hat{\mathcal{S}}$ we define the contraction map $f_B : B \to [0, \lambda(B)]$ by $f_B(t) = \lambda([0,t] \cap B)$, and we say $\boldsymbol{\mu}$ is contractable if $\boldsymbol{\mu}(f_B^{-1}[\cdot]) \overset{d}{=} \mathbf{1}_{[0,\lambda(B)]}\boldsymbol{\mu}$, for each $B \in \hat{\mathcal{S}}$, and where $\mathbf{1}_A\boldsymbol{\mu}$ denotes the restriction of $\boldsymbol{\mu}$ to A.

Proposition 2.18. *For $S = \mathbb{R}_+$ or $S = [0,1]$, and if λ stands for the Lebesgue measure, (i), (ii) and (iii) of Definition 2.20 are equivalent.*

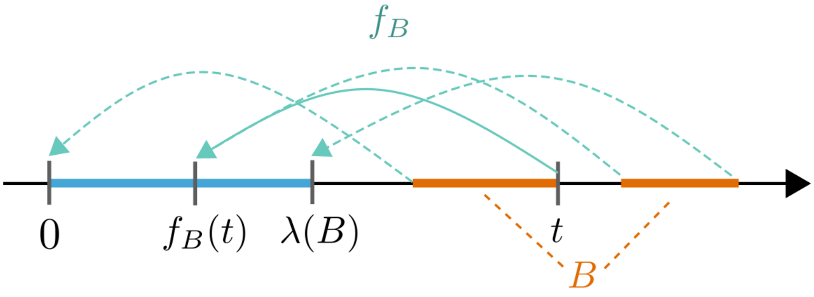


Figure 14: Illustration of a contraction map.

The proof of Proposition 2.18 is provided in Appendix B.14. The equivalence between (i) and (ii) of Definition 2.20 remains true for general Borel spaces $(S, \mathscr{B}_S)$ and arbitrary diffuse measures $\lambda \in \mathcal{M}(S)$. In the general case we recognize two scenarios, if $\lambda(S) = \infty$, through a suitable Borel bijection we can reduce to the case $S = \mathbb{R}_+$ and λ represents the Lebesgue measure over $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$. Alternatively, if $\lambda(S) < \infty$, $\boldsymbol{\mu}$ remains symmetric (exchangeable) with respect to $\lambda' = \lambda/\lambda(S)$ hence we may assume $\lambda(S) = 1$ and reduce to the case where $S = [0,1]$ and λ stands for the Lebesgue measure over $([0,1], \mathscr{B}_{[0,1]})$. Thus, the following result is a Corollary of Proposition 2.18.

Corollary 2.19. *A locally finite random measure $\boldsymbol{\mu}$ over a Borel space $(S, \mathscr{B}_S)$ is λ-symmetric if and only if it is λ-exchangeable, for $\lambda \in \mathcal{M}(S)$.*

Proposition 2.20. *Let $\boldsymbol{\mu}$ be a λ-symmetric random measure over the Borel space $(S, \mathscr{B}_S)$, set $\mu = \mathbb{E}[\boldsymbol{\mu}]$ then $\boldsymbol{\mu}$ is also μ-symmetric.*

The proof of Proposition 2.20 can be found in Appendix B.15. A first example of a λ-symmetric random measure is the diffuse measure, $\boldsymbol{\mu} = \boldsymbol{\beta}\lambda$, for some positive random variable, $\boldsymbol{\beta}$. For the case where $\lambda(S) = \infty$, another example of a random measure with λ-exchangeable increment is a mixed Poisson process, $\boldsymbol{\mu}$, with intensity $\boldsymbol{\kappa}\lambda$, for some positive random variable, $\boldsymbol{\kappa}$. Indeed, for every collection of disjoint sets $B_1, \ldots, B_n \in \mathscr{B}_S$, with $\lambda(B_i) = \lambda(B_j)$, $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$ is conditionally i.i.d., hence exchangeable. As for the case where $\lambda(S) < \infty$, an example of a λ-symmetric simple point process is a mixed binomial process, $\boldsymbol{\mu} = \sum_{j=1}^{\boldsymbol{\kappa}} \delta_{\boldsymbol{\xi}_j}$, based on $(\boldsymbol{\kappa}, \lambda')$, for some $\boldsymbol{\kappa}$ taking values in $\mathbb{N}$, and where $\lambda' = \lambda/\lambda(S)$. Recall that in this case, conditionally given $\boldsymbol{\kappa}$, $(\boldsymbol{\xi}_j)_{j=1}^{\boldsymbol{\kappa}} \overset{\text{iid}}{\sim} \lambda'$, from which is easy to see that for every measurable partition of S, $\{B_1, \ldots, B_n\}$, $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n)) \sim \mathsf{Multinomial}(\boldsymbol{\kappa}; \lambda'(B_1), \ldots, \lambda'(B_n))$, from which the exchangeability of the increments follows easily. As the following result states (see Appendix B.16 for a proof), these examples constitute the class of simple point processes and diffuse random measures with λ-exchangeable increments.

Lemma 2.21. *Let $(S, \mathscr{B}_S)$ be a Borel space with localizing ring $\hat{\mathcal{S}}$ and consider a diffuse measure $\lambda \in \mathcal{M}(S)$.*

i) *A locally finite simple point process, $\boldsymbol{\mu}$, is λ-symmetric if and only if it is mixed Poisson or a mixed binomial process based on $\boldsymbol{\kappa}$ and λ (or $\lambda/\lambda(S)$), for some random variable, $\boldsymbol{\kappa}$ taking values in $\mathbb{R}_+$ or $\mathbb{N}$, respectively.*

ii) *A locally finite difusse random measure, $\boldsymbol{\mu}$ is λ-symmetric if and only if $\boldsymbol{\mu} = \boldsymbol{\beta}\lambda$, for some non-negative random variable $\boldsymbol{\beta}$.*

More generally, if $\lambda(S) = \infty$, homogeneous completely random measures, with no diffuse component, are also examples of λ-symmetric random measures. Indeed, as explained by Theorem 1.12, such a measure, can be decomposed as $\boldsymbol{\mu} = \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$, where $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ defines a Poisson process whose intensity decomposes as $\lambda \otimes \varrho$, for some measure ϱ such that $\int_{\mathbb{R}_+} (x \wedge 1)\varrho(dx) < \infty$. This random measure, $\boldsymbol{\mu}$ satisfies that for every collection of disjoint sets $B_1, \ldots, B_n \in \mathscr{B}_S$, with $\lambda(B_i) = \lambda(B_j)$, $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$ is i.i.d., in particular it is exchangeable. Generalizing these processes, we might consider a random measure $\boldsymbol{\varrho}$ over $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$, where $\int_{\mathbb{R}_+} (x \wedge 1)\boldsymbol{\varrho}(dx) < \infty$ holds almost surely, and take $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ forming a Cox process directed by $\lambda \otimes \boldsymbol{\varrho}$ (so that conditionally given $\boldsymbol{\varrho}$, $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ defines a Poisson process). Then for $B_1, \ldots, B_n \in \mathscr{B}_S$ as above $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$ is conditionally i.i.d. (given $\boldsymbol{\varrho}$), so $\boldsymbol{\mu} = \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$ is also an example of a λ-symmetric random measure. As for the case where $\lambda(S) < \infty$, another example of a λ-exchangeable random measure is $\boldsymbol{\mu} = \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$, where the collections $(\boldsymbol{\alpha}_j)_{j\geq 1}$ and $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \lambda/\lambda(S)$ are independent. To see that the latter random measure has λ-symmetric increments, let $f : S \to S$, be a λ-preserving transformation, define $\boldsymbol{\nu} = \boldsymbol{\mu}(f^{-1}[\cdot])$ and consider a simple function $g = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$, where $A_1, \ldots, A_n \in \mathscr{B}_S$ are disjoint. Then

$$\boldsymbol{\mu}(g) = \sum_{j\geq 1} \boldsymbol{\alpha}_j g(\boldsymbol{\xi}_j) = \sum_{j\geq 1} \sum_{i=1}^n \boldsymbol{\alpha}_j a_i \delta_{\boldsymbol{\xi}_j}(A_i) \overset{d}{=} \sum_{j\geq 1} \sum_{i=1}^n \boldsymbol{\alpha}_j a_i \delta_{\boldsymbol{\xi}_j}(f^{-1}[A_i]) = \boldsymbol{\nu}(g),$$

and by the proof of Theorem 1.5, this show $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\mu}(f^{-1}[\cdot])$. Once again, the random measures discussed in this paragraph are not only simple examples of purely atomic random measures with λ-exchangeable increments, but characterize completely this class. Formally we have the following result whose proof appears in Appendix B.17.

Lemma 2.22. *Let $(S, \mathscr{B}_S)$ be a Borel space with localizing ring $\hat{S}$ and consider a diffuse measure $\lambda \in \mathcal{M}(S)$. Then a locally finite, purely atomic random measure, $\boldsymbol{\mu} = \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$, is λ-symmetric if and only*

i) *For $\lambda(S) < \infty$, $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{iid}{\sim} \lambda/\lambda(S)$ is independent of $(\boldsymbol{\alpha}_j)_{j\geq 1}$.*

ii) *For $\lambda(S) = \infty$, $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ defines a Cox process directed by $\lambda \otimes \boldsymbol{\varrho}$, for some random measure, $\boldsymbol{\varrho}$, over $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$ and satisfying $\int_{\mathbb{R}_+} (x \wedge 1)\boldsymbol{\varrho}(dx) < \infty$ almost surely.*

Putting together Lemma 2.22 and the second part of Lemma 2.21, we can derive the representation theorem for random measure with λ-exchangeable increments.

Theorem 2.23. *Let $(S, \mathscr{B}_S)$ be a Borel space with localizing ring $\hat{S}$ and consider a diffuse measure $\lambda \in \mathcal{M}(S)$. Then, a locally finite random measure $\boldsymbol{\mu}$ is λ-symmetric if and only if it can be decomposed as $\boldsymbol{\mu} = \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j} + \boldsymbol{\beta}\lambda$, for some positive random variable, $\boldsymbol{\beta}$, and where*

i) *For $\lambda(S) < \infty$, $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{iid}{\sim} \lambda/\lambda(S)$ is independent of $(\boldsymbol{\alpha}_j)_{j\geq 1}$.*

ii) *For $\lambda(S) = \infty$, $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ defines a Cox process directed by $\lambda \otimes \boldsymbol{\varrho}$, for some random measure, $\boldsymbol{\varrho}$, over $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$ and satisfying $\int_{\mathbb{R}_+} (x \wedge 1)\boldsymbol{\varrho}(dx) < \infty$ almost surely.*

Corollary 2.24. *Let $(S, \mathscr{B}_S)$ be a Borel space and consider a finite and diffuse measure λ over $(S, \mathscr{B}_S)$. Then, a random probability measure, $\boldsymbol{\mu}$, is λ-symmetric if and only if it takes the form*

$$\boldsymbol{\mu} = \sum_{j \geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j \geq 1} \mathbf{w}_j\right) \mu_0,$$

where $\mu_0 = \lambda/\lambda(S)$, and $(\boldsymbol{\xi}_j)_{j \geq 1} \overset{iid}{\sim} \mu_0$ is independent of $(\mathbf{w}_j)_{j \geq 1}$ which are non-negative random variables such that $\sum_{j \geq 1} \mathbf{w}_j \leq 1$ almost surely.

The random probability measures in Corollary 2.24 constitute the building blocks of Bayesian non-parametric models, and will be the objects of study of next section.

3 Species sampling processes

Random probability measures with exchangeable increments, better known as species sampling process are the building blocks for a wide range of Bayesian non-parametric models. This due to the fact that they are extremely flexible while remaining mathematically tractable. In comparison to other subjects of probability and statistics, the study of species sampling processes remains flourishing, and its has been mainly developed by means of concrete examples. The canonical example of species sampling processes in Bayesian non-parametric literature is the Dirichlet processes (Ferguson; 1973; Blackwell and MacQueen; 1973; Kingman; 1975; Sethuraman; 1994), for a complete compilation on this model see the monograph by Ghosal and van der Vaart (2017). Searching for generalizations and competitive alternatives to the canonical model, different constructions of species sampling processes have been developed. Some of the most notable are through the normalization of homogeneous completely random measures (Regazzini et al.; 2003; James et al.; 2009; Hjort et al.; 2010), through the prediction rule of exchangeable partitions (Blackwell and MacQueen; 1973; Pitman; 1996b; Hansen and Pitman; 2000; De Blasi et al.; 2015), by means of the stick-breaking decomposition of weights sequences (Sethuraman; 1994; Ishwaran and James; 2001; Pitman; 2006; Favaro et al.; 2012, 2016) and most recently by virtue of latent random subsets of $\mathbb{N}$ (Walker; 2007; Fuentes-García et al.; 2010; De Blasi et al.; 2020). Constructions of species sampling processes are extremely important because they allow us to specify the laws of random probability measures with exchangeable increments, which in general is not an easy task to do. This in turn creates a wide range of Bayesian non-parametric models.

In contrast to the natural development of Bayesian non-parametric literature, here we tackle the study of species sampling processes in the opposite direction, that is, we begin by deriving properties for the general class of species sampling processes and latter specialize the analysis to concrete examples. We start Section 3.1 with an overview of basic properties and the definition of important quantities related to the random probability measures in question. Section 3.2 is latter dedicated to analyse some convergence results of species sampling processes, these are extremely important for the new class of models we will define in Section 4. In Section 3.3, we characterize exchangeable sequences driven by species sampling processes. The results in Section 3.3 are specifically enriching because they make explicit the relation between exchangeable sequences, exchangeable partitions and random measures with exchangeable increments. While some of the content of Sections 3.1-3.3 is well known in literature, to best of our knowledge, it has not been described in the unified way presented here. Section 3.4 is then dedicated to study the support of species sampling processes, we do this following the work of Bissiri and Ongaro (2014) (also see Datta; 1991; Ghosal et al.; 1999; Wu and Ghosal; 2008, for more one this topic). As mentioned by Bissiri and Ongaro (2014), having a large support is the unique an essential requirement for a species sampling model to become a feasible Bayesian non-parametric prior. Hence, when we define the law of a species sampling process it is a priority to corroborate it has full support. In Section 3.5 we detail the methods previously mentioned to determine the distribution of a species sampling process. Finally, in Section 3.6 we review some of the most famous examples in Bayesian non-parametric statistics, including the celebrated Dirichlet processes.

3.1 Basic properties and definitions

Definition 3.1 (Species sampling process 1)**.** *A random probability measure $\boldsymbol{\mu}$, over a Borel space $(S, \mathscr{B}_S)$ is called a species sampling process (SSP) if it has μ_0-exchangeable increments with respect to a diffuse probability measure, μ_0 over $(S, \mathscr{B}_S)$.*

From Corollary 2.24 we have the following equivalent definition of species sampling processes, which is the best known in literature.

Definition 3.2 (Species sampling process 2)**.** *A random probability measure $\boldsymbol{\mu}$, over a Borel space $(S, \mathscr{B}_S)$ is called a species sampling process (SSP) if its atomic decomposition specializes to the form*

$$\boldsymbol{\mu} = \sum_{j \geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j \geq 1} \mathbf{w}_j\right) \mu_0, \tag{3.1}$$

where μ_0, which is called base measure of $\boldsymbol{\mu}$, is some diffuse probability measure over $(S, \mathscr{B}_S)$, and the collection of atoms $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j\geq 1} \overset{iid}{\sim} \mu_0$ is independent of the weights $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$ which are non-negative random variables such that $\sum_{j\geq 1} \mathbf{w}_j \leq 1$ almost surely. Whenever $\sum_{j\geq 1} \mathbf{w}_j = 1$ almost surely, so that $\boldsymbol{\mu}$ is purely atomic, we say $\boldsymbol{\mu}$ is proper, and otherwise we call $\boldsymbol{\mu}$ improper.

The term species sampling is due to Pitman, and it comes from the fact that we might think each atom, $\boldsymbol{\xi}_j$, represents an unknown species in a population, and $\mathbf{w}_j$ is the proportion of individuals in the population of species $\boldsymbol{\xi}_j$, if $\boldsymbol{\mu}$ is improper, $(1 - \sum_{j\geq 1} \mathbf{w}_j)$ represents the number of individuals in the population that come from a species with only one member. This way, if we were to sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, then $\mathbf{x}_i$ would be of species $\boldsymbol{\xi}_j$ with probability $\mathbf{w}_j$, for every $j \geq 1$, and $\mathbf{x}_i$ would be of a rare species, containing only one member, with probability $(1 - \sum_{j\geq 1} \mathbf{w}_j)$.

In Bayesian non-parametric statistics, the distribution of a SSP is often referred to as the prior distribution, or simply prior. We will adopt this terminology hereinafter. Clearly a prior is completely characterized by the distribution of the weights and the atoms. In other words a diffuse probability measure μ_0 over $(S, \mathscr{B}_S)$, together with a distribution, P, over $\overline{\Delta}_\infty = \{(w_1, w_2, \ldots) : w_j \geq 0, \sum_{j\geq 1} w_j \leq 1\}$, determine completely the distribution, say Q, of a SSP. However, the converse is not true, indeed there are infinitely many distributions over $\overline{\Delta}_\infty$ that lead to the exact same prior. This is explained by the following result.

Proposition 3.1. *Let $\boldsymbol{\mu}$ be a SSP as in (3.1), and let $\boldsymbol{\sigma}$ be a (possibly) random permutation of $\mathbb{N}$, independent of the atoms of $\boldsymbol{\mu}$. Then, $\boldsymbol{\mu}$ is equal in distribution to*

$$\sum_{j \geq 1} \mathbf{w}_{\boldsymbol{\sigma}(j)} \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j \geq 1} \mathbf{w}_{\boldsymbol{\sigma}(j)}\right) \mu_0.$$

In other words, the prior distribution is invariant under permutations of the weights.

The proof of Proposition 3.1 can be found in Appendix C.1. In practice, when defining prior distributions by means of choosing a base measure and a distribution over $\overline{\Delta}_\infty$, it is highly important to bethink of Proposition 3.1. Depending on the context, working with one ordering of the weights or another can have appealing advantages. Ideally one would have available the distribution of all possible permutations of a weights sequence, so that one could chose the most convenient, unfortunately this is generally not the case. When implementing these models, one usually works with the representation of the weights that is the most mathematically tractable, however one should keep in mind the considered representation of the weights is not unique for the prior in question.

One of the most important quantities related to SSPs is the tie probability,

$$\mathbb{P}[\mathbf{x}_i = \mathbf{x}_j] = \mathbb{E}\left[\mathbb{P}[\mathbf{x}_1 = \mathbf{x}_2 \mid \boldsymbol{\mu}]\right] = \mathbb{E}\left[\sum_{j\geq 1} (\mathbf{w}_j)^2\right] = \sum_{j\geq 1} \mathbb{E}\left[(\mathbf{w}_j)^2\right],$$

for $i \neq j$, and where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$. As Lemma 3.2 and Corollary 3.3 show, the first couple of moments of a prior are completely determined by the base measure and the tie probability. Formally we have the following definition.

Definition 3.3 (Tie probability)**.** *Let $\boldsymbol{\mu}$ be a species sampling process as in* (3.1)*. To $\rho = \sum_{j\geq 1} \mathbb{E}\left[(\mathbf{w}_j)^2\right]$ we call the tie probability of $\boldsymbol{\mu}$.*

Note that the tie probability does not depends on the ordering of the weights, in fact as $0 \leq \sum_{j\geq 1} \mathbf{w}_j \leq 1$ almost surely, we have that $\sum_{j\geq 1}(\mathbf{w}_j)^2 = \sum_{j\geq 1}(\mathbf{w}_{\boldsymbol{\sigma}(j)})^2$, almost surely, for any (possibly random) permutation, $\boldsymbol{\sigma}$, of $\mathbb{N}$. Before stating the first result that concerns this very important number, recall that for a suitable integrable function f, and a random (or deterministic) measure we denote $\boldsymbol{\mu}(f) = \int f d\boldsymbol{\mu} = \int f(s)\boldsymbol{\mu}(ds)$.

Lemma 3.2. *Let $(S, \mathscr{B}_S)$ be a Borel space and $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \mu_0$ be a species sampling process over S with base measure μ_0 and with tie probability ρ. Let $f, g : S \to \mathbb{R}$ be measurable and bounded functions. Then,*

i) $\mathbb{E}\left[\boldsymbol{\mu}(f)\right] = \mu_0(f)$.

ii) $\mathbb{E}\left[\boldsymbol{\mu}(f)^2\right] = \rho\, \mu_0(f^2) + (1-\rho)\mu_0(f)^2$

iii) $\mathbb{E}\left[\boldsymbol{\mu}(f)\boldsymbol{\mu}(g)\right] = \rho\, \mu_0(fg) + (1-\rho)\mu_0(f)\mu_0(g)$.

The proof of Lemma 3.2 appears in Appendix C.2. By making $f = \mathbf{1}_A$ and $g = \mathbf{1}_B$, we obtain the following straight-forward corollary of Lemma 3.2.

Corollary 3.3. *Let $(S, \mathscr{B}_S)$ be a Borel space and $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \mu_0$ be a species sampling process over S with base measure μ_0 and with tie probability ρ. Then, for any measurable sets A and B,*

i) $\mathbb{E}\left[\boldsymbol{\mu}(A)\right] = \mu_0(A)$.

ii) $\mathsf{Var}\left(\boldsymbol{\mu}(A)\right) = \rho\, \mu_0(A)(1 - \mu_0(A))$

iii) $\mathsf{Cov}\left(\boldsymbol{\mu}(A), \boldsymbol{\mu}(B)\right) = \rho(\mu_0(A \cap B) - \mu_0(A)\mu_0(B))$.

3.2 Limiting properties

Corollary 3.3, shows that the expectation, the variance and the self-covariance of a SSP are completely determined by the base measure and the tie probability. Realize that if ρ is close to zero, $\mathsf{Var}(\boldsymbol{\mu}(A)) \approx 0$ and $\mathsf{Cov}(\boldsymbol{\mu}(A), \boldsymbol{\mu}(B)) \approx 0$. Indeed, the first limiting result concerning a SSP is that, as $\rho \to 0$, a SSP converges weakly in distribution to its base measure, and conversely, as $\rho \to 1$, a SSP converges weakly in distribution to $\delta_{\boldsymbol{\xi}}$, for some $\boldsymbol{\xi} \sim \mu_0$.

Theorem 3.4. *Consider a Polish space S with Borel σ-algebra $\mathscr{B}_S$. Let $\mu_0, \mu_0^{(1)}, \mu_0^{(2)}, \ldots$ be diffuse probability measures over $(S, \mathscr{B}_S)$, such that $\mu_0^{(n)}$ converges weakly to μ_0 as $n \to \infty$. For $n \geq 1$ let $\rho^{(n)} \in (0,1)$, and let $\boldsymbol{\mu}^{(n)}$ be a SSP with base measure $\mu_0^{(n)}$ and tie probability $\rho^{(n)}$.*

i) *If $\rho^{(n)} \to 0$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to μ_0.*

ii) *If $\rho^{(n)} \to 1$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\delta_{\boldsymbol{\xi}}$, where $\boldsymbol{\xi} \sim \mu_0$.*

The proof of Theorem 3.4 is given in Appendix C.3.

Remark 3.1. *a) In Theorem 4.11, Kallenberg (2017), shows that the weak convergence in distribution of random probability measures is equivalent, to the weak convergence of the corresponding distributions. So in the context of Theorem 3.4, if we denote by $\mathsf{Q}^{(n)}$ the prior of $\boldsymbol{\mu}^{(n)}$, and by Q the distribution of $\delta_{\boldsymbol{\xi}}$, then the result also states that if $\rho^{(n)} \to 0$, then $\mathsf{Q}^{(n)}$ converges weakly to δ_{μ_0}, and if $\rho^{(n)} \to 1$, then $\mathsf{Q}^{(n)}$ converges weakly to Q as $n \to \infty$.*

b) If we require the SSP's in Theorem 3.4 to be defined in the same probability space, and $\rho^{(n)} \to 0$, then $\boldsymbol{\mu}^{(n)}$ also converges weakly in $\mathcal{L}_2$ to μ_0, as $n \to \infty$. As to the second part of the result, if we denote by $\boldsymbol{\xi}_1^{(n)}$ to the atom corresponding to the largest weight of $\boldsymbol{\mu}^{(n)}$, and we require $\boldsymbol{\xi}_1^{(n)} \to \boldsymbol{\xi}$ almost surely, we can assure that if $\rho^{(n)} \to 1$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in $\mathcal{L}_2$ to $\delta_{\boldsymbol{\xi}}$. The proof of this statement is contained in the proof of Theorem 3.4, in the Appendix.

c) The corresponding almost sure convergence can not be assured in general. Despite this, under the conditions stated in (b), there exist a subsequence $\left(\boldsymbol{\mu}^{(n_k)}\right)_{k\geq 1}$ such that (i) or (ii) in Theorem 3.4, holds almost surely.

To illustrate Theorem 3.4, in Figure 15 we show some simulations of different SSPs over $([0,1], \mathscr{B}_{[0,1]})$ all of them with base measure $\mu_0 = \mathsf{Unif}(0,1)$. The realization of the SSP in A, corresponds to a very small tie probability, whilst the one in F, has assigned a tie probability close to 1. Figure 16 shows the respective distribution functions, so for instance if B in Figure 15 illustrates a simulation of $\boldsymbol{\mu}^{(\mathsf{B})}$, then B in Figure 16 shows the same realization of $\mathbf{F}^{(\mathsf{B})}(x) = \boldsymbol{\mu}^{(\mathsf{B})}([0,x])$. Here, it can be appreciated that, when ρ is small the simulation of the random distribution function is very similar to the identity function on $[0,1]$, which coincides with the cumulative distribution function of a $\mathsf{Unif}(0,1)$. Conversely, for large values of ρ, the realization of the distribution function resembles that of $\delta_{\boldsymbol{\xi}}$ where $\boldsymbol{\xi}$ was drawn from a Uniform distribution.

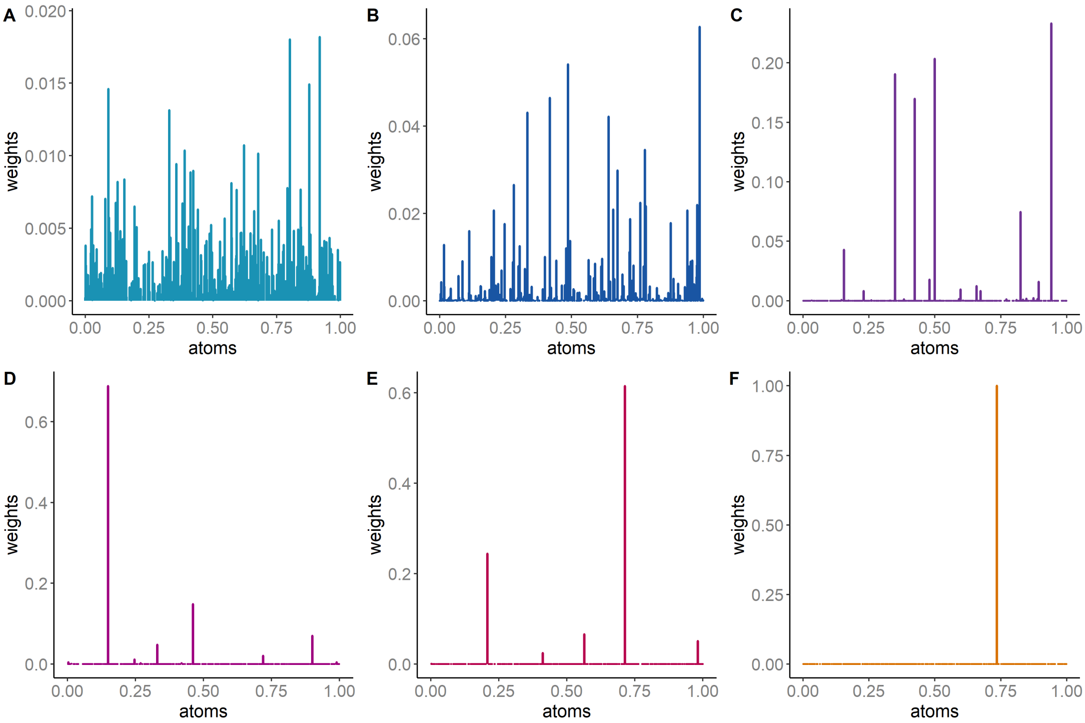


Figure 15: Simulations of six species sampling processes with base measure $\mathsf{Unif}(0, 1)$ and tie probability $\rho = 0.005, 0.02, 0.25, 0.5, 0.75, 0.95$, for A–F, respectively. In each sub-figure the height of the vertical lines indicate the weights and the intersection of each line the the x-axis indicates the corresponding atoms.

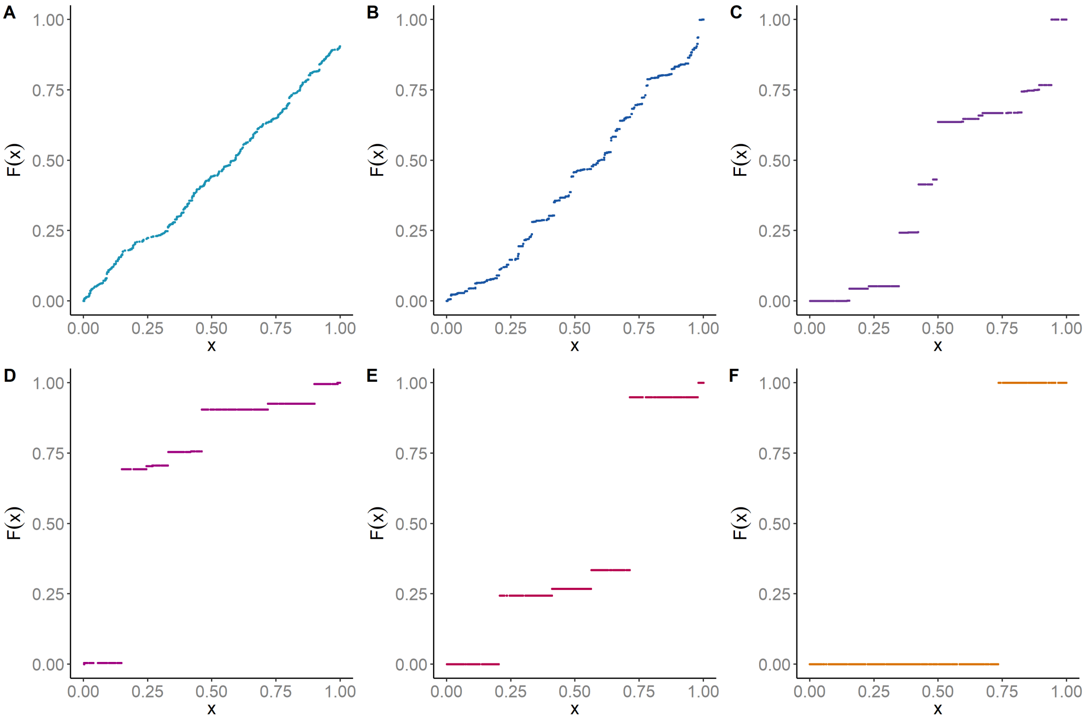


Figure 16: Simulations of the cumulative distribution functions, $\mathbf{F}(x) = \boldsymbol{\mu}((-\infty, x]) = \boldsymbol{\mu}([0, x])$, corresponding to the species sampling processes in Figure 15.

In Lemma 1.18, we showed that the mappings, $[(w_1, w_2, \ldots), (\mu_1, \mu_2, \ldots)] \mapsto \sum_{j\geq 1} w_j \mu_j$, from $\Delta_\infty \times \mathcal{P}(S)^\infty$ into $\mathcal{P}(S)$, and $[(w_1, w_2, \ldots), (s_1, s_2, \ldots)] \mapsto \sum_{j\geq 1} w_j \delta_{s_j}$ from $\Delta_\infty \times S^\infty$ into $\mathcal{P}(S)$, are continuous with respect to the weak topology, where Δ_∞ denotes the infinite dimensional simplex. The following Theorem, which is straight forward from Lemma 1.18, states another limiting property for SSPs.

Theorem 3.5. *Let $(S, \mathscr{B}_S)$ be a Polish space and consider the SSPs over $(S, \mathscr{B}_s)$, $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} + \mathbf{w}_0 \mu_0$, and $\boldsymbol{\mu}^{(n)} = \sum_{j\geq 1} \mathbf{w}_j^{(n)} \delta_{\boldsymbol{\xi}_j^{(n)}} + \mathbf{w}_0^{(n)} \mu_0^{(n)}$, for every $n \geq 1$, where $\mathbf{w}_0^{(n)} = 1 - \sum_{j\geq 1} \mathbf{w}_j^{(n)}$ and analogously $\mathbf{w}_0 = 1 - \sum_{j\geq 1} \mathbf{w}_j$. Let us denote $\mathbf{W}^{(n)} = \left(\mathbf{w}_0^{(n)}, \mathbf{w}_1^{(n)}, \ldots\right)$, $\mathbf{W} = (\mathbf{w}_0, \mathbf{w}_1, \ldots)$, $\Xi^{(n)} = \left(\boldsymbol{\xi}_j^{(n)}\right)_{j\geq 1}$ and $\Xi = (\boldsymbol{\xi}_j)_{j\geq 1}$, for $n \geq 1$.*

a) *If $\mathbf{W}^{(n)}$ and $\Xi^{(n)}$ converge almost surely to $\mathbf{W}$ and Ξ, respectively. Then $\boldsymbol{\mu}^{(n)}$ converges weakly almost surely to $\boldsymbol{\mu}$.*

b) *If $\mathbf{W}^{(n)}$ and $\Xi^{(n)}$ converge in distribution to $\mathbf{W}$ and Ξ, respectively. Then $\boldsymbol{\mu}^{(n)}$ converges in distribution to $\boldsymbol{\mu}$.*

3.3 Properties of samples from a species sampling process

In this section we will analyse the main properties of exchangeable sequences driven by a species sampling process. Recall that in general the law of a sequence $(\mathbf{x}_i)_{i\geq 1}$ is completely characterize by the finite dimensional distributions, $\mathbb{P}\left[\mathbf{x}_1 \in B_1, \ldots, \mathbf{x}_n \in B_n\right]$, for $n \geq 1$, and by the predictive distributions $\mathbb{P}\left[\mathbf{x}_{n+1} \in \cdot \mid \mathbf{x}_1, \ldots, \mathbf{x}_n\right]$, for $n \geq 1$ together with $\mathbb{P}[\mathbf{x}_1 \in \cdot]$. We start with a representation-like theorem, that describes the finite dimensional distributions and the predictive distributions of a sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, driven by an species sampling process, it also characterizes the law of the exchangeable partition of $\mathbb{N}$, $(\boldsymbol{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$ and the conditional law of $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ given $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$.

Theorem 3.6. *Let $(\mathbf{x}_i)_{i\geq 1}$ be an random sequence, taking values in a Polish space $(S, \mathscr{B}(S))$, and for $n \geq 1$, define $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$ as the random partition of $[n]$ generated by the random equivalence relation $i \sim j$ if and only if $\mathbf{x}_i = \mathbf{x}_j$. Let μ_0 be a diffuse probability measure over $(S, \mathscr{B}(S))$ and let π be an infinite EPPF. The following are equivalent in terms of the law of $(\mathbf{x}_i)_{i\geq 1}$.*

I. *$(\mathbf{x}_i)_{i\geq 1}$ is exchangeable and directed by a species sampling process $\boldsymbol{\mu}$ as in (3.1), with base measure μ_0, and whose size-biased pseudo-permuted weights $(\tilde{\mathbf{w}}_j)_{j\geq 1}$ satisfy $\pi(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^k \tilde{\mathbf{w}}_j^{n_j - 1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^j \tilde{\mathbf{w}}_j\right)\right]$.*

II. *Given the size-biased pseudo-permuted weights sequence, $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_j)_{j\geq 1}$, such that $\pi(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^k \tilde{\mathbf{w}}_j^{n_j - 1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^j \tilde{\mathbf{w}}_j\right)\right]$, $\mathbf{x}_1 \sim \mu_0$, and for every $n \geq 1$,*

$$\mathbb{P}[\mathbf{x}_{n+1} \in \cdot \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \tilde{\mathbf{W}}] = \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_1 \delta_{\mathbf{x}_j^*} + \left(1 - \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j\right) \mu_0,$$

where $\mathbf{x}_1^*, \ldots, \mathbf{x}_{\mathbf{K}_n}^*$ are the distinct values that $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ exhibits in order of appearance. That is $\mathbf{x}_j^* = \mathbf{x}_{\mathbf{k}_j}$ for every $j \geq 1$, with $\mathbf{k}_1 = 1$ and for $j \geq 1$, $\mathbf{k}_{j+1} = \min\{i \geq 1 : \mathbf{x}_i \notin \{\mathbf{x}_{\mathbf{k}_1}, \ldots, \mathbf{x}_{\mathbf{k}_j}\}\}$.

III. $\mathbf{x}_1 \sim \mu_0$*, and for every* $n \geq 1$,

$$\mathbb{P}[\mathbf{x}_{n+1} \in \cdot \mid \mathbf{x}_1, \ldots, \mathbf{x}_n] = \sum_{j=1}^{\mathbf{K}_n} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})} \delta_{\mathbf{x}_j^*} + \frac{\pi\left(\mathbf{n}^{(\mathbf{K}_n+1)}\right)}{\pi(\mathbf{n})} \mu_0,$$

where $\mathbf{x}_1^*, \ldots, \mathbf{x}_{\mathbf{K}_n}^*$ *are the distinct values that* $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ *exhibits,* $\mathbf{n} = (\mathbf{n}_1, \ldots \mathbf{n}_{\mathbf{K}_n})$ *is given by* $\mathbf{n}_j = |\{i \leq n : \mathbf{x}_i = \mathbf{x}_j^*\}|$, $\mathbf{n}^{(j)} = (\mathbf{n}_1, \ldots \mathbf{n}_{j-1}, \mathbf{n}_j + 1, \mathbf{n}_{j+1}, \ldots, \mathbf{n}_{\mathbf{K}_n})$ *and* $\mathbf{n}^{(\mathbf{K}_n+1)} = (\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}_n}, 1)$.

IV. *The law of* $(\mathbf{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$ *is described by the infinite EPPF,* π*, and for every* $n \geq 1$ *and* $B_1, \ldots, B_n \in \mathscr{B}(S)$

$$\mathbb{P}\left[\mathbf{x}_1 \in B_1, \ldots, \mathbf{x}_n \in B_n \mid \mathbf{\Pi}(\mathbf{x}_{1:n})\right] = \prod_{i=1}^{\mathbf{K}_n} \mu_0 \left(\bigcap_{j \in \mathbf{\Pi}_i} B_j \right)$$

where $\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_{\mathbf{K}_n}$ *are the blocks of* $\mathbf{\Pi}(\mathbf{x}_{1:n})$.

V. *For every* $n \geq 1$*, and any* $x_1, \ldots, x_n \in S$,

$$\mathbb{P}\left[\mathbf{x}_1 \in dx_1, \ldots \mathbf{x}_n \in dx_n\right] = \pi(n_1, \ldots, n_k) \prod_{i=1}^{k} \mu_0(dx_j^*)$$

where $x_1^*, \ldots, x_k^*$ *are the distinct values in* $\{x_1, \ldots, x_n\}$*, and* $n_j = |\{i : x_i = x_j^*\}|$.

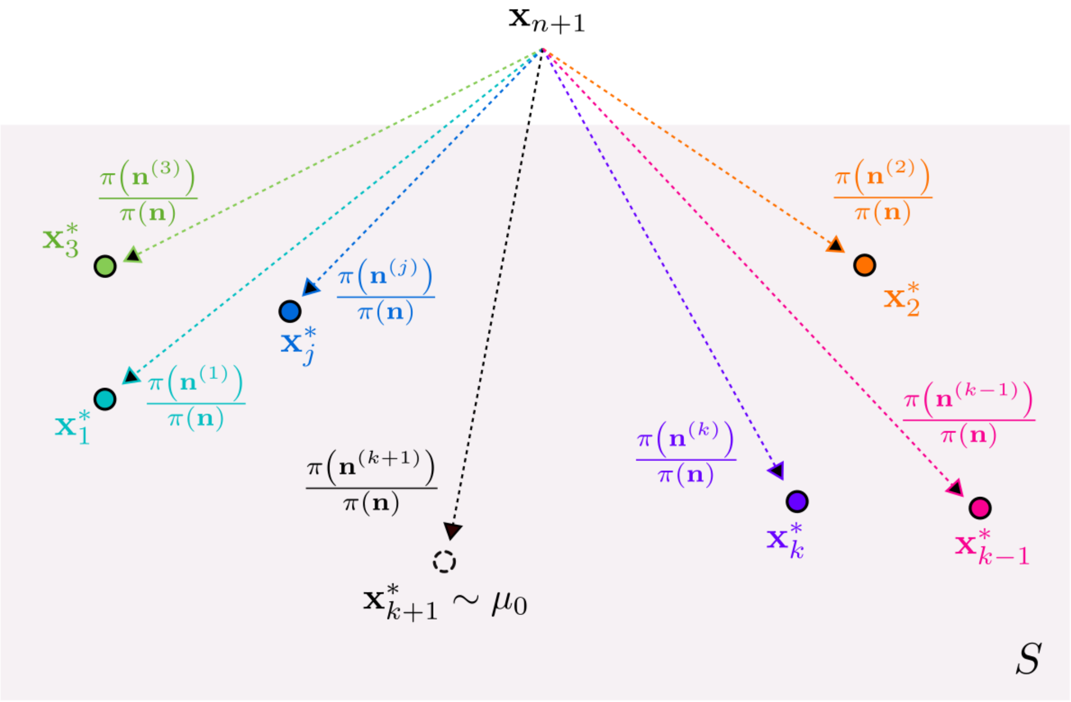

$\mathbf{x}_{n+1}$
$\frac{\pi(\mathbf{n}^{(3)})}{\pi(\mathbf{n})}$
$\mathbf{x}_3^*$
$\frac{\pi(\mathbf{n}^{(j)})}{\pi(\mathbf{n})}$
$\mathbf{x}_j^*$
$\frac{\pi(\mathbf{n}^{(1)})}{\pi(\mathbf{n})}$
$\mathbf{x}_1^*$
$\frac{\pi(\mathbf{n}^{(2)})}{\pi(\mathbf{n})}$
$\mathbf{x}_2^*$
$\frac{\pi(\mathbf{n}^{(k-1)})}{\pi(\mathbf{n})}$
$\frac{\pi(\mathbf{n}^{(k)})}{\pi(\mathbf{n})}$
$\frac{\pi(\mathbf{n}^{(k+1)})}{\pi(\mathbf{n})}$
$\mathbf{x}_k^*$
$\mathbf{x}_{k-1}^*$
$\mathbf{x}_{k+1}^* \sim \mu_0$
S

Figure 17: Prediction rule for an exchangeable sequence driven by a SSP. $\mathbf{K}_n = k$ denotes the number of distinct values that $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ exhibits, $\mathbf{x}_1^*, \ldots, \mathbf{x}_k^*$ are such distinct values, with corresponding frequencies $\mathbf{n} = (\mathbf{n}_1, \ldots \mathbf{n}_k)$, so that $\mathbf{n}_j = |\{i \leq n : \mathbf{x}_i = \mathbf{x}_j^*\}|$. We also denote $\mathbf{n}^{(j)} = (\mathbf{n}_1, \ldots \mathbf{n}_{j-1}, \mathbf{n}_j + 1, \mathbf{n}_{j+1}, \ldots, \mathbf{n}_k)$ and $\mathbf{n}^{(k+1)} = (\mathbf{n}_1, \ldots, \mathbf{n}_k, 1)$.

The proof of Theorem 3.6 can be found in Appendix C.4. Roughly speaking, the equivalence between I and III in Theorem 3.6, explains that for $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, the prediction rule of the sequence is given by a step in chinese restaurant process, with a determined prediction rule, and where the labels of the tables have been randomized independently according to μ_0. Indeed, if $\mathbf{x}_1^*, \mathbf{x}_2^* \ldots$ are the distinct values of $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$, in order of appearance, we might think that $\mathbf{x}_j^*$ represents the jth table to be open in a chinese restaurant process (see Figure 17 and compare it with Figure 10, recalling that in Section 2.2 we used the notation $\pi(j \mid \mathbf{n}) = \pi(\mathbf{n}^{(j)})/\pi(\mathbf{n})$). Analogously, the equivalence between I and II in Theorem 3.6 explains that $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$ may be constructed through the chinese restaurant process with random seating plan driven by a size-biased pseudo-permutation, and randomizing the labels of the tables according to μ_0. Theorem IV 3.6 explains that the law of the partition $\boldsymbol{\Pi} = (\boldsymbol{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$ is given by the infinite EPPF corresponding to the prediction rule in II and the expectation, $\pi(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \tilde{\mathbf{w}}_j^{n_j-1} \prod_{j=1}^{k-1}\left(1 - \sum_{i=1}^{j} \tilde{\mathbf{w}}_j\right)\right]$ in I and III, and that, conditionally given $\boldsymbol{\Pi}(\mathbf{x}_{1:n}) = \{\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_{\mathbf{K}_n}\}$, we specify $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by sampling $\mathbf{x}_1^*, \ldots, \mathbf{x}_{\mathbf{K}_n}^*$ independently from μ_0, and setting $\mathbf{x}_i = \mathbf{x}_j^*$ if and only if $i \in \boldsymbol{\Pi}_j$. That is to say, conditionally given $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$, the random vector $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ distributes as $\left(\mathbf{x}_{l_1}^*, \ldots, \mathbf{x}_{l_n}^*\right)$ with $l_r = j$ if and only if $r \in \boldsymbol{\Pi}_j$. For example, say that for some realization $\boldsymbol{\Pi}(\mathbf{x}_{1:6}) = \{\{1,4,5\},\{2,3\},\{6\}\}$, then under such event, $(\mathbf{x}_1, \ldots, \mathbf{x}_6)$ distributes as $(\mathbf{x}_1^*, \mathbf{x}_2^*, \mathbf{x}_2^*, \mathbf{x}_1^*, \mathbf{x}_1^*, \mathbf{x}_3^*)$, where $\{\mathbf{x}_1^*, \mathbf{x}_2^*, \mathbf{x}_3^*\} \overset{\text{iid}}{\sim} \mu_0$ independently of $\boldsymbol{\Pi}(\mathbf{x}_{1:6})$. With this in mind, the proof of the following Theorem becomes quite simple (see Appendix C.5 for details).

Theorem 3.7. *Let $(\mathbf{x}_i)_{i\geq 1}$ be an exchangeable sequence driven by a SSP, $\boldsymbol{\mu}$, with base measure, μ_0, and corresponding EPPF, π. Fix $n \geq 1$ and let $f : S^n \to \mathbb{R}$ be measurable function, then*

$$\begin{aligned}&\mathbb{E}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_n)\right] \\ &\quad = \sum_{A \in \mathcal{P}_{[n]}} \left\{ \int f(x_{l_1}, \ldots, x_{l_n}) \prod_{j=1}^{k} \prod_{r \in A_j} \mathbf{1}_{\{l_r = j\}}\, \mu_0(dx_1) \ldots \mu_0(dx_k) \right\} \pi(|A_1|, \ldots, |A_k|),\end{aligned} \tag{3.2}$$

whenever the integrals in the right side exist, and where, $k = |A|$ and $A_1, \ldots, A_k$ stand for the blocks of $A \in \mathcal{P}_{[n]}$. Moreover, if f is symmetric (and the integrals exist), Equation (3.2) reduces to

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_1, \ldots, \mathbf{x}_n)\right] = \sum_{k=1}^{n} \sum_{(m_1, \ldots, m_n) \in \mathcal{M}_n^k} & \frac{n!}{\prod_{i=1}^{n} (i!)^{m_i} (m_i!)} \pi(n_1, \ldots, n_k) \\ & \times \int f\left(x_{[n_1, \ldots, n_k]}\right) \mu_0(dx_1) \ldots \mu_0(dx_k),\end{aligned} \tag{3.3}$$

where $\mathcal{M}_n^k = \{(m_1, \ldots, m_n) \in \mathbb{Z}_+^n : \sum_{i=1}^{n} m_i = k, \sum_{i=1}^{n} i m_i = n\}$, and for $(m_1, \ldots, m_n) \in \mathcal{M}_n^k$ and $x = (x_1, \ldots, x_k) \in S^k$, $(n_1, \ldots, n_k)$ denotes the ranked composition of n into k parts such that $m_i = \sum_{j=1}^{k} \mathbf{1}_{\{n_j = i\}}$, and $x_{[n_1, \ldots, n_k]}$ denotes the vector of size n with the first n_1 entries equal to x_1, the next n_2 entries equal to x_2 and so on.

Example 3.1. *To make clearer the notation of Theorem 3.7 let us consider the case* $n = 3$, *so that* $f : S^3 \to \mathbb{R}$. *Note that the set of all partitions of* $[3]$ *is*

$$\mathcal{P}_{[3]} = [(\{1,2,3\}), (\{1\},\{2,3\}), (\{2\},\{1,3\}), (\{3\},\{1,2\}), (\{1\},\{2\},\{3\})] .$$

For $(\mathbf{x}_i^*)_{i=1}^3 \overset{iid}{\sim} \mu_0$ *we have that under the event* $(\mathbf{\Pi}(\mathbf{x}_{1:3}) = \{1,2,3\})$, $f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ *distributes like* $f(\mathbf{x}_1^*,\mathbf{x}_1^*,\mathbf{x}_1^*)$; *conditioning on* $(\mathbf{\Pi}(\mathbf{x}_{1:3}) = \{\{1\},\{2,3\}\})$, $f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ *is equal in distribution to* $f(\mathbf{x}_1^*,\mathbf{x}_2^*,\mathbf{x}_2^*)$; *given* $(\mathbf{\Pi}(\mathbf{x}_{1:3}) = \{\{2\},\{1,3\}\})$, $f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ *distributes identically as* $f(\mathbf{x}_1^*,\mathbf{x}_2^*,\mathbf{x}_1^*)$; *conditionally given* $(\mathbf{\Pi}(\mathbf{x}_{1:3}) = \{\{3\},\{1,2\}\})$, $f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$ *distributes like* $f(\mathbf{x}_1^*,\mathbf{x}_1^*,\mathbf{x}_2^*)$; *and conditioning on* $(\mathbf{\Pi}(\mathbf{x}_{1:3}) = \{\{1\},\{2\},\{3\}\})$, $f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) \overset{d}{=} f(\mathbf{x}_1^*,\mathbf{x}_2^*,\mathbf{x}_3^*)$. *This said, it is clear that*

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)\right] \\
&= \pi(3)\int f(x_1^*,x_1^*,x_1^*)\mu_0(dx_1^*) + \pi(1,2)\int f(x_1^*,x_2^*,x_2^*)\mu_0(dx_1^*)\mu_0(dx_2^*) \\
&+ \pi(1,2)\int f(x_1^*,x_2^*,x_1^*)\mu_0(dx_1^*)\mu_0(dx_2^*) + \pi(1,2)\int f(x_1^*,x_1^*,x_2^*)\mu_0(dx_1^*)\mu_0(dx_2^*) \\
&+ \pi(1,1,1)\int f(x_1^*,x_2^*,x_3^*)\mu_0(dx_1^*)\mu_0(dx_2^*)\mu_0(dx_3^*).
\end{aligned}$$

Furthermore, if f *is a symmetric function of its arguments, we obtain that*

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)\right] \\
&= \pi(3)\int f(x_1^*,x_1^*,x_1^*)\mu_0(dx_1^*) + 3\left\{\pi(1,2)\int f(x_1^*,x_2^*,x_2^*)\mu_0(dx_1^*)\mu_0(dx_2^*)\right\} \\
&\qquad + \pi(1,1,1)\int f(x_1^*,x_2^*,x_3^*)\mu_0(dx_1^*)\mu_0(dx_2^*)\mu_0(dx_3^*),
\end{aligned}$$

whenever the integrals exist.

For $n = 2$ we can write (3.2) in terms of the tie probability of the species sampling process. Indeed, for $\{\mathbf{x}_1,\mathbf{x}_2,\ldots \mid \boldsymbol{\mu}\} \sim \boldsymbol{\mu}$, with $\boldsymbol{\mu}$ a SSP with base measure μ_0 and corresponding EPPF π. The event $(\mathbf{x}_1 = \mathbf{x}_2)$ is identical to $(\mathbf{\Pi}(\mathbf{x}_{1:2}) = \{\{1,2\}\})$, so in terms of the EPPF, we can express the tie probability of $\boldsymbol{\mu}$ as, $\rho = \pi(2) = \mathbb{E}\left[\tilde{\mathbf{w}}_1\right]$, so that $1-\rho = \pi(1,1) = 1-\mathbb{E}\left[\tilde{\mathbf{w}}_1\right]$, where $\tilde{\mathbf{w}}_1$ is size-biased pick of the weights of $\boldsymbol{\mu}$. Thus,

$$\mathbb{E}[f(\mathbf{x}_1,\mathbf{x}_2)] = \rho\int f(s,s)\mu_0(ds) + (1-\rho)\int f(s_1,s_2)\mu_0(ds_1)\mu_0(ds_2).$$

Another quantity that can be written in terms of the tie probability is the prediction rule in Theorem 3.6 II, for $n = 1$,

$$\mathbb{P}[\mathbf{x}_2 \in \cdot \mid \mathbf{x}_1] = \frac{\pi(2)}{\pi(1)}\delta_{\mathbf{x}_1} + \frac{\pi(1,1)}{\pi(1)}\mu_0 = \rho\,\delta_{\mathbf{x}_1} + (1-\rho)\mu_0.$$

Noticing that the exchangeability of $(\mathbf{x}_i)_{i\geq 1}$ implies $(\mathbf{x}_1,\mathbf{x}_2) \overset{d}{=} (\mathbf{x}_i,\mathbf{x}_j)$ for every $i \neq j$, we trivially obtain the following corollary of Theorems 3.6 and 3.7.

Corollary 3.8. *Let* $\boldsymbol{\mu}$ *be a SSP with base measure* μ_0 *and tie probability* ρ. *Consider* $\{\mathbf{x}_1,\mathbf{x}_2,\ldots \mid \boldsymbol{\mu}\} \sim \boldsymbol{\mu}$. *Then, for every* $i \neq j$

a) $\mathbb{P}[\mathbf{x}_j \in \cdot \mid \mathbf{x}_i] = \rho\,\delta_{\mathbf{x}_i} + (1-\rho)\mu_0$.

b) *Conditioning on* $\mathbf{x}_i \neq \mathbf{x}_j$*,* $(\mathbf{x}_i, \mathbf{x}_j) \overset{iid}{\sim} \mu_0$.

c) *For any measurable function,* $f : S^2 \to \mathbb{R}$*, such that* $\int f(s_1, s_2)\mu_0(ds_1)\mu_0(ds_2)$ *and* $\int f(s,s)\mu_0(ds)$ *exist, we have that*

$$\mathbb{E}[f(\mathbf{x}_i, \mathbf{x}_j)] = \rho \int f(s,s)\mu_0(ds) + (1-\rho)\int f(s_1,s_2)\mu_0(ds_1)\mu_0(ds_2).$$

From Corollary 3.8, the following (conditional) moments are easy to compute, for details see Appendix C.6.

Corollary 3.9. *Let* $\boldsymbol{\mu}$ *be a SSP with base measure* μ_0 *and tie probability* ρ*. Consider* $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \sim \boldsymbol{\mu}$*. Then, for every* $i \neq j$

a) $\mathbb{E}[\mathbf{x}_j \mid \mathbf{x}_i] = \rho\,\mathbf{x}_i + (1-\rho)\mathbb{E}[\mathbf{x}_i]$

b) $\mathsf{Var}(\mathbf{x}_j \mid \mathbf{x}_i) = (1-\rho)\left\{\rho\left(\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i]\right)^2 + \mathsf{Var}(\mathbf{x}_i)\right\}$

c) $\mathsf{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \rho\,\mathsf{Var}(\mathbf{x}_i)$

d) $\mathsf{Corr}(\mathbf{x}_i, \mathbf{x}_j) = \rho$.

In the context of Corollary 3.9, for small values of ρ, $\mathbb{E}[\mathbf{x}_j \mid \mathbf{x}_i] \approx \mathbb{E}[\mathbf{x}_j]$, $\mathsf{Var}(\mathbf{x}_j \mid \mathbf{x}_i) \approx \mathsf{Var}(\mathbf{x}_j)$ and $\mathsf{Cov}(\mathbf{x}_i, \mathbf{x}_j) \approx 0$, alternatively for values of ρ close to 1, $\mathbb{E}[\mathbf{x}_j \mid \mathbf{x}_i] \approx \mathbf{x}_i$, $\mathsf{Var}(\mathbf{x}_j \mid \mathbf{x}_i) \approx 0$ and $\mathsf{Cov}(\mathbf{x}_i, \mathbf{x}_j) \approx \mathsf{Var}(\mathbf{x}_i)$. In fact for exchangeable sequences driven be SSPs we have the following version of Theorem 3.4.

Theorem 3.10. *Let* $(S, \mathscr{B}_S)$ *be a Polish spaces. For each* $n \geq 1$*, let* $\mathbf{X}^{(n)} = \left(\mathbf{x}_i^{(n)}\right)_{i \geq 1}$ *be an exchangeable sequence taking values in* $(S, \mathscr{B}_S)$ *and driven by a SSP,* $\boldsymbol{\mu}^{(n)}$*, with tie probability* $\rho^{(n)}$ *and base measure* $\mu_0^{(n)}$*. Say that as* $n \to \infty$*,* $\mu_0^{(n)}$ *converges weakly to* μ_0*.*

i) *If* $\rho^{(n)} \to 0$*, as* $n \to \infty$*, then* $\mathbf{X}^{(n)}$ *converges in distribution to* $(\mathbf{x}_i)_{i \geq 1} \overset{iid}{\sim} \mu_0$*.*

ii) *If* $\rho^{(n)} \to 1$*, as* $n \to \infty$*, then* $\mathbf{X}^{(n)}$ *converges in distribution to a sequence of identical random variables* $(\mathbf{x}, \mathbf{x}, \ldots)$*, where* $\mathbf{x} \sim \mu_0$*.*

The proof of Theorem 3.10 can be found in Appendix C.7, this result should already give us an idea of how flexible the class of SSPs is. At the beginning of Section 2, we mentioned that the two extrema of the mutual dependence between elements of an exchangeable sequence are complete independence, and identical random elements. These coincide with the limits of an exchangeable sequence driven by a SSP, as the tie probability approaches zero or one, respectively. Furthermore, from Corollary 3.9 (d) we get that using SSPs as drivers, for every $\rho \in (0,1)$, we can construct an exchangeable sequence with such correlation coefficient.

3.4 Full support

From Bayesian non-parametric perspective, a desirable property of the prior is that its support is as large as possible. In Section 1.4.3 we defined the notion of weak support of a random probability measure and in Proposition 1.19 we established and upper bound for it. For a species sampling process $\boldsymbol{\mu}$ over the Borel space $(S, \mathscr{B}_S)$, where S is Polish, we get that its weak topological support $\mathbb{WS}(\boldsymbol{\mu}) \subseteq \{\varphi \in \mathcal{P}(S) : \mathbb{S}(\varphi) \subseteq \mathbb{S}(\mu_0)\}$, where μ_0 is the base measure of $\boldsymbol{\mu}$. This motivates the following definition.

Definition 3.4. *Let $(S, \mathscr{B}_S)$ be a Borel space where S is Polish. A species sampling process $\boldsymbol{\mu}$, with base measure μ_0 is said to have full support, whenever $\mathbb{WS}(\boldsymbol{\mu}) = \{\varphi \in \mathcal{P}(S) : \mathbb{S}(\varphi) \subseteq \mathbb{S}(\mu_0)\}$. In this instance, its prior distribution is also said to have full support.*

In particular if the support of μ_0 is the whole space S, and $\boldsymbol{\mu}$ has full support, this assures $\mathbb{WS}(\boldsymbol{\mu}) = \mathcal{P}(S)$. As the next result explains, whether a species sampling process has full support or not is completely determined by the weights sequence. In effect, while the base measure sets up a candidate for the weak support of the species sampling process, the weights sequence dictate if the support fulfills the proposal.

Theorem 3.11. *Consider a species sampling process $\boldsymbol{\mu}$ as in (3.1) with distribution Q. The following are equivalent.*

I. Q *(equiv. $\boldsymbol{\mu}$) has full support.*

II. *For every $\varepsilon > 0$, and every $0 < \gamma < 1$ $\mathbb{P}\left[\max_{j\geq 1} \overline{\mathbf{w}}_j < \varepsilon, \sum_{j\geq 1} \mathbf{w}_j > \gamma\right] > 0$, where $\overline{\mathbf{w}}_j = \mathbf{w}_j / \sum_{j\geq 1} \mathbf{w}_j$.*

III. *For every $\varepsilon > 0$ there exists an integer m such that*

$$\mathbb{P}\left[\mathbf{w}_1 < \varepsilon, \ldots, \mathbf{w}_m < \varepsilon, \sum_{j=1}^{m} \mathbf{w}_j > 1 - \varepsilon\right] > 0.$$

Theorem 3.11 was proven by Bissiri and Ongaro (2014), for the sake of completeness we replicate their proof in Appendix C.8. For a proper species sampling processes, the weights, $(\mathbf{w}_j)_{j\geq 1}$, sum up to one almost surely, hence $\sum_{j\geq 1} \mathbf{w}_j > \gamma$ almost surely, for every $0 < \gamma < 1$ and $\overline{\mathbf{w}}_j = \mathbf{w}_j$ almost surely, for every $j \geq 1$. Thus the following corollary is straight-forward from Theorem 3.11.

Corollary 3.12. *A proper species sampling process, $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$, has full support if and only if for every $\varepsilon > 0$, $\mathbb{P}\left[\max_{j\geq 1} \mathbf{w}_j < \varepsilon\right] > 0$.*

3.5 Constructions

Although it is possible to derive structural properties for the general class of SSPs, from an operational viewpoint, it is vacuous as long as the distributions of these random probability measures remain unspecified. The first obvious way to determine the distribution of a SSP, $\boldsymbol{\mu}$, over $(S, \mathscr{B}_S)$, is through its finite dimensional distributions. That is, through the specification of the law of $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$ for every collection of disjoint sets, $B_1, \dots, B_n$. For some SSPs (see Section 3.6.1) it is possible to characterize the distributions of these vectors, but in general it this is very hard to do. To overcome this hurdle researchers have developed various methods to specify a prior distribution. Essentially, all of this methods exploit the atomic decomposition of a SSP, as described in Definition 3.2, and determine the prior by means of choosing a base measure and then constructing a distribution over $\overline{\Delta}_\infty$, which is the mathematically challenging task. Below we describe some of the most famous constructions.

3.5.1 Construction through a prediction rule

We have already described this construction indirectly. Indeed, given an infinite EPPF, π, and a diffuse probability measure, μ_0, over $(S, \mathscr{B}_S)$, we can build the exchangeable sequence, $(\mathbf{x}_i)_{i\geq 1}$, such that $\mathbf{x}_1 \sim \mu_0$, and for $n \geq 1$

$$\mathbb{P}[\mathbf{x}_{n+1} \in \cdot \mid \mathbf{x}_1, \dots, \mathbf{x}_n] = \sum_{j=1}^{\mathbf{K}_n} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})} \delta_{\mathbf{x}_j^*} + \frac{\pi\left(\mathbf{n}^{(\mathbf{K}_n+1)}\right)}{\pi(\mathbf{n})} \mu_0, \tag{3.4}$$

where $\mathbf{x}_1^*, \dots, \mathbf{x}_{\mathbf{K}_n}^*$ are the distinct values that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ exhibits, with corresponding frequencies $\mathbf{n} = (\mathbf{n}_1, \dots \mathbf{n}_{\mathbf{K}_n})$, and where $\mathbf{n}^{(j)} = (\mathbf{n}_1, \dots \mathbf{n}_{j-1}, \mathbf{n}_j + 1, \mathbf{n}_{j+1}, \dots, \mathbf{n}_{\mathbf{K}_n})$ and $\mathbf{n}^{(\mathbf{K}_n+1)} = (\mathbf{n}_1, \dots, \mathbf{n}_{\mathbf{K}_n}, 1)$. Theorem 3.6 together with Proposition 2.2 yield $\boldsymbol{\mu} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ is a species sampling process. The base measure of $\boldsymbol{\mu}$ is precisely the distribution of $\mathbf{x}_1$, μ_0, and as a consequence of Kingman's correspondence (see Theorem 2.10 or Proposition 2.11) π determines uniquely, up to permutations, the distribution of the weights. Furthermore, the correspondence between the EPPF and the distribution of the weights is made explicit through $\pi(n_1, \dots, n_k) = \mathbb{E}\left[\prod_{j=1}^k \tilde{\mathbf{w}}_j^{n_j-1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^j \tilde{\mathbf{w}}_j\right)\right]$, where $(\tilde{\mathbf{w}}_j)_{j\geq 1}$ is a size-biased pseudo-permutation. As to the tie probability of $\boldsymbol{\mu}$, it is given by $\rho = \pi(2)$. Clearly, every species sampling process can be constructed using this method.

An enormous advantage of this construction, specially from a practical perspective, is that for $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, the laws of $(\mathbf{x}_i)_{i\geq 1}$ and $(\boldsymbol{\Pi}(\mathbf{x}_{1:n}))_{n\geq 1}$ are automatically specified analytically. So, for example, if one wants to draw samples from this exchangeable sequence one could simply sample $\mathbf{x}_1 \sim \mu_0$ and sequentially $\mathbf{x}_2, \mathbf{x}_3, \dots$ from (3.4), without the need of sampling $\boldsymbol{\mu}$, which is determined by infinitely many random variables. A small drawback of this method is that, while it is true that the distribution of the weights is indirectly determined by the infinite EPPF, we may not be able to attain the law of $(\mathbf{w}_j)_{j\geq 1}$, explicitly. Hence, some structural properties, such as whether or not the SSP is proper or the prior has full support, can be rather complicated to corroborate.

3.5.2 Normalization of a random measure with independent increments

Recall from Section 1.3.3 that for a diffuse, locally finite measure, λ, over $(S, \mathscr{B}_S)$, a λ-homogeneous completely random measure, $\boldsymbol{\lambda}$, is one that can be decomposed as $\boldsymbol{\lambda} = \sum_{j\geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j} + c\lambda$, where c is some non-negative constant and $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ defines a Poisson process over $S \times \mathbb{R}_+$ with intensity $\lambda \otimes \varrho$, where ϱ satisfies, $\int_{\mathbb{R}_+} (x \wedge 1) \varrho(dx) < \infty$. Further, if λ is finite we get $\boldsymbol{\lambda}$ is finite almost surely, and if $c > 0$ or $\varrho(\mathbb{R}_+) = \int_{\mathbb{R}_+} \varrho(dx) = \infty$, we also have $\boldsymbol{\lambda}(S) > 0$, almost surely. Under these constrains we can define the random probability measure $\boldsymbol{\mu} = \boldsymbol{\lambda}/\boldsymbol{\lambda}(S)$. Now, for any disjoint measurable sets,

$B_1, \dots, B_n \in \mathscr{B}_S$ such that $\lambda(B_i) = \lambda(B_j)$, we have that $(\boldsymbol{\lambda}(B_1), \dots, \boldsymbol{\lambda}(B_n))$ forms an i.i.d. vector. When we normalize, $(\boldsymbol{\mu}(B_1), \dots, \boldsymbol{\mu}(B_n))$ losses the independence property, however it remains conditionally i.i.d. given $\boldsymbol{\lambda}(S)$. Thus $(\boldsymbol{\mu}(B_1), \dots, \boldsymbol{\mu}(B_n))$ is even exchangeable and from Definition 3.1 we obtain $\boldsymbol{\mu}$ is a SSP. Evidently, the base measure of $\boldsymbol{\mu}$ is given by $\mu_0 = \lambda/\lambda(S)$, and the distribution of the weights is indirectly characterized by ϱ. This kind of species sampling processes are also known as normalized random measures with independent increments (NRMI, for short Prünster; 2003; Regazzini et al.; 2003; James et al.; 2009). This construction can be regarded as the composition of two methods we have already described. In effect, for λ, ϱ and c as above, we can first build a partition, $\boldsymbol{\Pi}$, through the ordered paintbox (see Section 2.2.6) using a subordinator with intensity ϱ and drift c. Latter, given the EPPF, π, of $\boldsymbol{\Pi}$, and the diffuse random probability measure, $\mu_0 = \lambda/\lambda(S)$, construct the species sampling process $\boldsymbol{\mu}$ as in Section 3.5.1.

It is easy to see that an NRMI, $\boldsymbol{\mu}$, is proper if and only if $c = 0$. Indeed, the weights of $\boldsymbol{\mu}$, $(\mathbf{w}_j)_{j\geq 1}$, are precisely the normalized jumps of $\boldsymbol{\lambda}$, $(\boldsymbol{\alpha}_j/\boldsymbol{\lambda}(S))_{j\geq 1}$, and we have that $c = 0$ if and only if $\boldsymbol{\lambda}(S) = \sum_{j\geq 1} \boldsymbol{\alpha}_j$, which in turn is equivalent to

$$\sum_{j\geq 1} \mathbf{w}_j = \frac{1}{\boldsymbol{\lambda}(S)} \sum_{j\geq 1} \boldsymbol{\alpha}_j = 1.$$

For a proper NRMI, $\boldsymbol{\mu}$, the almost sure representation,

$$\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j^{\downarrow} \delta_{\boldsymbol{\xi}_j} = \sum_{j\geq 1} \frac{\boldsymbol{\alpha}_j^{\downarrow}}{\boldsymbol{\lambda}(S)} \delta_{\boldsymbol{\xi}_j},$$

where $\left(\mathbf{w}_j^{\downarrow}\right)_{j\geq 1}$ and $\left(\boldsymbol{\alpha}_j^{\downarrow}\right)_{j\geq 1}$ are the decreasing rearrangement of the weights and the jumps respectively, is also termed a Poisson-Kingman random probability measure. Using Corollary 3.12 and the Poisson-Kingman representation, Bissiri and Ongaro (2014) showed that a sufficient condition for a proper NRMI to have full support is that the distribution of $\boldsymbol{\lambda}(S)$ is absolutely continuous with respect to the Lebesgue measure.

As to the EPPF, if it is not already available as in Section 3.5.1, for this or other constructions, it is generally very hard to attain. For proper NRMI's there are methods to derive an expression for the EPPF and the tie probability (see for instance Lijoi et al.; 2005; James et al.; 2009). However, more often than not, these expressions are rather complicated, and some times not even in a closed form.

This method to define a prior is one of the most widely studied in literature. As for each known law of a subordinator and a finite measure over a Borel space, we can define the law of a SSP through this construction. Moreover, if we drop the hypothesis of the completely random measure being homogeneous, this method enables the characterization of laws of random probability measure with a more complex structure than a SSPs, that is, where the weights are not necessarily independent of the atoms. Despite this, a small disadvantage of this construction is that it is not exhaustive in the class of SSPs, meaning that there are SSPs that are not NRMI's.

3.5.3 Stick-breaking construction

In contrast to the previous methods to define a prior, this one consists in defining law of the atoms $(\boldsymbol{\xi}_j)_{j\geq 1}$ and that of the weights sequence $(\mathbf{w}_j)_{j\geq 1}$ directly. This is, choosing a the base measure, μ_0, over a Borel space $(S, \mathscr{B}_S)$, and construct a distribution over $(\overline{\Delta}_\infty, \mathscr{B}_{\overline{\Delta}_\infty})$, where $\overline{\Delta}_\infty = \{(w_1, w_2, \dots) \in [0,1]^\infty : \sum_{j\geq 1} w_j \leq 1\}$. Now, as the weights satisfy $\sum_{j\geq 1} \mathbf{w}_j \leq 1$, almost surely, they can not be i.i.d., exchangeable or even Markovian, unless they are deterministic, which makes notorious how challenging it can be to define distributions over $\overline{\Delta}_\infty$. Essentially, the stick-breaking construction translates the problem of defining a distribution over $\overline{\Delta}_\infty$ into the seemingly simpler one of defining a distribution over $[0,1]^\infty$, as explained in what follows.

Consider a stick of length one and a sequence of $[0,1]$-valued random variables, $(\mathbf{v}_i)_{i\geq 1}$. The stick breaking construction consists in sequentially and proportionally cutting the stick according to $(\mathbf{v}_i)_{i\geq 1}$. At the first step we are going to cut the stick into two parts, one of length $\mathbf{w}_1 = \mathbf{v}_1$ and the remainder of length $(1 - \mathbf{w}_1) = (1 - \mathbf{v}_1)$. Secondly, the leftover of length $(1 - \mathbf{v}_1)$, will be cut again into two parts, one proportional to $\mathbf{v}_2$, obtaining a stick of length $\mathbf{w}_2 = \mathbf{v}_2(1 - \mathbf{v}_1)$, and the remainder will then have length $1 - (\mathbf{w}_1 + \mathbf{w}_2) = (1 - \mathbf{v}_2)(1 - \mathbf{v}_1)$, see Figure 18 for an illustration. Continuing inductively, after the jth step, we will have j sub-sticks of lengths, $\mathbf{w}_1, \dots, \mathbf{w}_j$, where $\mathbf{w}_j = \mathbf{v}_j \prod_{i=1}^{j-1}(1 - \mathbf{v}_i)$, along with a remainder of length $1 - \sum_{j\geq 1} \mathbf{w}_j = \prod_{i=1}^{j}(1 - \mathbf{v}_i)$. By construction it is clear that $0 \leq \mathbf{w}_j \leq 1$ and $\sum_{j\geq 1} \mathbf{w}_j \leq 1$, hence the sequence $(\mathbf{w}_j)_{j\geq 1}$ takes values in $\overline{\Delta}_\infty$ and its distribution is completely characterized by that of $(\mathbf{v}_i)_{i\geq 1}$.

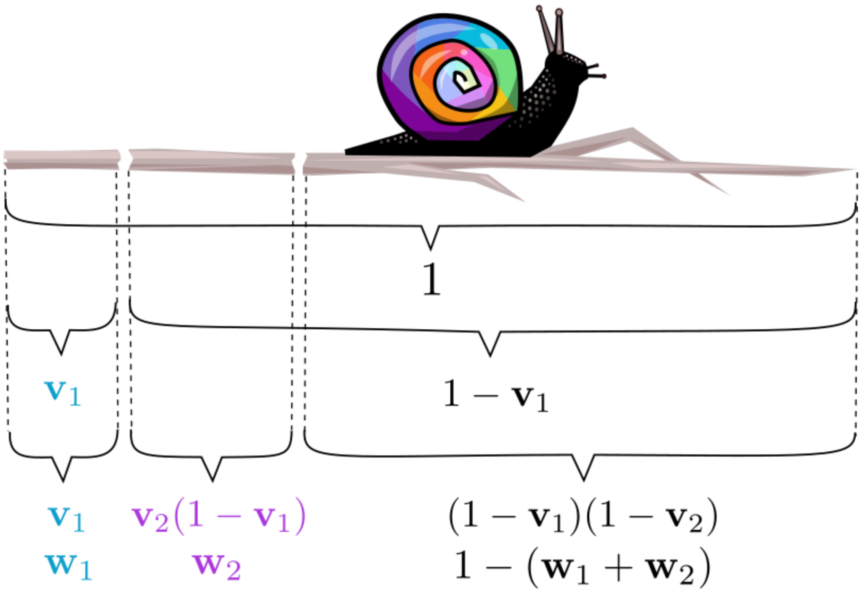

1
$\mathbf{v}_1$
$1 - \mathbf{v}_1$
$\mathbf{v}_1$
$\mathbf{v}_2(1 - \mathbf{v}_1)$
$(1 - \mathbf{v}_1)(1 - \mathbf{v}_2)$
$\mathbf{w}_1$
$\mathbf{w}_2$
$1 - (\mathbf{w}_1 + \mathbf{w}_2)$

Figure 18: First steps of the stick-breaking construction and an improbable snail.

The first thing to note is that any sequence of weights can be constructed this way, formally we have the following Proposition.

Proposition 3.13. *Let* $(\mathbf{w}_j)_{j\geq 1}$ *be a sequence of random variables that satisfy* $0 \leq \mathbf{w}_j \leq 1$, *for every* $j \geq 1$, *and* $\sum_{j\geq 1} \mathbf{w}_j \leq 1$, *almost surely. Then there exist a sequence of* $[0,1]$*-valued random variables,* $(\mathbf{v}_i)_{i\geq 1}$ *such that for every* $j \geq 1$, $\mathbf{w}_j = \mathbf{v}_j \prod_{i=1}^{j-1}(1 - \mathbf{v}_i)$, *almost surely.*

The proof of Proposition 3.13 can be found in Appendix C.9. A straight-forward consequence of this result is that the stick-breaking construction is exhaustive in the class of SSPs, meaning that the weights of any SSP can be stick-breaking constructed. Due to how much we will be exploiting this method to define a prior, we will dedicate it a formal definition.

Definition 3.5. *We call a stick-breaking process to any SSP, $\boldsymbol{\mu}$, as in (3.1), such that its weights, $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$, are decomposed as*

$$\mathbf{w}_1 = \mathbf{v}_1, \quad \mathbf{w}_j = \mathbf{v}_j \prod_{i=1}^{j-1} (1 - \mathbf{v}_i), \quad j \geq 2, \tag{3.5}$$

for some sequence, $\mathbf{V} = (\mathbf{v}_i)_{i\geq 1}$, of $[0,1]$-valued random variables. In this case we write $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$, or equivalently $\mathbf{W} = \mathsf{SB}[\mathbf{V}]$. To the weights, $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$, we call stick-breaking weights and to the the elements of $\mathbf{V} = (\mathbf{v}_i)_{i\geq 1}$ we call proportional length variables (or simply length variables) of $\boldsymbol{\mu}$ (or $\mathbf{W}$).

An important remark here is that in terms of the law of a SSP, the stick-breaking construction is not unique. This is a consequence of the fact that the prior is invariant under weights' permutations (see Proposition 3.1) and that the stick-breaking is an almost sure decomposition. In other words, to each distribution of a sequence of length variables over $[0,1]^\infty$, there corresponds a unique prior, but the converse is not true. To make this clearer consider the following example.

Example 3.2. *Fix $\mathbf{v}_1 = \mathbf{v}_2 = 1/2, \mathbf{v}_3 = 1$ and $\mathbf{v}_i = 0$ for every $i \geq 4$, also define $\mathbf{v}'_1 = 0, \mathbf{v}'_2 = 1/4, \mathbf{v}'_3 = 2/3, \mathbf{v}'_4 = 1$, and $\mathbf{v}'_i = 0$ for every $i \geq 5$. Trivially $(\mathbf{v}_i)_{i\geq 1} \overset{d}{\neq} (\mathbf{v}'_i)_{i\geq 1}$. Now, set $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ and $(\mathbf{w}'_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}'_i)_{i\geq 1}]$. It is easy to see $\mathbf{w}_1 = 1/2, \mathbf{w}_2 = \mathbf{w}_3 = 1/4$ and $\mathbf{w}_j = 0$ for every $j \geq 4$, also $\mathbf{w}'_1 = 0, \mathbf{w}_2 = \mathbf{w}_4 = 1/4, \mathbf{w}_3 = 1/2$ and $\mathbf{w}_j = 0$ for $j \geq 5$. Thus, $(\mathbf{w}'_j)_{j\geq 1}$ is a permutation of $(\mathbf{w}_j)_{j\geq 1}$, and clearly for every independent i.i.d. sequence $(\boldsymbol{\xi}_j)_{j\geq 1}$ we get $\sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} \overset{d}{=} \sum_{j\geq 1} \mathbf{w}'_j \delta_{\boldsymbol{\xi}_j}$.*

As the following result shows, important structural properties of a SSP and the corresponding prior can be restated in terms of the law of the length variables.

Lemma 3.14. *Let $\boldsymbol{\mu}$ be a species sampling process with stick-breaking weights $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$.*

i.a) *$\boldsymbol{\mu}$ is proper if and only if $\lim_{j\to\infty} \mathbb{E}\left[\prod_{i=1}^{j} (1 - \mathbf{v}_i)\right] = 0$.*

i.b) *$\boldsymbol{\mu}$ is proper if and only if*

$$\mathbb{P}\left[\left\{\sum_{i\geq 1} \mathbf{v}_i = \infty\right\} \cup \left\{\bigcup_{i\geq 1} [\mathbf{v}_i = 1]\right\}\right] = 1.$$

In particular, if $\sum_{i\geq 1} \mathbf{v}_i = \infty$ almost surely, we get $\boldsymbol{\mu}$ is proper.

ii) *If for every $\varepsilon > 0$ there exist $0 < \gamma < \varepsilon$ such that $\mathbb{P}\left[\bigcap_{i=1}^{n} (\gamma < \mathbf{v}_i < \varepsilon)\right] > 0$, for every $n \geq 1$, then $\boldsymbol{\mu}$ has full support.*

The proof of Lemma 3.14 appears in Appendix C.10, this result is extremely useful from an operational point of view (see Section 3.6 for illustrations). We shall highlight that the validity of Lemma 3.14 does not depends on the ordering of the weights. Thus, it is enough to have available the stick-breaking decomposition for one ordering of the weights.

Now we turn to analyse the relationship between the distribution of the length variables and tie probability, or more generally the EPPF. From Definition 3.3, it is obvious that for a stick-breaking process $\boldsymbol{\mu}$ with weights $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$, we can rewrite its tie probability as

$$\rho = \sum_{j\geq 1} \mathbb{E}\left[(\mathbf{v}_j)^2 \prod_{i=1}^{j-1}(1-\mathbf{v}_i)^2\right]. \tag{3.6}$$

As to the EPPF, from equation (2.19), we have that if the weights sum up to one, for any positive numbers, $n_1, \ldots, n_k$,

$$\pi(n_1, \ldots, n_k) = \sum_{(i_1,\ldots,i_k)} \mathbb{E}\left[\prod_{j=1}^{k} \left(\mathbf{v}_{i_j}\right)^{n_j-1} \prod_{l=1}^{i_j-1}(1-\mathbf{v}_l)^{n_j-1}\right], \tag{3.7}$$

where the sum ranges over all k-tuples of distinct positive integers. Unfortunately, more often than not, this equation does not lead to a closed expression for the EPPF. A quantity that is seemingly easier to compute is the pEPPF, π', corresponding to $\mathbf{W}$, as introduced in Section 2.2.7. Indeed from (2.17), we obtain that for any positive numbers, $n_1, \ldots, n_k$,

$$\pi'(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{v}_j^{n_j-1}(1-\mathbf{v}_j)^{\sum_{i>j} n_i}\right]. \tag{3.8}$$

In Section 2.2.7 we proved that π' is a symmetric function of its arguments if and only if $\mathbf{W}$ is in size-biased random order, in which case π' coincides with the EPPF, π, and particularly we obtain $\rho = \pi(2) = \pi'(2) = \mathbb{E}[\mathbf{v}_1]$.

One of the biggest downsides of this construction is that most distributions over $[0,1]^\infty$ will lead to stick-breaking weights that are not invariant under size-biased permutations, hence the EPPF will generally become defiant to attain. On the opposite side, an enormous strength of this construction is that it translates the complex problem of defining distributions of $\overline{\Delta}_\infty$ into the much easier one of finding distributions over $[0,1]^\infty$. Moreover the simplicity of this construction enables to prove structural properties of the SSP, in a simple and elegant form.

3.5.4 Random subsets of $\mathbb{N}$

In contrast to their counterparts, this method is relatively new (Fuentes-García et al.; 2010; De Blasi et al.; 2020; Gil-Leyva; 2021). The original motivation behind this method is mostly numerical, and it arises from the fact that most interesting SSPs depend on infinitely many random variables, which in practice makes them challenging to implement. In essence this construction consists in building a latent random set, say $\boldsymbol{\Psi}$, that takes values in $\mathcal{F}_{\mathbb{N}} = \{A \subseteq \mathbb{N} : 0 < |A| < \infty\}$, and that makes sampling $\mathbf{x}$ from the proper SSP, $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$, equivalent to sampling it from uniform random probability measure, $|\boldsymbol{\Psi}|^{-1} \sum_{j\in\boldsymbol{\Psi}} \delta_{\boldsymbol{\xi}_j}$. This approach not only solves the practical problem of dealing with infinitely many random variables (see Section 5.2.2), but also suggests a new form to construct distributions over the infinite dimensional simplex $\Delta_\infty = \{(w_1, w_2, \ldots) \in [0,1]^\infty : \sum_{j\geq 1} w_j = 1\}$.

So consider a random element, $\boldsymbol{\tau}$, taking values in some Borel space $(T, \mathscr{B}_T)$ and a mass probability kernel $\mathbb{p}_\Psi(\cdot \mid \cdot)$ from T into $\mathcal{F}_\mathbb{N}$, so that for each $t \in T$, $\mathbb{p}_\Psi(\cdot \mid t)$ is a mass probability function over $\mathcal{F}_\mathbb{N}$, and for each $A \in \mathcal{F}_\mathbb{N}$, $\mathbb{p}_\Psi(A \mid \cdot)$ is a measurable function from $(T, \mathscr{B}_T)$ into $([0,1], \mathscr{B}([0,1])$. Let $\boldsymbol{\Psi}$ be some random set satisfying $\{\boldsymbol{\Psi} \mid \boldsymbol{\tau}\} \sim \mathbb{p}_\Psi(\cdot \mid \boldsymbol{\tau})$, and say that given $\boldsymbol{\Psi}$, we uniformly pick one of its elements, $\mathbf{d}$. Define $\mathbf{w}_j$ as the conditional probability that $\mathbf{d} = j$ given $\boldsymbol{\tau}$, that is $\mathbf{w}_j = \mathbb{P}[\mathbf{d} = j \mid \boldsymbol{\tau}]$. Under the assumption that $\mathbf{d}$ is conditionally independent of $\boldsymbol{\tau}$ given $\boldsymbol{\Psi}$, and by the tower property of conditional expectation we obtain,

$$\mathbf{w}_j = \mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}|}\mathbf{1}_{\{j\in\boldsymbol{\Psi}\}} \,\middle|\, \boldsymbol{\tau}\right] = \sum_{A\in\mathcal{F}_\mathbb{N}} \frac{1}{|A|}\mathbf{1}_{\{j\in A\}}\mathbb{p}_\Psi(A \mid \boldsymbol{\tau}), \quad j \geq 1. \tag{3.9}$$

As the events $(\mathbf{d} = j)_{j\geq 1}$ are mutually disjoint and its union $(\mathbf{d} \in \mathbb{N})$ occurs almost surely we must have $\sum_{j\geq 1} \mathbf{w}_j = 1$. This is, by conditional monotone convergence theorem, and since $\sum_{j\geq 1} \mathbf{1}_{\{j\in\boldsymbol{\Psi}\}} = |\boldsymbol{\Psi}|$, we get

$$\sum_{j\geq 1} \mathbf{w}_j = \mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}|}\sum_{j\geq 1}\mathbf{1}_{\{j\in\boldsymbol{\Psi}\}} \,\middle|\, \boldsymbol{\tau}\right] = 1, \tag{3.10}$$

almost surely. Thus, each parametric distribution, $\mathbb{p}_\Psi$, over $\mathcal{F}_\mathbb{N}$, together with a randomization of its parameters, $\boldsymbol{\tau}$, and through (3.9), characterizes completely the law of $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$, which takes values in the infinite dimensional simplex. In other words, if we denote by $\mathbb{p}_\tau$ to the distribution of $\boldsymbol{\tau}$, the pair $(\mathbb{p}_\Psi, \mathbb{p}_\tau)$, defines a distribution over Δ_∞. Thus any species sampling process, $\boldsymbol{\mu}$, with weights $(\mathbf{w}_j)_{j\geq 1}$ as in (3.9) is proper. Now, to corroborate it has full support, from Corollary 3.12, it suffices to check that the largest weight is arbitrarily small with positive probability. This is relatively simple to do and fairly intuitive in terms of the random set. Note that for every $j \geq 1$,

$$\mathbf{w}_j = \mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}|}\mathbf{1}_{\{j\in\boldsymbol{\Psi}\}} \,\middle|\, \boldsymbol{\tau}\right] \leq \mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}|} \,\middle|\, \boldsymbol{\tau}\right].$$

Hence, if for every $\varepsilon > 0$, we have that the events, $\{|\boldsymbol{\Psi}| > 1/\varepsilon\} \subseteq \{\mathbf{w}_j < \varepsilon, j \geq 1\} = \{\max_{j\geq 1} \mathbf{w}_j < \varepsilon\}$, which yields

$$\mathbb{P}\left[\max_{j\geq 1} \mathbf{w}_j < \varepsilon\right] \geq \mathbb{P}\left[|\boldsymbol{\Psi}| > \varepsilon^{-1}\right],$$

and we obtain the following Corollary,

Corollary 3.15. *Let $\boldsymbol{\mu}$ be a SSP with weights, $(\mathbf{w}_j)_{j\geq 1}$, as in (3.9). If $\mathbb{P}[|\boldsymbol{\Psi}| > n] > 0$, for every $n \in \mathbb{N}$, then $\boldsymbol{\mu}$ has full support.*

Roughly speaking, if $\boldsymbol{\Psi}$ is allowed to contain arbitrarily many random elements, then the largest weight can be arbitrarily small, and $\boldsymbol{\mu}$ has full support. Next we prove that the weights of any proper species sampling process can be expressed as in (3.9) for some random set.

Proposition 3.16. *Let $\boldsymbol{\mu}$ be a proper species sampling process with weights $(\mathbf{w}_j)_{j\geq 1}$. Then there exist a random set $\boldsymbol{\Psi}$ taking values in $\mathcal{F}_{\mathbb{N}}$ and a random element $\boldsymbol{\tau}$ that takes values in a Borel space such that (3.9) holds.*

The proof of Proposition 3.16 can be found in Appendix C.11, this result guaranties that this construction holds for every SSP, however it is not unique. In fact, this construction is not even unique for a fixed weights sequence $(\mathbf{w}_j)_{j\geq 1}$. Meaning that there exists two distinct pairs $(\boldsymbol{\Psi}, \boldsymbol{\tau})$ and $(\boldsymbol{\Psi}', \boldsymbol{\tau}')$ such that (3.9) holds for $(\mathbf{w}_j)_{j\geq 1}$.

Our next result, see Appendix C.12 for a proof, embodies the original motivation behind this construction, and will be highlighted in Section 5.2.2 in an applied setting.

Proposition 3.17. *Let $(T, \mathscr{B}_T)$ be a Polish space and consider a mass probability kernel $\mathbb{p}_\Psi$ from T into $\mathcal{F}_{\mathbb{N}}$. Let $\boldsymbol{\tau}$ be a random element taking values in T, consider $\{\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n \mid \boldsymbol{\tau}\} \overset{iid}{\sim} \mathbb{p}_\Psi(\cdot \mid \boldsymbol{\tau})$ and define $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$ as in (3.9). Also consider a diffuse probability measure μ_0 over the Borel space $(S, \mathscr{B}_S)$, and an independent collection $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j\geq 1} \overset{iid}{\sim} \mu_0$. If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are conditionally independent given $\boldsymbol{\Xi}, \boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n$, with $\{\mathbf{x}_k \mid \boldsymbol{\Psi}_k, \boldsymbol{\Xi}\} \sim |\boldsymbol{\Psi}_k|^{-1} \sum_{j\in\boldsymbol{\Psi}_k} \delta_{\boldsymbol{\xi}_j}$, and we also assume that, $\mathbf{x}_k$ is conditionally independent of $\boldsymbol{\tau}$, given $\boldsymbol{\Psi}_k$, for $1 \leq k \leq n$, then $\{\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \mathbf{W}, \boldsymbol{\Xi}\} \overset{iid}{\sim} \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$.*

Deriving a closed expression for the EPPF in terms of these random sets does not seems feasible, however, this is not the main motivation for the method. In fact, this construction precisely arose as an alternative way to solve the practical implementation problem of SSPs for which a closed expression for the EPPF is not available (see Section 5.2.2). Putting this construction in a slightly distinct category than their counterparts.

3.6 Examples

In here we will present some of the most popular SSPs in Bayesian non-parametric literature. We will go through their constructions as well as basic structural properties.

3.6.1 Dirichlet process

The Dirichlet process has earned the title of the canonical example of SSPs in a Bayesian non-parametric context (Ferguson; 1973; Blackwell and MacQueen; 1973; Sethuraman; 1994), mainly due to its mathematical tractability and distinct representations. We will begin by defining this process in various ways and showing that these definitions are in fact equivalent.

Definition 3.6 (Dirichlet process, finite dimensional distributions)**.** *Let $(S, \mathscr{B}_S)$ be a Borel space and consider a diffuse finite measure μ over $(S, \mathscr{B}_S)$. We say $\boldsymbol{\mu}$ is a Dirichlet process with total mass parameter $\theta = \mu(S)$ and base measure $\mu_0 = \mu/\mu(S)$, if*

$$(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n)) \sim \mathsf{Dir}(\mu(B_1), \ldots, \mu(B_n))$$

for every measurable partition $\{B_i\}_{i=1}^n$ of S.

Definition 3.7 (Dirichlet process, prediction rule, EPPF)**.** *Let $(S, \mathscr{B}_S)$ be a Borel space and consider a diffuse probability measure μ_0 over $(S, \mathscr{B}_S)$. We say $\boldsymbol{\mu}$ is a Dirichlet* *process with total mass parameter $\theta > 0$ and base measure μ_0, if it is a species sampling process with base measure μ_0 and corresponding EPPF, π, given by*

$$\pi(n_1, \dots, n_k) = \frac{\theta^k \prod_{j=1}^{k} (n_j - 1)!}{(\theta)_n},$$

for every sequence of positive numbers $(n_1, \dots, n_k)$. Equivalently, we say $\boldsymbol{\mu}$ is a Dirichlet process with total mass parameter $\theta > 0$ and base measure μ_0, if it is the directing random measure of an exchangeable sequence $(\mathbf{x}_i)_{i \geq 1}$ such that $\mathbf{x}_1 \sim \mu_0$ and for every $n \geq 1$,

$$\mathbb{P}[\mathbf{x}_{n+1} \in \cdot \mid \mathbf{x}_1, \dots, \mathbf{x}_n] = \sum_{j=1}^{\mathbf{K}_n} \frac{\mathbf{n}_j}{\theta + n} \delta_{\mathbf{x}_j^*} + \frac{\theta}{\theta + n} \mu_0 = \frac{1}{\theta + n} \left(\sum_{i=1}^{n} \delta_{\mathbf{x}_i} + \mu_0 \right)$$

where $\mathbf{x}_1^*, \dots, \mathbf{x}_{\mathbf{K}_n}^*$ are the $\mathbf{K}_n$ distinct values that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ exhibits, and for every $1 \leq j \leq \mathbf{K}_n$, $\mathbf{n}_j = |\{1 \leq i \leq n : \mathbf{x}_i = \mathbf{x}_j^*\}|$.

Definition 3.8 (Dirichlet process, NRMI, Gamma process)**.** *Let $(S, \mathscr{B}_S)$ be a Borel space and consider a diffuse finite measure λ over $(S, \mathscr{B}_S)$. Let $\boldsymbol{\lambda} = \sum_{j \geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$ be a λ-homogeneous completely random measure with no diffuse component such that $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j \geq 1}$ defines a Poisson process with intensity $\lambda \otimes \varrho$, where*

$$\varrho(dx) = \frac{e^{-x}}{x} dx.$$

To the NRMI $\boldsymbol{\mu} = \boldsymbol{\lambda}/\boldsymbol{\lambda}(S)$ we call a Dirichlet process with total mass parameter $\theta = \lambda(S)$ and base measure $\mu_0 = \lambda/\lambda(S)$.

Definition 3.9 (Dirichlet process, stick-breaking)**.** *Let $(S, \mathscr{B}_S)$ be a Borel space and consider a diffuse probability measure μ_0 over $(S, \mathscr{B}_S)$. Fix $\theta > 0$, Let $(\mathbf{v}_i)_{i \geq 1} \overset{iid}{\sim} \mathsf{Be}(1, \theta)$ and set $(\mathbf{w}_j)_{j \geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i \geq 1}]$. To the stick-breaking process, $\boldsymbol{\mu}$, with base measure μ_0 and weights sequence $(\mathbf{w}_j)_{j \geq 1}$, we call a Dirichlet process with total mass parameter θ and base measure μ_0.*

Theorem 3.18. *The four definitions of the Dirichlet process are equivalent.*

Hereinafter we will denote by $\mathcal{D}_{(\theta, \mu_0)}$ to the distribution of a Dirichlet process with total mass parameter $\theta > 0$ and base measure μ_0. Our next result is a straight-forward consequence of the proof of Theorem 3.18 (see Appendix C.13).

Corollary 3.19. *Let $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta, \mu_0)}$ and consider $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mid \boldsymbol{\mu}\} \overset{iid}{\sim} \boldsymbol{\mu}$. Then, for every $m \geq 1$ and each measurable partition, $\{B_i\}_{i=1}^{n}$, of the corresponding space, we get*

$$\left\{ (\boldsymbol{\mu}(B_1), \dots, \boldsymbol{\mu}(B_n)) \mid \mathbf{x}^{(m)} \right\} \sim \mathsf{Dir} \left(\sum_{i=1}^{m} \delta_{\mathbf{x}_i}(B_1) + \theta \mu_0(B_1), \dots, \sum_{i=1}^{m} \delta_{\mathbf{x}_i}(B_n) + \theta \mu_0(B_n) \right),$$

where $\mathbf{x}^{(m)} = (\mathbf{x}_1, \dots, \mathbf{x}_m)$.

Structural properties of Dirichlet processes can be easily derived from their various representations. For example from Definition 3.8 it is direct that every Dirichlet process is proper. This property can also be obtained from Definition 3.9 together with Proposition

2.12 or Lemma 3.14, these also show that the representation of the Dirichlet weights given by Definition 3.9 is invariant under size biased permutations. Additionally, from Definition 3.9 we get that any $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$ has length variables $(\mathbf{v}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \mathsf{Be}(1,\theta)$, so evidently, for every $0 < \gamma < \varepsilon$, and $n \geq 1$,

$$\mathbb{P}\left[\bigcap_{i=1}^{n}(\gamma < \mathbf{v}_i < \varepsilon)\right] = \prod_{i=1}^{n}\mathbb{P}[\gamma < \mathbf{v}_i < \varepsilon] > 0.$$

Thus, Lemma 3.14 yields Dirichlet processes have full support. This proves the next Corollary.

Corollary 3.20. *Let $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$. Then $\boldsymbol{\mu}$ is proper and it has full support.*

As explained in Sections 3.1, 3.2 and 3.3 further basic properties of a SSP can be written in terms of the tie probability and the base measure. For $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$, from Definition 3.7, we easily compute

$$\rho = \pi(2) = \frac{1}{1+\theta}. \tag{3.11}$$

Note that as $\theta \to \infty$, $\rho \to 0$ and as $\theta \to 0$, $\rho \to 1$. While it would be vacuous to restate all the previously derived results for Dirichlet processes, we will highlight some of the most notable ones.

Corollary 3.21. *Let $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$. Then, for any measurable sets A and B,*

i) $\mathbb{E}\left[\boldsymbol{\mu}(A)\right] = \mu_0(A)$.

ii) $\mathsf{Var}\left(\boldsymbol{\mu}(A)\right) = \dfrac{\mu_0(A)(1-\mu_0(A))}{1+\theta}$.

iii) $\mathsf{Cov}\left(\boldsymbol{\mu}(A),\boldsymbol{\mu}(B)\right) = \dfrac{(\mu_0(A\cap B) - \mu_0(A)\mu_0(B))}{1+\theta}$.

Corollary 3.22. *Consider a Polish space S with Borel σ-algebra $\mathscr{B}_S$. Let $\mu_0, \mu_0^{(1)}, \mu_0^{(2)}, \ldots$ be diffuse probability measures over $(S,\mathscr{B}_S)$, such that $\mu_0^{(n)}$ converges weakly to μ_0 as $n \to \infty$. For $n \geq 1$ let $\rho^{(n)} > 0$, and let $\boldsymbol{\mu}^{(n)} \sim \mathcal{D}_{\left(\theta^{(n)},\mu_0^{(n)}\right)}$.*

i) *If $\theta^{(n)} \to \infty$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to μ_0.*

ii) *If $\theta^{(n)} \to 0$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\delta_{\boldsymbol{\xi}}$, where $\boldsymbol{\xi} \sim \mu_0$.*

Corollary 3.23. *Let $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{iid}{\sim} \boldsymbol{\mu}$ where $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$. Fix $n \geq 1$ and let $f : S^n \to \mathbb{R}$ be measurable function, then*

$$\begin{aligned}&\mathbb{E}\left[f(\mathbf{x}_1,\ldots,\mathbf{x}_n)\right]\\ &= \sum_{A\in\mathcal{P}_{[n]}}\left\{\int f(x_{l_1},\ldots,x_{l_n})\prod_{j=1}^{k}\prod_{r\in A_j}\mathbf{1}_{\{l_r=j\}}\,\mu_0(dx_1)\ldots\mu_0(dx_k)\right\}\frac{\theta^k\prod_{j=1}^{k}(|A_j|-1)!}{(\theta)_n},\end{aligned} \tag{3.12}$$

whenever the integrals in the right side exist, where $k = |A|$ and $A_1, \dots, A_k$ stand for the blocks of $A \in \mathcal{P}_{[n]}$. Moreover, if f is symmetric (and the integrals exist), equation 3.12 reduces to

$$\mathbb{E}\left[f(\mathbf{x}_1, \dots, \mathbf{x}_n)\right] = \frac{n!}{(\theta)_n} \sum_{k=1}^{n} \theta^{k-1} \sum_{(m_1,\dots,m_n)\in\mathcal{M}_n^k} \frac{1}{\prod_{i=1}^{n}(i)^{m_i}(m_i!)} \times$$
$$\times \int f\left(x_{[n_1,\dots,n_k]}\right)\, \mu_0(dx_1)\dots\mu_0(dx_k), \tag{3.13}$$

where $\mathcal{M}_n^k$, and $x_{[n_1,\dots,n_k]}$ are as in Theorem 3.7.

Corollaries 3.21 and 3.22 are also proven by Ghosal and van der Vaart (2017), whilst Corollary 3.23 was first proven by Yamato (1984), for the particular case of symmetric functions and through a very complicated induction argument.

Before moving on we shall comment that a very important random variable which summarizes relevant information about SSP is the number of distinct values, $\mathbf{K}_n$, that a sample from $\{\mathbf{x}_1, \dots, \mathbf{x}_n \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$ exhibits. We have already mentioned this random variable but we have not discussed it in detail. For the case where $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$, from Definition 3.7 it is easy to see that $(\mathbf{K}_n)_{n\geq 1}$ is a Markov chain such that $\mathbf{K}_1 = 1$ almost surely, and for $n \geq 1$

$$\mathbb{P}[\mathbf{K}_{n+1} = x \mid \mathbf{K}_n] = \frac{n}{\theta + n}\mathbf{1}_{\{x=\mathbf{K}_n\}} + \frac{\theta}{\theta+n}\mathbf{1}_{\{x=\mathbf{K}_n+1\}}. \tag{3.14}$$

The tower property of conditional expectation, this yields

$$\mathbb{E}[\mathbf{K}_{n+1}] = \frac{1}{n+\theta}\left(n\mathbb{E}[\mathbf{K}_n] + \theta\{\mathbb{E}[\mathbf{K}_n]+1\}\right) = \mathbb{E}[\mathbf{K}_n] + \frac{\theta}{n+\theta},$$

and by a simple induction argument we can compute

$$\mathbb{E}[\mathbf{K}_n] = \sum_{i=1}^{n} \frac{\theta}{i-1+\theta}.$$

For most SSPs characterizing analytically the distribution of $\mathbf{K}_n$ is generally more complicated or not feasible. In order to provide a sensible comparison between this and other models, for all examples of SSPs we will illustrate the distribution of $\mathbf{K}_n$ by means of numerical methods. That is, we will draw samples from $\mathbf{K}_n$, and latter analyse the frequency polygons (see Figure 19, below). For Dirichet processes, using equation 3.14, we can easily sample from the marginal distribution of $\mathbf{K}_n$, for $n \geq 1$ and $\theta > 0$, fixed.

Figure 19 illustrates how, for fixed n, the larger θ is, the bigger are the values favoured by the distribution of $\mathbf{K}_n$. This behaviour was already evident from Definition 3.7 and equation 3.14. Another way to understand this effect is by focussing on the stick-breaking construction. Indeed, larger values of θ translate to smaller length variables, which in turn implies that even the larger weights tend to be small, thus providing samples from $\{\mathbf{x}_1, \dots, \mathbf{x}_n \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$ that exhibit a wide variety of values.

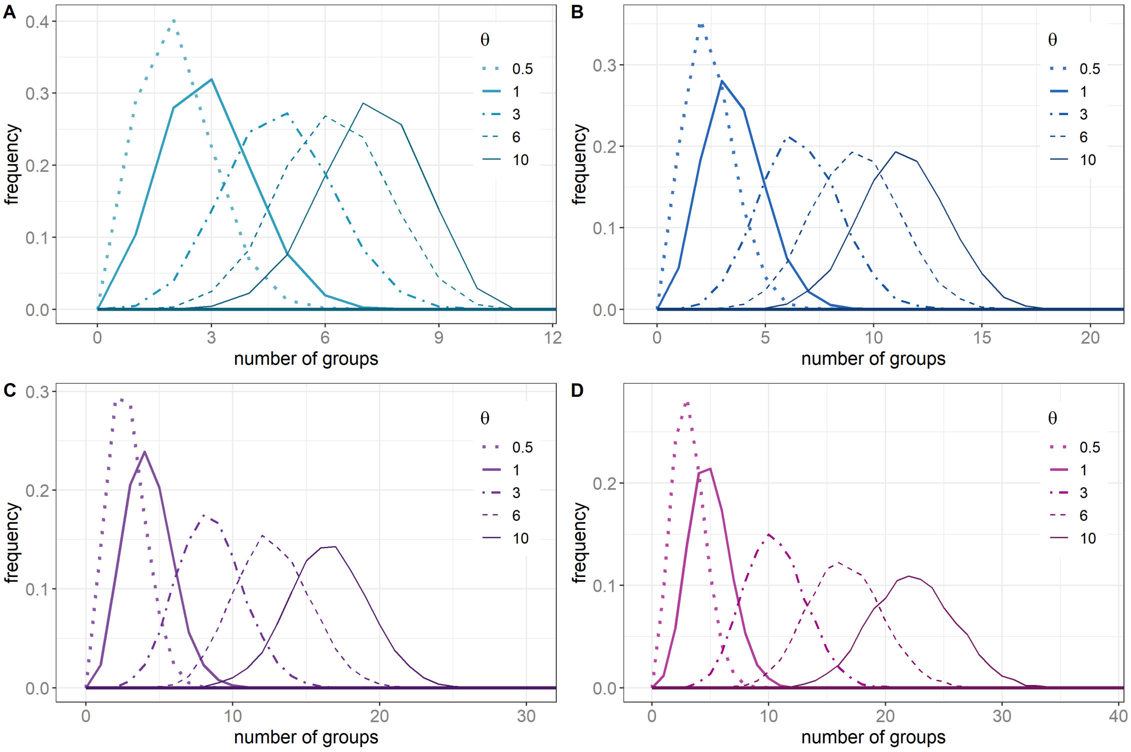


Figure 19: Frequency polygons of samples of size 10000 from the distribution of $\mathbf{K}_{10}$ (A), $\mathbf{K}_{20}$ (B), $\mathbf{K}_{40}$ (C) and $\mathbf{K}_{80}$ (D). For each fixed value of n we vary θ in the set $\{0.5, 1, 3, 6, 10\}$.

3.6.2 Pitman-Yor process

The Pitman-Yor process was first introduced by Perman et al. (1992) and Pitman and Yor (1992), and it is a generalization of the Dirichlet process. Pitman-Yor processes, that are not Dirichlet processes, inherit to some extent the tractability of the particular case. However, computations do tend to be more complicated, if feasible. For instance, to the best of our knowledge, the finite dimensional distributions of these random probability measures are not yet available. Other definitions provided in the former section can be adjusted to cover this more general class.

Definition 3.10 (Pitman-Yor process, prediction rule, EPPF)**.** *Fix $0 \leq \sigma < 1$ and $\theta > -\sigma$. Let $(S, \mathscr{B}_S)$ be a Borel space and consider a diffuse probability measure μ_0 over $(S, \mathscr{B}_S)$. We say $\boldsymbol{\mu}$ is a Pitman-Yor process with parameters (σ, θ, μ_0) if it is a species sampling process with base measure μ_0 and corresponding EPPF, π, given by*

$$\pi(n_1, \dots, n_k) = \frac{(\theta + \sigma)_{k-1\uparrow\sigma} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}}{(\theta + 1)_n},$$

for every sequence of positive numbers $(n_1, \dots, n_k)$, where $(x)_{m\uparrow\alpha} = \prod_{i=1}^{m}(x + i\alpha)$, and $(x)_m = (x)_{m\uparrow 1}$. Equivalently, we say $\boldsymbol{\mu}$ is a Pitman-Yor process with parameters (σ, θ, μ_0), if it is the directing random measure of an exchangeable sequence $(\mathbf{x}_i)_{i\geq 1}$ such that $\mathbf{x}_1 \sim \mu_0$ and for every $n \geq 1$,

$$\mathbb{P}[\mathbf{x}_{n+1} \in \cdot \mid \mathbf{x}_1, \dots, \mathbf{x}_n] = \sum_{j=1}^{\mathbf{K}_n} \frac{\mathbf{n}_j - \sigma}{\theta + n} \delta_{\mathbf{x}_j^*} + \frac{\theta + \sigma \mathbf{K}_n}{\theta + n} \mu_0$$

where $\mathbf{x}_1^*, \dots, \mathbf{x}_{\mathbf{K}_n}^*$ *are the* $\mathbf{K}_n$ *distinct values that* $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ *exhibits, and for every* $1 \leq j \leq \mathbf{K}_n$, $\mathbf{n}_j = |\{1 \leq i \leq n : \mathbf{x}_i = \mathbf{x}_j^*\}|$.

Definition 3.11 (Pitman-Yor process, stick-breaking)**.** *Fix* $0 \leq \sigma < 1$ *and* $\theta > -\sigma$. *Let* $(S, \mathscr{B}_S)$ *be a Borel space and consider a diffuse probability measure* μ_0 *over* $(S, \mathscr{B}_S)$. *Let* $(\mathbf{v}_i)_{i\geq 1}$ *be an independent collection of random variables with* $\mathbf{v}_i \sim \mathsf{Be}(1-\sigma, \theta + i\sigma)$, *and set* $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$. *To the stick-breaking process,* $\boldsymbol{\mu}$, *with base measure* μ_0 *and weights sequence* $(\mathbf{w}_j)_{j\geq 1}$, *we call a Pitman-Yor process with parameters* (σ, θ, μ_0).

The fact that Definitions 3.10 and 3.11 describe the same SSP is a direct consequence of Proposition 2.12. From now on, to the distribution of a Pitman-Yor process with parameters (σ, θ, μ_0) we denote by $\mathcal{PY}_{(\sigma,\theta,\mu_0)}$. As is the case of Dirichlet processes, structural properties of Pitman-Yor processes follow easily from their definitions. The proof of the following result is similar to that of Corollary 3.20.

Corollary 3.24. *Let* $\boldsymbol{\mu} \sim \mathcal{PY}_{(\sigma,\theta,\mu_0)}$. *Then* $\boldsymbol{\mu}$ *is proper and it has full support.*

From the Definition 3.10 we can derive the tie probability of Pitman-Yor process. Indeed, for $\boldsymbol{\mu} \sim \mathcal{PY}_{(\sigma,\theta,\mu_0)}$, its tie probability is given by

$$\rho = \pi(2) = \frac{1-\sigma}{\theta+1}. \tag{3.15}$$

Note that as $\theta \to \infty$, or for fixed θ, as $\sigma \to 1$, we get $\rho \to 0$. Alternatively, if $\theta \to -\sigma$, for fixed σ, we get $\rho \to 1$. So inserting this into the results in Sections 3.1, 3.2 and 3.3, one can rewrite for Pitman-Yor processes, all the properties we have already derived for general SSPs.

As to the random variables $\mathbf{K}_n = |\mathbf{\Pi}(\mathbf{x}_{1:n})|$ where $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu} \sim \mathcal{PY}_{(\sigma,\theta,\mu_0)}$, analogously as in the Dirichlet case, we get $(\mathbf{K}_n)_{n\geq 1}$ is a Markov chain with $\mathbf{K}_n = 1$ almost surely, and for $n \geq 1$,

$$\mathbb{P}[\mathbf{K}_{n+1} = x \mid \mathbf{K}_n] = \frac{n - \sigma\mathbf{K}_n}{\theta + n}\mathbf{1}_{\{x=\mathbf{K}_n\}} + \frac{\theta + \sigma\mathbf{K}_n}{\theta+n}\mathbf{1}_{\{x=\mathbf{K}_n+1\}}. \tag{3.16}$$

From this we obtain the recursion

$$\mathbb{E}[\mathbf{K}_{n+1}] = \frac{(n+\theta+\sigma)\mathbb{E}[\mathbf{K}_n] + \theta}{\theta+n},$$

which in turn yields

$$\mathbb{E}[\mathbf{K}_n] = \frac{\theta(\theta+\sigma)_n}{\sigma(\theta)_n} - \frac{\theta}{\sigma}.$$

For Pitman-Yor processes it is relatively easy to compute $\mathbb{P}[\mathbf{K}_n = x]$ for $n \geq 1$ and $x \in [n]$ (see for instance Pitman; 2006). Despite this, in other to compare Pitman-Yor processes to other models that do not enjoy this advantage, we illustrate the distribution of $\mathbf{K}_n$ by drawing samples from it. Analogously as in the Dirichlet case we can draw samples from this marginal distribution through equation 3.16. As observed in Figure 20, larger values of θ or σ impact the distribution of $\mathbf{K}_n$ with larger probabilities assigned to bigger values. Once again, this should already be evident from Definition 3.10 and equation 3.16.

The fact that clustering probabilities are so simple to compute for Pitman-Yor processes (including Dirichlet processes) is greatly influenced by the fact that their size-biased permuted weights have a simple stick-breaking representation. As shown by Pitman (1996a), the weights described in Definitions 3.9 and 3.11, are the only strictly positive stick-breaking weights with independent length variables that are invariant in distribution under size-biased permutations. This said it should not be surprising that for other stick-breaking processes, the clustering probabilities are much harder to characterize, whenever feasible.

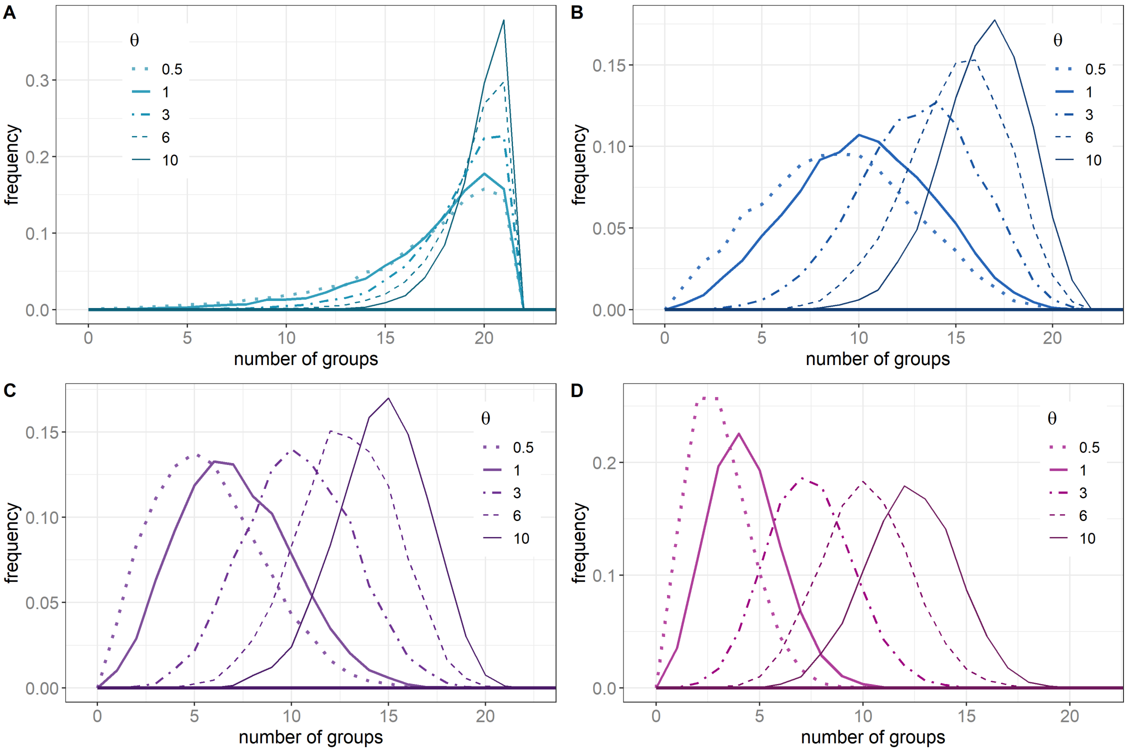


Figure 20: Frequency polygons of samples of size 10000 from the distribution of $\mathbf{K}_{20}$ for $\sigma = 0.9$ (A), $\sigma = 0.6$ (B), $\sigma = 0.4$ (C) and $\sigma = 0.1$ (D). For each fixed value of σ we vary θ in the set $\{0.5, 1, 3, 6, 10\}$.

3.6.3 Geometric process

In contrast to Dirichlet and Pitman-Yor processes, for Geometric processes it is their decreasingly ordered weights that have a simple stick-breaking representation. In fact it is the simplicity of their decreasing weights what has made this model popular. Geometric processes were introduced by Fuentes-García et al. (2010), here the authors presented a construction of these SSPs by means of random sets and derived the stick-breaking representation.

Definition 3.12 (Geometric process, stick-breaking)**.** *Let ν_0 be a probability measure over $([0,1], \mathscr{B}_{[0,1]})$ with $\nu_0(\{0\}) = 0$. Consider a Borel space $(S, \mathscr{B}_S)$ and a diffuse probability measure μ_0 over $(S, \mathscr{B}_S)$. Let $\mathbf{v} \sim \nu_0$ and for every $j \geq 1$ set $\mathbf{w}_j = \mathbf{v}(1-\mathbf{v})^{j-1}$, so that $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ where $\mathbf{v}_i = \mathbf{v}$ for every $i \geq 1$. To the stick-breaking process, $\boldsymbol{\mu}$, with base measure μ_0 and weights sequence $(\mathbf{w}_j)_{j\geq 1}$, we call a Geometric process with parameters (ν_0, μ_0).*

Definition 3.13 (Geometric process, random sets)**.** *Let ν_0 be a probability measure over $([0,1], \mathscr{B}_{[0,1]})$ with $\nu_0(\{0\}) = 0$. Consider a Borel space $(S, \mathscr{B}_S)$ and a diffuse probability measure μ_0 over $(S, \mathscr{B}_S)$. Let $\mathbf{v} \sim \nu_0$ and let $\mathbf{N}$ be a random variable taking values in $\mathbb{N}$ such that*

$$\mathbb{P}[\mathbf{N} = n \mid \mathbf{v}] = n\mathbf{v}^2(1-\mathbf{v})^{n-1} \tag{3.17}$$

that is, conditionally given $\mathbf{v}$, $\mathbf{N}$ is a displaced negative-binomial random variable with parameters $(2, \mathbf{v})$. Define the random set $\boldsymbol{\Psi} = \{1, \ldots, \mathbf{N}\}$ and $(\mathbf{w}_j)_{j\geq 1}$ through (3.9), with $\boldsymbol{\tau} = \mathbf{v}$. To the SSP, $\boldsymbol{\mu}$, with base measure μ_0 and weights sequence $(\mathbf{w}_j)_{j\geq 1}$, we call a Geometric process with parameters (ν_0, μ_0).

Note that if $(\mathbf{w}_j)_{j\geq 1}$ are as in Definition 3.13 then

$$\mathbf{w}_j = \mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}|}\mathbf{1}_{\{j\in\boldsymbol{\Psi}\}} \,\middle|\, \mathbf{v}\right] = \mathbb{E}\left[\frac{1}{\mathbf{N}}\mathbf{1}_{\{j\leq\mathbf{N}\}} \,\middle|\, \mathbf{v}\right] = \sum_{n\geq j} \mathbf{v}^2(1-\mathbf{v})^{n-1} = \mathbf{v}(1-\mathbf{v})^{j-1},$$

which shows that Definitions 3.12 and 3.13 describe the same SSP. To the distribution of a Geometric process with parameters ν_0 and μ_0 we denote by $\mathcal{G}_{(\nu_0,\mu_0)}$. Checking that Geometric processes are proper is trivial. Also, if there exists $0 < \varepsilon < 1$ such that $(0,\varepsilon)$ is contained in the support of ν_0, then, for $0 < \epsilon \leq \varepsilon$ we get that $(0,\epsilon)$ is as well contained in the support of ν_0, and for $\varepsilon < \epsilon < 1$ we have that $\nu_0((0,\epsilon)) > \nu_0((0,\varepsilon)) > 0$, either way for $(\mathbf{w}_j)_{j\geq 1}$ as in Definition 3.12 we obtain

$$\mathbb{P}\left[\max_{j\geq 1} \mathbf{w}_j < \epsilon\right] = \mathbb{P}[\mathbf{v} < \epsilon] = \nu_0((0,\epsilon)) > 0.$$

Thus Corollary 3.12 together with the definition of Geometric processes prove the following result.

Corollary 3.25. *Let $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0,\mu_0)}$. Then $\boldsymbol{\mu}$ is proper and, if there exists $0 < \varepsilon < 1$ such that $(0,\varepsilon)$ is contained in the support of ν_0, $\boldsymbol{\mu}$ has full support.*

In particular, if $\nu_0 = \mathsf{Be}(\alpha,\beta)$ for some $\alpha,\beta > 0$ we get $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0,\mu_0)}$ has full support. For Geometric processes it is very hard to compute the corresponding EPPF, while it is possible to derive an expression for the tie probability, even this one has a complicated expression in contrast to Pitman-Yor processes. Indeed, for $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0,\mu_0)}$, with weights as in Definition 3.12, its tie probability can be computed through

$$\rho = \sum_{j\geq 1} \mathbb{E}\left[(\mathbf{w}_j)^2\right] = \mathbb{E}\left[\sum_{j\geq 1} \mathbf{v}^2\left\{(1-\mathbf{v})^2\right\}^{j-1}\right] = \mathbb{E}\left[\frac{\mathbf{v}^2}{1-(1-\mathbf{v})^2}\right] = \mathbb{E}\left[\frac{\mathbf{v}}{2-\mathbf{v}}\right].$$

Hence

$$\rho = \frac{1}{2}\int_0^1 \frac{x}{1-(x/2)}\nu_0(dx), \tag{3.18}$$

and for $\nu_0 = \mathsf{Be}(\alpha,\beta)$, the above expression reduces to

$$\rho = \frac{\Gamma(\alpha+\beta)}{2\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \frac{x^\alpha(1-x)^{\beta-1}}{1-(x/2)}dx = \frac{{}_2F_1(1,\alpha+1;\alpha+\beta+1,1/2)}{2} \tag{3.19}$$

where ${}_2F_1$ denotes the Gauss-hypergeometric function. Once more, by inserting this quantity into previously derived results, one obtains a variety of basic properties for Geometric processes.

Naturally, for $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0,\mu_0)}$ the law of $(\mathbf{K}_n)_{n\geq 1}$ is also more challenging to characterize than for Pitman-Yor processes. Despite the mathematical hurdles to overcome, there have been advances in this topic recently. For instance, Mena and Walker (2012) showed that conditioning on the length variable, $\mathbf{v}$,

$$\mathbb{E}[\mathbf{K}_n \mid \mathbf{v}] = \sum_{r=1}^{n} (-1)^{r-1} \binom{n}{r} \frac{\mathbf{v}^r}{1-(1-\mathbf{v})^r},$$

taking expectations in the above equation one can derive an expression for $\mathbb{E}[\mathbf{K}_n]$. Also, De Blasi et al. (2020) analysed the asymptotic behaviour of $\mathbf{K}_n$, for some decreasing weights structures, including particular cases of Geometric processes. Figure 21 illustrates the distribution of $\mathbf{K}_n$ where $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0,\mu_0)}$. In this case we do not have available a prediction rule for $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, nor for $(\mathbf{K}_n)_{n\geq 1}$. However the construction through random sets together with Proposition 3.17 gives us a sampling method that does not requires truncating the weights sequence. To be more explicit, in other to draw a sample for $\mathbf{K}_n$ we can first sample the length variable $\mathbf{v} \sim \nu_0$, then given $\mathbf{v}$, sample independently $\mathbf{N}_1, \ldots, \mathbf{N}_n$ from (3.17). Finally, independently for $1 \leq i \leq n$, chose uniformly a random element of $\boldsymbol{\Psi}_i = \{1, \ldots, \mathbf{N}_i\}$, say $\mathbf{d}_i$. This way the number of distinct values in $\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ is precisely a sample of $\mathbf{K}_n$.

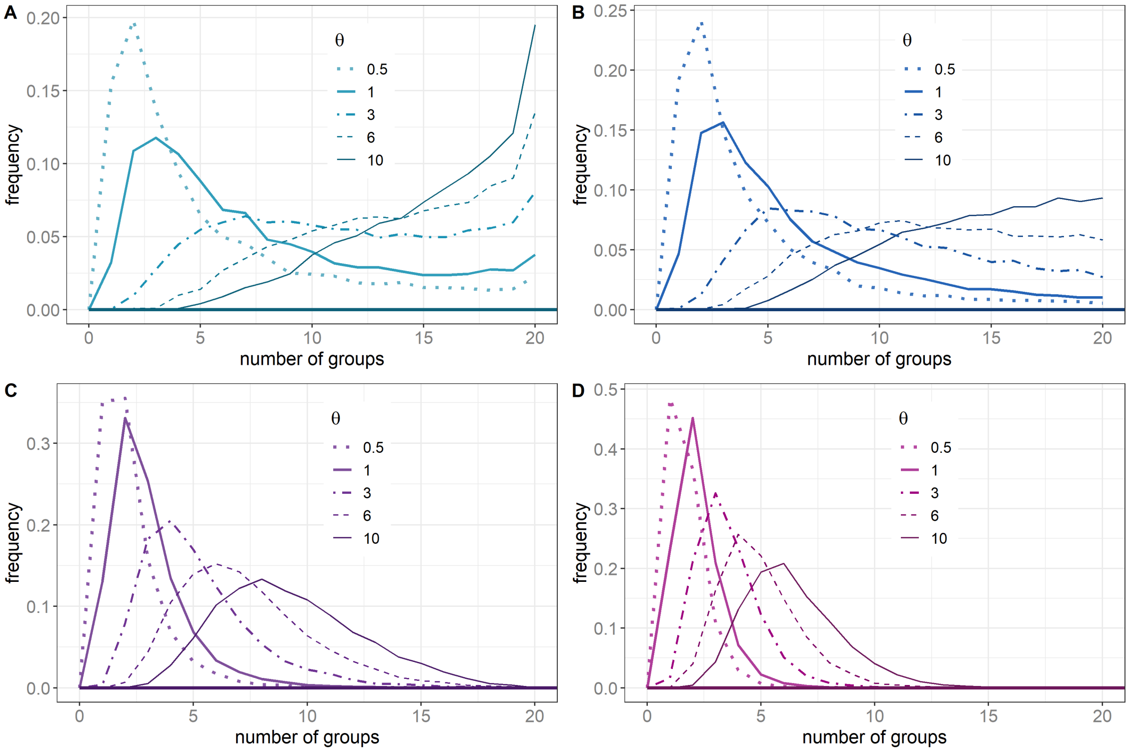


Figure 21: Frequency polygons of samples of size 10000 from the distribution of $\mathbf{K}_{20}$ corresponding to $\boldsymbol{\mu} \sim \mathcal{G}_{(\mathsf{Be}(\alpha,\theta),\mu_0)}$ for $\alpha = 0.7$ (A), $\alpha = 1$ (B), $\alpha = 3$ (C) and $\alpha = 6$ (D). For each fixed value of α we vary θ in the set $\{0.5, 1, 3, 6, 10\}$.

In Figure 21 we can observe that larger values of the second parameter, θ, of the Beta distribution, translate to a tendency of $\mathbf{K}_n$ to take bigger values. Conversely, smaller values of the first parameter, α, make the distribution of $\mathbf{K}_n$ tilt towards larger values. This behaviour is consistent with the one observed for Dirichlet and Pitman-Yor processes. In effect, small length variables lead to small values of the largest weights, hence the corresponding species sampling process, generates diverse samples.

In contrast to the size-biased permutation, decreasingly ordered weights do not allow us to compute clustering probabilities in a direct way. However, working with the decreasing representation of the weights do has important advantages, specially from an operational perspective. An obvious benefit for decreasing sequences is that the first weights are the ones that are the most representative, so for example, if truncation is necessary, working with this representation of the weights will simplify the problem. Another enormous advantage is that the decreasing ordering reduces identifiability issues that arise from Proposition 3.1, and the well-known label switching problem (see for instance Mena and Walker; 2015). Ideally one would have available the distribution for all permutations of the weights and be able to chose the ordering that performs better in a given situation. Unfortunately, this is generally not the case and one is forced to work with the representation of the weights that is the most accessible.

3.6.4 Other models

Each of the constructions described in Section 3.5, leads to generalizations or competitive alternatives to Dirichlet, Pitman-Yor and Geometric processes. For instance, Gibbs-type models are a natural generalization of the EPPF in Definition 3.10, as explained by De Blasi et al. (2015). Gibbs-type priors are characterized by having an EPPF of the form

$$\pi(n_1,\ldots,n_k) = V_{n,k}\prod_{i=1}^{k}(1-\sigma)_{n_i-1},$$

where $\sigma < 1$ and $V_{n,k}$ is a function of k and $n = \sum_{i=1}^{k} n_i$ that satisfies the recursive equation

$$V_{n,k} = (n-k\sigma)V_{n+1,k} + V_{n+1,k+1}.$$

These models share a close relation with product partition models as introduced by Hartigan (1990) and latter studied by Barry and Hartigan (1993) and Quintana and Iglesias (2003), among others. Namely, a product partition model refers to the distribution of a random partition, $\mathbf{\Pi}_n$, of $\{1,\ldots,n\}$ (not necessarily exchangeable) such that,

$$\mathbb{P}[\mathbf{\Pi}_n = \{A_1,\ldots,A_k\}] \propto \prod_{i=1}^{k} p(A_i),$$

for some positive function, p, called cohesion function. As consequence of the work by Gnedin and Pitman (2006), imposing the exchangeability constrain to a consistent family of product partition models, we are left with a Gibbs-type prior. This partially explains how complicated it can be to define a species sampling prior through its EPPF outside the class of Gibbs-type models.

One of the most widely studied constructions of SSPs is through the normalization of completely random measures. Perhaps the most famous example outside the Dirichlet process, is the normalized inverse-Gaussian process characterized by the intensity measure

$$\varrho(dx) = \frac{\exp\{-x/2\}}{x\sqrt{2x\pi}}dx.$$

For this random probability measure Favaro et al. (2012) derived a stick-breaking decomposition. Also, Lijoi et al. (2005) and James et al. (2009) analysed clustering probabilities related to normalized inverse-Gaussian processes. Further Bayesian non-parametric contributions based on NRMI's can be found in the work by Regazzini et al. (2003), Hjort et al. (2010) and Ghosal and van der Vaart (2017).

Generalizing Dirichlet and Pitman-Yor processes, stick-breaking processes featuring independent length variables have also been deeply studied (Pitman; 1996a; Ishwaran and James; 2001). There are only a handful of cases (and not general classes) of stick-breaking processes with explicit dependent length variables, the most notable example being the normalized inverse Gaussian process (Favaro et al.; 2012). General results concerning the dependent case have remained somehow elusive due to the mathematical hurdles to be overcome. In Section 4, we present the first general treatment of stick-breaking processes with dependent length variables, here we also define new Bayesian non-parametric stick-breaking priors and implement them latter in Section 5.

As to the construction using random sets, to the best of our knowledge, it has not yet been widely exploited. Outside the Geometric processes, further examples can be found in the work of De Blasi et al. (2020) and Gil-Leyva (2021). Before we move on to the main contribution of the present work, we shall mention that there are more interesting constructions and examples of SSPs apart from the ones discussed in here, some of them can be consulted in the work of Hjort et al. (2010), Phadia (2016), Castillo (2017) and Ghosal and van der Vaart (2017).

4 Stick-breaking processes with non-independent length variables

In this section we investigate the general classes of stick-breaking processes with exchangeable and Markovian length variables. Both of these classes generalize Dirichlet and Geometric processes. In fact, it is the following simple observation that motivate our models. Recall that for a Dirichlet process, its size-biased permuted weights can be decomposed as $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ where $(\mathbf{v}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \mathsf{Be}(1,\theta)$. As to Geometric process, a simple stick-breaking representation is available for its decreasing weights, $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v},\mathbf{v},\ldots)]$, where $\mathbf{v} \sim \nu_0$. Now, the sequences of Dirichlet length variables, $(\mathbf{v}_i)_{i\geq 1}$, and the one of Geometric length variables, $(\mathbf{v},\mathbf{v},\ldots)$, both have identically distributed elements, furthermore, both are exchangeable. As explained at the beginning of Section 2.1, for $(\mathbf{v}_i)_{i\geq 1}$ its directing random measure is the deterministic distribution $\mathsf{Be}(1,\theta)$, and for $(\mathbf{v},\mathbf{v},\ldots)$ it is directed by $\delta_{\mathbf{v}}$. Also, both sequences are homogeneous Markov processes, for the Dirichlet length variables we have that $\mathbf{v}_1 \sim \mathsf{Be}(1,\theta)$ and its one-step transition is given by $\mathbb{P}[\mathbf{v}_{n+1} \in dv \mid \mathbf{v}_n] = \mathsf{Be}(dv \mid 1,\theta)$, as for the Geometric length variables we have $\mathbf{v}_1 = \mathbf{v} \sim \nu_0$ and the one-step transition is precisely $\mathbb{P}[\mathbf{v}_{n+1} \in dv \mid \mathbf{v}_n] = \delta_{\mathbf{v}_n}(dv) = \delta_{\mathbf{v}}(dv)$. Other very important common features are that both sequences lead to stick-breaking weights that sum up to one, and both define stick-breaking processes with full support. A key difference between Dirichlet and Geometric length variables, is that they lay in opposite sides of the spectrum in terms of the dependence of their elements, while the elements in $(\mathbf{v}_i)_{i\geq 1}$ are completely independent, the entries of $(\mathbf{v},\mathbf{v},\ldots)$ are totally dependent. This annotation suggest that Dirichlet and Geometric processes can be generalized by means of exchangeable sequences and homogeneous Markov chains taking values in $[0,1]$. Moreover, this generalization will lead to general classes of priors for which Dirichlet and Geometric models are precisely the extreme points.

The rest of this section is organized as follows, in Section 4.1 we investigate the general class of stick-breaking processes with exchangeable length variables. Here we derive conditions on the directing random measure of the length variables to assure the SSP they define has full support and is proper. We also explain how to recover Dirichlet and Geometric processes as weak limits of stick-breaking processes with exchangeable length variables. For a wide subclass we compute the probability that consecutive weights are decreasingly ordered and explain how the stochastic ordering of the weights can be modulated by a single real-valued parameter, when the directing random measure of the length variables is a SSP as well. We finish Section 4.1 by specializing the analysis to the case where the length variables are directed by a Dirichlet process, and explain how the results developed can also be specialized to the case where the length variables are driven by a Pitman-Yor process or other interesting species sampling models. Section 4.2 is concerned with stick-breaking processes whose length variables form a Markov chain. Here we replicate the analysis of Section 4.1 for Markovian length variables. Informally the transition of the length variables plays similar role on Markov stick-breaking processes than that of the directing measure of the length variables on exchangeable stick-breaking processes. In Section 4.2.2 we derive conditions on the initial distribution and the transition of stationary Markov length variables so the SSP they define is propper and has full support. We also explain how Dirichlet and Geometric processes can be recovered as weak limits of stick-breaking processes with Markovian length variables. Latter, Sections 4.2.3 and 4.2.4, present different examples of stationary Markov processes of length variables that define proper SSPs with full support. We see that for these special cases, the stochastic ordering of the weights can be modulated by a single real-valued parameter, as well as how alike is the model to Dirichlet and Geometric process. In particular for the examples in Section 4.2.4 we will be able to compute the probability that consecutive weights are decreasingly ordered. To finalize the prior analysis of stick-breaking processes with non-independent length variables, in Section 4.3 we study non-stationary length variables, this automatically discards exchangeable length variables, but if we relax the hypothesis that the Markovian length variables are homogeneous and have a stationary distribution, will be able to study stick-breaking processes with non identically distributed length variables. In fact a motivation for the last kind of processes we study here is to generalize the models in Section 4.2 to contain or approximate Pitman-Yor process in addition to Dirichlet processes. Now, the analysis of stick-breaking processes with non-independent and non identically distributed length variables can become arduous. For this reason we simply focus in generalizing the examples in Section 4.2.4. In effect Section 4.3 is meant to illustrate how the models in Section 4.2 can be further generalized in interesting directions, and leave a path open for future research.

4.1 Stick-breaking processes with exchangeable length variables

Our objects of study here are stick-breaking processes whose weights' distribution remains invariant under permutations of the length variables (Gil-Leyva and Mena; 2021). We begin by defining this class of species sampling processes.

Definition 4.1. *We call an exchangeable stick-breaking process (ESB) to any SSP as in* (3.1) *whose weights* $(\mathbf{w}_j)_{j\geq 1}$ *decompose as* $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ *for some exchangeable sequence* $(\mathbf{v}_i)_{i\geq 1}$ *with* $[0,1]$*-valued elements. To the corresponding weights, we call an exchangeable stick-breaking weights sequence (ESBw).*

From a Bayesian perspective, one of the first properties one should analyse is whether a species sampling process has full support and if it is discrete almost surely. The following result, gives sufficient conditions on the directing random measure and the de Finetti measure of the exchangeable length variables so the respective ESB has full support and is proper.

Theorem 4.1. *Let* $\boldsymbol{\mu}$ *be an ESB with weights* $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$*, for some exchangeable sequence* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu}$*, with* $\nu_0 = \mathbb{E}[\boldsymbol{\nu}]$*.*

i) *If there exist* $\varepsilon > 0$ *such that* $(0,\varepsilon)$ *is contained in the support of* ν_0*, then* $\boldsymbol{\mu}$ *has full support.*

ii) $\boldsymbol{\mu}$ *is discrete almost surely if and only if* $\boldsymbol{\nu}(\{0\}) < 1$ *almost surely.*

The proof of Theorem 4.1, can be found in Appendix D.1. In this context, if $0 = \nu_0(\{0\}) = \mathbb{E}[\boldsymbol{\nu}(\{0\})]$, we have that $\boldsymbol{\nu}(\{0\}) = 0$ almost surely. Hence, a sufficient condition to assure $\sum_{j\geq 1}\mathbf{w}_j = 1$, is that 0 is not an atom of ν_0, that is to say, $\mathbb{P}[\mathbf{v}_i = 0] = 0$.

For example, if $\nu_0 = \mathsf{Be}(a, b)$ for some $a, b > 0$, so that marginally $\mathbf{v}_i \sim \mathsf{Be}(a, b)$, then 0 is not an atom of ν_0. Furthermore, the support of a $\mathsf{Be}(a, b)$ distribution is $[0, 1]$, which means that the conditions given in (i) and (ii) of Theorem 4.1 are satisfied and we have the following Corollary.

Corollary 4.2. *Let $\boldsymbol{\mu}$ be a species sampling process with weights collection, $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ for some exchangeable sequence $(\mathbf{v}_i)_{i\geq 1}$, such that marginally $\mathbf{v}_i \sim \mathsf{Be}(a, b)$. Then, $\boldsymbol{\mu}$ is proper and it has full support.*

Notice that Geometric processes with a Beta distributed length variable and stick-breaking processes featuring i.i.d. Beta lengths variables, including Dirichlet processes, are all particular cases of the SSPs in Corollary 4.2. Our next convergence result (see Appendix D.2 for a proof) explains how Geometric and Dirichlet processes can be recover as limits of non-trivial ESBs.

Theorem 4.3. *Let $(S, \mathscr{B}_S)$ be a Polish space. For each $n \geq 1$ consider a diffuse probability measure, $\mu_0^{(n)}$, over $(S, \mathscr{B}_S)$ and a random probability measure, $\boldsymbol{\nu}^{(n)}$, over $\big([0, 1], \mathscr{B}_{[0,1]}\big)$ such that $\boldsymbol{\nu}^{(n)}(\{0\}) < 1$ almost surely. Let $\boldsymbol{\mu}^{(n)}$ be an ESB with base measure $\mu_0^{(n)}$ and length variables $\left\{\mathbf{v}_1^{(n)}, \mathbf{v}_2^{(n)} \ldots \,\middle|\, \boldsymbol{\nu}^{(n)}\right\} \overset{iid}{\sim} \boldsymbol{\nu}^{(n)}$. Let us denote $\mathbf{V}^{(n)} = \left(\mathbf{v}_i^{(n)}\right)_{i\geq 1}$ and set $\mathbf{W}^{(n)} = \mathsf{SB}\left[\mathbf{V}^{(n)}\right]$. Say that as $n \to \infty$, $\mu_0^{(n)}$ converges weakly to the diffuse probability measure μ_0.*

i) *If $\boldsymbol{\nu}^{(n)}$ converges weakly in distribution to a deterministic probability measure ν_0 with $\nu_0 \neq \delta_0$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to a stick-breaking process, $\boldsymbol{\mu}$, with base measure μ_0 and featuring independent length variables $(\mathbf{v}_i)_{i\geq 1} \overset{iid}{\sim} \nu_0$, as $n \to \infty$. In particular if $\nu_0 = \mathsf{Be}(1, \theta)$, the limit $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta, \mu_0)}$, and $\mathbf{W}^{(n)}$ converges in distribution to the size-biased permuted weights of $\boldsymbol{\mu}$.*

ii) *If $\boldsymbol{\nu}^{(n)}$ converges weakly in distribution to $\delta_{\mathbf{v}}$ for some $[0, 1]$-valued random variable $\mathbf{v} \sim \nu_0$, where $\nu_0(\{0\}) = 0$. Then, as $n \to \infty$, $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0, \mu_0)}$, and $\mathbf{W}^{(n)}$ converges in distribution to the decreasingly ordered weights of $\boldsymbol{\mu}$.*

Before we move on to some examples and special sub-classes of ESBs, recall that some important quantities related to stick-breaking process can be written in terms of expectations of power products of length variables, $\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1 - \mathbf{v}_j)^{b_j}\right]$, for non-negative integers $(a_j, b_j)_{j=1}^k$. For instance, equations (3.6), (3.7) and (3.8) illustrate this for the tie probability, the EPPF and the pEPPF, respectively. This is why throughout this section we will be placing emphasis in computing the quantities $\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1 - \mathbf{v}_j)^{b_j}\right]$. In particular, if the length variables are exchangeable, for any non-negative integers $(a_j, b_i)_{j=1}^k$, we have that

$$\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1 - \mathbf{v}_j)^{b_j}\right] = \int \left\{\prod_{j=1}^k \int v^{a_j}(1 - v)^{b_j} \nu(dv)\right\} \mathsf{Q}(d\nu) \tag{4.1}$$

where Q is the distribution of directing random measure of $(\mathbf{v}_i)_{i\geq 1}$.

Example 4.1. *Fix $\alpha, \theta > 0$ and let $\mathbf{v} \sim \mathsf{Be}(\alpha, \theta)$. Also fix $\kappa \in \mathbb{N}$ and consider $\{\mathbf{z} \mid \mathbf{v}\} \sim \mathsf{Bin}(\kappa, \mathbf{v})$, so that marginally*

$$\mathbb{P}[\mathbf{z} = z] = \frac{(\alpha)_z(\theta)_{\kappa-z}}{(\alpha+\theta)_\kappa},$$

for every $z \in \{0, \dots, \kappa\}$. Now, let $\{\mathbf{v}_1, \mathbf{v}_2, \dots \mid \mathbf{z}\} \overset{iid}{\sim} \mathsf{Be}(\alpha + \mathbf{z}, \theta + \kappa - \mathbf{z})$. Clearly $(\mathbf{v}_i)_{i\geq 1}$ is exchangeable, and from the Beta-Binomial conjugate model we get that the marginal distribution of $\mathbf{v}_i$ is precisely a $\mathsf{Be}(\alpha, \theta)$ distribution. This means, by Corollary 4.2, that any SSP with length variables $(\mathbf{v}_i)_{i\geq 1}$, is proper and has full support. Moreover, by (4.1) we can compute for any $k \geq 1$ and non-negative integers $(a_j, b_j)_{j=1}^k$,

$$\begin{aligned}
&\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] \\
&= \sum_{z=0}^{\kappa} \left\{\prod_{j=1}^k \frac{\Gamma(\alpha+\theta)}{\Gamma(\alpha)\Gamma(\theta)} \int v^{\alpha+z+a_j-1}(1-v)^{\theta+\kappa-z+b_j-1} dv\right\} \frac{(\alpha)_z(\theta)_{\kappa-z}}{(\alpha+\theta)_\kappa} \\
&= \sum_{z=0}^{\kappa} \left\{\prod_{j=1}^k \frac{(\alpha)_{z+a_j}(\theta)_{\kappa-z+b_j}}{(\alpha+\theta)_{\kappa+a_j+b_j}}\right\} \frac{(\alpha)_z(\theta)_{\kappa-z}}{(\alpha+\theta)_\kappa}.
\end{aligned}$$

Hence for the corresponding pEPPF we obtain

$$\pi'(n_1, \dots, n_k) = \sum_{z=0}^{\kappa} \left\{\prod_{j=1}^k \frac{(\alpha)_{z+n_j-1}(\theta)_{\kappa-z+m_j}}{(\alpha+\theta)_{\kappa+m_{j-1}-1}}\right\} \frac{(\alpha)_z(\theta)_{\kappa-z}}{(\alpha+\theta)_\kappa}.$$

where $m_j = \sum_{i>j} n_i$. This shows that generally $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ is not invariant under size-biased permutations because π' is not a symmetric function.

To illustrate Theorem 4.3, consider $\mathbf{v} \sim \mathsf{Be}(\alpha, \theta)$ as above, and for each $\kappa \in \{0, 1, 2, \dots\}$ let $\left\{\mathbf{z}^{(\kappa)} \,\middle|\, \mathbf{v}\right\} \sim \mathsf{Bin}(\kappa, \mathbf{v})$, with the convention $\mathsf{Bin}(0, \mathbf{v}) = \delta_0$. Set $\left\{\mathbf{v}_1^{(\kappa)}, \mathbf{v}_2^{(\kappa)}, \dots \,\middle|\, \mathbf{z}^{(\kappa)}\right\} \overset{iid}{\sim} \mathsf{Be}\left(\alpha + \mathbf{z}^{(\kappa)}, \theta + \kappa - \mathbf{z}^{(\kappa)}\right)$, so that the directing random measure of $\mathbf{V}^{(\kappa)} = \left(\mathbf{v}_i^{(\kappa)}\right)_{i\geq 1}$ is precisely $\boldsymbol{\nu}^{(\kappa)} = \mathsf{Be}\left(\alpha + \mathbf{z}^{(\kappa)}, \theta + \kappa - \mathbf{z}^{(\kappa)}\right)$. It is easy to show (see Lemma 4.16 below) that $\boldsymbol{\nu}^{(\kappa)}$ converges weakly in distribution to $\delta_{\mathbf{v}}$ as $\kappa \to \infty$. Thus, the second part of Theorem 4.3 proves that the ESBs, $\boldsymbol{\mu}^{(\kappa)}$, with length variables $\mathbf{V}^{(\kappa)}$, and base measure $\mu_0^{(\kappa)}$ converge weakly in distribution to $\boldsymbol{\mu}^{(\infty)} \sim \mathcal{G}_{(\mathsf{Be}(\alpha,\theta),\mu_0)}$ as $\kappa \to \infty$, whenever the base measures $\left(\mu_0^{(\kappa)}\right)_{\kappa\geq 1}$ converge weakly μ_0. Also the choice $\kappa = 0$ yields $\boldsymbol{\nu}^{(\kappa)} = \boldsymbol{\nu}^{(0)} = \mathsf{Be}(\alpha, \theta)$. Particularly if $\alpha = 1$, we get that the ESB, $\boldsymbol{\mu}^{(0)}$, with length variables $\mathbf{V}^{(0)}$ is a Dirichlet process.

4.1.1 Exchangeable stick-breaking processes driven by species sampling models

As previously explained, SSPs are very flexible models, recall that these ones constitute the class of random probability measures with exchangeable increments with respect to diffuse finite measures. Moreover, depending on the distribution of the atoms and weights of an SSP, its weak topological support can be as wide as the complete space of probability measures over the measurable space where they are defined. Since we are interested in stick-breaking processes featuring exchangeable length variables, it seems sensible to analyse the special case where the length variables are driven by some SSP themselves. To be precise, we will study ESBs denoted by $\boldsymbol{\mu}$, with length variables $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{\text{iid}}{\sim} \boldsymbol{\nu}$, where $\boldsymbol{\nu}$ is a SSP over $\big([0,1], \mathscr{B}_{[0,1]}\big)$. In this instance, we simply say $\boldsymbol{\mu}$ is an ESB driven by the SSP $\boldsymbol{\nu}$ (see Figure 22 for an illustration of the underlying dependence structure of $\boldsymbol{\mu}$). To avoid confusion we denote the EPPF corresponding to $\boldsymbol{\nu}$ by π_ν, its tie probability by ρ_ν and its base measure by ν_0.

$$\begin{tikzcd}
0 \leq \mathbf{p}_j,\ \sum_{j\geq 1} \mathbf{p}_j \leq 1 \arrow[dr] & & \left(\mathbf{v}_j^*\right)_{j\geq 1} \overset{iid}{\sim} \nu_0 \arrow[dl] & \\
& \boldsymbol{\nu} = \sum_{j\geq 1} \mathbf{p}_j \delta_{\mathbf{v}_j^*} + \left(1 - \sum_{j\geq 1} \mathbf{p}_j\right) \nu_0 \arrow[d] & & \\
& \{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu} \arrow[d] & & \\
& (\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}] \arrow[dr] & & (\boldsymbol{\xi}_j)_{j\geq 1} \overset{iid}{\sim} \mu_0 \arrow[dl] \\
& & \boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} &
\end{tikzcd}$$

Figure 22: Underlying dependence structure of an ESB, $\boldsymbol{\mu}$, driven by a species sampling process, $\boldsymbol{\nu}$.

Note that in Figure 22 we assumed the ESB, $\boldsymbol{\mu}$, driven by the SSP, $\boldsymbol{\nu}$, is proper, this is justified by the second part of Theorem 4.1. Indeed, as $\boldsymbol{\nu}$ is a SSP, by definition its base measure, $\nu_0 = \mathbb{E}[\boldsymbol{\nu}]$ is required to be diffuse, thus $\boldsymbol{\nu}(\{0\}) = 0$ almost surely, and Theorem 4.1 (ii) yields $\boldsymbol{\mu}$ is proper. By Theorem 4.1 (i) we also know that if $(0, \varepsilon)$ belongs to the support of the base measure, ν_0, for some $0 < \varepsilon < 1$, then $\boldsymbol{\mu}$ has full support. This is summarized by the following result.

Corollary 4.4. *Let $\boldsymbol{\mu}$ be an ESB driven by a SSP $\boldsymbol{\nu}$ with base measure ν_0. Then*

i) *$\boldsymbol{\mu}$ is discrete almost surely.*

ii) *If there exist $0 < \varepsilon < 1$ such that $(0, \varepsilon)$ belongs to the support of ν_0, $\boldsymbol{\mu}$ has full support. In particular if $\nu_0 = \mathsf{Be}(\alpha, \theta)$, $\boldsymbol{\mu}$ has full support.*

Now, putting together Theorems 3.4 and 4.3 we get that modulating the tie probability, ρ_ν, of the underlying SSP, $\boldsymbol{\nu}$, we can arbitrarily approximate Dirichlet and Geometric processes by means of ESBs driven by SSPs. This is a very appealing result in contrast to the general version, while Theorem 4.3 modulates the convergence through random probability measures, which can be infinite dimensional, the next Corollary controls the convergence is terms of a single number, $\rho_\nu \in [0, 1]$.

Corollary 4.5. *Let $(S, \mathscr{B}_S)$ be a Polish space. Consider some diffuse probability measures $\mu_0, \mu_0^{(1)}, \mu_0^{(2)}, \ldots,$ over $(S, \mathscr{B}_S)$, and some diffuse probability measures, $\nu_0, \nu_0^{(1)}, \nu_0^{(2)}, \ldots,$ over $\left([0,1], \mathscr{B}_{[0,1]}\right)$. Say that as $n \to \infty$, $\mu_0^{(n)}$ converges weakly to μ_0 and $\nu_0^{(n)}$ converges weakly to ν_0. For $n \geq 1$, let $\rho_\nu^{(n)} \in (0,1)$, and consider the ESB $\boldsymbol{\mu}^{(n)}$ with base measure $\mu_0^{(n)}$, and length variables $\left\{\mathbf{v}_1^{(n)}, \mathbf{v}_2^{(n)} \ldots \,\middle|\, \boldsymbol{\nu}^{(n)}\right\} \overset{iid}{\sim} \boldsymbol{\nu}^{(n)}$, where $\boldsymbol{\nu}^{(n)}$ is a SSP with base measure $\nu_0^{(n)}$ and tie probability $\rho_\nu^{(n)}$. Also consider the stick-breaking weights of $\boldsymbol{\mu}^{(n)}$, $\mathbf{W}^{(n)} = \mathsf{SB}\left[\mathbf{V}^{(n)}\right]$, where $\mathbf{V}^{(n)} = \left(\mathbf{v}_i^{(n)}\right)_{i \geq 1}$.*

i) *If $\rho_\nu^{(n)} \to 0$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to a stick-breaking process, $\boldsymbol{\mu}$, with base measure, μ_0, and independent length variables $(\mathbf{v}_i)_{i \geq 1} \overset{iid}{\sim} \nu_0$. Particularly, if $\nu_0 = \mathsf{Be}(1, \theta)$, $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta, \mu_0)}$ and $\mathbf{W}^{(n)}$ converges in distribution to the size-biased permuted weights of $\boldsymbol{\mu}$.*

ii) *If $\rho_\nu^{(n)} \to 1$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0, \mu_0)}$, and $\mathbf{W}^{(n)}$ converges in distribution to the decreasingly ordered weights of $\boldsymbol{\mu}$.*

Corollary 4.5 has major consequences when it comes to understand the ordering of the weights of an ESB. For general ESB's with length variables, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{\text{iid}}{\sim} \boldsymbol{\nu}$, we have that

$$
\begin{aligned}
\mathbb{E}[\mathbf{w}_{j+1}] = \mathbb{E}\left[\mathbf{v}_j \prod_{i=1}^{j} (1 - \mathbf{v}_i)\right] &= \mathbb{E}\left[\mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}](1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^j\right] \\
&\leq \mathbb{E}\left[\mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}](1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^{j-1}\right] = \mathbb{E}[\mathbf{w}_j],
\end{aligned}
$$

that is the expected weights are decreasing. In most cases this is not true for the weights themselves. One of the most interesting features about ESBs driven by SSPs is that, for a fixed base measure, ν_0, by simply tuning the tie probability, we can modulate how likely are the weights to be decreasing. Moreover, as the following theorem shows (see Appendix D.3 for a proof) the probability that consecutive weights are decreasing, ranges from a quantity determined by the underlying base measure and one.

Theorem 4.6. *Fix a diffuse probability measure ν_0 over $\left([0,1], \mathscr{B}_{[0,1]}\right)$. For each $\rho_\nu \in (0,1)$, let $\left\{\mathbf{v}_1^{(\rho_\nu)}, \mathbf{v}_1^{(\rho_\nu)}, \ldots \mid \boldsymbol{\nu}^{(\rho_\nu)}\right\} \overset{iid}{\sim} \boldsymbol{\nu}^{(\rho_\nu)}$, for some SSP, $\boldsymbol{\nu}^{(\rho_\nu)}$, with base measure ν_0, and tie probability ρ_ν. Set $\left(\mathbf{w}_j^{(\rho_\nu)}\right)_{j \geq 1} = \mathsf{SB}\left[\left(\mathbf{v}_i^{(\rho_\nu)}\right)_{i \geq 1}\right]$. Then, for every $j \geq 1$,*

a) *$\mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right] = \rho_\nu + (1 - \rho_\nu)\mathbb{E}\left[\overrightarrow{\nu_0}(c(\mathbf{v}))\right]$, where $c(v) = 1 \wedge v(1-v)^{-1}$ for every $v \in [0,1]$, $\mathbf{v} \sim \nu_0$, and $\overrightarrow{\nu_0}$ is the distribution function of ν_0, that is $\overrightarrow{\nu_0}(x) = \nu_0([0,x])$.*

b) *The mapping $\rho_\nu \mapsto \mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right]$ is continuous and non-decreasing. Particularly, if $(0, \varepsilon)$ is contained in the support of ν_0, for some $\varepsilon > 0$, the mapping, $\rho_\nu \mapsto \mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right]$, is strictly increasing.*

c) *As $\rho_\nu \to 1$, $\mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right] \to 1$, and as $\rho_\nu \to 0$, $\mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right] \to \mathbb{E}\left[\overrightarrow{\nu_0}(c(\mathbf{v}))\right]$, where c, $\overrightarrow{\nu_0}$ and $\mathbf{v}$ are as in* (a)*.*

d) *For every* $\rho_\nu \in (0,1)$, $\mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right] \geq \mathbb{E}\left[\overrightarrow{\nu_0}(c(\mathbf{v}))\right]$, *where* c, $\overrightarrow{\nu_0}$ *and* $\mathbf{v}$ *are as in* (a).

Suprisingly, comparing (b) of Theorem 4.6 and (ii) in Corollary 4.4, we found that the requirement over the base measure, ν_0, that assures the ESB has full support, also guarantees the mapping, $\rho_\nu \mapsto \mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right]$, is strictly increasing. That is, for most ESBs of interest we will have that the probability that consecutive weights are decreasingly ordered grows with underlying tie probability, ρ_ν. As previously mentioned, this is the case of $\nu_0 = \mathsf{Be}(a,b)$. In Section 4.1.2 below, we will further specialize Theorem 4.6 for the case $\nu_0 = \mathsf{Be}(1,\theta)$. For now, we shall highlight that if $\nu_0 = \mathsf{Be}(1,\theta)$, as a consequence of Corollary 4.5, we know that as $\rho_\nu \to 0$ the corresponding weights converge in distribution to size-biased permuted weights. This is not true for other choices of ν_0, what remains true for any diffuse probability measure, ν_0 is that if $\rho_\nu \to 1$, the weights in question converge in distribution to decreasing weights. Recall that the ordering of the weights has important consequences when it comes to modelling, for example if the interest is in clustering, working with size-biased ordered weights can be advantageous, in contrast, working with decreasingly ordered weights reduces identifiability problems that arise from the invariance of the prior under permutations of its weights.

For the weights, $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$, where $(\mathbf{v}_i)_{i\geq 1}$ is driven by a SSP, another quantity that can be easily computed is the conditional probability

$$\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j \mid \mathbf{w}_1, \dots, \mathbf{w}_j] = \mathbb{P}[\mathbf{v}_{j+1} \leq c(\mathbf{v}_j) \mid \mathbf{v}_1, \dots, \mathbf{v}_j],$$

where c is as in Theorem 4.6 (a). First of all, to see that these probabilities are equal (almost surely), note that from the stick-breaking decomposition of the weights we get $\mathbf{w}_{j+1} \leq \mathbf{w}_j$ if and only if $\mathbf{v}_{j+1} \leq \mathbf{v}_j(1-\mathbf{v}_j)^{-1}$, which in turn is equivalent to $\mathbf{v}_{j+1} \leq c(\mathbf{v}_j)$. It is also straight forward from the stick-breaking construction that $(\mathbf{w}_1, \dots, \mathbf{w}_j)$ is $(\mathbf{v}_1, \dots, \mathbf{v}_j)$-measurable. Conversely, the proof Proposition 3.13 yields $(\mathbf{v}_1, \dots, \mathbf{v}_j)$ is $(\mathbf{w}_1, \dots, \mathbf{w}_j)$-measurable as well, when $0 < \mathbf{v}_i < 1$ almost surely for every $i \geq 1$, this is of course the case of ESBs, because the underlying base measure is diffuse. Now, since $(\mathbf{v}_i)_{i\geq 1}$ is sampled from a SSP, we already know from Theorem 3.6 how to compute $\mathbb{P}[\mathbf{v}_{j+1} \leq c(\mathbf{v}_j) \mid \mathbf{v}_1, \dots, \mathbf{v}_j]$, and we obtain our following result.

Theorem 4.7. *Let* $\boldsymbol{\nu}$ *be a species sampling process over* $\left([0,1], \mathscr{B}_{[0,1]}\right)$, *with base measure* ν_0 *and EPPF* π_ν. *Consider* $\{\mathbf{v}_1, \mathbf{v}_2, \dots \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu}$ *and define* $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$. *Set* c *and* $\overrightarrow{\nu_0}$ *as in Theorem 4.6. Then*

$$\begin{aligned}\mathbb{P}[\mathbf{w}_{n+1} \leq \mathbf{w}_n \mid \mathbf{w}_1, \dots, \mathbf{w}_n] &= \mathbb{P}[\mathbf{v}_{n+1} \leq c(\mathbf{v}_n) \mid \mathbf{v}_1, \dots, \mathbf{v}_n] \\ &= \sum_{j=1}^{\mathbf{K}_n} \frac{\pi_\nu\left(\mathbf{n}^{(j)}\right)}{\pi_\nu(\mathbf{n})} \mathbf{1}_{\{\mathbf{v}_j^* \leq c(\mathbf{v}_n)\}} + \frac{\pi_\nu\left(\mathbf{n}^{(\mathbf{K}_n+1)}\right)}{\pi_\nu(\mathbf{n})} \overrightarrow{\nu_0}(c(\mathbf{v}_n)),\end{aligned}$$

where $\mathbf{v}_1^*, \dots, \mathbf{v}_{\mathbf{K}_n}^*$ *are the distinct values that* $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ *exhibits,* $\mathbf{n} = (\mathbf{n}_1, \dots \mathbf{n}_{\mathbf{K}_n})$ *is given by* $\mathbf{n}_j = |\{i \leq n : \mathbf{v}_i = \mathbf{v}_j^*\}|$, $\mathbf{n}^{(j)} = (\mathbf{n}_1, \dots \mathbf{n}_{j-1}, \mathbf{n}_j + 1, \mathbf{n}_{j+1}, \dots, \mathbf{n}_{\mathbf{K}_n})$ *and* $\mathbf{n}^{(\mathbf{K}_n+1)} = (\mathbf{n}_1, \dots, \mathbf{n}_{\mathbf{K}_n}, 1)$.

A variety of properties of exchangeable length variables driven by SSPs are described in Section 3.3, for example from Corollary 3.9 we can compute some conditional moments. An specially important property is described in Theorem 3.7, as it allows to derive expectations of power products of length variables.

Theorem 4.8. *Consider the length variables* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu}$*, for some species sampling process,* $\boldsymbol{\nu}$*, with base measure* ν_0 *and EPPF* π_ν*. Then, for every sequence of non-negative integers* $(a_j, b_j)_{j=1}^k$*,*

$$\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{\{A_1,\ldots,A_m\}} \pi_v(|A_1|,\ldots,|A_m|)\times \\ \times \prod_{j=1}^m \left\{\int_{[0,1]} (v)^{\sum_{i\in A_j} a_i}(1-v)^{\sum_{i\in A_j} b_i}\,\nu_0(dv)\right\}, \tag{4.2}$$

where the sum ranges over the set of all partitions of $\{1,\ldots,k\}$*. In particular if* $\nu_0 = \mathsf{Be}(\alpha,\theta)$*, the integrals in* (4.2) *reduce to*

$$\int_{[0,1]} (v)^{\sum_{i\in A_j} a_i}(1-v)^{\sum_{i\in A_j} b_i}\,\nu_0(dv) = \frac{\Gamma(a+b)\Gamma(a+\sum_{i\in A_j} a_i)\Gamma(b+\sum_{i\in A_j} b_i)}{\Gamma(a)\Gamma(b)\Gamma(a+b+\sum_{i\in A_j}(a_i+b_i))}.$$

Inserting (4.2) into equations (3.6) and (3.7) we obtain expressions for the tie, ρ, probability and the EPPF , π, corresponding to the ESB $\boldsymbol{\mu}$. Unfortunately, these are rather hard to manage as are written in terms of infinitely many unordered sums. Also, inserting (4.2) into (3.8) we obtain an expression for the pEPPF corresponding to the weights. Although the pEPPF is also expressed through an unordered sum, it does not has infinitely many terms. As we will see in Section 5, this quantity has a nice interpretation in terms of the so-called latent allocation variables, which in turn provide an alternative to analyse clusters when the EPPF is not available (Fuentes-García et al.; 2019).

Clearly, Dirichlet and Geometric processes are two examples ESBs driven by SSPs. Indeed, the sequences of length variables $(\mathbf{v}_i)_{i\ge1} \overset{\text{iid}}{\sim} \mathsf{Be}(1,\theta)$ and $(\mathbf{v},\mathbf{v},\ldots)$ are both exchangeable and driven by the trivial SSPs $\boldsymbol{\nu} = \mathsf{Be}(1,\theta)$, with tie probability $\rho_\nu = 0$ and $\boldsymbol{\nu} = \delta_{\mathbf{v}}$, with tie probability $\rho_\nu = 1$, respectively. Next we focus in a sub-class of ESBs driven by a SSP, to which Dirichlet and Geometric processes do not belong, but can still be recovered as weak limits.

4.1.2 Dirichlet driven stick-breaking processes

ESBs driven by SSPs remain too general to specify a Bayesian non parametric prior. As mentioned in Section 3.6.1, the Dirichlet process is the canonical example of SSPs in Bayesian literature, mainly due to its mathematical tractability. So here we specialize to exchangeable length variables, $\mathbf{V} = (\mathbf{v}_i)_{i\ge1}$, driven by a Dirichlet process, $\boldsymbol{\nu} \sim \mathcal{D}_{(\beta,\nu_0)}$, with total mass parameter β and base measure ν_0. For the sake of a simpler analysis, and motivated by Corollary 4.2 and by the first part of Corollary 4.5, we will further concentrate in the case $\nu_0 = \mathsf{Be}(1,\theta)$.

Definition 4.2. *Let* $(S, \mathscr{B}_S)$ *be a Borel space and consider a diffuse probability measure,* μ_0*, over* $(S,\mathscr{B}_S)$*. Let* $\boldsymbol{\nu} \sim \mathcal{D}_{(\beta,\nu_0)}$ *where* $\nu_0 = \mathsf{Be}(1,\theta)$*. To the SSP,* $\boldsymbol{\mu}$*, with base measure* μ_0 *and exchangeable length variables,* $\{\mathbf{v}_1,\mathbf{v}_2,\ldots \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu}$*, we call a Dirichlet driven stick-breaking process (DSB) with parameter* (β,θ,μ_0)*, to its distribution we denote by* $\mathcal{DSB}_{(\beta,\theta,\mu_0)}$*. The weights sequence,* $\mathbf{W} = (\mathbf{w}_j)_{j\ge1} = \mathsf{SB}[(\mathbf{v}_i)_{i\ge1}]$*, will be referred to as Dirichlet driven stick-breaking weights sequence (DSBw) with parameters* (β,θ)*.*

The choice $\nu_0 = \mathsf{Be}(1,\theta)$, not only guaranties that the corresponding species sampling process is proper and has full support, but also allows us to recover Geometric and Dirichlet processes in the weak limits as $\beta \to 0$ ($\rho_\nu = 1/(\beta+1) \to 1$) and $\beta \to \infty$ ($\rho_\nu = 1/(\beta+1) \to 0$), respectively.

Corollary 4.9. *Let $(S, \mathscr{B}_S)$ be a Polish space and let $\mu_0, \mu_0^{(1)}, \mu_0^{(2)}$ be diffuse probability measures over $(S,\mathscr{B}_S)$ such that $\mu_0^{(n)}$ converges weakly to μ_0, as $n \to \infty$. For each $n \geq 1$ let $\beta^{(n)}, \theta^{(n)} \in (0,\infty)$, with $\theta^{(n)} \to \theta$ in $(0,\infty)$. Let $\boldsymbol{\mu}^{(n)} \sim \mathcal{DSB}_{\left(\beta^{(n)},\theta^{(n)},\mu_0^{(n)}\right)}$ and let $\mathbf{W}^{(n)}$ be the corresponding* DSBw *with parameters $\left(\beta^{(n)}, \theta^{(n)}\right)$.*

i) *If $\beta^{(n)} \to \infty$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\boldsymbol{\mu}^{(\infty)}$, where $\boldsymbol{\mu}^{(\infty)} \sim \mathcal{D}_{(\theta,\mu_0)}$, and $\mathbf{W}^{(n)}$ converges in distribution to the size-biased permutation of the weights of $\boldsymbol{\mu}^{(\infty)}$.*

ii) *If $\beta^{(n)} \to 0$, as $n \to \infty$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\boldsymbol{\mu}^{(0)}$, where $\boldsymbol{\mu}^{(0)} \sim \mathcal{G}_{(\mathsf{Be}(1,\theta),\mu_0)}$, and $\mathbf{W}^{(\beta)}$ converges in distribution to the decreasingly ordered weights of $\boldsymbol{\mu}^{(0)}$.*

Corollary 4.9 follows immediately from Corollary 4.5 by substituting $\rho_\nu^{(n)} = 1/\left(1+\beta^{(n)}\right)$. As to the ordering of DSBw's, we have the following Corollary of Theorems 4.6 and 4.7 (see Appendix D.4, for details in calculations).

Corollary 4.10. *Fix $\theta > 0$, and for each $\beta > 0$, consider a DSBw, $\left(\mathbf{w}_j^{(\beta)}\right)_{j\geq 1} = \mathsf{SB}\left[\left(\mathbf{v}_i^{(\beta)}\right)_{i\geq 1}\right]$, with parameters (β,θ). Let us denote by ${}_2F_1$ to the Gauss hypergeometric function. Then, for every $j \geq 1$,*

a) $\mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)}\right] = 1 - \dfrac{{}_2F_1(1,1;\theta+2,1/2)\beta\theta}{2(\beta+1)(\theta+1)}$, *for every $\beta > 0$.*

b) *The mapping $\beta \mapsto \mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)}\right]$ is continuous and strictly decreasing.*

c) *As $\beta \to \infty$, $\mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)}\right] \to 1 - \dfrac{{}_2F_1(1,1;\theta+2,1/2)\theta}{2(\theta+1)}$.*

d) *As $\beta \to 0$, $\mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)}\right] \to 1$.*

e) $\mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)}\right] \geq 1 - \dfrac{{}_2F_1(1,1;\theta+2,1/2)\theta}{2(\theta+1)}$, *for every $\beta > 0$.*

f) $$\mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)} \,\middle|\, \mathbf{w}_1^{(\beta)},\ldots,\mathbf{w}_j^{(\beta)}\right] = \frac{1}{\beta+j}\left\{\sum_{i\in\mathbf{A}_j}\mathbf{n}_i + \beta\left[1-\left(1-c\left(\mathbf{v}_j^{(\beta)}\right)\right)^\theta\right]\right\},$$
where $\mathbf{A}_j = \left\{i \leq \mathbf{K}_j : \mathbf{v}_i^* \leq c\left(\mathbf{v}_j^{(\beta)}\right)\right\}$, $c(v) = 1 \wedge v(1-v)^{-1}$, $\mathbf{v}_1^*,\ldots,\mathbf{v}_{\mathbf{K}_j}^*$ are the distinct values that $\{\mathbf{v}_1^{(\beta)},\ldots,\mathbf{v}_j^{(\beta)}\}$ exhibits, and $\mathbf{n}_i = \left|\left\{k \leq j : \mathbf{v}_k^{(\beta)} = \mathbf{v}_i^*\right\}\right|$.

In the context of Corollary 4.10, if the length variables of the DSBw's are marginally $\mathsf{Unif}(0,1)$ distributed, so that $\theta = 1$, we obtain

$$1 \geq \mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)}\right] = \frac{1+\beta\log(2)}{1+\beta} \geq \log(2)$$

and

$$\mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)} \,\middle|\, \mathbf{w}_1^{(\beta)},\dots,\mathbf{w}_j^{(\beta)}\right] = \frac{1}{\beta+j}\left\{\sum_{i\in\mathbf{A}_j}\mathbf{n}_i + \beta c\left(\mathbf{v}_j^{(\beta)}\right)\right\},$$

for every $\beta > 0$ and $j \geq 1$, where $\mathbf{A}_j$, $\mathbf{n}_i$ and c are as in (f).

Another quantities that simplifies nicely for DSBs are the expectations of power product of length variables, which in turn allow the derivation of an expression for clustering probabilities.

Corollary 4.11. *Consider the length variables* $\{\mathbf{v}_1,\mathbf{v}_2,\dots \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu}$*, where* $\boldsymbol{\nu} \sim \mathcal{D}_{(\beta,\mathsf{Be}(1,\theta))}$*. Then, for every sequence of non-negative integers* $(a_j,b_j)_{j=1}^k$*,*

$$\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{\{A_1,\dots,A_m\}}\frac{(\beta\theta)^m}{(\beta)_k}\prod_{j=1}^m\frac{(|A_j|-1)!\,(\sum_{i\in A_j}a_i)!}{(\theta+\sum_{i\in A_j}b_i)_{1+\sum_{i\in A_j}a_i}}, \tag{4.3}$$

where the sum ranges over the set of all partitions of $\{1,\dots,k\}$*. In particular, if* $\theta = 1$*,*

$$\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{\{A_1,\dots,A_m\}}\frac{\beta^m}{(\beta)_k}\prod_{j=1}^m\frac{(|A_j|-1)!\,(\sum_{i\in A_j}a_i)!\,(\sum_{i\in A_j}b_i)!}{(1+\sum_{i\in A_j}(a_i+b_i))!}.$$

To illustrate Corollaries 4.9 and 4.10, in Figure 23 we show some simulations of Dirichlet driven length variables and the corresponding stick-breaking weights sequences for distinct values of $\rho_\nu = 1/(1+\beta)$. In A.v and B.v we can observe that for larger values of ρ_ν (denoted by ρ in the images), the length variables tend to repeat values more often, this then leads to weights (A.w and B.w) than are more likely to be decreasingly ordered. Conversely smaller values of the tie probability make the length variables more likely to be sampled independently of previous length variables, thus the corresponding weights tend to behave similarly to the Dirichlet size-biased weights. Figure 24 also illustrates the convergence in Corollary 4.9 through the distribution the number of distinct values, $\mathbf{K}_n = |\boldsymbol{\Pi}(\mathbf{x}_{1:n})|$, that a sample from $\{\mathbf{x}_1,\dots,\mathbf{x}_n \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$ exhibits, where $\boldsymbol{\mu}$ is a DSB. As is the case of most priors for which the distribution of the size-biased permuted weights is not available, for DSBs the distribution of $\mathbf{K}_n$ is very hard to characterize analytically. Despite, whenever one can sample from the finite dimensional distributions of the weights, $(\mathbf{w}_1,\dots,\mathbf{w}_m)$, drawing samples from $\mathbf{K}_n$ is relatively simple, as explained here. First sample $(\mathbf{u}_k)_{k=1}^n \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1)$, and $\mathbf{w}_1,\dots\mathbf{w}_m$ up the first index, m, that satisfies $\sum_{j=1}^m \mathbf{w}_j > \max_k \mathbf{u}_k$. Define $\mathbf{d}_k = i$, if and only if $\sum_{j=1}^{i-1}\mathbf{w}_j \leq \mathbf{u}_k < \sum_{j=1}^{i}\mathbf{w}_j$, with the convention that the empty sum equals zero. Finally, note that the number of distinct values in $\{\mathbf{d}_1,\dots,\mathbf{d}_n\}$, is precisely a sample from $\mathbf{K}_n$. Evidently, to obtain a sample of $\mathbf{w}_1,\dots,\mathbf{w}_m$, it suffices to sample the length variables, $\mathbf{v}_1,\dots\mathbf{v}_m$. For a DSBw this can be easily done using the prediction rule of the Dirichlet driven length variables.

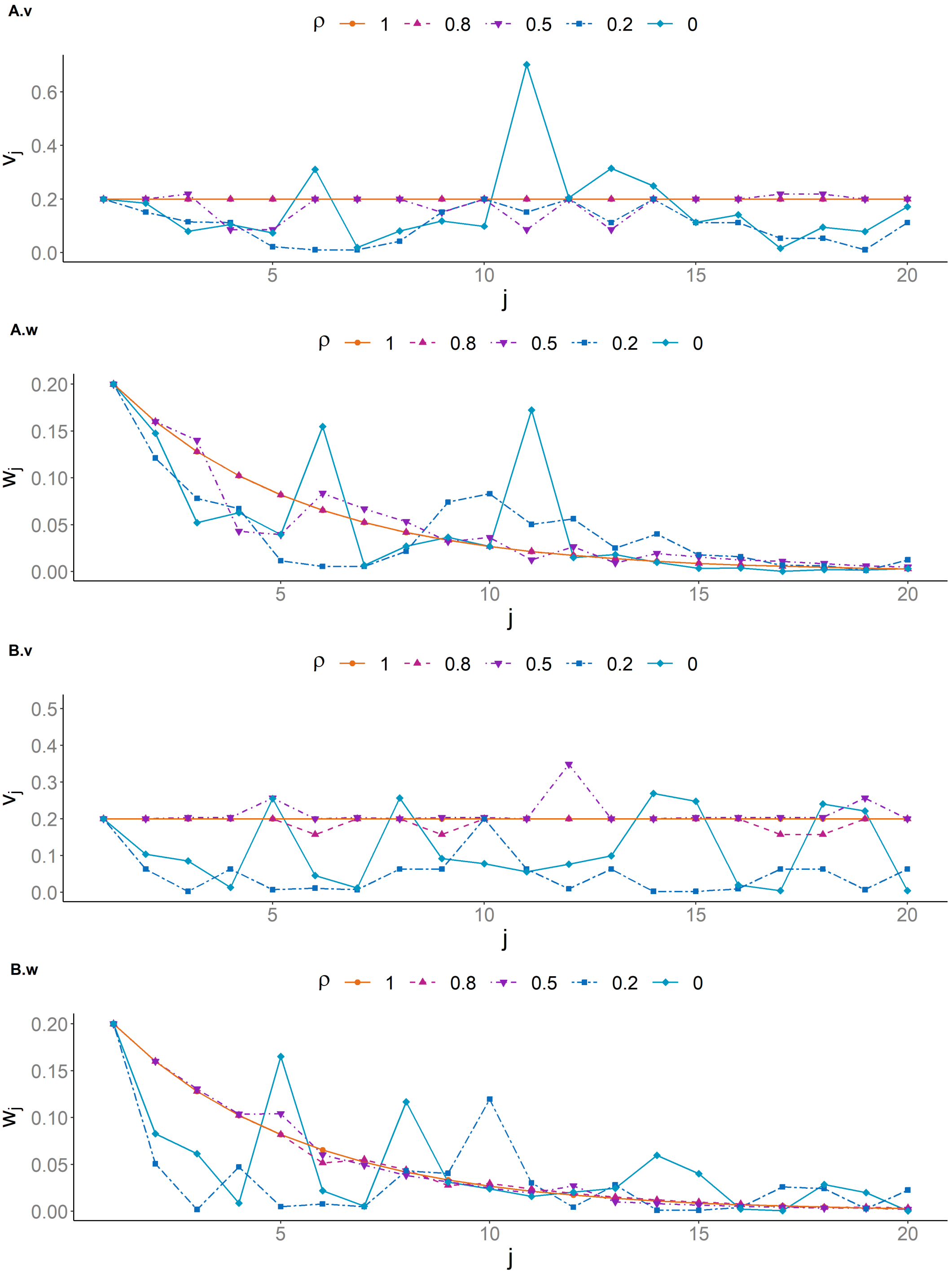


Figure 23: Simulations (A.v,B.v) of $\{\mathbf{v}_1, \ldots, \mathbf{v}_{20} \mid \boldsymbol{\nu}\} \overset{\text{iid}}{\sim} \boldsymbol{\nu} \sim \mathcal{D}_{(\beta,\nu_0)}$ with $\nu_0 = \mathsf{Be}(1,\theta)$ for distinct values of $\rho_\nu = 1/(\beta+1)$. ρ_ν, denoted by ρ, in the image was fixed to $0, 0.2, 0.5, 0.8$ and 1. A.w and B.w show the corresponding DSBw's $(\mathbf{w}_j)_{j=1}^{20} = \mathsf{SB}[(\mathbf{v}_j)_{j=1}^{20}]$.

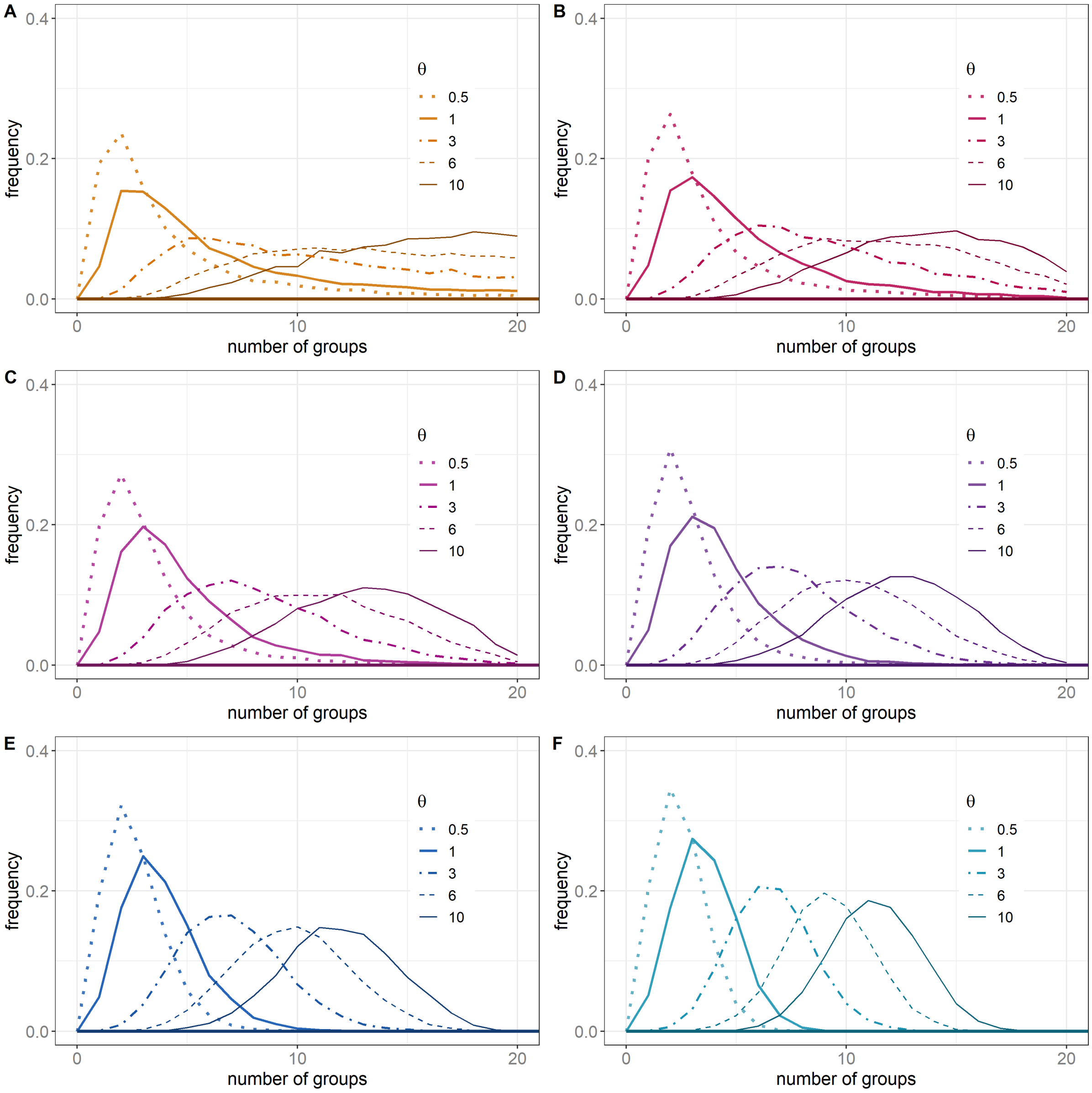


Figure 24: Frequency polygons of samples of size 10000 from the distribution of $\mathbf{K}_{20}$ corresponding to a DSB prior with $\rho_\nu = 1$ (A), $\rho_\nu = 0.8$ (B), $\rho_\nu = 0.6$ (C), $\rho_\nu = 0.4$ (D), $\rho_\nu = 0.2$ (E) and $\rho_\nu = 0$ (F). For each fixed value of ρ_ν we vary θ in the set $\{0.5, 1, 3, 6, 10\}$.

Let us denote by $\mathbf{K}_n^{(\beta,\theta)}$ to the random variable, $\mathbf{K}_n$, corresponding to a DSB with parameters (β, θ, μ_0), for simplicity when $\beta = 0$ and $\beta = \infty$ we refer to a Geometric and Dirichlet process, respectively. In Figure 24, we exhibit the distribution of $\mathbf{K}_n^{(\beta,\theta)}$ for different choices of β and θ and for $n = 20$, recall that $\beta = (1-\rho_\nu)/\rho_\nu$. Inhere, we observe a graphical representation of how $\mathbf{K}_n^{(\beta,\theta)} \xrightarrow{d} \mathbf{K}_n^{(0,\theta)}$ as $\beta \to 0$ and $\mathbf{K}_n^{(\beta,\theta)} \xrightarrow{d} \mathbf{K}_n^{(\infty,\theta)}$ as $\beta \to \infty$. In the same figure we see that an increment on β contributes to the distribution of $\mathbf{K}_n^{(\beta,\theta)}$ with a smaller mean and variance. Conversely, decreasing the value of β, impacts the prior distribution of $\mathbf{K}_n^{(\beta,\theta)}$ with a larger mean and variance, and a heavier right tail, say less informative. Consistently with the observations for other models (see Figures 19, 20 and 21), for DSBs we appreciate that for a fixed value of

$\rho_\nu = 1/(1+\beta)$, an increment on θ makes the distribution of $\mathbf{K}_n^{(\beta,\theta)}$ favour larger values. This due to the fact that marginally $\mathbf{v}_i \sim \mathsf{Be}(1,\theta)$, so bigger values θ suggest smaller length variables, which translate to small values even for the largest weights, this in turn means that samples from the corresponding DSBs are diverse. That is to say, if θ is large, samples from $\{\mathbf{x}_1, \mathbf{x}_2, \dots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu} \sim \mathcal{DSB}_{(\beta,\theta,\mu_0)}$ take a wide variety of distinct values.

4.1.3 Models beyond Dirichlet driven stick-breaking processes

The generality of Section 4.1.1 allows us to construct a great variety of new Bayesian nonparametric priors. In fact, for every known species sampling process, $\boldsymbol{\nu}$, with available EPPF, π_ν and tie probability, ρ_ν, the analysis in Section 4.1.1 can be carried out, thus leading to a complete analysis of a new model. This is the case of Gibbs-type priors (e.g. De Blasi et al.; 2015), and some normalized random measures with independent increments (James et al.; 2009).

To provide another concrete example, let us consider the case where $\boldsymbol{\nu}$ is a Pitman-Yor process (see Section 3.6.2) with parameters $\alpha \in [0,1)$ and $\beta > -\sigma$. Recall that for this species sampling process its tie probability is

$$\rho_\nu = \pi_\nu(2) = \frac{1-\alpha}{\beta+1}. \tag{4.4}$$

Hence, for stick-breaking weights $(\mathbf{w}_j)_{j\geq1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq1}]$, where $\{\mathbf{v}_1, \mathbf{v}_2, \dots \mid \boldsymbol{\nu}\} \overset{\text{iid}}{\sim} \boldsymbol{\nu} \sim \mathcal{PY}_{(\alpha,\beta,\nu_0)}$ by substituting (4.4) into Theorem 4.6 we obtain,

$$\mathbb{P}[\mathbf{w}_j \geq \mathbf{w}_{j+1}] = \frac{1-\alpha+(\beta+\alpha)\mathbb{E}\left[\overrightarrow{\nu_0}(c(\mathbf{v}))\right]}{\beta+1},$$

for every $j \geq 1$, where $\overrightarrow{\nu_0}$ and c are as in the same theorem. For the special case $\nu_0 = \mathsf{Be}(1,\theta)$, the probability in question simplifies to

$$\mathbb{P}[\mathbf{w}_j \geq \mathbf{w}_{j+1}] = 1 - \frac{{}_2F_1(1,1;\theta+2,1/2)(\beta+\alpha)\theta}{2(\beta+1)(\theta+1)},$$

and if $\theta = 1$ we even get

$$\mathbb{P}[\mathbf{w}_j \geq \mathbf{w}_{j+1}] = \frac{1-\alpha+(\beta+\alpha)\log(2)}{\beta+1}.$$

Note that, by (4.4), as $\beta \to \infty$ or $\alpha \to 1$ and $\beta \to \beta' \in (-1,\infty)$ we get $\rho_\nu \to 0$. Alternatively, as $\alpha \to \alpha' \in [0,1)$ and $\beta \to -\alpha'$, the tie probability $\rho_\nu \to 1$. Therefore, Corollary 4.5 assures that by means of Pitman-Yor driven ESBs we can approximate (weakly in distribution) Dirichlet and Geometric process. This last assertion is not true for all sub-classes of ESBs with underlying species sampling process. In fact, if $\boldsymbol{\nu}$ represents a normalized inverse-Gaussian random measure, with total mass parameter $\beta > 0$, as proved by Lijoi et al. (2005), its tie probability is

$$\rho_\nu = \frac{1}{2}[1+\beta^2 e^\beta E_1(\beta) - \beta],$$

where $E_1(\beta) = \int_\beta^\infty x^{-1}e^{-x}dx$ is the exponential integral. Using the inequality

$$\frac{e^{-\beta}}{2}\log\left(1+\frac{2}{\beta}\right) < E_1(\beta) < e^{-\beta}\log\left(1+\frac{1}{\beta}\right),$$

it can be shown as $\beta \to \infty$, $\rho_\nu \to 0$, and as $\beta \to 0$, $\rho_\nu \to c \le 1/2$. Thus Geometric processes can not be recovered as weak limits of ESBs with normalized inverse-Gaussian processes as the directing random measure of the length variables.

Back to Pitman-Yor driven ESBs, using the prediction rule of Pitman-Yor processes and Theorem 4.7, for $(\mathbf{w}_j)_{j\ge1} = \mathsf{SB}[(\mathbf{v}_i)_{i\ge1}]$, where $\{\mathbf{v}_1,\mathbf{v}_2,\dots \mid \boldsymbol{\nu}\} \overset{\text{iid}}{\sim} \boldsymbol{\nu} \sim \mathcal{PY}_{(\alpha,\beta,\nu_0)}$, we can compute the conditional probability

$$\begin{aligned}\mathbb{P}[\mathbf{w}_{n+1} \le \mathbf{w}_n \mid \mathbf{w}_1,\dots,\mathbf{w}_n] &= \sum_{j=1}^{\mathbf{K}_n}\frac{\mathbf{n}_j-\alpha}{\theta+n}\mathbf{1}_{\{\mathbf{v}_j^*\le c(\mathbf{v}_n)\}} + \frac{\theta+\alpha\mathbf{K}_n}{\theta+n}\overrightarrow{\nu_0}(c(\mathbf{v}_n))\\ &= \frac{1}{\theta+n}\left\{\sum_{j\in\mathbf{A}_n}\mathbf{n}_j - |\mathbf{A}_n|\alpha + (\theta+\alpha\mathbf{K}_n)\,\overrightarrow{\nu_0}(c(\mathbf{v}_n))\right\},\end{aligned}$$

where c, $\overrightarrow{\nu_0}$, $\mathbf{K}_n$, $\mathbf{v}_1^*,\dots,\mathbf{v}_{\mathbf{K}_n}^*$ and $\mathbf{n}_1,\dots,\mathbf{n}_{\mathbf{K}_n}$ are as is the same theorem, and $\mathbf{A}_n = \{j \le \mathbf{K}_n : \mathbf{v}_j^* \le c(\mathbf{v}_n)\}$. For these length variables, recalling the EPPF of $\boldsymbol{\nu}$,

$$\pi_\nu(n_1,\dots,n_k) = \frac{(\beta+\alpha)_{k-1\uparrow\alpha}\prod_{j=1}^k(1-\alpha)_{n_j-1}}{(\beta+1)_{n-1}},$$

and using (4.2), we can also compute for any non-negative integers $(a_j,b_j)_{j=1}^k$,

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{\{A_1,\dots,A_m\}} &\frac{(\beta+\alpha)_{m-1\uparrow\alpha}\prod_{j=1}^m(1-\alpha)_{|A_j|-1}}{(\beta+1)_{k-1}}\times\\ &\times\prod_{j=1}^m\left\{\int_{[0,1]}(v)^{\sum_{i\in A_j}a_i}(1-v)^{\sum_{i\in A_j}b_i}\,\nu_0(dv)\right\},\end{aligned}$$

where the sum ranges over the set of all partitions of $\{1,\dots,k\}$. In particular, if $\nu_0 = \mathsf{Be}(1,\theta)$, the above even reduces to

$$\mathbb{E}\left[\prod_{j=1}^k \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{\{A_1,\dots,A_m\}}\frac{\theta^m(\beta+\alpha)_{m-1\uparrow\alpha}}{(\beta+1)_{k-1}}\prod_{j=1}^m\frac{(1-\alpha)_{|A_j|-1}\left(\sum_{i\in A_j}a_i\right)!}{\left(\theta+\sum_{i\in A_j}b_i\right)_{1+\sum_{i\in A_j}a_i}}.$$

There are also interesting choices of $\boldsymbol{\nu}$, outside Bayesian non-parametric priors. For example, one might consider the species sampling process with finitely many atoms

$$\boldsymbol{\nu} = \boldsymbol{\alpha}\sum_{j=1}^{\kappa}\mathbf{p}_j\delta_{\mathbf{v}_j^*} + (1-\boldsymbol{\alpha})\,\nu_0,$$

for some $\kappa \in \mathbb{N}$ and where $\mathbf{p}_j \ge 0$, $\sum_{j=1}^\kappa \mathbf{p}_j = 1$ and $\boldsymbol{\alpha}$ is an independent random variable taking values in $[0,1]$. The advantage of such a measure is that, depending on the distribution of $(\mathbf{p}_j)_{j=1}^\kappa$ and $\boldsymbol{\alpha}$, the EPPF, π_ν, could be relatively simple to derive, as well as the tie probability which can be computed through $\rho_\nu = \mathbb{E}\left[\boldsymbol{\alpha}^2\right]\sum_{j=1}^{\kappa}\mathbb{E}\left[\mathbf{p}_j^2\right]$. Noting that at no point did we required the underlying species sampling process, $\boldsymbol{\nu}$, to be proper, the choice $\boldsymbol{\alpha} < 1$ almost surely, would imply that $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{\text{iid}}{\sim} \boldsymbol{\nu}$ exhibits exactly κ values that repeat infinitely often and that there are infinitely many indexes i that contribute to $(\boldsymbol{\Pi}(\mathbf{v}_{1:n}))_{n\geq 1}$ as a singleton. For example if

$$\boldsymbol{\nu} = \boldsymbol{\alpha}\delta_{\mathbf{v}^*} + (1-\boldsymbol{\alpha})\,\mathsf{Be}(1,\theta),$$

where $\mathbf{v}^* \sim \mathsf{Be}(1,\theta)$, there would be exactly one value, $\mathbf{v}^*$ in $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{\text{iid}}{\sim} \boldsymbol{\nu}$ that repeats infinitely often and infinitely many $\mathbf{v}_i$'s that are sampled independently from a $\mathsf{Be}(1,\theta)$ distribution. In other words, with probability $\boldsymbol{\alpha}$, $\mathbf{v}_i$ is a length variable of a Geometric process, and with probability $1-\boldsymbol{\alpha}$ it is sampled as if it were a length variable of a Dirichlet process. Hence, even this extremely simple species sampling process, $\boldsymbol{\nu}$, can lead to interesting ESBs that are some kind of hybrids between Dirichlet and Geometric process. The general analysis in Section 4.1.1, also covers this simple scenario.

4.2 Stick-breaking processes with Markovian length variables

As mentioned at the begining of Section 4, Dirichlet length variables, $(\mathbf{v}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \mathsf{Be}(1,\theta)$, and Geometric length variables $(\mathbf{v}, \mathbf{v}, \ldots)$ are not only exchangeable, but are also trivial Markov processes. The objective here is to replicate the analysis we did for ESBs but this time considering stick-breaking processes with Markovian length variables. For the exchangeable case, in Section 2.1 we had already analysed sequence with such symmetry, we have not yet discussed Markov processes whatsoever, so we start with a quick review of Markov chains.

4.2.1 Preliminaries of discrete-time Markov processes

Let $(S, \mathscr{B}_S)$ be a Borel space and $(\mathbf{v}_i)_{i\geq 1}$ a random sequence taking values in S. We say that $(\mathbf{v}_i)_{i\geq 1}$ is a Markov chain whenever

$$\mathbb{P}[\mathbf{v}_{i+1} \in B \mid \mathbf{v}_1, \ldots, \mathbf{v}_i] = \mathbb{P}[\mathbf{v}_{i+1} \in B \mid \mathbf{v}_i],$$

(almost surely) for every $B \in \mathscr{B}_S$ and $i \geq 1$, equivalently

$$\mathbb{E}[f(\mathbf{v}_{i+1}) \mid \mathcal{F}_i] = \mathbb{E}[f(\mathbf{v}_{i+1}) \mid \mathbf{v}_i],$$

for every measurable $f : S \to \mathbb{R}_+$ and where $\mathcal{F}_i$ denotes the σ-algebra generated by $\mathbf{v}_1, \ldots, \mathbf{v}_i$. In other words, $\mathbf{v}_{i+1}$ is conditionally independent of $(\mathbf{v}_1, \ldots, \mathbf{v}_{i-1})$, given $\mathbf{v}_i$. Another characterization of Markov chains is that we can write $\mathbf{v}_{i+1} = g_i(\mathbf{v}_i, \mathbf{u}_i)$ for every $i \geq 1$, where $g_i : S \times [0,1] \to S$ is a measurable function and $(\mathbf{u}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \mathsf{Unif}(0,1)$ is independent of $(\mathbf{v}_i)_{i\geq 1}$. In particular if $g_i = g_1$ for every $i \geq 1$ so that

$$\mathbb{P}[\mathbf{v}_{i+1} \in B \mid \mathbf{v}_i] = \mathbb{P}[\mathbf{v}_2 \in B \mid \mathbf{v}_1],$$

we say that $(\mathbf{v}_i)_{i\geq 1}$ is an homogeneous Markov chain. In general, to the distribution of $\mathbf{v}_1$, ν_0, we call initial distribution of the chain $(\mathbf{v}_i)_{i\geq 1}$, and to the probability kernels from S into S, $\boldsymbol{\nu}_i : S \to S$, that satisfy

$$\boldsymbol{\nu}_i(\mathbf{v}_i; B) = \mathbb{P}[\mathbf{v}_{i+1} \in B \mid \mathbf{v}_i],$$

we call one-step transition probability kernels. If $(\mathbf{v}_i)_{i\geq 1}$ is homogeneous $\boldsymbol{\nu}_i = \boldsymbol{\nu}_1$ for every $i \geq 1$, for simplicity, in this instance we dismiss the index and write $\boldsymbol{\nu} = \boldsymbol{\nu}_1 = \boldsymbol{\nu}_i$. Whether $(\mathbf{v}_i)_{i\geq 1}$ is homogeneous or not, we can compute the joint distribution of $(\mathbf{v}_1, \ldots, \mathbf{v}_n)$, through

$$\begin{aligned}\mathbb{P}\left[\mathbf{v}_1 \in B_1, \ldots, \mathbf{v}_n \in B_n\right] &= (\nu_0 \circ \boldsymbol{\nu}_1 \circ \cdots \circ \boldsymbol{\nu}_{n-1})\left(\prod_{i=1}^n \mathbf{1}_{B_i}\right)\\ &= \int \ldots \int \left\{\prod_{i=1}^n \mathbf{1}_{B_i}(v_i)\right\} \boldsymbol{\nu}_{n-1}(v_{n-1}; dv_n)\cdots \boldsymbol{\nu}_1(v_1; dv_2)\nu_0(dv_1),\end{aligned} \tag{4.5}$$

for every $n \geq 1$ and $B_1, \ldots, B_n \in \mathscr{B}_S$. Similarly, we can obtain the marginal distribution

$$\begin{aligned}\mathbb{P}\left[\mathbf{v}_n \in B\right] &= (\nu_0 \boldsymbol{\nu}_1 \cdots \boldsymbol{\nu}_{n-1})(\mathbf{1}_B)\\ &= \int \ldots \int \mathbf{1}_B(v_n)\boldsymbol{\nu}_{n-1}(v_{n-1}; dv_n)\cdots \boldsymbol{\nu}_1(v_1; dv_2)\nu_0(dv_1),\end{aligned} \tag{4.6}$$

(see Definition 1.2, for the composition and product of kernels).

Now, for an homogenous Markov chain, $(\mathbf{v}_i)_{i\geq 1}$ with one-step transition probability kernel, $\boldsymbol{\nu}$, we say the probability measure, λ, is an stationary distribution whenever $\lambda\boldsymbol{\nu} = \lambda$, this is

$$\lambda(f) = \int f(s)\lambda(ds) = \int\int f(v)\boldsymbol{\nu}(s; dv)\lambda(ds) = (\lambda\boldsymbol{\nu})(f),$$

for every measurable function $f : S \to \mathbb{R}_+$. In this instance we also call λ invariant for $\boldsymbol{\nu}$. Note that if λ is an stationary distribution for $(\mathbf{v}_i)_{i\geq 1}$ and $\mathbf{v}_i \sim \lambda$, then $\mathbf{v}_{i+1} \sim \lambda$ marginally. Further, if the initial distribution, ν_0, is an stationary distribution for the chain, then $\mathbf{v}_n \sim \nu_0$, for every $n \geq 1$, in this case we also call $(\mathbf{v}_i)_{i\geq 1}$ an stationary Markov chain. Let us consider the following trivial examples.

Example 4.2. *Consider the probability kernel $\boldsymbol{\nu} : S \to S$ given by $\boldsymbol{\nu}(s; \cdot) = \delta_s$, for each $s \in S$. Then any probability measure λ over $(S, \mathscr{B}_S)$ is invariant for $\boldsymbol{\nu}$, as*

$$(\lambda\boldsymbol{\nu})(f) = \int\int f(v)\delta_s(dv)\lambda(ds) = \int f(s)\lambda(ds) = \lambda(f),$$

for every measurable function $f : S \to \mathbb{R}_+$. This means that an homogeneous Markov chain $(\mathbf{v}_i)_{i\geq 1}$ with one-step transition, $\boldsymbol{\nu}$, has infinitely many stationary distributions. Indeed, if $\mathbf{v}_i \sim \lambda$ and $\mathbf{v}_{i+1} = \mathbf{v}_i$ almost surely, then $\mathbf{v}_{i+1} \sim \lambda$, for every $i \geq 1$ and any probability measure λ over $(S, \mathscr{B}_S)$. Moreover, any choice of the initial distribution ν_0, yields an stationary Markov chain.

Example 4.3. *Consider the constant probability kernel $\boldsymbol{\nu} : S \to S$ given by $\boldsymbol{\nu}(s; \cdot) = \nu$ for each $s \in S$. Then, for any probability measure λ over $(S, \mathscr{B}_S)$ we get*

$$(\lambda\boldsymbol{\nu})(f) = \int\int f(v)\nu(dv)\lambda(ds) = \int f(v)\nu(dv) = \nu(f),$$

for every integrable function $f : S \to \mathbb{R}_+$. Thus, for $\boldsymbol{\nu}$ there exist exactly one invariant probability measure, and it is itself. In other words, the homogeneous Markov chain $(\mathbf{v}_i)_{i\geq 1}$ with one-step transition, $\boldsymbol{\nu} = \nu$, is stationary if and only if its initial distribution $\nu_0 = \nu$.

In general for a fixed one-step transition probability kernel, $\boldsymbol{\nu}$, it is possible that there exist no invariant probability measures, that there exists exactly one invariant probability measure or that there are infinitely many invariant probability measures. It is easy to see that if the probability measures λ_1 and λ_2 are both invariant for $\boldsymbol{\nu}$, then for every $t \in [0,1]$, the mixture, $t\lambda_1 + (1-t)\lambda_2$, is also invariant for $\boldsymbol{\nu}$, hence the set of invariant probability measures with respect to $\boldsymbol{\nu}$, denoted by $\mathcal{I}[\boldsymbol{\nu}]$, is convex. We say that a probability measure $\lambda \in \mathcal{I}[\boldsymbol{\nu}]$ is $\boldsymbol{\nu}$-ergodic if it is extremal in $\mathcal{I}[\nu]$, that is, it can not be decomposed as a mixture of other probability measures λ_i's belonging to $\mathcal{I}[\boldsymbol{\nu}]$. It can be shown that two invariant $\boldsymbol{\nu}$-ergodic probability measures, λ and ν are either identical or mutually singular (meaning that the exist disjoint $A, B \in \mathscr{B}_S$ such that $A \cup B = S$, $\lambda(A) = 0$, and $\nu(B) = 0$). To conclude this preliminary section we state the ergodic theorem for stationary Markov chains.

Theorem 4.12 (Birkhoff; Ergodic theorem for stationary Markov chains)**.** *Consider an stationary Markov chain, $(\mathbf{v}_i)_{i\geq 1}$, with initial and stationary distribution, ν_0, and one-step transition probability kernel, $\boldsymbol{\nu}$. If ν_0 is $\boldsymbol{\nu}$-ergodic, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i\leq n} f(\mathbf{v}_i) = \int f(v)\nu_0(dv) = \mathbb{E}[f(\mathbf{v}_1)], \tag{4.7}$$

almost surely, for every measurable function $f : S \to \mathbb{R}$. In particular, if $\mathcal{I}[\boldsymbol{\nu}] = \{\nu_0\}$, 4.7 holds.

A detailed review of Markov processes can be found in the work by Feller (1968); Karlin and Taylor (1975); Kallenberg (2002).

4.2.2 Stick-breaking processes featuring stationary Markovian length variables

Just as in the instance where the length variables are exchangeable, here we begin by analysing general properties of stick-breaking processes featuring stationary and Markovian length variables.

Definition 4.3. *We call a Markov stick-breaking process (MSB) to any SSP as in (3.1) whose weights $(\mathbf{w}_j)_{j\geq 1}$ decompose as $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$, where $(\mathbf{v}_i)_{i\geq 1}$ is a Markov process that takes values in $[0,1]$. To the corresponding weights, we call an Markov stick-breaking weights sequence (MSBw).*

In this section we focus in MSBs with stationary length variables $(\mathbf{v}_i)_{i\geq 1}$. Further, we will assume that the initial and stationary distribution, ν_0, is $\boldsymbol{\nu}$-ergodic, for the one-step transition probability kernel, $\boldsymbol{\nu}$, of the chain. Recall that if the invariant measure of $\boldsymbol{\nu}$ is unique then it is also $\boldsymbol{\nu}$-ergodic. This said, we begin by giving conditions on $\boldsymbol{\nu}$ and ν_0 under which the corresponding MSB is proper and has full support.

Theorem 4.13. *Let $\boldsymbol{\mu}$ be an MSB with weights $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$, for some stationary Markov chain $(\mathbf{v}_i)_{i\geq 1}$ with one-step transition $\boldsymbol{\nu}$, and initial and stationary distribution, ν_0.*

i) *If for every $\epsilon > 0$ there exist $0 < \delta < \epsilon$ such that for every $n \geq 1$*

$$(\nu_0 \circ \boldsymbol{\nu} \circ \cdots \circ \boldsymbol{\nu})\left((\delta,\epsilon)^n\right) = \int_\delta^\epsilon \ldots \int_\delta^\epsilon \boldsymbol{\nu}(v_{n-1}; dv_n) \cdots \boldsymbol{\nu}(v_1; dv_2)\nu_0(dv_1) > 0,$$

then $\boldsymbol{\mu}$ has full support. In particular, if there exist $\varepsilon > 0$ such that $(0,\varepsilon)$ is contained in the support of ν_0, and for every $v \in (0,\varepsilon)$, $(0,\varepsilon)$ is also contained in the support of $\boldsymbol{\nu}(v;\cdot)$, then $\boldsymbol{\mu}$ has full support.

ii) *If ν_0 is $\boldsymbol{\nu}$-ergodic, then $\boldsymbol{\mu}$ is proper if and only if $\nu_0 \neq \delta_0$. In particular if ν_0 is the only invariant measure for $\boldsymbol{\nu}$, then $\boldsymbol{\mu}$ is proper if and only if $\nu_0 \neq \delta_0$.*

The proof of Theorem 4.13 can be found in Appendix D.5. The second part of this result shows that most stationary Markov chains $(\mathbf{v}_i)_{i\geq 1}$ will lead to a proper MSB, $\boldsymbol{\mu}$. Indeed, if the $\boldsymbol{\nu}$-ergodic measure $\nu_0 = \delta_0$, necessarily $(\mathbf{v}_i)_{i\geq 1} = (0,0,\dots)$ almost surely, in which, non-interesting, case $\boldsymbol{\mu}$ is even diffuse almost surely. Otherwise, if $\nu_0 \neq \delta_0$, which covers most cases of interest, $\boldsymbol{\mu}$ will be purely atomic almost surely. In particular if the initial distribution $\nu_0 = \mathsf{Be}(a,b)$ is $\boldsymbol{\nu}$-ergodic, then $\boldsymbol{\mu}$ has full support.

As can be done by means of ESBs, Dirichlet and Geometric process can be recovered in the weak limits of MSBs, as our following result explains (see Appendix D.6 for a proof).

Theorem 4.14. *Let $(S,\mathscr{B}_S)$ be a Polish space. For each $n \geq 1$ consider a diffuse probability measure, $\mu_0^{(n)}$, over $(S,\mathscr{B}_S)$, a probability kernel, $\boldsymbol{\nu}^{(n)} : [0,1] \to [0,1]$, from $[0,1]$ into itself, and $\boldsymbol{\nu}^{(n)}$-ergodic and invariant probability measure $\nu_0^{(n)} \neq \delta_0$. Let $\boldsymbol{\mu}^{(n)}$ be an MSB with base measure $\mu_0^{(n)}$ and length variables, $\mathbf{V}^{(n)} = \left(\mathbf{v}_i^{(n)}\right)_{i\geq 1}$, that form an stationary Markov chain with one-step transition probability kernel, $\boldsymbol{\nu}^{(n)}$, and initial distribution, $\nu_0^{(n)}$. Also set $\mathbf{W}^{(n)} = \mathsf{SB}\left[\mathbf{V}^{(n)}\right]$. Say that as $n \to \infty$, $\mu_0^{(n)}$ converges weakly to the diffuse probability measure μ_0, and $\nu_0^{(n)}$ converges weakly to some probability measure ν_0 with $\nu_0(\{0\}) = 0$.*

i) *If for every $v_n \to v$ in $[0,1]$, $\boldsymbol{\nu}^{(n)}(v_n;\cdot)$ converges weakly to $\nu_0(v;\cdot) = \nu_0$, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to a stick-breaking process $\boldsymbol{\mu}$ with base measure μ_0 and featuring independent length variables $(\mathbf{v}_i)_{i\geq 1} \overset{iid}{\sim} \nu_0$, as $n \to \infty$. In particular if $\nu_0 = \mathsf{Be}(1,\theta)$, the limit $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta,\mu_0)}$, and $\mathbf{W}^{(n)}$ converges in distribution to the size-biased permuted weights of $\boldsymbol{\mu}$.*

ii) *If for every $v_n \to v$ in $[0,1]$, $\boldsymbol{\nu}^{(n)}(v_n;\cdot)$ converges weakly to δ_v, then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0,\mu_0)}$, as $n \to \infty$, and $\mathbf{W}^{(n)}$ converges in distribution to the decreasingly ordered weights of $\boldsymbol{\mu}$.*

In the following couple of Sections we will be providing examples using parametrized one-step transition probability kernels. In particular we will specialize Theorem 4.14 to make it more clearer. For the general case, some remaining important quantities can be expressed in terms of the transition, $\boldsymbol{\nu}$, and the stationary distribution ν_0, of the length variables $(\mathbf{v}_i)_{i\geq 1}$. For example for the MSBw $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$, we have that the probability that consecutive weights are decreasingly ordered is

$$\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j] = \mathbb{E}[\mathbb{P}[\mathbf{v}_{j+1} \leq c(\mathbf{v}_j) \mid \mathbf{v}_j]] = \mathbb{E}\left[\boldsymbol{\nu}\left(\mathbf{v}_j; [0, c(\mathbf{v}_j)]\right)\right],$$

where $c(v) = 1 \wedge v(1-v)^{-1}$. Recalling that $\mathbf{v}_j \sim \nu_0$, because ν_0 is a stationary, this yields

$$\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j] = \int \boldsymbol{\nu}(v;[0,c(v)])\nu_0(dv) = \int_{[0,1]}\int_{[0,c(v)]} \boldsymbol{\nu}(v;dx)\nu_0(dv). \tag{4.8}$$

Another important quantity regarding the ordering of the weights is the conditional probability $\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j \mid \mathbf{w}_1, \dots, \mathbf{w}_j]$, as explained before Theorem 4.7, for general stick-breaking weights sequences, such that $0 < \mathbf{v}_i < 1$ almost surely, we have that

$$\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j \mid \mathbf{w}_1, \dots, \mathbf{w}_j] = \mathbb{P}[\mathbf{v}_{j+1} \leq c(\mathbf{v}_j) \mid \mathbf{v}_1, \dots, \mathbf{v}_j].$$

In contrast to the case where the length variables are exchangeable (see for instance Theorem 4.7), for Markovian length variables, this conditional probability only depends on the last length variable. Indeed, the Markov property of the length variables of MSBw's implies

$$\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j \mid \mathbf{w}_1, \dots, \mathbf{w}_j] = \mathbb{P}[\mathbf{v}_{j+1} \leq c(\mathbf{v}_j) \mid \mathbf{v}_j] = \boldsymbol{\nu}\left(\mathbf{v}_j; [0, c(\mathbf{v}_j)]\right), \tag{4.9}$$

where c is as above. Note that (4.9), is conditionally independent of $\mathbf{v}_1, \dots, \mathbf{v}_{j-1}$ given $\mathbf{v}_j$, opposed to the exchangeable counterpart where $\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j \mid \mathbf{w}_1, \dots, \mathbf{w}_j]$ depends measurably on each length variable $(\mathbf{v}_1, \dots, \mathbf{v}_j)$.

Before we move, note that the expectations of power products of Markovian length variables, which are important to characterize clustering related probabilities, can also we described in terms of the transition, $\boldsymbol{\nu}$, and the stationary distribution ν_0. From (4.5) it is easy to see

$$\mathbb{E}\left[\prod_{j=1}^{k} \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \int \dots \int \prod_{j=1}^{k} v_j^{a_j}(1-v_j)^{b_j}\, \boldsymbol{\nu}(v_{k-1}; dv_k) \cdots \boldsymbol{\nu}_1(v_1; dv_2)\nu_0(dv_1) \tag{4.10}$$

for any non-negative integers $(a_j, b_j)_{j=1}^{k}$.

4.2.3 Beta-Binomial stick-breaking processes

Beta-Binomial stick-breaking processes were introduced by Gil-Leyva et al. (2020). These processes are special because they were the first class of stick-breaking processes with dependent length variables to be introduced, that recover Dirichlet and Geometric processes as weak limits. We begin this section by defining the Beta-Binomial transition and study some properties of this probability kernel.

Definition 4.4 (Beta-Binomial transition)**.** *Consider $\alpha, \theta > 0$ and $\kappa \in \{0, 1, \dots\}$. To the probability kernel $\boldsymbol{\nu} : [0,1] \to [0,1]$ given by*

$$\boldsymbol{\nu}(p; dv) = \sum_{z=0}^{\kappa} \mathsf{Be}(dv \mid \alpha + z, \theta + \kappa - z)\mathsf{Bin}(z \mid \kappa, p) \tag{4.11}$$

we call a Beta-Binomial transition, where $\mathsf{Bin}(x \mid \kappa, v)$ denotes the mass probability function at x of a Binomial distribution, $\mathsf{Bin}(\kappa, p)$, and $\mathsf{Be}(v \mid \alpha, \theta)$ denotes the density function at v of a $\mathsf{Be}(\alpha, \theta)$ distribution. Here we use the convention that $\mathsf{Bin}(0, p) = \delta_0$ for every $p \in [0, 1]$.

The first thing to note about the Beta-Binomial transition is that it is a mixture of Beta distributions. In effect, if $\{\mathbf{v}_{j+1} \mid \mathbf{v}_j\} \sim \boldsymbol{\nu}(\mathbf{v}_j; \cdot)$, where $\boldsymbol{\nu}$ is as in (4.11), this simply means that given $\mathbf{v}_j$, $\mathbf{v}_{j+1}$ is sampled from a $\mathsf{Be}(\alpha + z, \theta + \kappa - z)$ distribution with probability $\mathsf{Bin}(z \mid \kappa, \mathbf{v}_j)$, for every $z \in \{0, \dots, \kappa\}$. Thus, if we were to sample $\mathbf{v}_{j+1}$ from

$\mathbb{P}[\mathbf{v}_{j+1} \in \cdot \mid \mathbf{v}_j]$, we can first sample a latent random variable $\{\mathbf{z}_j \mid \mathbf{v}_j\} \sim \mathsf{Bin}(\kappa, \mathbf{v}_j)$ and then sample $\{\mathbf{v}_{j+1} \mid \mathbf{z}_j\} \sim \mathsf{Be}(\alpha+\mathbf{z}_j, \theta+\kappa-\mathbf{z}_j)$. Recall from the Beta-Binomial conjugate model that if $\mathbf{v} \sim \mathsf{Be}(\alpha, \theta)$ and $\{\mathbf{z} \mid \mathbf{v}\} \sim \mathsf{Bin}(\kappa, \mathbf{v})$ then the conditional density of $\mathbf{v}$ given $\mathbf{z}$ is proportional to $(\mathbf{v})^{\alpha+\mathbf{z}}(1-\mathbf{v})^{\theta+\kappa-\mathbf{z}}$, hence $\{\mathbf{v} \mid \mathbf{z}\} \sim \mathsf{Be}(\alpha+\mathbf{z}, \theta+\kappa-\mathbf{z})$. This said, it is clear that if $\mathbf{v}_j \sim \mathsf{Be}(\alpha, \theta)$ marginally, then we also have $\mathbf{v}_{j+1} \sim \mathsf{Be}(\alpha, \theta)$ marginally, meaning that $\nu_0 = \mathsf{Be}(\alpha, \theta)$ is an invariant distribution for $\boldsymbol{\nu}$ as in (4.11). Further, since the support of ν_0 and $\boldsymbol{\nu}(p; \cdot)$ is $[0,1]$ for every $p \in [0,1]$ we even get ν_0 is the only invariant distribution for $\boldsymbol{\nu}$, therefore it is $\boldsymbol{\nu}$-ergodic. This proves the following Lemma.

Lemma 4.15. *Consider a Beta-Binomial transition, $\boldsymbol{\nu}$, as in (4.11). Then,*

i) *$\nu_0 = \mathsf{Be}(\alpha, \theta)$ is the only invariant distribution for $\boldsymbol{\nu}$, in particular we get ν_0 is $\boldsymbol{\nu}$-ergodic.*

ii) *For every $\varepsilon > 0$ and $p \in [0,1]$, $(0, \varepsilon)$ is contained in the support of ν_0 and $\boldsymbol{\nu}(p; \cdot)$.*

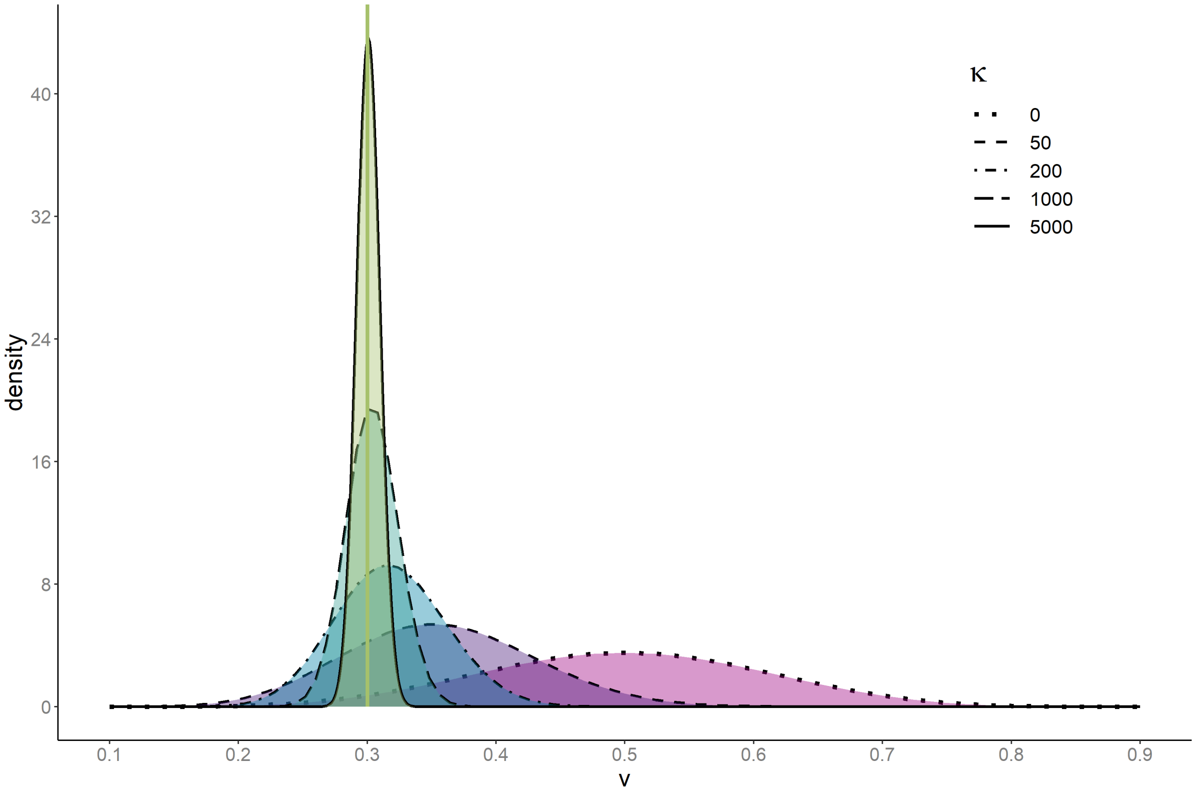


Figure 25: Density function of the Beta-Binomial transition $\boldsymbol{\nu}^{(\kappa)}(p; \cdot) = \sum_{z=0}^{\kappa} \mathsf{Be}(dv \mid \alpha+z, \theta+\kappa-z)\mathsf{Bin}(z \mid \kappa, p)$ for $\kappa \in \{0, 50, 200, 1000, 5000\}$. In all cases we fixed $p = 0.3$ and $\alpha = \theta = 10$.

Our next result concerning the Beta-Binomial transition, same that is proved in Appendix D.7, is one of the main motivations behind this model.

Lemma 4.16. *For each $\kappa \in \{0, 1, \ldots\}$ let $\alpha^{(\kappa)}, \theta^{(\kappa)} > 0$ and consider the Beta-Binomial transition, $\boldsymbol{\nu}^{(\kappa)}$, given by $\boldsymbol{\nu}^{(\kappa)}(p; dv) = \sum_{z=0}^{\kappa} \mathsf{Be}\left(dv \mid \alpha^{(\kappa)}+z, \theta^{(\kappa)}+\kappa-z\right) \mathsf{Bin}(z \mid \kappa, p)$, for every $p \in [0,1]$. Also consider the invariant distribution $\nu_0^{(\kappa)} = \mathsf{Be}\left(\alpha^{(\kappa)}, \theta^{(\kappa)}\right)$ of $\boldsymbol{\nu}^{(\kappa)}$.*

i) *For the choice $\kappa = 0$ we get $\boldsymbol{\nu}^{(0)}(p;\cdot) = \mathsf{Be}\left(\alpha^{(0)}, \theta^{(0)}\right) = \nu_0^{(0)}$, for every $p \in [0,1]$.*

ii) *Say that as $\kappa \to \infty$, and $p_\kappa \to p$ in $[0,1]$, $\alpha^{(\kappa)} \to \alpha$ and $\theta^{(\kappa)} \to \theta$ in $(0,\infty)$. Then $\boldsymbol{\nu}^{(\kappa)}(p_\kappa;\cdot)$ converges weakly to δ_p, and $\nu_0^{(\kappa)}$ converges weakly to $\mathsf{Be}(\alpha,\theta)$, as $\kappa \to \infty$.*

In Figure 25 we illustrate the convergence of the Beta-Binomial transitions, mentioned in Lemma 4.16, by means of the densities of $\boldsymbol{\nu}^{(k)}(p_\kappa;\cdot)$. To better show the effect of κ we fixed $\alpha^{(\kappa)} = \theta^{(\kappa)} = 10$ and $p_\kappa = 0.3$ for each κ. Here it can be appreciated that as κ grows, the mass concentrates in a smaller interval centered around $p = 0.3$. So we have a graphical illustration of how $\boldsymbol{\nu}^{(k)}(p;\cdot)$ converges weakly to δ_p.

Naturally we will be using the Beta-Binomial transition and its invariant distribution as a driver of MSBs, formally we define these stick-breaking processes below.

Definition 4.5. *Let $(S, \mathscr{B}_S)$ be Borel space and consider a diffuse probability measure, μ_0, over $(S, \mathscr{B}_S)$. Let $\boldsymbol{\nu}$ be a Beta-Binomial transition with parameters (α, θ, κ), as in (4.11). We call a Beta-Binomial stick-breaking process (BBSB) with parameters $(\alpha, \theta, \kappa, \mu_0)$ to any MSB, $\boldsymbol{\mu}$, with base measure μ_0 and Markovian length variables $(\mathbf{v}_i)_{i\geq 1}$ with one-step transition probability kernel, $\boldsymbol{\nu}$, and initial distribution $\nu_0 = \mathsf{Be}(\alpha,\theta)$. The stick-breaking weights $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ will be referred to as Beta-Binomial stick-breaking weights sequence (BBSBw) with parameters (α, θ, κ).*

With all the results we have developed thus far, it is fairly simple to show that BBSBs constitute a feasible class Bayesian non-parametric priors. Putting together Theorem 4.13 and Lemma 4.15 we obtain the following result.

Corollary 4.17. *Any BBSB is proper and has full support.*

Another immediate result arises from Theorem 4.14 and Lemma 4.16, this one proves that by tuning κ and α one approximate Dirichlet and Geometric processes by means of BBSBs

Corollary 4.18. *Let $(S, \mathscr{B}_S)$ be a Polish space, for every $\kappa \in \{0, 1, \ldots\}$ let $\alpha^{(\kappa)}, \theta^{(\kappa)} > 0$ and let $\mu_0^{(\kappa)}$ be a diffuse probability measure over $(S, \mathscr{B}_S)$. Consider a BBSB $\boldsymbol{\mu}^{(\kappa)}$ with parameters $\left(\alpha^{(\kappa)}, \theta^{(\kappa)}, \kappa, \mu_0^{(\kappa)}\right)$. Say that as $\kappa \to \infty$, $\alpha^{(\kappa)} \to \alpha$, $\beta^{(\kappa)} \to \beta$ in $(0,\infty)$ and $\mu_0^{(\kappa)}$ converges weakly to the diffuse probability measure μ_0. Then*

i) *For $\kappa = 0$ we get that $\boldsymbol{\mu}^{(0)}$ is a stick-breaking process with base measure $\mu_0^{(0)}$ and featuring i.i.d. length variables $(\mathbf{v}_i)_{i\geq 1} \overset{iid}{\sim} \mathsf{Be}\left(\alpha^{(0)}, \theta^{(0)}\right)$. In particular, if $\alpha^{(0)} = 1$, $\boldsymbol{\mu}^{(0)} \sim \mathcal{D}_{\left(\theta^{(0)}, \mu_0^{(0)}\right)}$ and the corresponding BBSBw, $\mathbf{W}^{(0)}$, is invariant under size-biased permutations.*

ii) *As $\kappa \to \infty$, $\boldsymbol{\mu}^{(\kappa)}$ converges weakly in distribution to $\boldsymbol{\mu}^{(\infty)} \sim \mathcal{G}_{(\mathsf{Be}(\alpha,\theta),\mu_0)}$, and the corresponding BBSBw's, $\mathbf{W}^{(\kappa)}$ converge in distribution to the decreasing ordered weights of $\boldsymbol{\mu}^{(\infty)}$.*

Realize that if $\mu_0^{(\kappa)} = \mu_0$, $\alpha^{(\kappa)} = \alpha$ and $\theta^{(\kappa)} = \theta$ for every $\kappa \in \{0, 1, \ldots\}$, a simpler version of Corollary 4.18 is attained. The dependence of these parameters on κ is simply to have the most general version of the result in question. This said, note that just as can be done using DSBs, as analysed in Section 4.1.2, by simply tuning a single real-valued parameter, we can approximate Dirichlet and Geometric processes through BBSBs. In contrast to DSBs, where the modulating parameter, denoted by β, takes values in $(0,\infty)$ (or $\rho_\nu = 1/(1+\beta)$, which takes values in $(0,1)$), for BBSBs the tuning parameter, κ, takes values in $\{0,1,\ldots\}$. This might be regarded as a small disadvantage of BBSBs against DSBs. For example Dirichlet processes can be approximated with arbitrary precision by means of DSBs that are not a Dirichlet process, by making $\rho_\nu = 1/(1+\beta)$ arbitrarily small. This can not be done using BBSBs, if $\kappa \neq 0$, so that the BBSB is not a Dirichlet process, the smaller value κ can take is one, which is not arbitrarily close to zero. A more significant disadvantage of BBSBs against DSBs, or more generally stick-breaking processes driven by SSPs, is that computing the probability that consecutive weights are decreasing is much harder to do for BBSBs. In fact, for BBSBw's, $\left(\mathbf{w}_j^{(\kappa)}\right)_{j\geq 1}$, with parameters (α,θ,κ), from (4.8), we obtain

$$\begin{aligned}\mathbb{P}\left[\mathbf{w}_{j+1}^{(\kappa)} \leq \mathbf{w}_j^{(\kappa)}\right] &= \int_0^1 \int_0^{c(p)} \sum_{z=0}^{\kappa} \mathsf{Be}(dv \mid \alpha+z, \theta+\kappa-z)\mathsf{Bin}(z \mid \kappa, p)\mathsf{Be}(dp \mid \alpha,\theta)\\ &= \sum_{z=0}^{\kappa}\binom{\kappa}{z}\frac{\Gamma(\alpha+\theta)}{\Gamma(\alpha)\Gamma(\theta)}\int_0^1 \mathcal{I}_{c(p)}(\alpha+z,\theta+\kappa-z)p^{\alpha+z}(1-p)^{\theta+\kappa-z}dp\\ &= \sum_{z=0}^{\kappa}\binom{\kappa}{z}\frac{(\alpha)_z(\theta)_{\kappa-z}}{(\alpha+\theta)_\kappa}\mathbb{E}\left[\mathcal{I}_{c(\mathbf{v})}(\alpha+z,\theta+\kappa-z)\right]\end{aligned}$$

where $c(v) = 1\wedge v(1-v)^{-1}$, $\mathbf{v} \sim \mathsf{Be}(\alpha+z,\theta+\kappa-z)$ and $\mathcal{I}_x(a,b) = \int_0^x \mathsf{Be}(dv \mid a,b)$ denotes the regularized Beta function. Comparing this quantity with Theorem 4.6 or Corollary 4.10, we see that $\mathbb{P}\left[\mathbf{w}_{j+1}^{(\kappa)} \leq \mathbf{w}_j^{(\kappa)}\right]$ is a much more complicated function of κ than its counterparts of the tuning parameters ρ_ν or β. For BBSBw's it is even hard to show that the mapping $\kappa \mapsto \mathbb{P}\left[\mathbf{w}_{j+1}^{(\kappa)} \leq \mathbf{w}_j^{(\kappa)}\right]$ is non-decreasing, or derive an optimal lower bound for $\mathbb{P}\left[\mathbf{w}_{j+1}^{(\kappa)} \leq \mathbf{w}_j^{(\kappa)}\right]$. However, what can be shown as a consequence of Corollary 4.18 is that $\mathbb{P}\left[\mathbf{w}_{j+1}^{(\kappa)} \leq \mathbf{w}_j^{(\kappa)}\right] \to 1$ as $\kappa\to\infty$, meaning that if κ is sufficiently large, the probability in question is close to one. Now, using equation (4.9) we can compute the conditional probability

$$\mathbb{P}\left[\mathbf{w}_{j+1}^{(\kappa)} \leq \mathbf{w}_j^{(\kappa)} \,\middle|\, \mathbf{w}_1^{(\kappa)},\ldots,\mathbf{w}_j^{(\kappa)}\right] = \sum_{z=0}^{\kappa}\mathcal{I}_{c\left(\mathbf{v}_j^{(\kappa)}\right)}(\alpha+z,\theta+\kappa-z)\mathsf{Bin}\left(z \,\middle|\, \kappa, \mathbf{v}_j^{(\kappa)}\right)$$

where $\left(\mathbf{v}_i^{(\kappa)}\right)_{i\geq 1}$ are the length variables of $\left(\mathbf{w}_j^{(\kappa)}\right)_{j\geq 1}$. In contrast to $\mathbb{P}\left[\mathbf{w}_{j+1}^{(\kappa)} \leq \mathbf{w}_j^{(\kappa)}\right]$, the conditional probability a less complex function of κ. Nonetheless, when compared against Theorem 4.7 and Corollary 4.10 (f), it remains a complicated function of the tuning parameter.

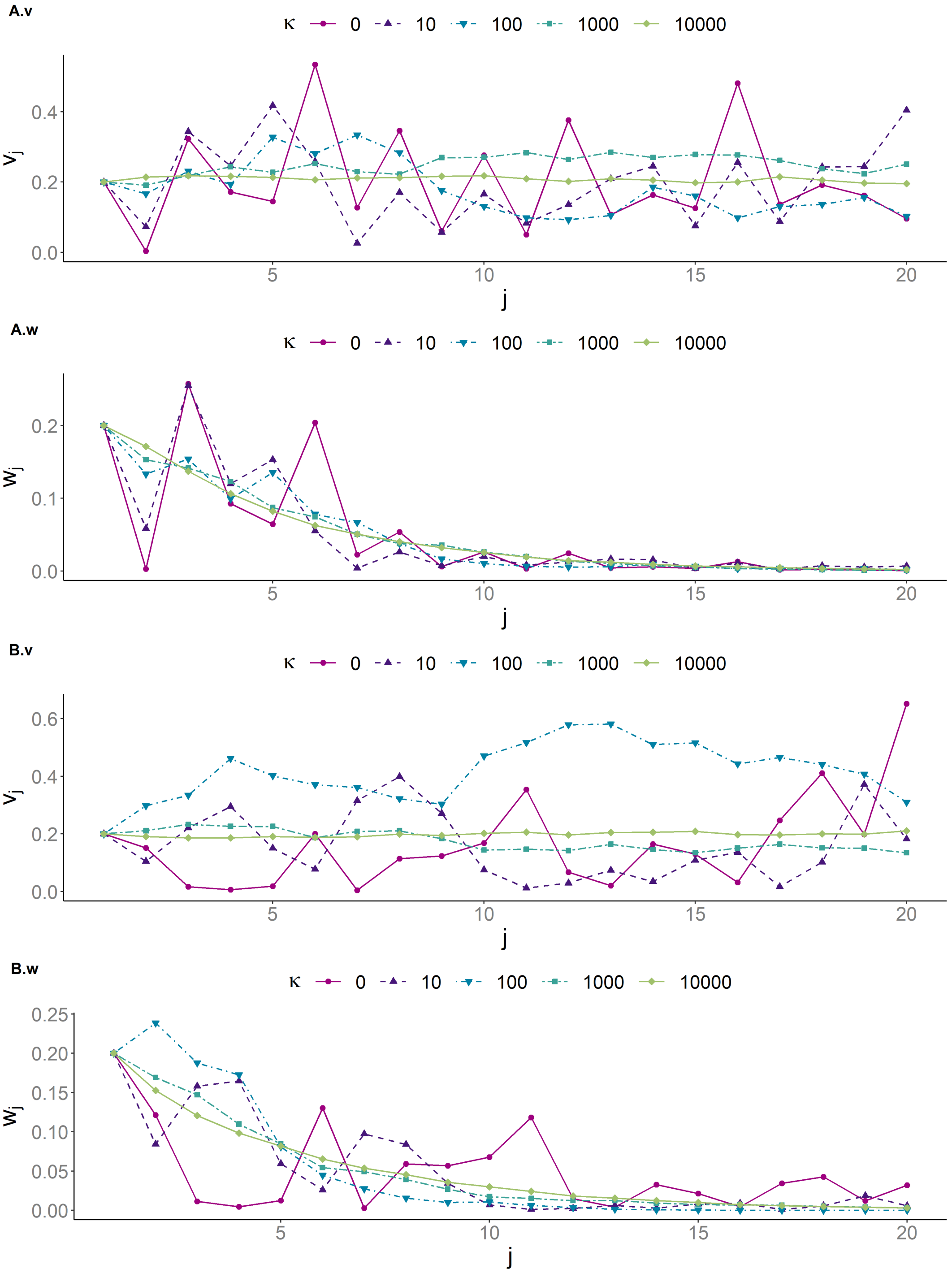


Figure 26: Simulations (A.w,B.w) BBSBw's $(\mathbf{w}_j)_{j\geq1}$ with parameters $(\alpha = 1, \theta, \kappa)$ for $\kappa \in \{0, 10, 100, 1000, 10000\}$. A.v and B.v show the corresponding underlying length variables $(\mathbf{v}_i)_{i=1}^{20}$.

In Figure 26 we show some simulations of BBSBw's for distinct values of κ. In A.v and B.v we observe that for a larger value of κ, the length variable $\mathbf{v}_{j+1}$ takes values closer to $\mathbf{v}_j$, alternative a smaller value of κ allows $\mathbf{v}_{j+1}$ to take values that are farther from $\mathbf{v}_j$ with a larger probability. Informally, we might think of $\mathbf{v}_{j+1}$ as a noisy observation of $\mathbf{v}_j$, and the larger κ is, the smaller is the noise. Now, if $\mathbf{v}_{j+1}$ takes a value that is very close to $\mathbf{v}_j$, so that $\mathbf{v}_{j+1} \approx \mathbf{v}_j \leq \mathbf{v}_j(1-\mathbf{v}_j)^{-1}$, this yields $\mathbf{w}_{j+1} \leq \mathbf{w}_j$, which explains why as κ grows the weights are more likely to be decreasingly ordered. If we compare Figure 26 with Figure 23 we see that the tuning parameters, κ and ρ_ν, of BBSBs and DSBs, respectively, modulate the ordering of the weights in very distinct ways. While in DSBs the tie probability, ρ_ν, controls the probability that $\mathbf{v}_{j+1}$ takes a value previously observed in $(\mathbf{v}_1, \ldots, \mathbf{v}_j)$, in particular the probability that $\mathbf{v}_{j+1} = \mathbf{v}_j$, for BBSBs $\mathbf{v}_{j+1} = \mathbf{v}_j$ occurs with probability zero, however κ modulates how likely $\mathbf{v}_{j+1} \approx \mathbf{v}_j$. Also, for DSBs, under the event $\{\mathbf{v}_j \neq \mathbf{v}_{j+1}\}$ we have that the value of $\mathbf{v}_{j+1}$ is chosen independently of the value $\mathbf{v}_j$ takes, in contrast, for BBSBs under the event $\{\mathbf{v}_j \neq \mathbf{v}_{j+1}\}$, which occurs almost surely, the value $\mathbf{v}_j$ takes always affects that of $\mathbf{v}_{j+1}$ unless $\kappa = 0$. This said, at least empirically, it seems that the way the stochastic ordering of the weights is controlled by BBSBs is neater than that of DSBs.

To better understand the effect of tuning parameter on the length variables, hence the weights, we can compute conditional moments of length variables. For DSBs these moments can be attained from Corollary 3.9. For BBSBs we have our next result (see Appendix D.8 for details on computations).

Proposition 4.19. *Let $(\mathbf{v}_i)_{i\geq 1}$ be a stationary Markov chain with Beta-Binomial transition $\boldsymbol{\nu}$ as in (4.11), and initial distribution $\mathsf{Be}(\alpha, \theta)$. Then, for every $i \geq 1$.*

a) $\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{v}_i] = \dfrac{\alpha + \kappa \mathbf{v}_i}{\alpha + \theta + \kappa}$.

b) $\mathsf{Var}(\mathbf{v}_{i+1} \mid \mathbf{v}_i) = \dfrac{(\alpha + \kappa\mathbf{v}_i)(\theta + \kappa(1-\mathbf{v}_i)) + \kappa\mathbf{v}_i(1-\mathbf{v}_i)(\alpha+\theta+\kappa)}{(\alpha+\theta+\kappa)^2(\alpha+\theta+\kappa+1)}$.

c) $\mathsf{Cov}(\mathbf{v}_i, \mathbf{v}_{i+1}) = \dfrac{\kappa\alpha\theta}{(\alpha+\theta)^2(\alpha+\theta+1)(\alpha+\theta+\kappa)}$.

d) $\mathsf{Corr}(\mathbf{v}_i, \mathbf{v}_{i+1}) = \dfrac{\mathsf{Cov}(\mathbf{v}_i, \mathbf{v}_{i+1})}{\sqrt{\mathsf{Var}(\mathbf{v}_i)}\sqrt{\mathsf{Var}(\mathbf{v}_{i+1})}} = \dfrac{\kappa}{\alpha+\theta+\kappa}$.

Consistently with the analysis we have been carrying out for BBSBs, if for a fixed value of κ, we increase α and θ, we get $\mathsf{Corr}(\mathbf{v}_i, \mathbf{v}_{i+1}) \approx 0$. Alternatively, if we fix α and θ, for larger values of κ, $\mathsf{Corr}(\mathbf{v}_i, \mathbf{v}_{i+1}) \approx 1$. Also, if α and θ are very small with respect to κ, $\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{v}_i] \approx \mathbf{v}_i$ and $\mathsf{Var}(\mathbf{v}_{i+1} \mid \mathbf{v}_i) \approx 2\mathbf{v}_i(1-\mathbf{v}_i)/(\kappa+1)$. So this yields an alternative path to understanding how, for $\kappa = 0$ the length variables are independent, and as $\kappa \to \infty$, $\mathbb{P}[\mathbf{v}_{i+1} \in \cdot \mid \mathbf{v}_i] \to \delta_{\mathbf{v}_i}$. An extra piece of information we obtain from Proposition 4.19 is that for larger values of α or θ, we require an even larger value of κ for $(\mathbf{v}_i)_{i\geq 1}$ to approximate $(\mathbf{v}, \mathbf{v}, \ldots)$ in distribution. That is, if either α or θ take a big value, we require an even bigger of κ to approximate a Geometric process using BBSBs.

Another quantity of interest determined by the length variables is computed in the following result (See Appendix D.9 for details in calculations).

Proposition 4.20. *Let $(\mathbf{v}_i)_{i\geq 1}$ be a stationary Markov chain with Beta-Binomial transition $\boldsymbol{\nu}$ as in (4.11), and initial distribution $\mathsf{Be}(\alpha, \theta)$. Then for any non-negative integers $(a_j, b_j)_{j=1}^n$,*

$$\mathbb{E}\left[\prod_{j=1}^n \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{z_1=0}^{\kappa} \cdots \sum_{z_n=0}^{\kappa} \left\{ \prod_{i=1}^n \binom{\kappa}{z_i} \frac{(\alpha_i)_{a_i+z_i}(\theta_i)_{b_i+\kappa-z_i}}{(\alpha_i+\theta_i)_{a_i+b_i+\kappa}} \right\}$$

where $\alpha_1 = \alpha$, $\theta_1 = \theta$ and for $2 \leq i \leq k$, $\alpha_i = \alpha + z_{i-1}$ and $\theta_i = \theta + \kappa - z_{i-1}$.

Recall that the expectations of power products of length variables allows us to compute clustering-related probabilities through (3.6), (3.7) and (3.8). In particular Proposition 4.20 together with (3.8) show that more often that not, BBSBw's will not be in size-biased random order, making it hard to study analytically clustering probabilities. For our last prior analysis of BBSBs we study empirically the distribution of $\mathbf{K}_n$.

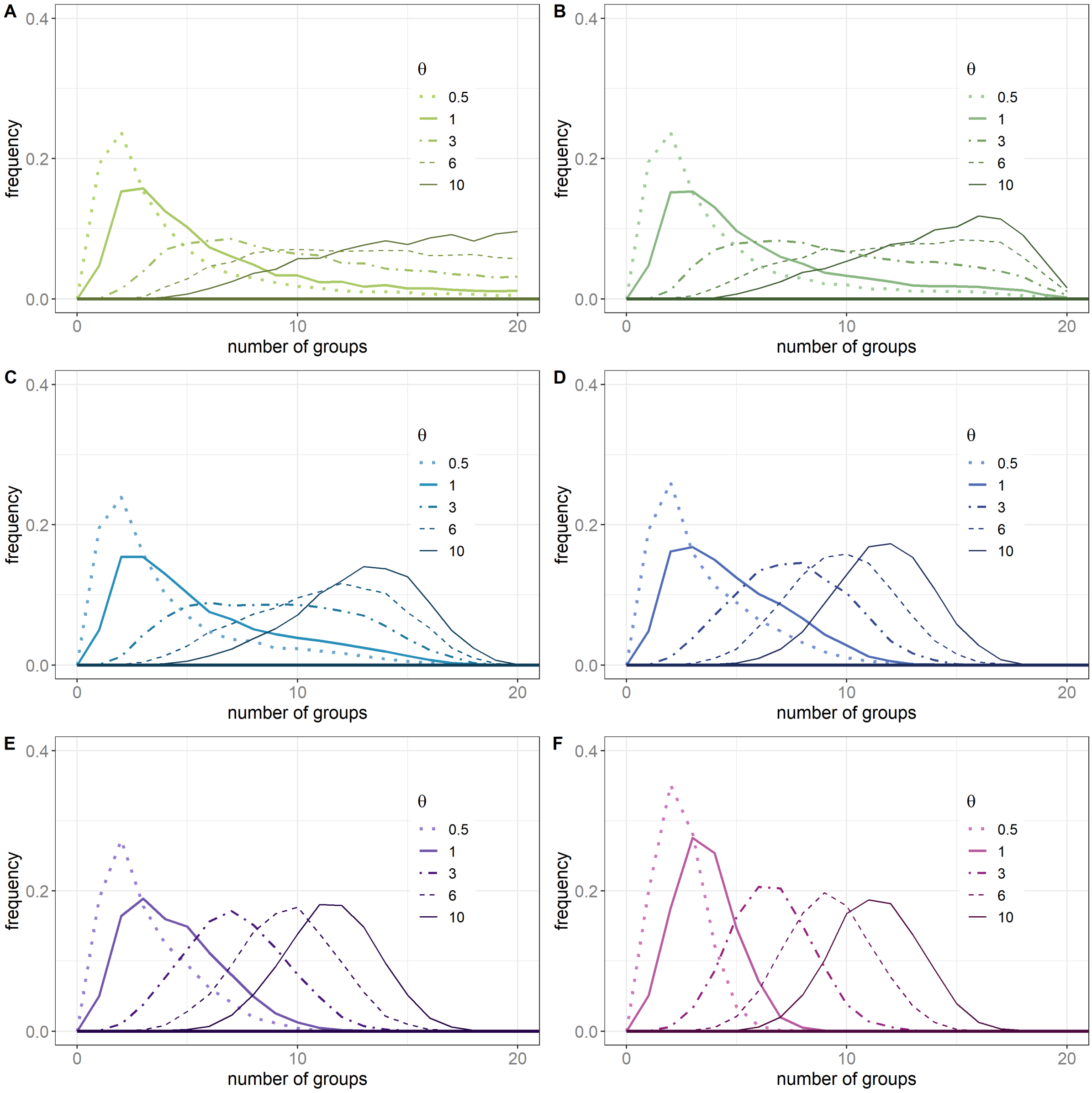


Figure 27: Frequency polygons of samples of size 10000 from the distribution of $\mathbf{K}_{20}$ corresponding to a BBSB prior with $\kappa = 0$ (A), $\kappa = 5$ (B), $\kappa = 10$ (C), $\kappa = 100$ (D), $\kappa = 1000$ (E) and $\kappa = \infty$ (F). For each fixed value of κ we vary θ in the set $\{0.5, 1, 3, 6, 10\}$. In all cases we fixed $\alpha = 1$.

In Figure 27 we illustrate the distribution of $\mathbf{K}_n^{(\kappa,\theta)} = |\mathbf{\Pi}(\mathbf{x}_{1:n})|$ where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, and $\boldsymbol{\mu}$ is a BBSB with paramaters $\alpha = 1$, $\theta \in \{0.5, 1, 3, 6, 10\}$ and $\kappa \in \{0, 5, 10, 100, 1000, \infty\}$. Here $\mathbf{K}_n^{(\infty,\theta)}$ refers to the case here $\boldsymbol{\mu} \sim \mathcal{G}_{(\mathsf{Be}(1,\theta),\mu_0)}$, also note that the choice $\alpha = 1$ allows to recover Dirichlet processes when $\kappa = 0$. Figure 27 is graphical representation of Corollary 4.18, here we can appreciate that $\mathbf{K}_n^{(\kappa,\theta)} \xrightarrow{d} \mathbf{K}_n^{(0,\theta)}$ as $\kappa \to 0$ and $\mathbf{K}_n^{(\kappa,\theta)} \xrightarrow{d} \mathbf{K}_n^{(\infty,\theta)}$ as $\kappa \to \infty$. Note that for a fixed value of θ, an increment of κ contributes to the distribution of $\mathbf{K}_n^{(\kappa,\theta)}$ with a larger mean in variance. In other words the distribution of $\mathbf{K}_n^{(\kappa,\theta)}$ becomes less informative as κ grows. In Figure 27 we also see that parameter the θ affects the rates of convergence. Explicitly, if θ is small we see that the distribution of $\mathbf{K}_n^{(\kappa,\theta)}$ approximates well that of $\mathbf{K}_n^{(\infty,\theta)}$, even is κ is not that large. However, we require a really small value of κ in order to approximate the distribution of $\mathbf{K}_n^{(0,\theta)}$ through that of $\mathbf{K}_n^{(\infty,\theta)}$. Conversely if θ is big, we observe that the convergence rate of $\mathbf{K}_n^{(\kappa,\theta)} \xrightarrow{d} \mathbf{K}_n^{(0,\theta)}$, as $\kappa \to 0$, is fast, meaning that we do not need an extremely small value of κ to attain, $\mathbf{K}_n^{(\kappa,\theta)} \approx \mathbf{K}_n^{(0,\theta)}$ in distribution. In this case, where θ is large, the convergence rate $\mathbf{K}_n^{(\kappa,\theta)} \xrightarrow{d} \mathbf{K}_n^{(\infty,\theta)}$, as $\kappa \to \infty$, is slow, so an very large value of κ is required to approximate the distribution of $\mathbf{K}_n^{(\infty,\theta)}$ through that of $\mathbf{K}_n^{(\kappa,\theta)}$. Comparing Figures 24 and 27 we see that this is not the case for DSBs, where θ does not seems to affect the convergence of DSBs to the limiting processes, at least when we analyse the distribution of $\mathbf{K}_n$. In this sense, one might think that the way DSBs approximate Dirchlet and Geometric processes is neater than the way BBSBs do.

4.2.4 Spike and slab stick-breaking processes (SSBs)

To motivate this second example of MSBs recall that if $\{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots \mid \hat{\boldsymbol{\nu}}\} \overset{\text{iid}}{\sim} \hat{\boldsymbol{\nu}}$ where $\hat{\boldsymbol{\nu}}$ is a species sampling process over $\big([0,1], \mathscr{B}_{[0,1]}\big)$ with base measure $\hat{\nu}_0$ and tie probability ρ_ν then for every $i \geq 1$,

$$\mathbb{P}\left[\hat{\mathbf{v}}_{i+1} \in \cdot \mid \hat{\mathbf{v}}_i\right] = \rho_\nu\, \delta_{\hat{\mathbf{v}}_i} + (1-\rho_\nu)\hat{\nu}_0,$$

(see for instance Corollary 3.8). Here we will work with Markovian length variables $(\mathbf{v}_i)_{i\geq 1}$ with one-step transition probability kernel $\boldsymbol{\nu} : [0,1] \to [0,1]$ such that

$$\boldsymbol{\nu}(\mathbf{v}_i; \cdot) = \mathbb{P}\left[\mathbf{v}_{i+1} \in \cdot \mid \mathbf{v}_i\right] = \mathrm{p}\delta_{\mathbf{v}_i} + (1-\mathrm{p})\nu_0,$$

for some $\mathrm{p} \in [0,1]$, and a suitable probability measure ν_0 over $[0,1]$. First of all note that for any probability measure λ over $[0,1]$, and measurable function $f : [0,1] \to \mathbb{R}_+$,

$$\begin{aligned}\lambda\boldsymbol{\nu}(f) &= \int\int f(x)\boldsymbol{\nu}(v; dx)\lambda(dv) \\ &= \mathrm{p}\int f(v)\lambda(dv) + (1-\mathrm{p})\nu_0(f) = \mathrm{p}\lambda(f) + (1-\mathrm{p})\nu_0(f),\end{aligned}$$

which implies that ν_0 is invariant for $\boldsymbol{\nu}$, moreover, if $\mathrm{p} < 1$, ν_0 is only invariant measure for $\boldsymbol{\nu}$. This said, we will further assume $\mathbf{v}_1 \sim \nu_0$, so that marginally $\mathbf{v}_i \sim \nu_0$, for every $i \geq 1$, and $(\mathbf{v}_i)_{i\geq 1}$ becomes a stationary Markov chain. With this considerations taken into account, note that if $\rho_\nu = \mathrm{p}$, and $\hat{\nu}_0 = \nu_0$, then for every $i \geq 1$, $(\hat{\mathbf{v}}_i, \hat{\mathbf{v}}_{j+1})$ is equal in distribution to $(\mathbf{v}_i, \mathbf{v}_{j+1})$. So in a certain way the following models are the Markovian version of ESBs driven by SSPs. Formally we define this processes below.

Definition 4.6. *Let $(S, \mathscr{B}_S)$ be Borel space and consider a diffuse probability measure, μ_0, over $(S, \mathscr{B}_S)$. Fix $\mathrm{p} \in [0,1]$ and consider a probability measure, ν_0, over $\big([0,1], \mathscr{B}_{[0,1]}\big)$ such that $\nu_0(\{0\}) = 0$. To the probability kernel $\boldsymbol{\nu} : [0,1] \to [0,1]$ given by*

$$\boldsymbol{\nu}(v; \cdot) = \mathrm{p}\delta_v + (1-\mathrm{p})\nu_0, \tag{4.12}$$

for every $v \in [0,1]$, we call a spike and slab transition probability kernel with parameters (p, ν_0). Also, to any MSB, $\boldsymbol{\mu}$, with base measure μ_0 and Markovian length variables, $(\mathbf{v}_i)_{i\geq 1}$, with initial distribution ν_0 and one-step transition probability kernel, $\boldsymbol{\nu}$, as in (4.12), we call a spike and slab Markov stick-breaking process (SSB, for short) with parameters $(\mathrm{p}, \nu_0, \mu_0)$. The corresponding stick-breaking weights $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ will be referred to as spike and slab Markov stick-breaking weights sequence (SSBw) with parameters (p, ν_0).

As we have mentioned, for $\boldsymbol{\nu}$ as in (4.12), ν_0 is the only invariant measure whenever $\mathrm{p} < 1$. In this case ν_0 is clearly $\boldsymbol{\nu}$-ergodic, and the requirement $\nu_0(\{0\}) = 0$, trivially implies $\nu_0 \neq \delta_0$. Thus, Theorem 4.13 shows that any SSB, $\boldsymbol{\mu}$, with $\mathrm{p} < 1$ is proper. For the case $\mathrm{p} = 1$, the SSB becomes a Geometric process, which we know is proper as long as the distribution, ν_0, of the length variable, satisfies $\nu_0(\{0\}) = 0$, as required. This means that any SSB as introduced in Definition 4.6 is proper. Now, if there exist $0 < \varepsilon < 1$, such that $(0, \varepsilon)$ is contained in the support of ν_0, then for $\mathrm{p} < 1$, we also have that $(0, \varepsilon)$ is contained in the support of $\boldsymbol{\nu}(v; \cdot)$ for every $v \in [0,1]$, in which case we get the corresponding SSB has full support. Under the same assumption over ν_0, for the case $\mathrm{p} = 1$, Corollary 3.25 shows this SSB also has full support. This discussion is summarized in the following result.

Corollary 4.21. *Let $\boldsymbol{\mu}$ be a SSB with parameters $(\mathrm{p}, \nu_0, \mu_0)$. Then $\boldsymbol{\mu}$ is proper, and if there exist $0 < \varepsilon < 1$, such that $(0, \varepsilon)$ is contained in the support of ν_0, $\boldsymbol{\mu}$ has full support. In particular, the choice $\nu_0 = \mathsf{Be}(\alpha, \theta)$ yields $\boldsymbol{\mu}$ has full support.*

Now, fix μ_0 and ν_0 as in Definition 4.6, and for $\mathrm{p} \in [0,1]$, consider the SSB, $\boldsymbol{\mu}^{(\mathrm{p})}$, with parameters $(\mathrm{p}, \nu_0, \mu_0)$. As we have mentioned $\boldsymbol{\mu}^{(1)} \sim \mathcal{G}_{(\nu_0, \mu_0)}$, and evidently $\boldsymbol{\mu}^{(0)}$ is a stick-breaking process featuring i.i.d. length variables, in particular, if $\nu_0 = \mathsf{Be}(1, \theta)$, $\boldsymbol{\mu}^{(0)} \sim \mathcal{D}_{(\theta, \mu_0)}$. By simply looking at equation (4.12) and using Theorem 4.14 it is very easy to show that $\boldsymbol{\mu}^{(\mathrm{p})}$ converges weakly in distribution to $\boldsymbol{\mu}^{(1)}$ as $\mathrm{p} \to 1$ and that $\boldsymbol{\mu}^{(\mathrm{p})}$ converges weakly in distribution to $\boldsymbol{\mu}^{(0)}$ as $\mathrm{p} \to 0$. The next result rephrases this in a slightly more general scenario (see Appendix D.10 for the details regarding the convergence of the probability kernels).

Corollary 4.22. *Let $(S, \mathscr{B}_S)$ be a Polish space. For each $n \geq 1$ let $\mathrm{p}_n \in (0,1)$, and consider a diffuse probability measure, $\mu_0^{(n)}$, over $(S, \mathscr{B}_S)$, and a probability measure, $\nu_0^{(n)}$, over $\left([0,1], \mathscr{B}_{[0,1]}\right)$ with $\nu_0^{(n)}(\{0\}) = 0$. Also, let $\boldsymbol{\mu}^{(n)}$ be SSB with parameters $\left(\mathrm{p}_n, \nu_0^{(n)}, \mu_0^{(n)}\right)$. Say that, as $n \to \infty$, $\mu_0^{(n)}$ converges weakly to the diffuse probability measure μ_0, and $\nu_0^{(n)}$ converges weakly to a probability measure ν_0, with $\nu_0(\{0\}) = 0$.*

i) *If $\mathrm{p}_n \to 0$ then $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to a stick-breaking process, $\boldsymbol{\mu}$, featuring i.i.d. length variables $(\mathbf{v}_i)_{i\geq 1} \overset{iid}{\sim} \nu_0$. In particular, if $\nu_0 = \mathsf{Be}(1, \theta)$, $\boldsymbol{\mu} \sim \mathcal{D}_{(\theta, \mu_0)}$, and the corresponding SSBw, $\mathbf{W}^{(n)}$, of $\boldsymbol{\mu}^{(n)}$, converge in distribution to the size-biased permuted weights of $\boldsymbol{\mu}$.*

ii) *If $\mathrm{p}_n \to 1$, $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\boldsymbol{\mu} \sim \mathcal{G}_{(\nu_0, \mu_0)}$, and the corresponding SSBw, $\mathbf{W}^{(n)}$ converge in distribution to the decreasing ordered weights of $\boldsymbol{\mu}$.*

So far we have developed three concrete examples of classes of stick-breaking processes with dependent length variables, by means of which, by tuning a single real-valued parameter we can recover Dirichlet and Geometric processes. Namely for DSBs we can tune the parameter β, which takes values in $(0,\infty)$, (or $\rho_\nu = 1/(1+\beta) \in (0,1)$), for BBSBs the modulating parameter is $\kappa \in \{0,1,\ldots\}$, and for SSBs we the tuning parameter is $\mathrm{p} \in [0,1]$. A common advantage DSBs and SSBs over BBSBs is that the tuning parameter lays in a continuous space, which in turn translates to the fact that Dirichlet and Geometric processes can be approximated arbitrarily, this is of course not true for BBSBs. Another common advantage of SSBs and DSBs, or more generally ESBs driven by SSPs, is that we can easily compute the probability that consecutive weights are decreasing. In fact for these classes of processes, the probability in question is very similar. As mentioned at the beginning of this section, for $\{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots \mid \hat{\boldsymbol{\nu}}\} \overset{\text{iid}}{\sim} \hat{\boldsymbol{\nu}}$ where $\hat{\boldsymbol{\nu}}$ is a species sampling process over $\big([0,1], \mathscr{B}_{[0,1]}\big)$ with base measure $\hat{\nu}_0$ and tie probability ρ_ν, and for the stationary Markov chain $(\mathbf{v}_i)_{i\geq 1}$, with initial distribution ν_0 and one-step transition probability kernel, $\boldsymbol{\nu}$, as in (4.12), we get $(\hat{\mathbf{v}}_i, \hat{\mathbf{v}}_{j+1})$ is equal in distribution to $(\mathbf{v}_i, \mathbf{v}_{j+1})$ whenever $\rho_\nu = \mathrm{p}$ and $\hat{\nu}_0 = \nu_0$. This means that if we consider $(\hat{\mathbf{w}}_j)_{j\geq 1} = \mathsf{SB}[(\hat{\mathbf{v}}_i)_{i\geq 1}]$ and $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$, then $\mathbb{P}[\hat{\mathbf{w}}_{j+1} \leq \hat{\mathbf{w}}_j] = \mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j]$. Hence by substituting ρ_ν by p in Theorem 4.6 we already have available important properties of $\mathbb{P}[\mathbf{w}_{j+1} \leq \mathbf{w}_j]$. Explicitly, if $\left(\mathbf{w}_j^{(\mathrm{p})}\right)_{j\geq 1}$ is a SSBw with parameters (p, ν_0), we have

$$\mathbb{P}\left[\mathbf{w}_j^{(\mathrm{p})} \geq \mathbf{w}_{j+1}^{(\mathrm{p})}\right] = \mathrm{p} + (1-\mathrm{p})\mathbb{E}\left[\overrightarrow{\nu_0}(c(\mathbf{v}))\right],$$

where $c(v) = 1 \wedge v(1-v)^{-1}$ for every $v \in [0,1]$, $\mathbf{v} \sim \nu_0$, and $\overrightarrow{\nu_0}$ is the distribution function of ν_0, that is $\overrightarrow{\nu_0}(x) = \nu_0([0,x])$. In particular, the choice $\nu_0 = \mathsf{Be}(1,\theta)$ yields

$$\mathbb{P}\left[\mathbf{w}_j^{(\mathrm{p})} \geq \mathbf{w}_{j+1}^{(\mathrm{p})}\right] = 1 - \frac{{}_2F_1(1,1;\theta+2,1/2)(1-\mathrm{p})\theta}{2(\theta+1)},$$

and if $\theta = 1$, we get $\mathbb{P}\left[\mathbf{w}_j^{(\mathrm{p})} \geq \mathbf{w}_{j+1}^{(\mathrm{p})}\right] = \mathrm{p} + (1-\mathrm{p})\log(2)$. Now, to highlight the difference between SSBw's and ESBw's where the length variables are driven by SSPs, we can use (4.9) to compute the conditional probability

$$\mathbb{P}\left[\mathbf{w}_n^{(\mathrm{p})} \geq \mathbf{w}_{n+1}^{(\mathrm{p})} \,\middle|\, \mathbf{w}_1^{(\mathrm{p})}, \ldots, \mathbf{w}_n^{(\mathrm{p})}\right] = \mathrm{p} + (1-\mathrm{p})\overrightarrow{\nu_0}\left(c\left(\mathbf{v}_n^{(\mathrm{p})}\right)\right),$$

for the case $\nu_0(\{0\}) = \nu_0(\{1\}) = 0$, which simplifies to

$$\mathbb{P}\left[\mathbf{w}_n^{(\mathrm{p})} \geq \mathbf{w}_{n+1}^{(\mathrm{p})} \,\middle|\, \mathbf{w}_1^{(\mathrm{p})}, \ldots, \mathbf{w}_n^{(\mathrm{p})}\right] = 1 - (1-\mathrm{p})\left[1 - c\left(\mathbf{v}_n^{(\mathrm{p})}\right)\right]^{\theta},$$

if $\nu_0 = \mathsf{Be}(1,\theta)$. Comparing these equations to Theorem 4.7, we see that for SSBw's the conditional probability depends only on the nth length variable, meanwhile for the exchangeable counterpart, it depends on every length variable up to index n. This due to the fact that exchangeable length variables, $(\hat{\mathbf{v}}_i)_{i\geq 1}$, are allowed to take previously observed values, in contrast, the length variables of SSBw's, $(\mathbf{v}_i)_{i\geq 1}$, either take the last observed value, or choose a new one independently of past observations. Note that as a function of the indexes $\{1,2,\ldots\}$, both $(\hat{\mathbf{v}}_i)_{i\geq 1}$ and $(\mathbf{v}_i)_{i\geq 1}$ are piecewise constant functions. Further, if for some index $j \geq 1$ we observe $\mathbf{v}_j = \mathbf{v}_{j+1} = \mathbf{v}_{j+2}$, then $\mathbf{w}_{j+1}/\mathbf{w}_j = (1-\mathbf{v}_j) = (1-\mathbf{v}_{j+1}) = \mathbf{w}_{j+2}/\mathbf{w}_{j+1}$. Thus, $(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ is piecewise decreasing and the rate at which the weights decrease is also piecewise constant. In this sense, the parameter p controls how often do the weight decrease and how often do the decreasing rates change. In the extreme case p = 1 the decreasing rate never changes and in the opposite instance, p = 0, the decreasing rate always changes. A similar behaviour is observed for $(\hat{\mathbf{w}}_j)_{j\geq 1} = \mathsf{SB}[(\hat{\mathbf{v}}_i)_{i\geq 1}]$ in terms of their tie probability ρ_ν. See Figures 23 and 28 for an illustration.

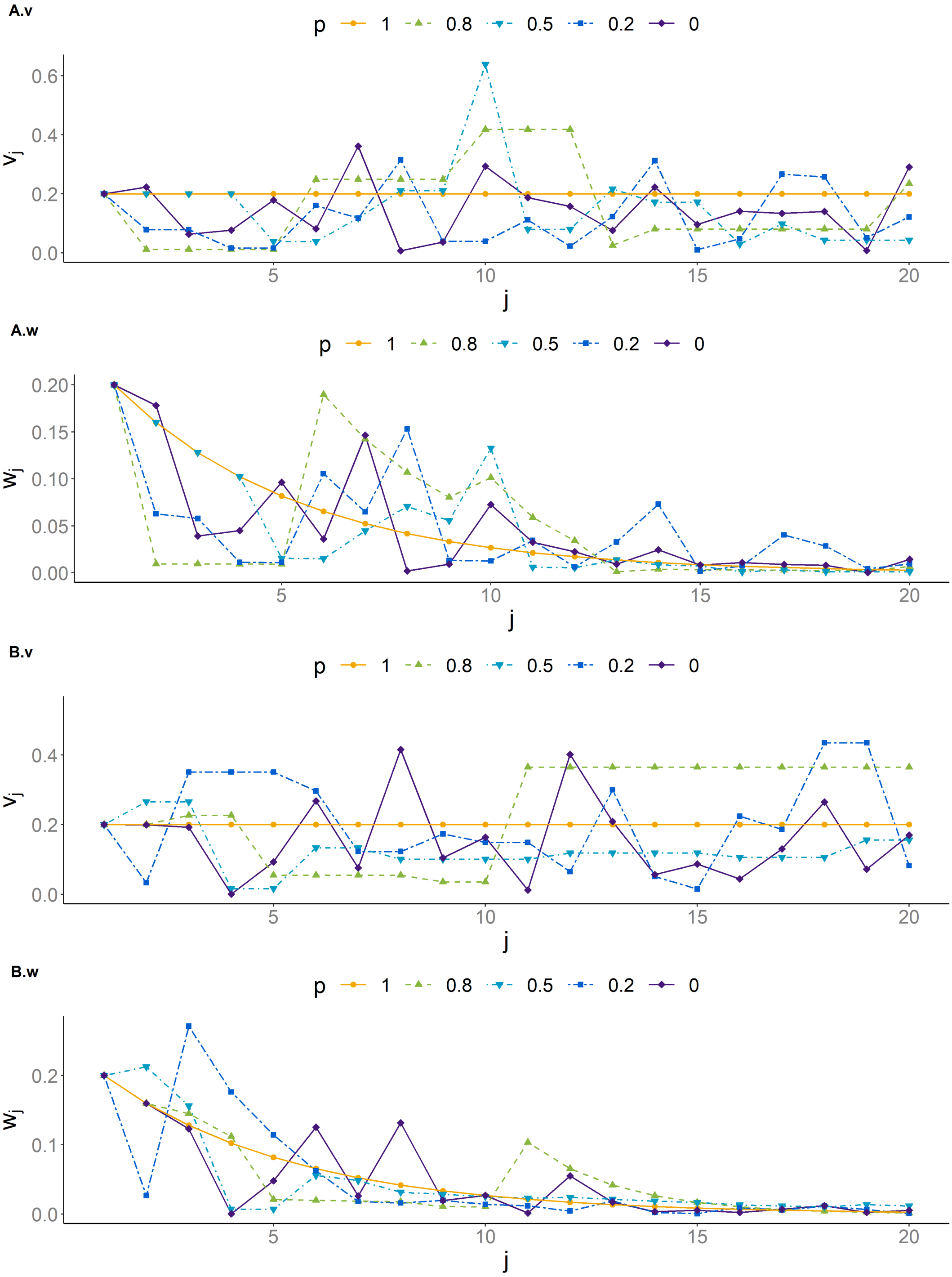


Figure 28: Simulations (A.v,B.v) of a ν_0-Bernoulli Markov chain $(\mathbf{v}_j)_{j=1}^{20}$ with $\nu_0 = \mathsf{Be}(1,\theta)$ for distinct values of p, this probability was fixed to $0, 0.2, 0.5, 0.8$ and 1. A.w and B.w show the corresponding stick-breaking weights $(\mathbf{w}_j)_{j=1}^{20} = \mathsf{SB}[(\mathbf{v}_j)_{j=1}^{20}]$.

Another property that the length variables of SSBs, $(\mathbf{v}_i)_{i\geq 1}$, share with exchangeable length variables, $(\hat{\mathbf{v}}_i)_{i\geq 1}$, driven by species sampling processes, are some conditional moments. Recalling that $(\mathbf{v}_i, \mathbf{v}_{i+1}) \overset{d}{=} (\hat{\mathbf{v}}_i, \hat{\mathbf{v}}_{i+1})$, whenever $\rho_\nu = \mathrm{p}$ and $\hat{\nu}_0 = \nu_0$, and using Corollary 3.9 we obtain.

Corollary 4.23. *Let $(\mathbf{v}_i)_{i\geq 1}$ be an stationary Markov chain with initial distribution ν_0 and one-step transition probability kernel $\boldsymbol{\nu}$ as in (4.12). Then for every $i \geq 1$,*

a) $\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{v}_i] = \mathrm{p}\, \mathbf{v}_i + (1-\rho)\mathbb{E}[\mathbf{v}_i]$,

b) $\mathsf{Var}(\mathbf{v}_{i+1} \mid \mathbf{v}_i) = (1-\mathrm{p})\left\{\mathrm{p}\left(\mathbf{v}_i - \mathbb{E}[\mathbf{v}_i]\right)^2 + \mathsf{Var}(\mathbf{v}_i)\right\}$,

c) $\mathsf{Cov}(\mathbf{v}_i, \mathbf{v}_{i+1}) = \mathrm{p}\, \mathsf{Var}(\mathbf{v}_i)$,

d) $\mathsf{Corr}(\mathbf{v}_i, \mathbf{v}_{i+1}) = \mathrm{p}$.

In terms of the length variables an important difference between SSBs and ESBs driven by SSPs are the expectations of power products of length variables. For ESBs these are computed in Theorem 4.8, and Corollary 4.11 for the special case of DSBs, for SSBs we compute these expectations below.

Proposition 4.24. *Let ν_0 be a diffuse probability measure over $\left([0,1], \mathscr{B}_{[0,1]}\right)$ and let $(\mathbf{v}_i)_{i\geq 1}$ be an stationary Markov chain with initial distribution ν_0 and one-step transition probability kernel $\boldsymbol{\nu}$ as in (4.12). Then for any non-negative integers $(a_j, b_j)_{j=1}^k$,*

$$\mathbb{E}\left[\prod_{j=1}^{k} \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{(\tau_0,\ldots,\tau_m)} \mathrm{p}^{m-1}(1-\mathrm{p})^{k-m} \times \\ \times \prod_{j=0}^{m-1}\left\{\int_{[0,1]} (v)^{\sum_{i\in A_j} a_i}(1-v)^{\sum_{i\in A_j} b_i}\, \nu_0(dv)\right\},$$

where $A_j = \{\tau_j, \ldots, \tau_{j+1}-1\}$, and the sum ranges over all sequences $(\tau_0, \ldots, \tau_m)$ with $\tau_0 = 1 < \tau_1 < \cdots < \tau_m = k+1$. In particular if $\nu_0 = \mathsf{Be}(1,\theta)$

$$\mathbb{E}\left[\prod_{j=1}^{k} \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{(\tau_0,\ldots,\tau_m)} \mathrm{p}^{m-1}(1-\mathrm{p})^{k-m}\theta^m \prod_{j=0}^{m-1} \frac{\left(\sum_{i\in A_j} a_i\right)!}{\left(\theta + \sum_{i\in A_j} b_i\right)_{1+\sum_{i\in A_j} a_i}}.$$

See Appendix D.11 for a proof. To finish the prior analysis of SSBs in Figure 29 we illustrate the distribution of $\mathbf{K}_n^{(\mathrm{p},\theta)} = |\boldsymbol{\Pi}(\mathbf{x}_{1:n})|$ where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$ and $\boldsymbol{\mu}$ is a SSB with parameters $(\mathrm{p}, \mathsf{Be}(1,\theta), \mu_0)$. Consistently with the notation of the present section, and for the sake of a sensible comparison between models here we denote $\hat{\mathbf{K}}_n^{(\rho_\nu,\theta)} = |\boldsymbol{\Pi}(\hat{\mathbf{x}}_{1:n})|$ where $\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots \mid \hat{\boldsymbol{\mu}}\} \overset{\text{iid}}{\sim} \hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}$ is an ESB driven by a SSP with underlying tie probability ρ_ν and base measure $\hat{\nu}_0 = \mathsf{Be}(1,\theta)$. In particular if the SSP of the length variables is a Dirichlet process, Figure 24 illustrates the distribution of $\hat{\mathbf{K}}_n^{(\rho_\nu,\theta)}$ for distinct values of ρ_ν and θ. Analogously as for other models, Figure 29 illustrates the convergence in distribution stated in Corollary 4.22. When we compare Figures 24, 27 and 29 we observe that, seemingly, the way SSBs approximate Dirichlet and Geometric process is much more similar to that of DSBs than BBSBs, in spite of the fact that BBSBs and SSBs share Markovian length variables. However, if we only focus in Figures 24 and 29 we see that the convergence rates to the limit processes are different. For instance for DSBs we see that $\hat{\mathbf{K}}^{(\rho_\nu,\theta)}$ approximates much better $\hat{\mathbf{K}}^{(1,\theta)} \overset{d}{=} \mathbf{K}^{(1,\theta)}$, for $\rho_\nu = 0.8$ than $\mathbf{K}^{(\mathrm{p},\theta)}$ does for $\mathrm{p} = 0.8$. For SSBs we see that even if $\mathrm{p} = 0.95$, $\mathbf{K}^{(\mathrm{p},\theta)}$ struggles to approximate $\mathbf{K}^{(1,\theta)}$. Conversely, we have that $\mathbf{K}^{(\mathrm{p},\theta)} \approx \mathbf{K}^{(0,\theta)}$, for $\mathrm{p} = 0.4$, while DSBs we requires $\rho_\nu < 0.2$ to attain $\hat{\mathbf{K}}^{(\rho_\nu,\theta)} \approx \hat{\mathbf{K}}^{(0,\theta)}$ with sufficient precision. This suggest that the convergence rate of SSBs to Dirichlet processes is faster than that DSBs. In contrast the converge rate of SSBs to Geometric processes is very slow when compared to DSBs.

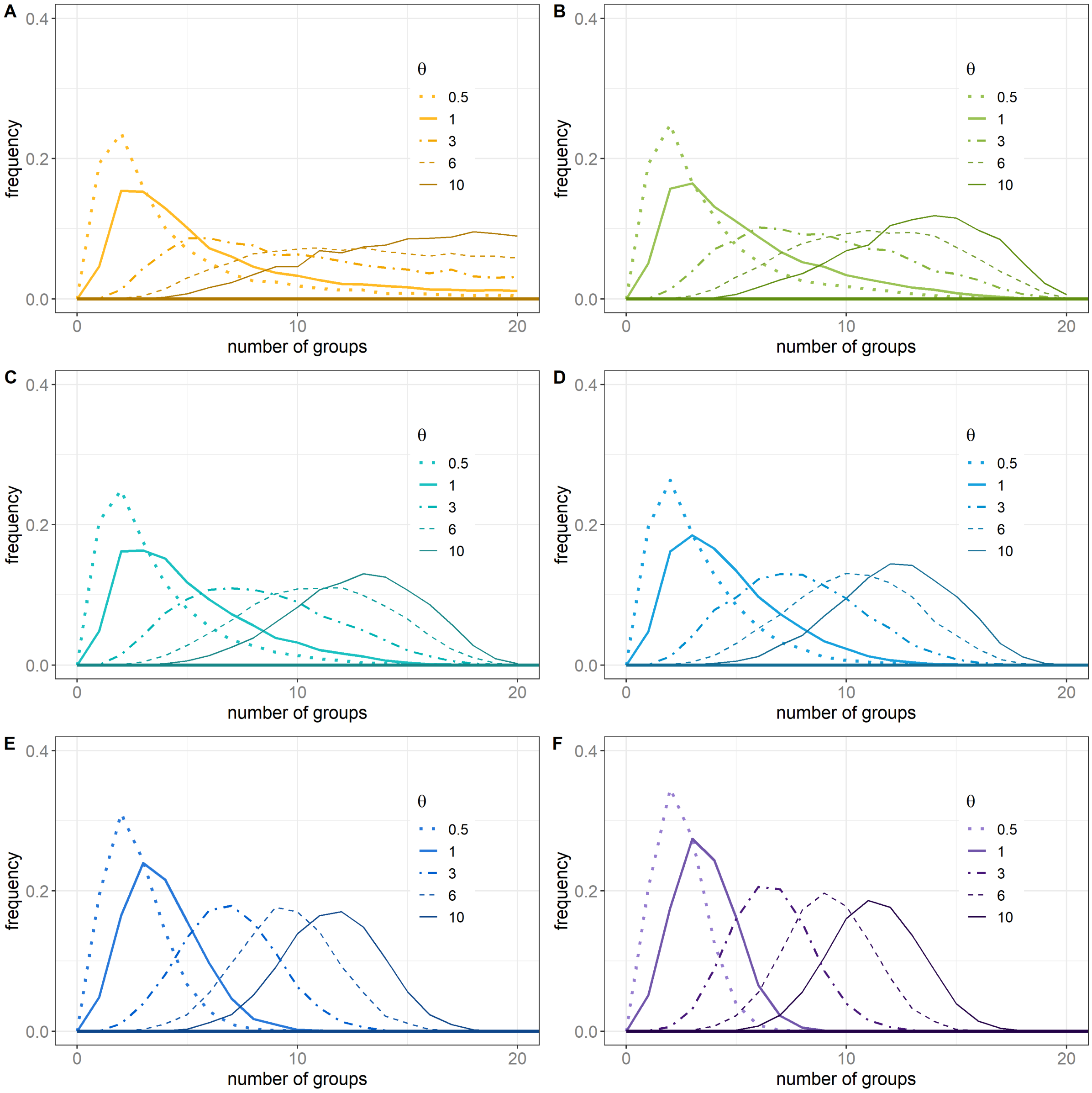


Figure 29: Frequency polygons of samples of size 10000 from the distribution of $\mathbf{K}_{20}$ corresponding to a DSB prior with $\mathrm{p} = 1$ (A), $\mathrm{p} = 0.95$ (B), $\mathrm{p} = 0.9$ (C), $\mathrm{p} = 0.8$ (D), $\mathrm{p} = 0.4$ (E) and $\mathrm{p} = 0$ (F). For each fixed value of ρ_ν we vary θ in the set $\{0.5, 1, 3, 6, 10\}$.

4.3 Stick-breaking processes with non-homogeneous length variables

After analysing stick-breaking processes with stationary length variables, either Markovian or exchangeable, one of the first question that might come to mind is whether it is possible to generalize these models to include or approximate stick-breaking processes with non identically distributed length variables, for example Pitman-Yor processes. The first thing we note is that if length variables are not equally distributed, then they are not stationary. This automatically discards exchangeable length variables, however, if we loosen the homogeneity and stationarity assumptions, we will be able to recover more general processes through Markovian length variables. As noted in past sections there are various choices of transitions that will allow us to recover Dirichlet and Geometric processes as weak limits, this is also the case of these more general models, however the analysis and the notation is much more complicated due to the lack of stationarity. To keep the study simple, here we will be focussing on generalizing SSBs processes, albeit one should keep in mind that for other transitions, such as the Beta-Binomial transition, a similar analysis can be carried on.

4.3.1 Non-homogeneous SSBs

To set up the framework fix $\mathrm{p} \in [0,1]$, and consider a collection of probability measures, $\nu = (\nu_i)_{i\geq 1}$, over $\big([0,1], \mathscr{B}_{[0,1]}\big)$, with $\nu_i(\{0\}) = \nu_i(\{1\}) = 0$, and a collection of continuous functions $\Upsilon = (\Upsilon_i : [0,1] \to [0,1])_{i\geq 1}$ such that for every $0 < v < 1$ and $i \geq 1$, $0 < \Upsilon_i(v) < 1$. Here we will be working with non-homogeneous Markovian length variables, $(\mathbf{v}_i)_{i\geq 1}$, with initial distribution $\nu_0 = \nu_1$ and one-step transition probability kernel at time i, $\boldsymbol{\nu}_i : [0,1] \to [0,1]$, given by

$$\boldsymbol{\nu}_i(\mathbf{v}_i; \cdot) = \mathbb{P}\left[\mathbf{v}_{i+1} \in \cdot \mid \mathbf{v}_i\right] = \mathrm{p}\,\delta_{\Upsilon_i(\mathbf{v}_i)} + (1-\mathrm{p})\nu_{i+1}, \tag{4.13}$$

so that $\mathbf{v}_{i+1} = \Upsilon_i(\mathbf{v}_i)$ with probability p and with probability $(1-\mathrm{p})$ $\mathbf{v}_{i+1}$ is sampled independently from ν_{i+1}. To the collection of transitions $(\boldsymbol{\nu}_i)_{i\geq 1}$ we call non-homogeneous SSB transitions with parameters $(\mathrm{p}, \nu, \Upsilon)$. When we used stationary length variables, we continuously required that the stationary distribution does not has an atom in zero, here due to Theorem 4.25 below, we also require that the distribution of $\mathbf{v}_i$ does not has an atom in one. This is done by asking that the functions Υ_i's map $(0,1)$ into $(0,1)$ and imposing the condition $\nu_i(\{1\}) = \nu_i(\{0\}) = 0$. Mainly this is done to avoid technical subtleties, but there is not much lost in generality by asking these conditions, indeed if marginally the distribution of $\mathbf{v}_i$ is diffuse, for example $\mathbf{v}_i \sim \mathsf{Be}(a_i, b_i)$, this conditions are satisfied. The restriction that the functions Υ_i's are continuous is due to Theorem 4.26.

Definition 4.7. *Let $(S, \mathscr{B}_S)$ be Borel space and consider a diffuse probability measure, μ_0, over $(S, \mathscr{B}_S)$. Fix $\mathrm{p} \in [0,1]$, and consider a collection of probability measure, $\nu = (\nu_i)_{i\geq 1}$, over $\big([0,1], \mathscr{B}_{[0,1]}\big)$, with $\nu_i(\{0\}) = \nu_0(\{1\}) = 0$, and a collection of continuous functions $\Upsilon = (\Upsilon_i : [0,1] \to [0,1])_{i\geq 1}$ such that Υ_i maps $(0,1)$ into $(0,1)$. To any MSB, $\boldsymbol{\mu}$, with base measure μ_0 and Markovian length variables, $(\mathbf{v}_i)_{i\geq 1}$, with initial distribution $\nu_0 = \nu_1$ and one-step transitions $(\boldsymbol{\nu}_i)_{i\geq 1}$, as in (4.13), we call a non-homogeneous SSB with parameters $(\mathrm{p}, \nu, \Upsilon, \mu_0)$. The corresponding stick-breaking weights*

$(\mathbf{w}_j)_{j\geq 1} = \mathsf{SB}[(\mathbf{v}_i)_{i\geq 1}]$ *will be referred to as non-homogeneous SSBw's with parameters* $(\mathrm{p}, \nu, \Upsilon)$.

In the above definition, note that if Υ_i is the identity function and $\nu_i = \nu_0$ for every $i \geq 1$ we recover a SSB with parameters $(\mathrm{p}, \nu_0, \mu_0)$ as in Definition 4.6. Also note that the choice $\mathrm{p} = 1$, yields $\mathbf{v}_{i+1} = \Upsilon_i(\mathbf{v}_i)$ almost surely, so that the non-homogeneous SSB has length variables $\left(\Upsilon^{(i)}(\mathbf{v})\right)_{i\geq 0}$ where $\Upsilon^{(0)}$ denotes the identity function and for every $i \geq 1$, $\Upsilon^{(i)} = \Upsilon_i \circ \cdots \circ \Upsilon_1$. Just as in the stick-breaking decomposition of Geometric processes, these length variables are completely determined by a single random variable, $\mathbf{v}$, and clearly, if Υ_i denotes the identity function for every $i \geq 1$, we recover a Geometric process. At the other end, where $\mathrm{p} = 0$, we get the length variables $(\mathbf{v}_i)_{i\geq 1}$ are independent with $\mathbf{v}_i \sim \nu_i$ for every $i \geq 1$, in particular if $\nu_i = \mathsf{Be}(1-\sigma, \theta + i\sigma)$, we recover a Pitman-Yor process. These non-homogeneous SSBs with either independent or completely dependent length variables will be referred to as limiting processes. In general due to the lack of stationarity of the length variables, it is more complicated to prove that non-homogeneous SSBs are proper. In fact, not all of these processes are proper, and this will be determined by the collections of functions, Υ, and probability measures, ν. However, a very nice result about non-homogeneous SSBs is that if one of the limiting processes is proper, we can assure the corresponding non-homogeneous SSB with parameter $\mathrm{p} \in (0,1)$ is proper. Formally, we have the following result.

Theorem 4.25. *Let* μ_0, ν *and* Υ *be as in Definition 4.7. For each* $\mathrm{p} \in [0,1]$, *let* $\boldsymbol{\mu}^{(\mathrm{p})}$ *be a non-homogeneous SSB with parameters* $(\mathrm{p}, \nu, \Upsilon, \mu_0)$. *If either* $\boldsymbol{\mu}^{(0)}$ *or* $\boldsymbol{\mu}^{(1)}$ *is proper, then* $\boldsymbol{\mu}^{(\mathrm{p})}$ *is proper for each* $\mathrm{p} \in (0,1)$.

The proof of Theorem 4.25 can be found in Appendix D.12. Generally it is much more easier to prove non-homogeneous SSBs with parameter $\mathrm{p} \in (0,1)$ are proper by using the limiting process that features independent length variables, $\boldsymbol{\mu}^{(0)}$, than to do this by using the limiting processes with completely dependent length variables, $\boldsymbol{\mu}^{(1)}$. Indeed, as can be seen in the proof of Theorem 4.25 to show that $\boldsymbol{\mu}^{(0)}$ is proper it suffices to prove

$$\sum_{i\geq 1} \int x\, \nu_i(dx) = \sum_{i\geq 1} \mathbb{E}\left[\mathbf{v}_i\right] = \infty$$

where $\mathbf{v}_i \sim \nu_i$. In contrast to prove that $\boldsymbol{\mu}^{(1)}$ is proper we must show $\sum_{i\geq 0} \Upsilon^{(i)}(\mathbf{v}) = \infty$ almost surely, where $\mathbf{v} \sim \nu_1$, $\Upsilon^{(0)}$ is the identity function and for every $i \geq 1$, $\Upsilon^{(i)} = \Upsilon_i \circ \cdots \circ \Upsilon_1$. So generally we will use $\boldsymbol{\mu}^{(0)}$ to prove that for every $\mathrm{p} \in (0,1)$, $\boldsymbol{\mu}^{(\mathrm{p})}$ is proper. The following result shows that the limiting processes can be recovered as weak limits of non-homogeneous SSBs with $\mathrm{p} \in (0,1)$, as long as every process involved is proper.

Theorem 4.26. *Let* μ_0, ν *and* Υ *be as in Definition 4.7, with the additional assumptions that* $\sum_{i\geq 1} \mathbb{E}\left[\mathbf{v}_i\right] = \infty$, *with* $\mathbf{v}_i \sim \nu_i$, $\sum_{i\geq 0} \Upsilon^{(i)}(\mathbf{v}) = \infty$ *almost surely, where* $\mathbf{v} \sim \nu_1$, $\Upsilon^{(0)}$ *is the identity function and for every* $i \geq 1$, $\Upsilon^{(i)} = \Upsilon_i \circ \cdots \circ \Upsilon_1$. *For each* $\mathrm{p} \in (0,1)$, *let* $\boldsymbol{\mu}^{(\mathrm{p})}$ *be a non-homogeneous SSB with parameters* $(\mathrm{p}, \nu, \Upsilon, \mu_0)$, *and consider the corresponding non-homogeneous SSBw,* $\mathbf{W}^{(\mathrm{p})}$, *with parameters* $(\mathrm{p}, \nu, \Upsilon)$. *Then*

i) *As* $\mathrm{p} \to 0$, $\boldsymbol{\mu}^{(\mathrm{p})}$ *converges weakly in distribution to a stick-breaking process,* $\boldsymbol{\mu}^{(0)}$, *with independent length variables* $(\mathbf{v}_i)_{i\geq 1}$ *where* $\mathbf{v}_i \sim \nu_i$. *In particular if* $\nu_i =$

$\mathsf{Be}(1-\sigma,\theta+i\sigma)$ *for some* $0\leq\sigma<1$ *and* $\theta>-\sigma$*, then* $\boldsymbol{\mu}^{(0)}$ *is a Pitman-Yor process and* $\mathbf{W}^{(\mathrm{p})}$ *converges in distribution to the size-biased permuted weights of* $\boldsymbol{\mu}^{(0)}$*.*

ii) *As* $\mathrm{p}\to 1$*,* $\boldsymbol{\mu}^{(\mathrm{p})}$ *converges weakly in distribution to a stick-breaking process,* $\boldsymbol{\mu}^{(1)}$*, with completely dependent length variables,* $\left(\Upsilon^{(i)}(\mathbf{v})\right)_{i\geq 0}$*. In particular, if for every* $i\geq 1$*, and* $v\in[0,1]$*,* $\Upsilon_i(v)\leq v$*, then* $\mathbf{W}^{(\mathrm{p})}$ *converges in distribution to the decreasingly ordered weights of* $\boldsymbol{\mu}^{(1)}$*.*

See Appendix D.13 for a proof. Theorem 4.26 generalizes Corollary 4.22, for the special case $\nu_0^{(n)}=\nu_0$ and $\mu_0^{(n)}=\mu_0$ for every $n\geq 1$ (in the notation of Corollary 4.22). In this sense a more general version of Theorem 4.26 still holds if for every $n\geq 1$ we consider a non-homogeneous SSB, $\boldsymbol{\mu}^{(n)}$, with parameters $\left(\mathrm{p}_n,\left(\nu_i^{(n)}\right)_{i\geq 1},\Upsilon,\mu_0^{(n)}\right)$, and let $\mathrm{p}_n\to 0$ or $\mathrm{p}_n\to 1$, $\mu_0^{(n)}\overset{w}{\to}\mu_0$ and for every $i\geq 1$, $\nu_i^{(n)}\overset{w}{\to}\nu_i$. The proof of this more general version is practically identical to that of Theorem 4.26. The main reason we stated the simpler variant is to simplify the notation and due to Theorem 4.25. Now, in the context of Theorem 4.26, as the second part states, if $\Upsilon_i(v)\leq v$ for $i\geq 1$ and $v\in[0,1]$, then as $\mathrm{p}\to 1$, $\mathbf{W}^{(\mathrm{p})}$ converges in distribution to a weights sequence that is decreasingly ordered. Under this assumption we can even show that as p grows, the weights become more likely to be decreasingly ordered. Indeed if $\mathbf{w}_j^{(\mathrm{p})}$ is the jth element of $\mathbf{W}^{(\mathrm{p})}$ then

$$\mathbb{P}\left[\mathbf{w}_{j+1}^{(\mathrm{p})}\leq\mathbf{w}_j^{(\mathrm{p})}\right]=\mathrm{p}+(1-\mathrm{p})\mathbb{E}\left[\overrightarrow{\nu_{j+1}}(c(\mathbf{v}_j))\right],$$

where $\overrightarrow{\nu_{j+1}}$ denotes the distribution function of ν_{j+1}, $\mathbf{v}_j\sim\nu_j$ and $c(v)=1\wedge v(1-v)^{-1}$ for $v\in[0,1]$. Evidently, if $\mathbb{E}\left[\overrightarrow{\nu_{j+1}}(c(\mathbf{v}_j))\right]<1$, the mapping

$$\mathrm{p}\mapsto\mathbb{P}\left[\mathbf{w}_{j+1}^{(\mathrm{p})}\leq\mathbf{w}_j^{(\mathrm{p})}\right]$$

is strictly increasing, and otherwise, $\mathbb{P}\left[\mathbf{w}_{j+1}^{(\mathrm{p})}\leq\mathbf{w}_j^{(\mathrm{p})}\right]=1$. Note that unlike the stationary case, for non-homogeneous SSBs the probability in question depends on j.

The last essential characteristic we must study for non-homogeneous SSBs is their support. Fortunately, for $\mathrm{p}\in[0,1)$ it is easy to derive sufficient conditions for the SSP to have full support. In fact, the following result is straight-forward generalization of Theorem 4.13 and Corollary 4.21, for this reason we exclude the proof.

Theorem 4.27. *For each* $\mathrm{p}\in[0,1)$ *let* $\boldsymbol{\mu}^{(\mathrm{p})}$ *be an non-homogeneous SSB with parameters* $(\mathrm{p},\nu,\Upsilon,\mu_0)$*, where* μ_0*,* ν *and* Υ *are as in Definition 4.7. If there exists* $\varepsilon>0$ *such that for every* $i\geq 1$*,* $(0,\varepsilon)$ *is contained in the support of* ν_i *then* $\boldsymbol{\mu}^{(\mathrm{p})}$ *has full support for each* $\mathrm{p}<1$*.*

The last Theorem does not gives sufficient conditions under which the limit process $\boldsymbol{\mu}^{(1)}$ has full support. In general for these processes it is not trivial to derive the result. However, if $\boldsymbol{\mu}^{(1)}$ is proper and for each $i\geq 1$, $\Upsilon_i(v)\leq v$ for every $v\in[0,1]$, then as shown in the proof of Theorem 4.26, the corresponding SSBw, $\mathbf{W}^{(1)}=\left(\mathbf{w}_j^{(1)}\right)_{j\geq 1}$, is decreasing. This yields $\mathbf{w}_j^{(1)}$, which equals the first length variable, is the largest weight. Hence, if we require $\nu_1((0,\varepsilon))>0$, for some $\varepsilon>0$, we get $\mathbb{P}\left[\mathbf{w}_j^{(1)}<\epsilon\right]>0$, for every $\epsilon>0$, and

Theorem 3.12 implies $\boldsymbol{\mu}^{(1)}$ is has full support. This small discussion is summarized by the following proposition.

Proposition 4.28. *Let $\boldsymbol{\mu}^{(1)}$ be a proper non-homogeneous SSB with parameters $(1, \nu, \Upsilon, \mu_0)$, where μ_0, ν and Υ are as in Definition 4.7. If there exists $\varepsilon > 0$, such that $\nu_1((0, \varepsilon)) > 0$ and for each $i \geq 1$, $\Upsilon_i(v) \leq v$ for all $v \in [0,1]$, then $\boldsymbol{\mu}^{(1)}$ has full support.*

4.3.2 Non-homogeneous SSBs with $\mathsf{Be}(1-\sigma, \theta + i\sigma)$ length variables

To specialize the analysis to a very interesting case, let us consider $0 \leq \sigma < 1$ and $\theta > -\sigma$, and define the collection of functions $\Upsilon = (\Upsilon_i)_{i\geq 1}$ given by

$$\Upsilon_i(x) = \mathcal{I}^{-1}_{\mathcal{I}_x(1-\sigma,\theta+i\sigma)}(1-\sigma, \theta + (i+1)\sigma), \tag{4.14}$$

for every $i \geq 1$, where $\mathcal{I}_x(\alpha, \beta)$ denotes the regularized Beta function

$$\mathcal{I}_x(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x v^{\alpha-1}(1-v)^{\beta-1} dv$$

and $\mathcal{I}^{-1}_y(\alpha, \beta)$ denotes the inverse regularized Beta function, so that $\mathcal{I}_x(\alpha, \beta) = y$ if and only if $\mathcal{I}^{-1}_y(\alpha, \beta) = x$. Theorem 4.29 below presents a list of important properties of Υ_i, meanwhile, Figure 30 illustrates Υ_i for distinct values of θ and σ.

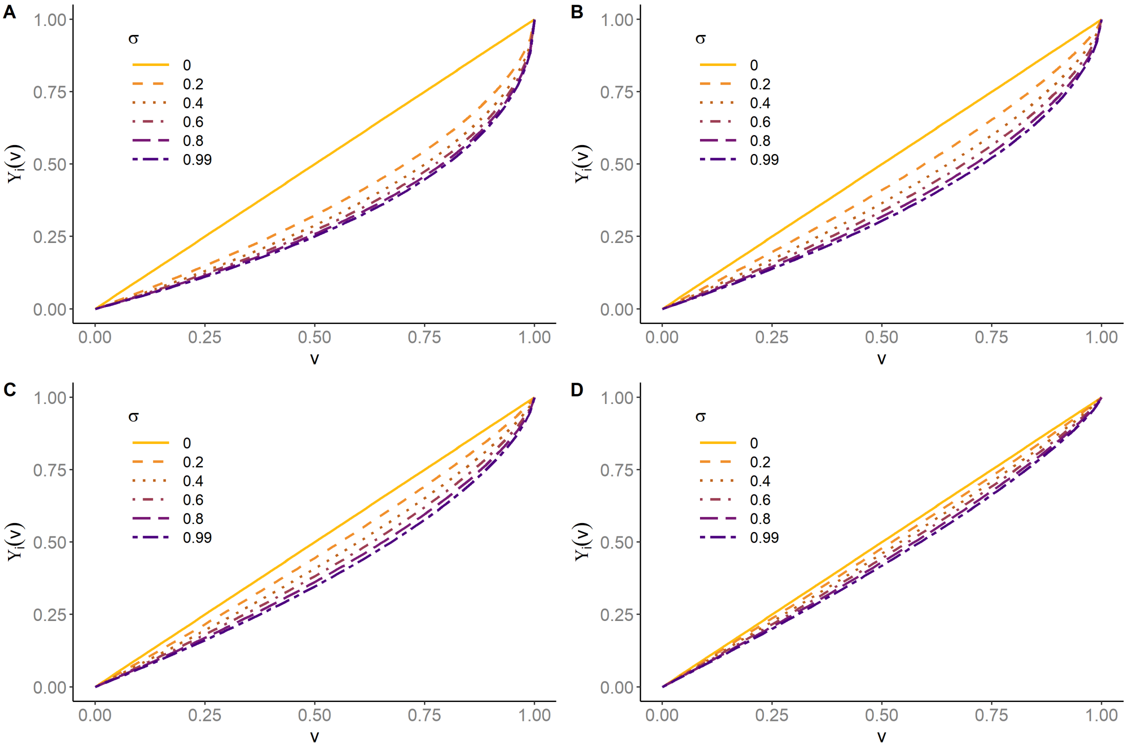


Figure 30: Graph of $\mathcal{I}^{-1}_{\mathcal{I}_x(1-\sigma,\theta)}(1-\sigma, \theta+\sigma)$ for $\sigma \in \{0.2, 0.4, 0.6, 0.8, 0.99\}$ and $\theta \in \{0.1, 0.5, 1, 3\}$ (A–D) respectively.

Theorem 4.29. *Let $\Upsilon = (\Upsilon_i)_{i\geq 1}$ be as in equatioin (4.14). Then*

a) Υ_i *is continuous, increasing and maps* $(0,1)$ *into* $(0,1)$*.*

b) *The inverse function of* Υ_i *is* $\Upsilon_i^{-1}(y) = \mathcal{I}^{-1}_{\mathcal{I}_y(1-\sigma,\theta+(i+1)\sigma)}(1-\sigma,\theta+i\sigma)$*.*

c) *For every* $n \geq 1$*,* $\Upsilon^{(n)}(x) = (\Upsilon_1 \circ \cdots \circ \Upsilon_n)(x) = \mathcal{I}^{-1}_{\mathcal{I}_x(1-\sigma,\theta+\sigma)}(1-\sigma,\theta+(n+1)\sigma)$*.*

d) *If* $\mathbf{v}_i \sim \mathsf{Be}(1-\sigma,\theta+i\sigma)$ *then* $\Upsilon_i(\mathbf{v}_i) \sim \mathsf{Be}(1-\sigma,\theta+(i+1)\sigma)$*. This yields that if* $\mathbf{v}_1 \sim \mathsf{Be}(1-\sigma,\theta+\sigma)$*, then* $\mathbf{v}_n = \Upsilon^{(n)}(\mathbf{v}_1) \sim \mathsf{Be}(1-\sigma,\theta+(n+1)\sigma)$*.*

e) $\Upsilon_i(v) \leq v$*, for* $i \geq 1$ *and* $v \in [0,1]$*.*

f) $\sum_{n\geq 0} \Upsilon^{(n)}(v) = \infty$*, for every* $v \in (0,1)$*, where* $\Upsilon^{(0)}$ *is the identity function, and* $\Upsilon^{(n)}$ *is as in* (c)*, for* $n \geq 1$*,.*

The proof of Theorem 4.29 can be found in Appendix D.14. The main motivation behind studying these functions is, of course, to use them to define the transition of the length variables, $\left(\mathbf{v}_i^{(\mathrm{p})}\right)_{i\geq 1}$, of a non-homogeneous SSBw, $\left(\mathbf{w}_j^{(\mathrm{p})}\right)_{j\geq 1}$, or the non-homogeneous SSB, $\boldsymbol{\mu}^{(\mathrm{p})}$. So for $\mathrm{p} \in [0,1]$ consider a non-homogeneous Markov chain $\left(\mathbf{v}_i^{(\mathrm{p})}\right)_{i\geq 1}$ with initial distribution $\nu_0 = \nu_1 = \mathsf{Be}(1-\sigma,\theta+\sigma)$ and one-step transition probability kernel at time i,

$$\boldsymbol{\nu}_i\left(\mathbf{v}_i^{(\mathrm{p})};\cdot\right) = \mathbb{P}\left[\mathbf{v}_{i+1}^{(\mathrm{p})} \in \cdot \,\middle|\, \mathbf{v}_i^{(\mathrm{p})}\right] = \mathrm{p}\,\delta_{\Upsilon_i\left(\mathbf{v}_i^{(\mathrm{p})}\right)} + (1-\mathrm{p})\nu_{i+1},$$

where Υ_i is as in (4.14) and $\nu_{i+1} = \mathsf{Be}(1-\sigma,\theta+(i+1)\sigma)$ for every $i \geq 1$. As a consequence of Theorem 4.29 (d), we know that $\mathbf{v}_i^{(\mathrm{p})} \sim \mathsf{Be}(1-\sigma,\theta+i\sigma)$, marginally, despite the value of p. At the extreme case $\mathrm{p} = 0$ we find the elements of $\left(\mathbf{v}_i^{(0)}\right)_{i\geq 1}$ are independent, the corresponding stick-breaking weights, $\left(\mathbf{w}_j^{(0)}\right)_{j\geq 1}$, are invariant under size-biased permutations and the non-homogeneous SSB, $\boldsymbol{\mu}^{(0)}$, is a Pitman-Yor processes. At the other end, where $\mathrm{p} = 1$, we discover a sequence of length variables, $\left(\mathbf{v}_i^{(1)}\right)_{i\geq 1} = \left(\Upsilon^{(n)}(\mathbf{v})\right)_{n\geq 0}$, that are completely determined by a single random variable $\mathbf{v}_1^{(1)} = \mathbf{v} \sim \mathsf{Be}(1-\sigma,\theta+\sigma)$. In this case, as a consequence of the proof of Theorem 4.26 (ii) and Theorem 4.29 (e), we know the corresponding stick-breaking weights $\left(\mathbf{w}_j^{(1)}\right)_{j\geq 1}$ are decreasingly ordered and the non-homogeneous SSB, $\boldsymbol{\mu}^{(1)}$, generalizes a Geometric process $\hat{\boldsymbol{\mu}}^{(1)} \sim \mathcal{G}_{(\mathsf{Be}(1,\theta),\mu_0)}$ in an analogous way that the Pitman-Yor process, $\boldsymbol{\mu}^{(0)}$, generalizes a Dirichlet process, $\hat{\boldsymbol{\mu}}^{(0)} \sim \mathcal{D}_{(\theta,\mu_0)}$. In fact, the choice $\sigma = 0$ yields that $\left(\mathbf{v}_i^{(\mathrm{p})}\right)_{i\geq 1}$ is an stationary Markov chain with $\mathsf{Be}(1,\theta)$ marginals, for every $\mathrm{p} \in [0,1]$, so that $\boldsymbol{\mu}^{(1)} \sim \mathcal{G}_{(\mathsf{Be}(1,\theta),\mu_0)}$, $\boldsymbol{\mu}^{(0)} \sim \mathcal{D}_{(\theta,\mu_0)}$, and for $\mathrm{p} \in (0,1)$, $\boldsymbol{\mu}^{(\mathrm{p})}$ is a SSB as introduced in Definition 4.6. So we have fulfilled the objective of generalizing (stationary) SSBs to include interesting processes with length variables that are not identically distributed, such as Pitman-Yor processes. Now, for these non-homogeneous SSBs, using the results we have developed thus far, it is straight-forward to prove essential properties. For instance since the Pitman-Yor process $\boldsymbol{\mu}^{(0)}$ is proper, using Theorem 4.25, we find that for every $\mathrm{p} \in (0,1)$, the non-homogeneous SSB, $\boldsymbol{\mu}^{(\mathrm{p})}$ is also proper. Further, from the proof Theorem 4.25 and Theorem 4.29 (f) we also know the generalized Geometric process, $\boldsymbol{\mu}^{(1)}$, is proper.

Hence, Theorem 4.26 proves that as $\mathrm{p} \to 0$, $\boldsymbol{\mu}^{(\mathrm{p})}$ converges weakly in distribution to $\boldsymbol{\mu}^{(0)}$, and as $\mathrm{p} \to 1$, $\boldsymbol{\mu}^{(\mathrm{p})}$ converges weakly in distribution to $\boldsymbol{\mu}^{(1)}$, as long as all these SSPs share the same base measure. Finally, from Theorem 4.27 and Proposition 4.28 we know that for each $\mathrm{p} \in [0,1]$ and each $0 \leq \sigma < 1$, $\theta > -\sigma$, $\boldsymbol{\mu}^{(\mathrm{p})}$ has full support, meaning that this special king of non-homogeneous SSBs lead to feasible Bayesian-non parametric priors.

4.3.3 Final remarks on stick-breaking processes with non-homogeneous length variables

We will not be studying MSBs with non-stationary length variables in further detail, but before we move on, there are a couple of quick remarks to make:

i) In addition to $\Upsilon = (\Upsilon_i)_{i\geq 1}$ as in (4.14), there are other interesting choices of Υ. For example, if we consider the positive numbers $(\alpha_i, \beta_i)_{i\geq 1}$ and set

$$\Upsilon_i = \mathcal{I}^{-1}_{\mathcal{I}_x(\alpha_i,\beta_i)}(\alpha_{i+1}, \theta_{i+1}), \tag{4.15}$$

and $\nu_i = \mathsf{Be}(\alpha_i, \beta_i)$ we can construct non-homogeneous Markov chains of length variables with the given Beta marginals. Of course, if we set $\alpha_i = 1 - \sigma$ and $\beta_i = \theta + i\sigma$, we recover the models in Section 4.3.2. For Υ_i as in (4.15), analogous properties to that of Theorem 4.29 (a)–(d) hold, but depending on $(\alpha_i, \beta_i)_{i\geq 1}$, properties such as the ones in (e)–(f), might not be true. If we are not interested in length variables with Beta marginals, a wide variety of interesting choices for Υ might come to mind, the one thing we require is that $\Upsilon_i : [0,1] \to [0,1]$ is continuous and maps $(0,1)$ into $(0,1)$.

ii) If instead of considering the transitions of length variables, $\boldsymbol{\nu}_i$'s, as in (4.13) we set

$$\boldsymbol{\nu}_i(p; dv) = \sum_{z=0}^{\kappa} \mathsf{Be}(dv \mid 1 - \sigma + z, \theta + (i+1)\sigma + \kappa - z)\mathsf{Bin}(z \mid \kappa, \Upsilon_i(p)),$$

where Υ_i is as in (4.14), then we can generalize BBSBs to contain Pitman-Yor processes, in the same way we generalized SSBs, and a similar analysis to that of Sections 4.3.1 and 4.3.2 can be carried on.

5 Gibbs sampling methods

The main motivation behind Bayesian non-parametric statistics is to avoid restrictive parametric assumptions about the distribution that generates the data. This is done by constructing random probability measures with large support, see Datta (1991); Ghosal et al. (1999); Wu and Ghosal (2008); Bissiri and Ongaro (2014) for an explanation of why the large support is an essential requirement for Bayesian non-parametric priors. In general, constructing an arbitrary random probability measure with a wide support can be extremely complicated, which is why SSPs are the building blocks for the vast majority of Bayesian non-parametric models. Indeed, SSPs are sufficiently tractable so that many theoretical properties can be derived, this allows a deep comprehension of the model. Furthermore, as analysed in Section 3.4, it is possible to construct SSPs with (weak topological) supports as large as possible. Last but not least, models based on SSPs are feasible to implement with the aid of Markov chain Monte Carlo (MCMC) methods. To illustrate the usefulness of Bayesian non-parametric methods based on SSPs, here we will use them to estimate density of data and/or cluster data points that present no repetitions.

The first thing we do in Section 5.1 is to explain the canonical methodology of how to transform proper SSPs into a random mixtures of diffuse kernels. These diffuse random probability measures will in turn allow modelling continuous density functions. In here we find the main reason of why we have been particularly interested in proving that certain SSPs are proper. In fact, for density estimation purposes through mixture modelling, the essential requirement we need to impose on the SSP are that it is proper and has full support. While weights summing up to one will allow a simple construction of a diffuse random probability measure, the full support warranties sufficient flexibility of the Bayesian non-parametric model. Now, depending on the available construction or representation of the SSP, there exist distinct MCMC methods to implement the models, some of them we review in Section 5.2. For instance, if the SSP is a Dirichlet process, we can exploit the prediction rule to derive a Gibbs sampler algorithm, as developed by Escobar (1988, 1994); MacEachern (1994) and Escobar and West (1995), this algorithm can be adapted for any species sampling prior with EPPF obtainable in closed form. If this is not the case we can exploit the construction through random sets of SSPs (see Section 3.5.4) to derive a Gibbs sampling method. For most stick-breaking processes the EPPF nor the latent random sets are available, for this type of priors, Walker (2007) derive a general slicer sampler algorithm, latter modified by Kalli et al. (2011). Here we are particular interested in adjusting Bayesian non-parametric models based on the new priors introduced in Section 4, hence we will be using Walker's slice sampler or its modification as driver for the implementation of our models. Finally in Section 5.3 we will design small experiments to test the performance of priors based on stick-breaking processes with dependent length variable and contrast the results with the ones obtained with Dirichlet and Geometric priors.

5.1 Random mixtures

In order to estimate the density of a dataset, the first thing we need to show is how to transform a proper SSP, which is a purely atomic random probability measure into a diffuse random probability measure. So consider a couple of Borel spaces, $(S, \mathscr{B}_S)$ and

$(T, \mathscr{B}_T)$, a proper SSP, $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$, over $(S, \mathscr{B}_S)$, and a diffuse probability kernel $\mathcal{K} : S \to T$ (so that $\mathcal{K} : S \times \mathscr{B}_T \to [0,1]$, for each $s \in A$, $\mathcal{K}(s, \cdot)$ is a diffuse probability measure over $(T, \mathscr{B}_T)$ and for every $B \in \mathscr{B}_T$, $\mathcal{K}(\cdot, B)$ is a measurable function). Usually, we consider $\{\mathcal{K}(s, \cdot) : s \in S\}$ to be a parametric family and S its parameter space. For example if $T = \mathbb{R}$ and $S = \mathbb{R} \times \mathbb{R}_+$ we might choose $\mathcal{K}$ to denote a Gaussian kernel so that $\mathcal{K}((s_1, s_2), \cdot)$ is a Normal distribution with mean $s_1 \in \mathbb{R}$ and variance $s_2 \in \mathbb{R}_+$. For this reason, throughout this section we will use the notation $\mathcal{K}(B \mid s) = \mathcal{K}(s, B)$ for every diffuse probability kernel $\mathcal{K} : S \to T$. With these objects in mind, we can define the measurable transformation of $\boldsymbol{\mu}$

$$\boldsymbol{\mu} \mapsto \int \mathcal{K}(\cdot \mid s)\boldsymbol{\mu}(ds) = \sum_{j\geq 1} \mathbf{w}_j \mathcal{K}(\cdot \mid \boldsymbol{\xi}_j). \tag{5.1}$$

First of all note that $\boldsymbol{\phi} = \sum_{j\geq 1} \mathbf{w}_j \mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$ is a diffuse random probability measure over $(T, \mathscr{B}_T)$ as

$$\boldsymbol{\phi}(T) = \sum_{j\geq 1} \mathbf{w}_j \mathcal{K}(T \mid \boldsymbol{\xi}_j) = \sum_{j\geq 1} \mathbf{w}_j = 1,$$

and for each $t \in T$,

$$\boldsymbol{\phi}(\{t\}) = \sum_{j\geq 1} \mathbf{w}_j \mathcal{K}(\{t\} \mid \boldsymbol{\xi}_j) = 0.$$

The measurability of $\boldsymbol{\phi}$ is direct from (5.1).

Remark 5.1. *In particular if S and T, endowed with suitable metrics, are Polish, and for each $s_n \to s$ in S we have that $\mathcal{K}(\cdot \mid s_n)$ converges weakly to $\mathcal{K}(\cdot \mid s)$ in $\mathcal{P}(T)$, from Lemma 1.18 we get that the mapping (5.1) is even continuous with respect to the weak topology. This assures that whenever the SSPs $\boldsymbol{\mu}^{(n)}$ converge weakly in distribution (or almost surely) to $\boldsymbol{\mu}$, then $\boldsymbol{\phi}^{(n)} = \int \mathcal{K}(\cdot \mid s)\boldsymbol{\mu}^{(n)}(ds)$ also converges weakly in distribution (or almost surely) to $\boldsymbol{\phi} = \int \mathcal{K}(\cdot \mid s)\boldsymbol{\mu}(ds)$. In other words, versions of Theorems 4.3 and 4.14 and their corollaries also hold for these diffuse transformations of species sampling processes. Given that we will be interested in implementing models based on stick-breaking processes with exchangeable and Markovian length variables, and compare the results with the limiting processes the Dirichlet and Geometric priors, we will continue under the assumptions that this extra continuity condition is imposed on the kernel $\mathcal{K}$.*

Formally we have the following definition.

Definition 5.1 (Random mixtures/diffuse kernels)**.** *Consider a couple of Borel spaces $(S, \mathscr{B}_S)$ and $(T, \mathscr{B}_T)$, a diffuse probability kernel $\mathcal{K} : S \to T$, a sequence of non-negative random variables $(\mathbf{w}_j)_{j\geq 1}$ with $\sum_{j\geq 1} \mathbf{w}_j = 1$ almost surely, and an independent sequence $(\boldsymbol{\xi}_j)_{j\geq 1}$ of i.i.d. random objects that take values in S. To the diffuse random probability measure $\boldsymbol{\phi} = \sum_{j\geq 1} \mathbf{w}_j \mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$ we call a random mixture.*

Note that just as a SSP, $\boldsymbol{\mu}$, the law of a random mixture $\boldsymbol{\phi}$ is completely characterized by the distribution of the atoms, $(\boldsymbol{\xi}_j)_{j\geq 1}$, and that of the weights, $(\mathbf{w}_j)_{j\geq 1}$. So it is natural to wonder, why is it important to construct random mixtures through measurably transforming a SSP? Well, the main reason is that random mixtures are harder to analyse than SSPs, perhaps due to the lack of exchangeable increments. Regarding random mixtures as a measurable transformation of SSPs allows us to capture important structural properties through the latent discrete structures. For example, it is clear from (5.1) that the law of a random mixture is completely determined by that of a SSP, and some methods to define the law of a SSP consist on integrating over the weights sequence, such as the construction of SSPs through EPPF. Another important issue is that analysing the support of a SSPs is much easier, and under mild conditions on the kernel, $\mathcal{K}$, we can assure that if the SSP has full support, then the Bayesian non-parametric model, that utilizes $\boldsymbol{\phi}$ as the directing random measure of exchangeable observations, will be flexible enough. In what follows we will present a quick overview of i.i.d. samples from a random mixture, here the advantages of taking into consideration the underlying SSP become more obvious.

5.1.1 Samples from random mixtures

In Bayesian non-parametric statistics it is common to model the law of exchangeable sequences $(\mathbf{y}_i)_{i\geq 1}$ such that $\mathbb{P}[\mathbf{y}_i = \mathbf{y}_k] = 0$, for each pair of indexes $i \neq k$, by means of a random mixture, that is to consider $\{\mathbf{y}_1, \mathbf{y}_2, \ldots \mid \boldsymbol{\phi}\} \overset{\text{iid}}{\sim} \boldsymbol{\phi} = \sum_{j\geq 1} \mathbf{w}_j \mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$. Indeed, the diffuseness of $\boldsymbol{\phi}$ yields $\mathbb{P}[\mathbf{y}_i = \mathbf{y}_k] = 0$, for $i \neq k$, this is why random mixtures are sensible tools to model the directing random measure of exchangeable data that present no repetitions, whereas SSPs are not due to their discrete nature. However, even in this setting the study of SSP is crucial. Namely, in terms of the law of $(\mathbf{y}_i)_{i\geq 1}$ the following sampling schemes are equivalent:

I. $\{\mathbf{y}_1, \mathbf{y}_2, \ldots \mid \boldsymbol{\phi}\} \overset{\text{iid}}{\sim} \boldsymbol{\phi}$.

II. $\{\mathbf{x}_1, \mathbf{x}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$, and independently for $i \geq 1$, $\{\mathbf{y}_i \mid \mathbf{x}_i\} \sim \mathcal{K}(\cdot \mid \mathbf{x}_i)$.

In effect, under the second scheme, by the tower property of conditional expectation, since $\boldsymbol{\mu}$ is $(\mathbf{x}_i)_{i\geq 1}$-measurable, and from (5.1), we get that for any measurable sets $B_1, \ldots, B_n$,

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{i=1}^{n} (\mathbf{y}_i \in B_i) \,\middle|\, \boldsymbol{\mu}\right] &= \mathbb{E}\left[\mathbb{P}\left[\bigcap_{i=1}^{n} (\mathbf{y}_i \in B_i) \,\middle|\, (\mathbf{x}_i)_{i\geq 1}\right] \,\middle|\, \boldsymbol{\mu}\right] \\
&= \prod_{i=1}^{n} \mathbb{E}\left[\mathcal{K}(B_i \mid \mathbf{x}_i) \mid \boldsymbol{\mu}\right] \\
&= \prod_{i=1}^{n} \int \mathcal{K}(B_i \mid s) \boldsymbol{\mu}(ds) \\
&= \prod_{i=1}^{n} \boldsymbol{\phi}(B_i),
\end{aligned}$$

as $\boldsymbol{\phi}$ is $\boldsymbol{\mu}$-measurable, this implies $\{\mathbf{y}_1, \mathbf{y}_2, \ldots \mid \boldsymbol{\phi}\} \overset{\text{iid}}{\sim} \boldsymbol{\phi}$. An important gain from the second sampling scheme is that we can compute the joint distribution of $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$. Using the tower property of conditional expectation once more, we attain that for any measurable sets, $B_1, \ldots, B_n$,

$$\mathbb{P}\left[\bigcap_{i=1}^{n} (\mathbf{y}_i \in B_i)\right] = \mathbb{E}\left[\prod_{i=1}^{n} \mathcal{K}(B_i \mid \mathbf{x}_i)\right].$$

Being that $f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \mathcal{K}(B_i \mid \mathbf{x}_i)$ is a measurable function of a sample from a SSP, Theorem 3.7 provides an expression for $\mathbb{P}[\mathbf{y}_1 \in B_1, \ldots, \mathbf{y}_n \in B_n]$. Note that sampling $\{\mathbf{y}_1, \mathbf{y}_2, \ldots \mid \boldsymbol{\phi}\} \overset{\text{iid}}{\sim} \boldsymbol{\phi}$ simply means that with probability $\mathbf{w}_j$, $\mathbf{y}_i$ is sampled from $\mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$. This said, another major advantage of the second sampling scheme over the first one is that $\mathbf{x}_i$ records the information about which component of the mixture, $\{\mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)\}_{j\geq 1}$, was $\mathbf{y}_i$ sampled from. More precisely, since the elements in $(\boldsymbol{\xi}_j)_{j\geq 1}$ are distinct almost surely, whenever $\mathbf{x}_i = \boldsymbol{\xi}_j$ we know that with probability one $\mathbf{y}_i$ was sampled from $\mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$. This then allow us to cluster elements in $(\mathbf{y}_i)_{i\geq 1}$ according to the ties exhibited in $(\mathbf{x}_i)_{i\geq 1}$, in other words, according to the component of the mixture from which they were ultimately sampled. In this setting the number of distinct values $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ exhibits, denoted by $\mathbf{K}_n$ in previous sections, can be interpreted as the number of significant components in $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, that is, from how many elements in $\{\mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)\}_{j\geq 1}$ do we have at least one sample point. Notice that a priori this means we are modelling the clusters of the data points $\mathbf{y}_1, \ldots, \mathbf{y}_n$ through the EPPF corresponding to $\boldsymbol{\mu}$. In some cases, specially if the EPPF of the model is not available, it is convenient to restate sampling scheme II, by introducing the so called latent allocation variables $(\mathbf{d}_i)_{i\geq 1}$, defined by $\mathbf{d}_i = j$ if and only if $\mathbf{x}_i = \boldsymbol{\xi}_j$. Note that since $\boldsymbol{\xi}_j \neq \boldsymbol{\xi}_k$ almost surely for every $j \neq k$, the latent allocation variables are almost surely well defined, and we can rewrite the second sampling scheme as:

III. Given $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$, $\{\mathbf{d}_1, \mathbf{d}_2, \ldots \mid \mathbf{W}\} \overset{\text{iid}}{\sim} \sum_{j\geq 1} \mathbf{w}_j \delta_j$, given $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j\geq 1}$ and $(\mathbf{d}_i)_{i\geq 1}$, $\{\mathbf{x}_i \mid \mathbf{d}_i, \boldsymbol{\Xi}\} \sim \delta_{\boldsymbol{\xi}_{\mathbf{d}_i}}$, and $\{\mathbf{y}_i \mid \mathbf{x}_i\} \sim \mathcal{K}(\cdot \mid \mathbf{x}_i)$, independently for $i \geq 1$.

or equivalently

III. $\{\mathbf{d}_1, \mathbf{d}_2, \ldots \mid \mathbf{W}\} \overset{\text{iid}}{\sim} \sum_{j\geq 1} \mathbf{w}_j \delta_j$, and $\{\mathbf{y}_i \mid \mathbf{d}_i, \boldsymbol{\Xi}\} \sim \mathcal{K}(\cdot \mid \boldsymbol{\xi}_{\mathbf{d}_i})$, independently for $i \geq 1$.

Sampling schemes II and III contain exactly the same amount of information, the difference is that in one of them the clustering of the observations is recorded by $(\mathbf{x}_i)_{i\geq 1}$ whilst in the other it is recorded through $(\mathbf{d}_i)_{i\geq 1}$. Indeed, if we consider the random partition $\boldsymbol{\Pi}(\mathbf{d}_{1:n})$ of $\{1, \ldots, n\}$, generated by the random equivalence relation $i \boldsymbol{\sim} k$ if and only if $\mathbf{d}_i = \mathbf{d}_k$, and analogously for $(\mathbf{x}_i)_{i\geq 1}$, then $\boldsymbol{\Pi}(\mathbf{d}_{1:n})$ is equal almost surely to $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$, because outside a $\mathbb{P}$-null set $\mathbf{d}_i = \mathbf{d}_k$ if an only if $\mathbf{x}_i = \mathbf{x}_k$. This clearly means that both, $\boldsymbol{\Pi}(\mathbf{d}_{1:n})$ and $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$ are exchangeable, and share same the EPPF. The key difference between modelling the clusters through $(\mathbf{d}_i)_{i\geq 1}$ and $(\mathbf{x}_i)_{i\geq 1}$ is that in order to compute the finite dimensional distributions of $(\mathbf{x}_i)_{i\geq 1}$ we require the EPPF (see Theorem 3.6 V). As to $(\mathbf{d}_i)_{i\geq 1}$, it is easy to compute for any $d_1, \ldots, d_n \in \mathbb{N}$

$$\mathbb{P}[\mathbf{d}_1 = d_1, \ldots, \mathbf{d}_n = d_n \mid \mathbf{W}] = \prod_{j=1}^{k} \mathbf{w}_j^{m_j},$$

where $m_j = |\{i \leq n : d_i = j\}| = \sum_{i=1}^{n} \mathbf{1}_{\{d_i = j\}}$ and $k = \max\{d_1, \ldots, d_n\}$, and taking expectations in the last equation we obtain

$$\mathbb{P}[\mathbf{d}_1 = d_1, \ldots, \mathbf{d}_n = d_n] = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{w}_j^{m_j}\right]. \tag{5.2}$$

Moreover, if the stick-breaking decomposition of the weights $\mathbf{W} = \mathsf{SB}(\mathbf{V})$ where $\mathbf{V} = (\mathbf{v}_i)_{i\geq 1}$, is available, we can rewrite

$$\mathbb{P}[\mathbf{d}_1 = d_1, \dots, \mathbf{d}_n = d_n] = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{v}_j^{m_j}(1-\mathbf{v}_j)^{\sum_{i>j} m_i}\right]. \tag{5.3}$$

Thus, to compute the joint distribution of $(\mathbf{d}_1, \dots, \mathbf{d}_n)$ we do not need to have the EPPF in explicit form. For example, for the models introduced in Section 4 most of the times we will be able to attain an expression for the expectations of power products of length variables, hence the joint distribution of $(\mathbf{d}_1, \dots, \mathbf{d}_n)$, in contrast to the EPPF which is much harder to compute. In fact, if we compare equation (5.3) with (3.8) we see that the pEPPF corresponding to $\mathbf{W}$ is similar to the joint distribution of $(\mathbf{d}_1, \dots, \mathbf{d}_n)$ shifted, while the pEPPF is evaluated in $n_1, \dots, n_k \geq 1$, the joint distribution of the latent allocation variables is written in terms of $m_1, \dots, m_k \geq 0$. The reason behind the fact that the EPPF is not required to compute the finite dimensional distributions of $(\mathbf{d}_i)_{i\geq 1}$, is that, contrary to the directing random measure, $\boldsymbol{\mu}$, of $(\mathbf{x}_i)_{i\geq 1}$, the law of $\sum_{j\geq 1} \mathbf{w}_j\delta_j$, which is the directing random measure of $(\mathbf{d}_i)_{i\geq 1}$, is not invariant under permutations of the weights. So although for the overall model the ordering of the weights is irrelevant, if we focus on the latent allocation variables, it does have an impact which can be used to our advantage if necessary.

In general when the EPPF is available and manageable it might be more convenient to assume sampling scheme II, otherwise, most of the times it is necessary to consider the latent allocation variables and even expand sampling scheme III in order to implement the models. This small discussion will become clearer in the subsequent section.

5.2 Gibbs sampling methods

Throughout this section, for the sake of simplicity we will be working with the Bayesian notation. So for a couple of random elements, $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$, that take values in Borel spaces, we will denote by $\mathbb{p}(\boldsymbol{\eta})$ to the marginal density or mass probability function of $\boldsymbol{\eta}$ and by $\mathbb{p}(\boldsymbol{\eta} \mid \boldsymbol{\zeta})$ to the conditional density or mass probability function of $\boldsymbol{\eta}$ given $\boldsymbol{\zeta}$. For example

$$\mathbb{p}(\boldsymbol{\eta}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha),\Gamma(\beta)}\boldsymbol{\eta}^{\alpha-1}(1-\boldsymbol{\eta})^{\beta-1}\mathbf{1}_{[0,1]}(\boldsymbol{\eta}), \quad \text{or} \quad \mathbb{p}(\boldsymbol{\eta}) \propto \boldsymbol{\eta}^{\alpha-1}(1-\boldsymbol{\eta})^{\beta-1}\mathbf{1}_{[0,1]}(\boldsymbol{\eta}),$$

means that $\boldsymbol{\eta} \sim \mathsf{Be}(\alpha, \beta)$, and

$$\mathbb{p}(\boldsymbol{\zeta} \mid \boldsymbol{\eta}) = \binom{n}{\boldsymbol{\zeta}}\boldsymbol{\eta}^{\boldsymbol{\zeta}}(1-\boldsymbol{\eta})^{n-\boldsymbol{\zeta}}\,\mathbf{1}_{\{0,\dots,n\}}(\boldsymbol{\zeta})$$

refers to $\{\boldsymbol{\zeta} \mid \boldsymbol{\eta}\} \sim \mathsf{Bin}(n, \boldsymbol{\zeta})$. We will also be excluding the indicator functions that express the support of the distribution when they are obvious.

Before we discuss how to implement some Bayesian non-parametric models let us present a quick introduction into Bayesian modelling, so that we can appreciate the usefulness and in some cases necessity of MCMC methods. As previously mentioned, in Bayesian statistics one usually models presumably exchangeable data $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ as i.i.d. sampled from a determined distribution given a random element $\boldsymbol{\eta}$, that is

$$\mathbb{p}(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} \mathbb{p}(\mathbf{y}_i \mid \boldsymbol{\eta}).$$

To fully specify the Bayesian model one has to pick the prior (marginal) distribution of $\boldsymbol{\eta}$, $\mathbb{p}(\boldsymbol{\eta})$, so that we can try to compute the posterior distribution

$$\mathbb{p}(\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) \propto \mathbb{p}(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\eta})\mathbb{p}(\boldsymbol{\eta}).$$

Whenever feasible we can then estimate quantities of interest, which are commonly measurable functions of $\boldsymbol{\eta}$, say $f(\boldsymbol{\eta})$, through the expected a posteriori (EAP)

$$\mathbb{E}\left[f(\boldsymbol{\eta}) \mid \mathbf{y}_1, \ldots, \mathbf{y}_n\right] = \int f(\boldsymbol{\eta})\mathbb{p}(d\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n),$$

or through the maximum a posteriori (MAP) $f(\hat{\boldsymbol{\eta}})$, where

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \mathbb{p}(\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n).$$

Example 5.1. *Say we encounter $\{0,1\}$-valued data $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, and decide to model them as i.i.d. sampled from $\{\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\eta}\} \overset{iid}{\sim} \mathsf{Ber}(\boldsymbol{\eta})$, where $\boldsymbol{\eta} \sim \mathsf{Be}(\alpha, \beta)$, that is*

$$\mathbb{p}(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} \mathbb{p}(\mathbf{y}_i \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} \mathbb{p}(\mathbf{y}_i \mid \boldsymbol{\eta}) = \prod_{i=1}^{n} \boldsymbol{\eta}^{\mathbf{y}_i}(1-\boldsymbol{\eta})^{1-\mathbf{y}_i}.$$

and $\mathbb{p}(\boldsymbol{\eta}) = \mathsf{Be}(\boldsymbol{\eta} \mid \alpha, \beta)$. Then we get

$$\mathbb{p}(\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) \propto \boldsymbol{\eta}^{\alpha + \sum_{i=1}^{n} \mathbf{y}_i - 1}(1-\boldsymbol{\eta})^{\beta + n - \sum_{i=1}^{n} \mathbf{y}_i - 1}$$

which means $\{\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n\} \sim \mathsf{Be}\left(\alpha + \sum_{i=1}^{n} \mathbf{y}_i, \beta + n - \sum_{i=1}^{n} \mathbf{y}_i\right)$. If we were interested in the probability that $\mathbf{y}_i = 1$ for an extra data point, we could estimate this probability through the EAP estimator

$$\mathbb{E}[\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n] = \frac{\alpha + \sum_{i=1}^{n} \mathbf{y}_i}{\alpha + \beta + n}, \tag{5.4}$$

or through the MAP estimator

$$\hat{\boldsymbol{\eta}} = \frac{\alpha + \sum_{i=1}^{n} \mathbf{y}_i - 1}{\alpha + \beta + n - 2}, \tag{5.5}$$

provided that $\alpha + \sum_{i=1}^{n} \mathbf{y}_i > 1$ and $\beta + n - \sum_{i=1}^{n} \mathbf{y}_i > 1$. Notice that if there exist a true number, p, such that $(\mathbf{y}_i)_{i \geq 1} \overset{iid}{\sim} \mathsf{Ber}(p)$, then both estimators the one in (5.4) and the one in (5.5) converge to p as the sample size, $n \to \infty$, this is a simple consequence of the law of large numbers. Whenever this holds, we say the estimators are consistent, and this is clearly a desirable property of models. Now, a key factor that allowed us to compute the posterior distribution in the current example is that $\mathbb{p}(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\eta})$ and $\mathbb{p}(\boldsymbol{\eta})$ form a conjugate pair, that is, the prior distribution of $\boldsymbol{\eta}$ and its posterior distribution belong to the same parametric family, in this situation, the Beta distribution. In general, when conjugacy is not attained, which is commonly the case of more complex models, it is much harder to compute the posterior distribution.

If it is not possible to compute $\mathbb{p}(\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)$ we can try to draw i.i.d. samples, $\left(\boldsymbol{\eta}^{(t)}\right)_{t=1}^{T}$ from the posterior distribution (Gilks et al.; 1995, e.g. by means of rejection sampling) and then, by the strong law of large numbers, we can estimate the quantity of interest through the EAP

$$\mathbb{E}\left[f(\boldsymbol{\eta}) \mid \mathbf{y}_1, \ldots, \mathbf{y}_n\right] \approx \frac{1}{T} \sum_{t=1}^{T} f\left(\boldsymbol{\eta}^{(t)}\right)$$

or through the MAP $f(\hat{\boldsymbol{\eta}}) \approx f\left(\boldsymbol{\eta}^{(\hat{t})}\right)$, where

$$\hat{t} = \underset{t \leq T}{\arg\max}\, \mathbb{p}\left(\boldsymbol{\eta}^{(t)} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n\right).$$

In the case that direct sampling from $\mathbb{p}(\boldsymbol{\eta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)$ is not feasible, one can recur to (MCMC) method, such as the Gibbs sampler, to draw correlated samples. This algorithm is broadly described below.

Say that after possibly re-parametrizing the model through $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m)$, where $\boldsymbol{\eta} = g(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m)$ for some measurable function g, it is possible to compute the conditional distributions

$$\mathbb{p}(\boldsymbol{\eta}_j \mid \ldots) = \mathbb{p}(\boldsymbol{\eta}_j \mid \mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\eta}_{-j}) \propto \mathbb{p}(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m)\mathbb{p}(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m),$$

where $\boldsymbol{\eta}_{-j} = \{\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m\} \backslash \{\boldsymbol{\eta}_j\}$, for every $j \in \{1, \ldots, m\}$. The Gibbs sampler algorithm consists in choosing some initial values $\boldsymbol{\eta}_1^{(0)}, \ldots, \boldsymbol{\eta}_m^{(0)}$ for $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m$ in the support of $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m)$ and sequentially for $t \geq 0$, sample:

- $\boldsymbol{\eta}_1^{(t+1)}$ from $\mathbb{p}(\boldsymbol{\eta}_1 \mid \ldots)$ given $\boldsymbol{\eta}_j = \boldsymbol{\eta}_j^{(t)}$ for $j \in \{2, \ldots, m\}$,
- $\boldsymbol{\eta}_2^{(t+1)}$ from $\mathbb{p}(\boldsymbol{\eta}_2 \mid \ldots)$ given $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_1^{(t+1)}$ and $\boldsymbol{\eta}_j = \boldsymbol{\eta}_j^{(t)}$ for $j \in \{3, \ldots, m\}$,
- $\vdots$
- $\boldsymbol{\eta}_m^{(t+1)}$ from $\mathbb{p}(\boldsymbol{\eta}_m \mid \ldots)$ given $\boldsymbol{\eta}_j = \boldsymbol{\eta}_j^{(t+1)}$ for $j \in \{1, \ldots, m-1\}$,

as illustrated in Figure 31. Then $\left(\boldsymbol{\eta}_1^{(t)}, \ldots, \boldsymbol{\eta}_m^{(t)}\right)_{t \geq 0}$ is a Markov chain with unique stationary distribution $\mathbb{p}(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)$, so the Ergodic theorem for stationary Markov chains assures that we can find $T_0 \in \mathbb{N}$ such that for every $t > T_0$, $\left(\boldsymbol{\eta}_1^{(t)}, \ldots, \boldsymbol{\eta}_m^{(t)}\right)$ is precisely a sample from $\mathbb{p}(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)$. When implementing the Gibbs sampler one usually samples from the Markov chain up to some time $T > T_0$ and disregards the samples $\left(\boldsymbol{\eta}_1^{(t)}, \ldots, \boldsymbol{\eta}_m^{(t)}\right)_{t=0}^{T_0}$ obtained during the so called burn-in period. This way, we can estimate the quantity of interest using the EAP

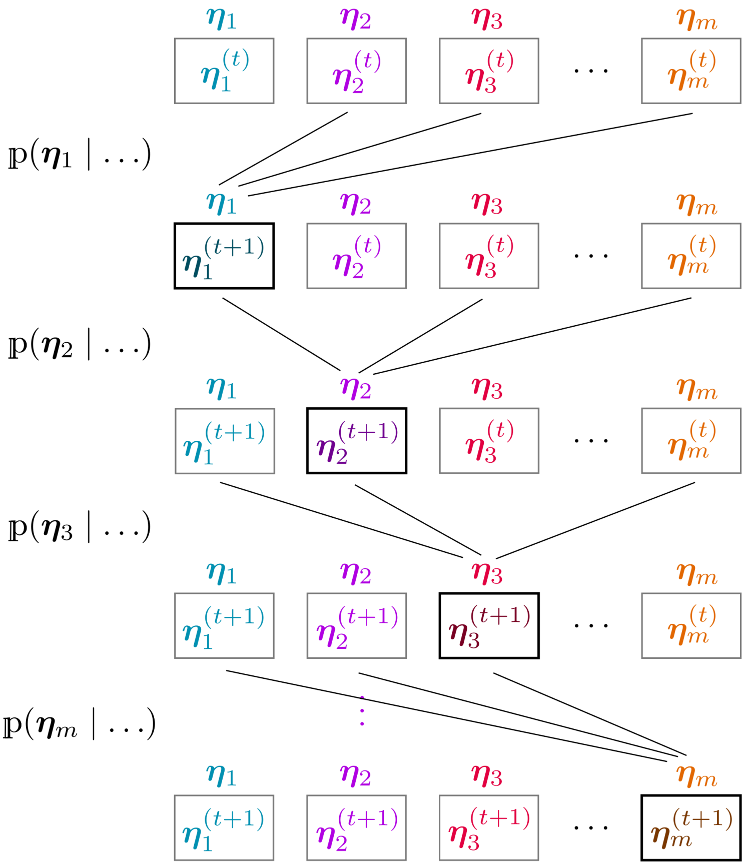
$$\mathbb{E}\left[f(g(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m)) \mid \mathbf{y}_1, \ldots, \mathbf{y}_n\right] \approx \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} f\left(g\left(\boldsymbol{\eta}_1^{(t)}, \ldots, \boldsymbol{\eta}_m^{(t)}\right)\right),$$

or alternatively through the MAP

$$f(\hat{\boldsymbol{\eta}}) = f\left(g\left(\boldsymbol{\eta}_1^{(\hat{t})}, \ldots, \boldsymbol{\eta}_m^{(\hat{t})}\right)\right)$$

where

$$\hat{t} = \underset{T_0 < t \leq T}{\arg\max}\, \mathbb{p}\left(\boldsymbol{\eta}_1^{(t)}, \dots, \boldsymbol{\eta}_m^{(t)} \,\middle|\, \mathbf{y}_1, \dots, \mathbf{y}_n \right).$$



$\boldsymbol{\eta}_1$ $\boldsymbol{\eta}_2$ $\boldsymbol{\eta}_3$ $\boldsymbol{\eta}_m$
$\boldsymbol{\eta}_1^{(t)}$ $\boldsymbol{\eta}_2^{(t)}$ $\boldsymbol{\eta}_3^{(t)}$ . . . $\boldsymbol{\eta}_m^{(t)}$
$\mathbb{p}(\boldsymbol{\eta}_1 \mid \dots)$
$\boldsymbol{\eta}_1$ $\boldsymbol{\eta}_2$ $\boldsymbol{\eta}_3$ $\boldsymbol{\eta}_m$
$\boldsymbol{\eta}_1^{(t+1)}$ $\boldsymbol{\eta}_2^{(t)}$ $\boldsymbol{\eta}_3^{(t)}$ . . . $\boldsymbol{\eta}_m^{(t)}$
$\mathbb{p}(\boldsymbol{\eta}_2 \mid \dots)$
$\boldsymbol{\eta}_1$ $\boldsymbol{\eta}_2$ $\boldsymbol{\eta}_3$ $\boldsymbol{\eta}_m$
$\boldsymbol{\eta}_1^{(t+1)}$ $\boldsymbol{\eta}_2^{(t+1)}$ $\boldsymbol{\eta}_3^{(t)}$ . . . $\boldsymbol{\eta}_m^{(t)}$
$\mathbb{p}(\boldsymbol{\eta}_3 \mid \dots)$
$\boldsymbol{\eta}_1$ $\boldsymbol{\eta}_2$ $\boldsymbol{\eta}_3$ $\boldsymbol{\eta}_m$
$\boldsymbol{\eta}_1^{(t+1)}$ $\boldsymbol{\eta}_2^{(t+1)}$ $\boldsymbol{\eta}_3^{(t+1)}$ . . . $\boldsymbol{\eta}_m^{(t)}$
$\mathbb{p}(\boldsymbol{\eta}_m \mid \dots)$
$\boldsymbol{\eta}_1$ $\boldsymbol{\eta}_2$ $\boldsymbol{\eta}_3$ $\boldsymbol{\eta}_m$
$\boldsymbol{\eta}_1^{(t+1)}$ $\boldsymbol{\eta}_2^{(t+1)}$ $\boldsymbol{\eta}_3^{(t+1)}$ . . . $\boldsymbol{\eta}_m^{(t+1)}$

Figure 31: One iteration of the Gibbs sampler.

Returning to the Bayesian non-parametric problem we are interested in, say we model data points $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$, that present no repetitions as i.i.d. sampled from a random mixture $\boldsymbol{\phi} = \sum_{j \geq 1} \mathbf{w}_j \mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$. Hereinafter we will work under the assumption that the base measure, μ_0, of the proper SSP, $\boldsymbol{\mu} = \sum_{j \geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$, and $\mathcal{K}(\cdot \mid s)$, for each $s \in S$, have a density with respect to suitable measures, this then implies $\boldsymbol{\phi}$ also has a density and is well defined. For the sake of a simpler notation we will be denoting the corresponding densities with the same letters, so that

$$\mathbb{p}(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\phi}) = \prod_{i=1}^{n} \boldsymbol{\phi}(\mathbf{y}_i),$$

or alternatively

$$\mathbb{p}(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \boldsymbol{\Xi}, \mathbf{W}) = \prod_{i=1}^{n} \sum_{j \geq 1} \mathbf{w}_j \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_j),$$

where $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j \geq 1}$ and $\mathbf{W} = (\mathbf{w}_j)_{j \geq 1}$. Specifying the prior distributions $\mathbb{p}(\boldsymbol{\phi})$ or $\mathbb{p}(\boldsymbol{\Xi}, \mathbf{W})$ can be done by constructing a SSP through one of the methods described in Section

3.5. In general it is not possible to compute $\mathbb{p}(\boldsymbol{\phi} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)$ or $\mathbb{p}(\boldsymbol{\Xi}, \mathbf{W} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)$ explicitly, nor to draw i.i.d. samples from them. Furthermore, the random objects we are interested in are infinitely dimensional, which means that even if we want to recur to a Gibbs sampler algorithm we will need to re-parametrize the model, so that its implementation is feasible. Depending on the available representation or construction of the species sampling prior, there are various well know methods to overcome this obstacle. In what follows we describe some of them.

5.2.1 Using the prediction rule

As mentioned in Section 5.1.1, modelling $\{\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\phi}\} \overset{\text{iid}}{\sim} \boldsymbol{\phi}$ is equivalent to

$$\mathbb{p}(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_i) \tag{5.6}$$

where $\{\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$. Now, if the EPPF, π, or the prediction rule for $(\mathbf{x}_i)_{i \geq 1}$ are available, using the exchangeability of $(\mathbf{x}_i)_{i \geq 1}$ and Theorem 3.6, we know that a priori, for every $i \leq n$,

$$\mathbb{p}(\mathbf{x}_i \mid \mathbf{X}_{-i}) = \sum_{j=1}^{\mathbf{K}} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})} \delta_{\mathbf{x}_j^*}(\mathbf{x}_i) + \frac{\pi\left(\mathbf{n}^{(\mathbf{K}+1)}\right)}{\pi(\mathbf{n})} \mu_0(\mathbf{x}_i),$$

where $\mathbf{X}_{-i} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \setminus \{\mathbf{x}_i\}$, $\mathbf{x}_1^*, \ldots, \mathbf{x}_{\mathbf{K}}^*$ are the distinct values in $\mathbf{X}_{-i}$, $\mathbf{n} = (\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}})$ is given by $\mathbf{n}_j = |\{l \neq i : \mathbf{x}_l = \mathbf{x}_j^*\}|$, $\mathbf{n}^{(j)} = (\mathbf{n}_1, \ldots \mathbf{n}_{j-1}, \mathbf{n}_j + 1, \mathbf{n}_{j+1}, \ldots, \mathbf{n}_{\mathbf{K}})$ and $\mathbf{n}^{(\mathbf{K}+1)} = (\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}_n}, 1)$. Under this approach we want to update $\mathbf{x}_1, \ldots, \mathbf{x}_n$ at each iteration of the Gibbs sampler. In order to do so it suffices to compute

$$\begin{aligned} \mathbb{p}(\mathbf{x}_i \mid \ldots) &\propto \mathbb{p}(\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) \mathbb{p}(\mathbf{x}_i \mid \mathbf{X}_{-i}) \\ &\propto \sum_{j=1}^{\mathbf{K}} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})} \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_j^*) \delta_{\mathbf{x}_j^*}(\mathbf{x}_i) + \frac{\pi\left(\mathbf{n}^{(\mathbf{K}+1)}\right)}{\pi(\mathbf{n})} \mu_0(\mathbf{x}_i) \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_i) \\ &\propto \sum_{j=1}^{\mathbf{K}} \mathbf{q}_j \delta_{\mathbf{x}_j^*}(\mathbf{x}_i) + \mathbf{q}_{\mathbf{K}+1} \frac{\mu_0(\mathbf{x}_i) \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_i)}{\int \mathcal{K}(\mathbf{y}_i \mid s) \mu_0(ds)} \end{aligned} \tag{5.7}$$

where

$$\mathbf{q}_j = \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})} \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_j^*), \tag{5.8}$$

for every $j \in \{1, \ldots, \mathbf{K}\}$ and

$$\mathbf{q}_{\mathbf{K}+1} = \frac{\pi\left(\mathbf{n}^{(\mathbf{K}+1)}\right)}{\pi(\mathbf{n})} \int \mathcal{K}(\mathbf{y}_i \mid s) \mu_0(ds). \tag{5.9}$$

For example, if π is the EPPF of a Dirichlet process with total mass parameter θ get have that

$$\mathbf{q}_j = \frac{\mathbf{n}_j}{\theta + n - 1} \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_j^*),$$

for every $j \in \{1, \dots, \mathbf{K}\}$ and

$$\mathbf{q}_{\mathbf{K}+1} = \frac{\theta}{\theta + n - 1} \int \mathcal{K}(\mathbf{y}_i \mid s)\mu_0(ds).$$

In general to sample $\mathbf{x}_i$ from (5.7) we set $\mathbf{x}_i = \mathbf{x}_j^*$ with probability proportional to $\mathbf{q}_j$ or with probability proportional to $\mathbf{q}_{\mathbf{K}+1}$ we sample $\mathbf{x}_i$ from

$$\mathbb{p}(\mathbf{x}_i \mid \mathbf{y}_i) \propto \mu_0(\mathbf{x}_i)\mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_i),$$

which can be easily done whenever μ_0 and $\mathcal{K}$ form a conjugate pair, in this case the integral

$$\int \mathcal{K}(\mathbf{y}_i \mid s)\mu_0(ds)$$

is generally feasible to compute or approximate. The algorithm can be stated as follows:

Algorithm 1: First T iterations of a Gibbs sampler algorithm using the EPPF

```
Pick some values $\mathbf{x}_1^{(0)}, \dots, \mathbf{x}_n^{(0)}$ in the support of $\mu_0$ and initialize
 $(\mathbf{x}_1, \dots, \mathbf{x}_n) = \left(\mathbf{x}_1^{(0)}, \dots, \mathbf{x}_n^{(0)}\right)$.
for $t \in \{1, \dots, T\}$ do
    for $i \in \{1, \dots, n\}$ do
        Sample $\mathbf{x}_i$ from (5.7);
        Set $\mathbf{x}_i^{(t)} = \mathbf{x}_i$;
Result: The Markov chain $\left(\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_n^{(t)}\right)_{t=1}^{T}$.
```

Unfortunately, this method can suffer from a slow convergence to the stationary distribution $\mathbb{p}(\mathbf{x}_1, \dots, \mathbf{x}_n \mid \mathbf{y}_1, \dots, \mathbf{y}_n)$, when the values of $\mathbf{q}_j$ as in (5.8) become much larger than $\mathbf{q}_{\mathbf{K}+1}$ as in (5.9), which can cause that many iterations of the Gibbs sampler occur before a new value $\mathbf{x}_j^*$ is sampled. To overcome this problem West et al. (1994) and MacEachern (1994) proposed to re-express the sampling of $\mathbf{x}_1, \dots, \mathbf{x}_n$ in terms of the distinct values $\mathbf{x}_1^*, \dots, \mathbf{x}_{\mathbf{K}_n}^*$ that $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ exhibits and the partition structure $\{\mathbf{\Pi}_1, \dots, \mathbf{\Pi}_{\mathbf{K}_n}\}$ where $\mathbf{\Pi}_j = \{i \leq n : \mathbf{x}_i = \mathbf{x}_j^*\}$, essentially this results in adding a step to Algorithm 1 where $\mathbf{x}_j^*$ is updated for every $j \in \{1, \dots, \mathbf{K}_n\}$. To spell this out first note that we can rewrite (5.6) as

$$\mathbb{p}(\mathbf{y}_1, \dots, \mathbf{y}_n \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{j=1}^{\mathbf{K}_n} \prod_{i \in \mathbf{\Pi}_j} \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_j^*).$$

Also recall from Theorem 3.6 that

$$\mathbb{p}(\mathbf{x}_1, \dots, \mathbf{x}_n) = \pi(\mathbf{n}_1, \dots, \mathbf{n}_{\mathbf{K}_n}) \prod_{j=1}^{\mathbf{K}_n} \mu_0(\mathbf{x}_j^*),$$

where $\mathbf{n}_j = |\mathbf{\Pi}_j|$. Hence to update $\mathbf{x}_j^*$ for $\{1, \dots, \mathbf{K}_n\}$, we have to sample it from

$$\mathbb{p}(\mathbf{x}_j^* \mid \dots) \propto \mu_0(\mathbf{x}_j^*) \prod_{i \in \mathbf{\Pi}_j} \mathcal{K}(\mathbf{y}_i \mid \mathbf{x}_j^*) \tag{5.10}$$

which is easy to do if μ_0 and $\mathcal{K}$ constitute a conjugate pair. Now, to update the random partition $\{\mathbf{\Pi}_1, \ldots, \mathbf{\Pi}_{\mathbf{K}_n}\}$ we can update one a time for, $i \in \{1, \ldots, n\}$, to which block does i belongs. Noting that $i \in \mathbf{\Pi}_j$ if and only if $\mathbf{x}_i = \mathbf{x}_j^*$, to do this we can sample $\mathbf{x}_i$ from $\mathbb{p}(\mathbf{x}_i \mid \ldots)$ and this way update the membership of i. Since $\mathbf{x}_1^*, \ldots, \mathbf{x}_{\mathbf{K}_n}^*$ provide no more information about $\mathbf{x}_i$ than $\mathbf{X}_{-i}$ does, it is straightforward to see that conditioning on $\mathbf{x}_1^*, \ldots, \mathbf{x}_{\mathbf{K}_n}^*$, $\mathbb{p}(\mathbf{x}_i \mid \ldots)$ remains as in (5.7). With this considerations taken into account we can modify Algorithm 1 as follows:

Algorithm 2: First T iterations of a modified Gibbs sampler algorithm using the EPPF

```
Initialize the partition of $\{1,\ldots,n\}$, $\{\mathbf{\Pi}_1,\ldots,\mathbf{\Pi}_{\mathbf{K}_n}\} = \left\{\mathbf{\Pi}_1^{(0)},\ldots,\mathbf{\Pi}_{\mathbf{K}_n}^{(0)}\right\}$
for $t \in \{1,\ldots,T\}$ do
    for $j \in \{1,\ldots,\mathbf{K}_n\}$ do
        Sample $\mathbf{x}_j^*$ from (5.10);
    for $i \in \{1,\ldots,n\}$ do
        Sample $\mathbf{x}_i$ from (5.7);
        Set $\mathbf{x}_i^{(t)} = \mathbf{x}_i$;
    Define $\mathbf{\Pi}_j = \{i \leq n : \mathbf{x}_i = \mathbf{x}_j^*\}$, where $\mathbf{x}_1^*,\ldots,\mathbf{x}_{\mathbf{K}_n}^*$ are the distinct values
     $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ exhibits.
Result: The Markov chain $\left(\mathbf{x}_1^{(t)},\ldots,\mathbf{x}_n^{(t)}\right)_{t=1}^{T}$.
```

Despite whether we choose to implement Algorithm 1 or 2, given the samples $\left(\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}\right)_{t=1}^{T}$ obtained from the Gibbs sampler, we can use the EAP to estimate the density of the data at y by means of

$$\phi_{\text{EAP}}(y) \approx \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \frac{1}{n} \sum_{j=1}^{\mathbf{K}_n^{(t)}} \mathbf{n}_j^{(t)} \mathcal{K}\left(y \,\middle|\, \mathbf{x}_j^{*(t)}\right)$$

where T_0 is the last iteration of the burn-in period, $\mathbf{K}_n^{(t)}$ is number the distinct values in $\left\{\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}\right\}$, $\mathbf{x}_1^{*(t)}, \ldots, \mathbf{x}^{*(t)}_{\mathbf{K}_n{}^{(t)}}$ are such values, and $\mathbf{n}_j^{(t)} = \left|\left\{i \leq n : \mathbf{x}_i^{(t)} = \mathbf{x}_j^{*(t)}\right\}\right|$. By means of the EAP, we can also estimate the posterior distribution of the number of significant components, $\mathbf{K}_n$, in the data set $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ through

$$\mathbb{P}[\mathbf{K}_n = j \mid \mathbf{y}_1, \ldots, \mathbf{y}_n] \approx \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} \mathbf{1}_{\left\{\mathbf{K}_n^{(t)}=j\right\}}.$$

Alternatively, we can find

$$\begin{aligned}\hat{t} &= \operatorname*{arg\,max}_{T_0<t\leq T} \mathbb{p}\left(\mathbf{y}_1, \ldots, \mathbf{y}_n \,\middle|\, \mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}\right) \mathbb{p}\left(\mathbf{x}_1^{(t)}, \ldots, \mathbf{x}_n^{(t)}\right) \\ &= \operatorname*{arg\,max}_{T_0<t\leq T} \pi\left(\mathbf{n}_1^{(t)}, \ldots, \mathbf{n}_{\mathbf{K}_n^{(t)}}^{(t)}\right) \prod_{j=1}^{\mathbf{K}_n^{(t)}} \mu_0\left(\mathbf{x}_j^{*(t)}\right) \prod_{i \in \mathbf{\Pi}_j^{(t)}} \mathcal{K}\left(\mathbf{y}_i \,\middle|\, \mathbf{x}_j^{*(t)}\right).\end{aligned}$$

and use the MAP to estimate the density at y, through

$$\phi_{\text{MAP}}(y) \approx \frac{1}{n} \sum_{j=1}^{\mathbf{K}_n^{(\hat{t})}} \mathbf{n}_j^{(\hat{t})} \mathcal{K}\left(y \,\middle|\, \mathbf{x}_j^{*(\hat{t})}\right).$$

By means of the MAP we can also estimate the number of significant components by $\mathbf{K}_n^{(\hat{t})}$, and the clusters of the data points $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, by putting $\mathbf{y}_i$ and $\mathbf{y}_k$ in the same block if and only if $\mathbf{x}_i^{(\hat{t})} = \mathbf{x}_k^{(\hat{t})}$. Another way to estimate the clusters is via the mixture components,

$$\left\{\frac{\mathbf{n}_j^{(\hat{t})}}{n}\mathcal{K}\left(\cdot \,\middle|\, \mathbf{x}_j^{*(\hat{t})}\right)\right\}_{j=1}^{\mathbf{K}_n^{(\hat{t})}},$$

by defining

$$\mathbf{c}_i = \operatorname*{arg\,max}_{1\le j\le \mathbf{K}_n^{(\hat{t})}} \left\{\frac{\mathbf{n}_j^{(\hat{t})}}{n}\mathcal{K}\left(\mathbf{y}_i \,\middle|\, \mathbf{x}_j^{*(\hat{t})}\right)\right\},$$

for every $i \in \{1, \ldots, n\}$, and putting $\mathbf{y}_i$ and $\mathbf{y}_k$ in the same cluster if and only if $\mathbf{c}_i = \mathbf{c}_k$.

Notice that, although $\boldsymbol{\phi}$ is infinitely dimensional, this Gibbs sampler allows us to estimate the density by only sampling finitely many random elements, this we were able to achieve by writing the model in terms of a sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of the SSP, $\boldsymbol{\mu}$, and exploiting their distinct representations as stated in Theorem 3.6. Furthermore, if μ_0 and $\mathcal{K}$ are conjugate the algorithm is very easy to implement, for the cases where conjugacy is not attained various solutions have been proposed (e.g. West et al.; 1994; Walker and Damien; 1998; MacEachern and Müller; 1998; Neal; 2000). Perhaps the major drawback of this method is that we require the EPPF or the prediction rule, which is unattainable in closed forms for most SSP priors. Usually, the canonical examples of the present algorithm are the Dirichlet and the Pitman-Yor processes because their prediction rules are simple enough that the sampler is feasible to implement. Now for some models it is possible to characterize the partition structure given a latent random variable (e.g. James et al.; 2009) in these cases the algorithm can be adapted to cover a wider range of SSP priors. For the priors without this characteristic there are other paths to implement the model.

5.2.2 Using latent random sets

As before, say we model data as i.i.d. sampled from $\{\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\phi}\} \overset{\text{iid}}{\sim} \boldsymbol{\phi} = \sum_{j\ge1} \mathbf{w}_j \mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$. This time assume that the EPPF corresponding to $\boldsymbol{\mu} = \sum_{j\ge1}\mathbf{w}_j\delta_{\boldsymbol{\xi}_j}$ is note available but we can construct $\boldsymbol{\mu}$ by means of latent random sets as explained in Section 3.5.4. That is we can write

$$\mathbf{w}_j = \mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}|}\mathbf{1}_{\{j\in\boldsymbol{\Psi}\}} \,\middle|\, \boldsymbol{\tau}\right], \quad j \ge 1, \tag{5.11}$$

for a random element $\boldsymbol{\tau} \sim \mathbb{p}(\boldsymbol{\tau})$ that takes values in a Borel space, and some random set $\{\boldsymbol{\Psi} \mid \boldsymbol{\tau}\} \sim \mathbb{p}(\boldsymbol{\Psi} \mid \boldsymbol{\tau})$ that takes values in $\mathcal{F}_{\mathbb{N}} = \{A \subseteq \mathbb{N} : 0 < |A| < \infty\}$. As a consequence of Proposition 3.17 and the discussion in Section 5.1.1, assuming $\{\mathbf{y}_1, \ldots, \mathbf{y}_n \mid \boldsymbol{\phi}\} \overset{\text{iid}}{\sim} \boldsymbol{\phi}$, is equivalent to consider $\{\mathbf{y}_i \mid \mathbf{x}_i\} \sim \mathcal{K}(\cdot \mid \mathbf{x}_i)$, $\{\mathbf{x}_i \mid \boldsymbol{\Xi}, \boldsymbol{\Psi}_i\} \sim |\boldsymbol{\Psi}_i|^{-1}\sum_{j\in\boldsymbol{\Psi}_i}\delta_{\boldsymbol{\xi}_j}$, independently for $i \in \{1, \ldots, n\}$, and $\{\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n \mid \boldsymbol{\tau}\} \overset{\text{iid}}{\sim} \mathbb{p}(\boldsymbol{\Psi} \mid \boldsymbol{\tau})$, where evidently $\boldsymbol{\tau} \sim \mathbb{p}(\boldsymbol{\tau})$ and $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j\ge1} \overset{\text{iid}}{\sim} \mu_0$. By further introducing the latent allocation variables

$\mathbf{d}_i = j$ if and only if $\mathbf{x}_i = \boldsymbol{\xi}_j$ we can re-express

$$\begin{aligned}
\{\mathbf{y}_i \mid \Xi, \mathbf{d}_i\} &\sim \mathcal{K}(\cdot \mid \boldsymbol{\xi}_{\mathbf{d}_i}) \text{ indep. for } i \in \{1, \dots, n\} \\
\{\mathbf{d}_i \mid \boldsymbol{\Psi}_i\} &\sim \mathsf{Unif}(\boldsymbol{\Psi}_i) \text{ indep. for } i \in \{1, \dots, n\} \\
\{\boldsymbol{\Psi}_1, \dots, \boldsymbol{\Psi}_n \mid \boldsymbol{\tau}\} &\overset{\text{iid}}{\sim} \mathbb{p}(\boldsymbol{\Psi} \mid \boldsymbol{\tau}) \\
\boldsymbol{\tau} &\sim \mathbb{p}(\boldsymbol{\tau}) \\
(\boldsymbol{\xi}_j)_{j\geq 1} &\overset{\text{iid}}{\sim} \mu_0.
\end{aligned}$$

that is

$$\mathbb{p}(\mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{d}_1, \dots, \mathbf{d}_n, \boldsymbol{\Psi}_1, \dots, \boldsymbol{\Psi}_n \mid \Xi, \boldsymbol{\tau}) = \prod_{i=1}^{n} \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i}) \frac{1}{|\boldsymbol{\Psi}_i|} \mathbf{1}_{\{\mathbf{d}_i \in \boldsymbol{\Psi}_i\}} \mathbb{p}(\boldsymbol{\Psi}_i \mid \boldsymbol{\tau}). \tag{5.12}$$

Under this approach, the full conditional distributions required to update the random elements involved, at each iteration of the Gibbs sampler, are proportional to (5.12) multiplied by the prior of $\boldsymbol{\tau}$ and Ξ, and computed below.

Updating $\boldsymbol{\xi}_j$ for $j \geq 1$:

$$\mathbb{p}(\boldsymbol{\xi}_j \mid \dots) \propto \mu_0(\boldsymbol{\xi}_j) \prod_{i \in \mathbf{D}_j} \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_j) \tag{5.13}$$

where $\mathbf{D}_j = \{i \leq n : \mathbf{d}_i = j\}$. Drawing samples from (5.13) is easy when conjugacy is attained for μ_0 and $\mathcal{K}$. Note that if $\mathbf{D}_j = \emptyset$ the $\boldsymbol{\xi}_j$ is simply sampled form its prior distribution μ_0.

Updating $\mathbf{d}_i$ for $i \in \{1, \dots, n\}$:

$$\mathbb{p}(\mathbf{d}_i \mid \dots) \propto \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i}) \mathbf{1}_{\{\mathbf{d}_i \in \boldsymbol{\Psi}_i\}}$$

which yields

$$\mathbb{p}(\mathbf{d}_i \mid \dots) = \frac{\mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i}) \mathbf{1}_{\{\mathbf{d}_i \in \boldsymbol{\Psi}_i\}}}{\sum_{j \in \boldsymbol{\Psi}_i} \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i})}. \tag{5.14}$$

To sample $\mathbf{d}_i$ from this distribution simply means that we set $\mathbf{d}_i = j$ with probability proportional to $\mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_j)$ for each $j \in \boldsymbol{\Psi}_i$. Being that $\boldsymbol{\Psi}$ is non-empty and finite almost surely drawing samples from (5.14) is easy.

Updating $\boldsymbol{\Psi}_i$ for $i \in \{1, \dots, n\}$:

$$\mathbb{p}(\boldsymbol{\Psi}_i \mid \dots) \propto \frac{\mathbb{p}(\boldsymbol{\Psi}_i \mid \boldsymbol{\tau})}{|\boldsymbol{\Psi}_i|} \mathbf{1}_{\{\mathbf{d}_i \in \boldsymbol{\Psi}_i\}}. \tag{5.15}$$

Updating $\boldsymbol{\tau}$:

$$\mathbb{p}(\boldsymbol{\tau} \mid \dots) \propto \mathbb{p}(\boldsymbol{\tau}) \prod_{i=1}^{n} \mathbb{p}(\boldsymbol{\Psi}_i \mid \boldsymbol{\tau}). \tag{5.16}$$

Whether it is simple or possible to sample from equations (5.14) and (5.16) depends on $\mathbb{p}(\boldsymbol{\Psi}_i \mid \boldsymbol{\tau})$ and $\mathbb{p}(\boldsymbol{\tau})$. An example of a species sampling prior where this is very easy to do is the Geometric prior. Recall from Section 3.6.3 that if $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} \sim \mathcal{G}(\nu_0, \mu_0)$, then we can write $\mathbf{w}_j$ as in (5.11) where $\boldsymbol{\tau} \sim \nu_0$ and $\boldsymbol{\Psi} = \{1, \dots, \mathbf{N}\}$, with

$$\mathbb{p}(\mathbf{N} \mid \boldsymbol{\tau}) = \mathbf{N}\boldsymbol{\tau}^2 (1 - \boldsymbol{\tau})^{\mathbf{N}-1}.$$

Hence, if the Geometric prior is considered, by setting $\boldsymbol{\Psi}_i = \{1,\dots,\mathbf{N}_i\}$, where $\{\mathbf{N}_1,\dots,\mathbf{N}_n \mid \boldsymbol{\tau}\} \overset{\text{iid}}{\sim} \mathbb{p}(\mathbf{N} \mid \boldsymbol{\tau})$, (5.12) becomes

$$\mathbb{p}(\mathbf{y}_1,\dots,\mathbf{y}_n,\mathbf{d}_1,\dots,\mathbf{d}_n,\mathbf{N}_1,\dots,\mathbf{N}_n \mid \Xi,\boldsymbol{\tau}) = \prod_{i=1}^{n} \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i})\mathbf{1}_{\{\mathbf{d}_i\le\mathbf{N}_i\}}\boldsymbol{\tau}^2(1-\boldsymbol{\tau})^{\mathbf{N}_i-1}. \tag{5.17}$$

In this particular case, the full conditionals of $\boldsymbol{\xi}_j$ and $\mathbf{d}_i$ are left unchanged. Clearly, to update $\boldsymbol{\Psi}_i$ it suffices to update $\mathbf{N}_i$ and latter define $\boldsymbol{\Psi}_i = \{1,\dots,\mathbf{N}_i\}$, as follows:

Updating $\mathbf{N}_i$ for every $i \in \{1,\dots,n\}$:

$$\mathbb{p}(\mathbf{N}_i \mid \dots) \propto (1-\boldsymbol{\tau})^{\mathbf{N}_i-1}\mathbf{1}_{\{\mathbf{d}_i\le\mathbf{N}_i\}}, \tag{5.18}$$

thus, a posteriori $\mathbf{N}_i$ follows a truncated Geometric distribution, which is very easy to sample from using the memory loss property of the distribution in question. It only remains to compute the full conditional of $\boldsymbol{\tau}$, which is also very simple to do if we assign a $\mathsf{Be}(\alpha,\beta)$ prior to this random variable.

Updating $\boldsymbol{\tau}$:

$$\mathbb{p}(\boldsymbol{\tau} \mid \dots) \propto \mathbb{p}(\boldsymbol{\tau})\prod_{i=1}^{n}\boldsymbol{\tau}^2(1-\boldsymbol{\tau})^{\mathbf{N}_i-1} \propto \boldsymbol{\tau}^{\alpha+2n-1}(1-\boldsymbol{\tau})^{\beta+\sum_{i=1}^n \mathbf{N}_i-n-1}, \tag{5.19}$$

thus to update $\boldsymbol{\tau}$ we simply sample it from a $\mathsf{Be}\left(\alpha+2n, \beta+\sum_{i=1}^n \mathbf{N}_i - n\right)$. For more details on the implementation of this algorithm for the Geometric process, see Fuentes-García et al. (2010), other examples of Bayesian non-parametric priors for which this algorithm is particularly useful can be found in De Blasi et al. (2020) and Gil-Leyva (2021).

Returning to the general scenario, realize that we do not require to sample $\boldsymbol{\xi}_j$ for every $j \ge 1$ it suffices to sample enough of them so that the updating of $(\mathbf{d}_i)_{i=1}^n$ can take place. This is, it is enough to sample $\boldsymbol{\xi}_j$ for every $j \in \bigcup_{i=1}^n \boldsymbol{\Psi}_i$ which is a finite set, hence feasible to do. The complete Gibbs sampler algorithm is shown below:

Algorithm 3: First T iterations of a Gibbs sampler algorithm using latent random sets

```
Initialize $\mathbf{d}_i = \mathbf{d}_i^{(0)} \in \mathbb{N}$ for every $i \in \{1,\dots,n\}$
for $t \in \{1,\dots,T\}$ do
    for $i \in \{1,\dots,n\}$ do
        Sample $\boldsymbol{\Psi}_i$ from (5.15);
        Set $\boldsymbol{\Psi}_i^{(t)} = \boldsymbol{\Psi}_i$;
    Sample $\boldsymbol{\tau}$ from (5.16);
    Set $\boldsymbol{\tau}^{(t)} = \boldsymbol{\tau}$;
    for $j \in \bigcup_{i=1}^n \boldsymbol{\Psi}_i$ do
        Sample $\boldsymbol{\xi}_j$ from (5.13);
        Set $\boldsymbol{\xi}_j^{(t)} = \boldsymbol{\xi}_j$;
    for $i \in \{1,\dots,n\}$ do
        Sample $\mathbf{d}_i^{(t)}$ from (5.14);
        Set $\mathbf{d}_i^{(t)} = \mathbf{d}_i$;
Result: The Markov chain
    $\left(\mathbf{d}_1^{(t)},\dots,\mathbf{d}_n^{(t)},\boldsymbol{\Psi}_1^{(t)},\dots,\boldsymbol{\Psi}_n^{(t)},\boldsymbol{\tau}^{(t)},\left(\boldsymbol{\xi}_j^{(t)}\right)_{j\in\bigcup_{i=1}^n \boldsymbol{\Psi}_i^{(t)}}\right)_{t=1}^{T}$.
```

Given the samples

$$\left(\mathbf{d}_1^{(t)}, \dots, \mathbf{d}_n^{(t)}, \boldsymbol{\Psi}_1^{(t)}, \dots, \boldsymbol{\Psi}_n^{(t)}, \boldsymbol{\tau}^{(t)}, \left(\boldsymbol{\xi}_j^{(t)}\right)_{j \in \bigcup_{i=1}^n \boldsymbol{\Psi}_i^{(t)}}\right)_{t=1}^T$$

obtained from the Gibbs sampler, we can use the EAP to estimate the density of the data at y by means of

$$\phi_{\mathrm{EAP}}(y) \approx \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\left|\boldsymbol{\Psi}_i^{(t)}\right|} \sum_{j \in \boldsymbol{\Psi}_i^{(t)}} \mathcal{K}\left(y \,\middle|\, \boldsymbol{\xi}_j{}^{(t)}\right),$$

where T_0 is the last iteration of the burn-in period. We can also estimate the posterior distribution of the number of significant components, $\mathbf{K}_n$, in the data set $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ through

$$\mathbb{P}[\mathbf{K}_n = j \mid \mathbf{y}_1, \dots, \mathbf{y}_n] \approx \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \mathbf{1}_{\left\{\mathbf{K}_n^{(t)} = j\right\}}.$$

where $\mathbf{K}_n^{(t)}$ is the number of distinct values in $\left\{\mathbf{d}_1^{(t)}, \dots, \mathbf{d}_n^{(t)}\right\}$. Alternatively, we can find

$$\begin{aligned}\hat{t} &= \operatorname*{arg\,max}_{T_0 < t \le T} \left\{ \mathbb{p}(\mathbf{y}_1, \dots, \mathbf{y}_n, \mathbf{d}_1, \dots, \mathbf{d}_n, \boldsymbol{\Psi}_1, \dots, \boldsymbol{\Psi}_n \mid \boldsymbol{\Xi}, \boldsymbol{\tau}) \mathbb{p}\left(\boldsymbol{\tau}^{(t)}\right) \prod_{j \in \bigcup_i \boldsymbol{\Psi}_i^{(t)}} \mu_0\left(\boldsymbol{\xi}_j^{(t)}\right) \right\} \\ &= \operatorname*{arg\,max}_{T_0 < t \le T} \left\{ \prod_{i=1}^{n} \mathcal{K}\left(\mathbf{y}_i \,\middle|\, \boldsymbol{\xi}_{\mathbf{d}_i^{(t)}}^{(t)}\right) \frac{\mathbb{p}\left(\boldsymbol{\Psi}_i^{(t)} \,\middle|\, \boldsymbol{\tau}^{(t)}\right)}{\left|\boldsymbol{\Psi}_i^{(t)}\right|} \mathbf{1}_{\left\{\mathbf{d}_i^{(t)} \in \boldsymbol{\Psi}_i^{(t)}\right\}} \mathbb{p}\left(\boldsymbol{\tau}^{(t)}\right) \prod_{j \in \bigcup_i \boldsymbol{\Psi}_i^{(t)}} \mu_0\left(\boldsymbol{\xi}_j^{(t)}\right) \right\} \\ &.\end{aligned}$$

and use the MAP to estimate the density at y, through

$$\phi_{\mathrm{MAP}}(y) \approx \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\left|\boldsymbol{\Psi}_i^{(\hat{t})}\right|} \sum_{j \in \boldsymbol{\Psi}_i^{(\hat{t})}} \mathcal{K}\left(y \,\middle|\, \boldsymbol{\xi}_j{}^{(\hat{t})}\right).$$

By means of the MAP we can also estimate the number of significant components by $\mathbf{K}_n^{(\hat{t})}$, and the clusters of the data points $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$, by putting $\mathbf{y}_i$ and $\mathbf{y}_k$ in the same cluster if and only if $\mathbf{d}_i^{(\hat{t})} = \mathbf{d}_k^{(\hat{t})}$, or via the mixture components

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}_{\left\{j \in \boldsymbol{\Psi}_i^{(\hat{t})}\right\}}}{\left|\boldsymbol{\Psi}_i^{(\hat{t})}\right|} \mathcal{K}\left(\cdot \,\middle|\, \boldsymbol{\xi}_j^{(\hat{t})}\right) \right\}_{j \in \bigcup_{i=1}^n \boldsymbol{\Psi}_i^{(\hat{t})}},$$

by defining

$$\mathbf{c}_i = \operatorname*{arg\,max}_{j \in \bigcup_{i=1}^n \boldsymbol{\Psi}_i^{(\hat{t})}} \left\{ \frac{1}{n} \sum_{k=1}^{n} \frac{\mathbf{1}_{\left\{j \in \boldsymbol{\Psi}_k^{(\hat{t})}\right\}}}{\left|\boldsymbol{\Psi}_k^{(\hat{t})}\right|} \mathcal{K}\left(\mathbf{y}_i \,\middle|\, \boldsymbol{\xi}_j^{(\hat{t})}\right) \right\},$$

for every $i \in \{1, \dots, n\}$, and putting $\mathbf{y}_i$ and $\mathbf{y}_k$ in the same cluster if and only if $\mathbf{c}_i = \mathbf{c}_k$.

5.2.3 Slice sampler

This Gibbs sampler algorithm was derived by Walker (2007), and is particularly useful for those species sampling priors where the distribution of the weights is available in closed form but the EPPF is not. This is the case for many stick-breaking processes such as the models introduced in Section 4. Essentially this algorithm consists in defining latent random sets using slices and then apply the algorithm is Section 5.2.2. The construction of these latent random sets is precisely the proof of Proposition 3.16, despite we will replicate it here. First of all consider a latent random variable $\mathbf{u}$ with marginal support $[0,1]$, and such that

$$\mathbb{p}(\mathbf{u} \mid \mathbf{W}) = \sum_{j\geq 1} \mathbf{1}_{\{\mathbf{u}<\mathbf{w}_j\}} = |\{j \geq 1 : \mathbf{u} < \mathbf{w}_j\}|, \tag{5.20}$$

where $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$ (see Figure 49 in Appendix C.12 for an illustration of $\mathbb{p}(\mathbf{u} \mid \mathbf{W})$). Now, define $\boldsymbol{\Psi} = \{j \geq 1 : \mathbf{u} < \mathbf{w}_j\}$, and note that from (5.20), $\mathbf{u} < \max_j \mathbf{w}_j$, so $\boldsymbol{\Psi}$ is non-empty almost surely. Moreover, being that $\sum_{j\geq 1} \mathbf{w}_j = 1$, we also have that $\boldsymbol{\Psi}$ is finite with probability one. Finally if we set $\boldsymbol{\tau} = \mathbf{W}$, we get

$$\mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}|}\mathbf{1}_{\{j\in\boldsymbol{\Psi}\}}\,\middle|\,\boldsymbol{\tau}\right] = \int_0^1 \frac{1}{|\boldsymbol{\Psi}|}\mathbf{1}_{\{\mathbf{u}<\mathbf{w}_j\}}\mathbb{p}(\mathbf{u} \mid \boldsymbol{\tau})d\mathbf{u} = \int_0^1 \mathbf{1}_{\{\mathbf{u}<\mathbf{w}_j\}}d\mathbf{u} = \mathbf{w}_j$$

for every $j \geq 1$. Thus, $\boldsymbol{\Psi} = \{j \geq 1 : \mathbf{u} < \mathbf{w}_j\}$ and $\boldsymbol{\tau} = \mathbf{W}$ satisfy equation (5.11) meaning that we have constructed a random set that allows a re-parametrization or augmentation of model in such way that only finitely many random elements have to be updated at each iteration of the Gibbs sampler. Namely, under analogous arguments to those of Section 5.2.2, modelling $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ as i.i.d. sampled from the random mixture $\boldsymbol{\phi}$ is equivalent to

$$\begin{aligned}
\{\mathbf{y}_i \mid \boldsymbol{\Xi}, \mathbf{d}_i\} &\sim \mathcal{K}(\cdot \mid \boldsymbol{\xi}_{\mathbf{d}_i}) \text{ indep. for } i \in \{1,\ldots,n\}\\
\{\mathbf{d}_i \mid \boldsymbol{\Psi}_i\} &\sim \mathsf{Unif}(\boldsymbol{\Psi}_i) \text{ indep. for } i \in \{1,\ldots,n\}\\
\boldsymbol{\Psi}_i &= \{j \geq 1 : \mathbf{u}_i < \mathbf{w}_j\} \text{ indep. for } i \in \{1,\ldots,n\}\\
\{\mathbf{u}_1,\ldots,\mathbf{u}_n \mid \mathbf{W}\} &\overset{\text{iid}}{\sim} \mathbb{p}(\mathbf{u} \mid \mathbf{W})\\
\mathbf{W} &\sim \mathbb{p}(\mathbf{W})\\
(\boldsymbol{\xi}_j)_{j\geq 1} &\overset{\text{iid}}{\sim} \mu_0.
\end{aligned}$$

Noting that $j \in \boldsymbol{\Psi}_i$ if and only if $\mathbf{u}_i < \mathbf{w}_j$ and $|\boldsymbol{\Psi}_i| = \mathbb{p}(\mathbf{u}_i \mid \mathbf{W})$, for every $i \in \{1,\ldots,n\}$, this yields $\mathbb{p}(\mathbf{y}_1,\ldots,\mathbf{y}_n \mid \mathbf{W}, \boldsymbol{\Xi})$ can be augmented as

$$\mathbb{p}(\mathbf{y}_1,\ldots,\mathbf{y}_n,\mathbf{d}_1,\ldots,\mathbf{d}_n,\mathbf{u}_1,\ldots,\mathbf{u}_n \mid \mathbf{W}, \boldsymbol{\Xi}) = \prod_{i=1}^n \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i})\mathbf{1}_{\{\mathbf{u}_i<\mathbf{w}_{\mathbf{d}_i}\}}. \tag{5.21}$$

The full conditionals required to update the random elements at each iteration of the Gibbs sampler are proportional to

$$\mathbb{p}(\mathbf{y}_1,\ldots,\mathbf{y}_n,\mathbf{d}_1,\ldots,\mathbf{d}_n,\mathbf{u}_1,\ldots,\mathbf{u}_n \mid \mathbf{W}, \boldsymbol{\Xi})\mathbb{p}(\boldsymbol{\Xi})\mathbb{p}(\mathbf{W})$$

and described below.

Updating $\boldsymbol{\xi}_j$ for $j \geq 1$:

$$\mathbb{p}(\boldsymbol{\xi}_j \mid \ldots) \propto \mu_0(\boldsymbol{\xi}_j) \prod_{i \in \mathbf{D}_j} \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_j), \tag{5.22}$$

where $\mathbf{D}_j = \{i \leq n : \mathbf{d}_i = j\}$.

Updating $\mathbf{d}_i$ for $i \in \{1, \ldots, n\}$:

$$\mathbb{p}(\mathbf{d}_i \mid \ldots) \propto \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i}) \mathbf{1}_{\{\mathbf{u}_i < \mathbf{w}_{\mathbf{d}_i}\}} \propto \mathcal{K}(\mathbf{y}_i \mid \boldsymbol{\xi}_{\mathbf{d}_i}) \mathbf{1}_{\{\mathbf{d}_i \in \boldsymbol{\Psi}_i\}}. \tag{5.23}$$

Hence the updating of $(\boldsymbol{\xi}_j)_{j \geq 1}$ and $(\mathbf{d}_i)_{i=1}^n$ in this Gibbs sampler remains identical to the updating of them in the algorithm described in Section 5.2.2.

Updating $\mathbf{u}_i$ for $i \in \{1, \ldots, n\}$:

$$\mathbb{p}(\mathbf{u}_i \mid \ldots) \propto \mathbf{1}_{\{\mathbf{u}_i < \mathbf{w}_{\mathbf{d}_i}\}} \propto \mathsf{Unif}(\mathbf{u}_i \mid 0, \mathbf{w}_{\mathbf{d}_i}). \tag{5.24}$$

that is we simply have to sample $\mathbf{u}_i \sim \mathsf{Unif}(\mathbf{u}_i \mid 0, \mathbf{w}_{\mathbf{d}_i})$ independently for $i \in \{1, \ldots, n\}$.

Updating $\mathbf{W}$:

$$\mathbb{p}(\mathbf{W} \mid \ldots) \propto \mathbb{p}(\mathbf{W}) \prod_{i=1}^{n} \mathbf{1}_{\{\mathbf{u}_i < \mathbf{w}_{\mathbf{d}_i}\}}. \tag{5.25}$$

For instance if the stick-breaking representation of the weights is available $\mathbf{W} = \mathsf{SB}(\mathbf{V})$, for a sequence of length variables $\mathbf{V} = (\mathbf{v}_j)_{j \geq 1}$ with mathematically tractable prior distribution, we can the weights sequence via the length variables.

Updating $\mathbf{V}$:

$$\mathbb{p}(\mathbf{V} \mid \ldots) \propto \mathbb{p}(\mathbf{V}) \prod_{i=1}^{n} \mathbf{1}\left\{\mathbf{u}_i < \mathbf{v}_{\mathbf{d}_i} \prod_{l < \mathbf{d}_i} (1 - \mathbf{v}_l)\right\}. \tag{5.26}$$

If we denote by $\mathbf{V}_{-j}$, to the collection of available $\mathbf{v}_l$'s excluding $\mathbf{v}_j$, after some algebra it is easy to obtain

$$\mathbb{p}(\mathbf{v}_j \mid \ldots) \propto \mathbb{p}(\mathbf{v}_j \mid \mathbf{V}_{-j}) \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\}. \tag{5.27}$$

where

$$\mathbf{a}_j = \max_{\{i : \mathbf{d}_i = j\}} \left\{\frac{\mathbf{u}_i}{\prod_{l < \mathbf{d}_i} (1 - \mathbf{v}_l)}\right\} \tag{5.28}$$

and

$$\mathbf{b}_j = 1 - \max_{\{i : \mathbf{d}_i > j\}} \left\{\frac{\mathbf{u}_i}{\mathbf{v}_{\mathbf{d}_i} \prod_{l < \mathbf{d}_i, l \neq j} (1 - \mathbf{v}_l)}\right\}, \tag{5.29}$$

with the convention that $\max \emptyset = 0$. Evidently, for $j > \mathbf{k} = \max\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ the posterior distribution $\mathbb{p}(\mathbf{v}_j \mid \ldots)$ coincides with the prior $\mathbb{p}(\mathbf{v}_j \mid \mathbf{V}_{-j})$.

Remark 5.2. *Once we have updated $\mathbf{v}_l$ for every $l < j$ we can set $\mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_l)$ and latter compute the random sets $\boldsymbol{\Psi}$'s. Note that we do not require to update $\boldsymbol{\xi}_j$, $\mathbf{v}_j$ and $\mathbf{w}_j$ for every $j \geq 1$, it suffices to sample them for $j \leq \mathbf{m}$, where $\mathbf{m}$ is the first natural number that satisfies $\sum_{j=1}^{\mathbf{m}} \mathbf{w}_j \geq \max_i(1 - \mathbf{u}_i)$, then is not possible that $\mathbf{u}_i < \mathbf{w}_j$ for any $i \leq n$ and $j > m$. This way we can completely define the random sets $\boldsymbol{\Psi}_i$'s and the updating of $(\mathbf{d}_i)_{i=1}^n$ can take place.*

Algorithm 4: First T iterations of a Gibbs sampler algorithm for stick-breaking priors using slices

```
Initialize $\mathbf{d}_i = \mathbf{d}_i^{(0)} \in \mathbb{N}$ for every $i \in \{1, \ldots, n\}$ and $\mathbf{w}_j = \mathbf{w}_j^{(0)}$ for
  $j \leq \max\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$
for $t \in \{1, \ldots, T\}$ do
    for $i \in \{1, \ldots, n\}$ do
        Sample $\mathbf{u}_i$ from (5.24);
        Set $\mathbf{u}_i^{(t)} = \mathbf{u}_i$;
    Set $\mathbf{m} = \max\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$;
    for $j \leq \mathbf{m}$ do
        Sample $\boldsymbol{\xi}_j$ from (5.22);
        Sample $\mathbf{v}_j$ from (5.27);
        Set $\mathbf{w}_j^{(t)} = \mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_l)$ and $\boldsymbol{\xi}_j^{(t)} = \boldsymbol{\xi}_j$;
    while $\sum_{j=1}^{\mathbf{m}} \mathbf{w}_j < \max_i(1 - \mathbf{u}_i)$ do
        Set $\mathbf{m} = \mathbf{m} + 1$;
        Sample $\boldsymbol{\xi}_{\mathbf{m}}$ from $\mu_0$;
        Sample $\mathbf{v}_{\mathbf{m}}$ from $\mathbb{p}(\mathbf{v}_{\mathbf{m}} \mid \mathbf{V}_{-\mathbf{m}})$;
        Set $\mathbf{w}_{\mathbf{m}} = \mathbf{w}_{\mathbf{m}}^{(t)} = \mathbf{v}_{\mathbf{m}} \prod_{l<\mathbf{m}}(1 - \mathbf{v}_l)$, and $\boldsymbol{\xi}_{\mathbf{m}}^{(t)} = \boldsymbol{\xi}_{\mathbf{m}}$;
    Set $\mathbf{m}^{(t)} = \mathbf{m}$;
    if a prior distribution has been assigned to an hyper-parameter $\boldsymbol{\lambda}$: then
        Sample $\boldsymbol{\lambda}$ from $\mathbb{p}(\boldsymbol{\lambda} \mid \ldots)$;
        Set $\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}$;
    for $i \in \{1, \ldots, n\}$ do
        Define $\boldsymbol{\Psi}_i = \{j : \mathbf{u}_i < \mathbf{w}_j\}$;
        Sample $\mathbf{d}_i$ from (5.23);
        Set $\mathbf{d}_i^{(t)} = \mathbf{d}_i$;
Result: The Markov chain
    $\left(\mathbf{d}_1^{(t)}, \ldots, \mathbf{d}_n^{(t)}, \mathbf{u}_1^{(t)}, \ldots, \mathbf{u}_n^{(t)}, \left(\mathbf{w}_j^{(t)}\right)_{j \leq \mathbf{m}^{(t)}}, \left(\boldsymbol{\xi}_j^{(t)}\right)_{j \leq \mathbf{m}^{(t)}}, \boldsymbol{\lambda}^{(t)}\right)_{t=1}^{T}$.
```

Let us consider some examples that are extremely relevant to us because they specialize this Gibbs sampler algorithm to some models studied in Section 4.

Example 5.2 (Updating the length variables of a Dirichlet process)**.** *If the species sampling process $\boldsymbol{\mu} = \sum_{j \geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$ is a Dirichlet process with total mass parameter θ. We know, from Definition 3.9, that we may decompose $\mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_j)$ for the length variables $(\mathbf{v}_j)_{j \geq 1} \overset{iid}{\sim} \mathsf{Be}(1, \theta)$. In this case (5.27) simplifies to*

$$\mathbb{p}(\mathbf{v}_j \mid \ldots) \propto \mathsf{Be}(\mathbf{v}_j \mid 1, \theta) \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\}.$$

So to update $\mathbf{v}_j$, for $j \leq \mathbf{k} = \max\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$, we simply sample it from a truncated Beta distribution. To do this we can compute the posterior cumulative distribution function of $\mathbf{v}_j$,

$$F_j(\mathbf{v}_j) = \frac{(1 - \mathbf{a}_j)^{\theta} - (1 - \mathbf{v}_j)^{\theta}}{(1 - \mathbf{a}_j)^{\theta} - (1 - \mathbf{b}_j)^{\theta}}$$

and use the inverse-sampling technique, that is we first sample a uniform random variable

$\mathbf{z}_j \sim \mathsf{Unif}(0,1)$ *and then set* $\mathbf{v}_j = F_j^{-1}(\mathbf{z}_j)$*, where* F_j^{-1} *is the inverse function of* F_j*. Clearly for* $j > \mathbf{k}$*, to update* $\mathbf{v}_j$ *we sample it from its prior distribution* $\mathsf{Be}(\mathbf{v}_j \mid 1, \theta)$*.*

Example 5.3 (Updating the length variables of a Dirichlet driven stick-breaking process (DSB))**.** *If* $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j} \sim \mathcal{DSB}(\beta, \theta, \mu_0)$*, is a DSB (see Definition 4.2) then we may write* $\mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_j)$ *for some exchangeable length variables* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu}$ *where* $\boldsymbol{\nu} \sim \mathcal{D}(\beta, \mathsf{Be}(1,\theta))$ *is a Dirichlet process with total mass parameter* β *and base measure* $\mathsf{Be}(1,\theta)$*. By Theorem 3.6, Definition 3.7, and exploiting the exchangeability of the length variables, we know that a priori, for any finite subset,* $\mathbf{V}_{-j}$*, of the length variables that does not contain* $\mathbf{v}_j$*,*

$$\mathbb{p}(\mathbf{v}_j \mid \mathbf{V}_{-j}) = \sum_{l=1}^{\mathbf{K}} \frac{\mathbf{n}_l}{\mathbf{n}+\beta}\delta_{\mathbf{v}_l^*}(\mathbf{v}_j) + \frac{\beta}{\mathbf{n}+\beta}\mathsf{Be}(\mathbf{v}_j \mid 1, \theta), \tag{5.30}$$

where $\mathbf{n} = |\mathbf{V}_{-j}|$*,* $\mathbf{K}$ *is the number of distinct values in* $\mathbf{V}_{-j}$*,* $\mathbf{v}_1^*, \ldots, \mathbf{v}_{\mathbf{K}}^*$ *are such distinct values and* $\mathbf{n}_l = |\{\mathbf{v}_i \in \mathbf{V}_{-j} : \mathbf{v}_i = \mathbf{v}_l^*\}|$*. Now, by Remark 5.2 below, we know that when updating* $\mathbf{v}_j$*, the set of available length variables,* $\mathbf{V}_{-j}$*, is finite at each iteration, hence* (5.27) *specializes to*

$$\mathbb{p}(\mathbf{v}_j \mid \ldots) \propto \sum_{l\in\mathbf{C}_j} \mathbf{n}_l\, \delta_{\mathbf{v}_l^*} + \beta\left[(1-\mathbf{a}_j)^\theta - (1-\mathbf{b}_j)^\theta\right] f_j(\mathbf{v}_j) \tag{5.31}$$

where $\mathbf{C}_j = \{l : \mathbf{a}_j < \mathbf{v}_l^* < \mathbf{b}_j\}$ *and* f_j *denotes the density*

$$f_j(\mathbf{v}_j) = \frac{\theta(1-\mathbf{v}_j)^{\theta-1}}{(1-\mathbf{a}_j)^\theta - (1-\mathbf{b}_j)^\theta}\mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\}.$$

This way, to update $\mathbf{v}_j$*, we set* $\mathbf{v}_j = \mathbf{v}_l^*$ *for* $l \in \mathbf{C}_j$ *with probability*

$$\mathbf{q}_l = \frac{\mathbf{n}_j}{\sum_{l\in\mathbf{C}_j} \mathbf{n}_l + \beta\left[(1-\mathbf{a}_j)^\theta - (1-\mathbf{b}_j)^\theta\right]},$$

or we sample $\mathbf{v}_j$ *from the truncated Beta distribution, with probability*

$$\mathbf{q}_0 = \frac{\beta\left[(1-\mathbf{a}_j)^\theta - (1-\mathbf{b}_j)^\theta\right]}{\sum_{l\in\mathbf{C}_j} \mathbf{n}_l + \beta\left[(1-\mathbf{a}_j)^\theta - (1-\mathbf{b}_j)^\theta\right]}.$$

When implementing DSB priors, motivated by Corollaries 4.5 and 4.9, it might be of interest to estimate the underlying tie probability $\rho_\nu = \mathbb{P}[\mathbf{v}_j = \mathbf{v}_l] = 1/(\beta+1)$*, in order to do this we can consider this quantity random and assign it a prior,* $\mathbb{p}(\boldsymbol{\rho_\nu})$*. By doing so, roughly speaking, we allow the model to choose between DSB priors that behaves arbitrarily similar to a Dirichlet prior, to a Geometric prior or somewhere in between. If we decide to assign a prior distribution to this hyper-parameter, it is straightforward to check that the full conditionals of* $\boldsymbol{\xi}_j$*,* $\mathbf{d}_i$ *and* $\mathbf{u}_i$ *will remain unchanged, as to the full conditional of* $\mathbf{v}_j$*, it will be as described above conditionally given* $\beta = \boldsymbol{\beta} = (1-\boldsymbol{\rho_\nu})/\boldsymbol{\rho_\nu}$*. In this circumstance we will require to update* $\boldsymbol{\rho_\nu}$ *at each iteration of the Gibbs sampler. Since* $\boldsymbol{\rho_\nu}$ *only affects directly the length variables, it is easy to see that the full conditional distribution of this random variable is*

$$\mathbb{p}(\boldsymbol{\rho_\nu} \mid \ldots) \propto \mathbb{p}(\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\rho_\nu})\mathbb{p}(\boldsymbol{\rho_\nu})$$

Being that at each iteration of the Gibbs sampler we only sample finitely many length variables (see Remark 5.2) say $\mathbf{v}_1, \dots, \mathbf{v}_m$ *and from Theorem 3.6 we obtain*

$$\mathbb{p}(\boldsymbol{\rho_\nu} \mid \dots) \propto \pi(\mathbf{n}_1, \dots \mathbf{n}_{\mathbf{K}_m})\mathbb{p}(\boldsymbol{\rho_\nu})$$

where $\mathbf{K}_m$ *is the number of distinct values* $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ *exhibits,* $\mathbf{n}_1, \dots, \mathbf{n}_{\mathbf{K}_m}$ *are the frequencies of the distinct values and* π *is the EPPF of the Dirichlet process. Hence,*

$$\mathbb{p}(\boldsymbol{\rho_\nu} \mid \dots) \propto \frac{(1-\boldsymbol{\rho_\nu})^{\mathbf{K}_m - 1}\, \boldsymbol{\rho_\nu}^{m-\mathbf{K}_m}}{\prod_{i=0}^{m-2}(1+i\boldsymbol{\rho_\nu})}\mathbb{p}(\boldsymbol{\rho_\nu}).$$

Drawing samples from the full conditional of $\boldsymbol{\rho_\nu}$ *is possible with the aid of a rejection sampling method such as Adaptive Rejection Metropolis Sampling (ARMS) (Gilks et al.; 1995).*

Example 5.4 (Updating the length variables of a spike and slab stick-breaking process (SSB))**.** *If* $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$ *is a spike and slab stick-breaking process with parameters* $(\mathrm{p}, \mathsf{Be}(1,\theta), \mu_0)$ *(see Definition 4.6) we can decompose* $\mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1-\mathbf{v}_l)$ *for every* $j \geq 1$ *where* $(\mathbf{v}_j)_{j\geq 1}$ *is a Markov chain with initial and stationary distribution* $\mathsf{Be}(1,\theta)$ *and transition*

$$\mathbb{p}(\mathbf{v}_{j+1} \mid \mathbf{v}_j) = \mathrm{p}\, \delta_{\mathbf{v}_j}(\mathbf{v}_{j+1}) + (1-\mathrm{p})\mathsf{Be}(\mathbf{v}_{j+1} \mid 1, \theta).$$

Exploiting the Markov property of the length variables it is easy to see that for $j = 1$*,* (5.27) *specializes to*

$$\begin{aligned}
\mathbb{p}(\mathbf{v}_1 \mid \dots) &\propto \mathbb{p}(\mathbf{v}_1)\mathbb{p}(\mathbf{v}_2 \mid \mathbf{v}_1)\mathbf{1}\{\mathbf{a}_1 < \mathbf{v}_1 < \mathbf{b}_1\} \\
&\propto \mathsf{Be}(\mathbf{v}_1 \mid 1,\theta)\left\{\mathrm{p}\, \mathbf{1}_{\{\mathbf{v}_1 = \mathbf{v}_2\}} + (1-\mathrm{p})\mathsf{Be}(\mathbf{v}_2 \mid 1,\theta)\right\}\mathbf{1}\{\mathbf{a}_1 < \mathbf{v}_1 < \mathbf{b}_1\} \\
&\propto \mathrm{p}\mathbf{1}\{\mathbf{a}_1 < \mathbf{v}_2 < \mathbf{b}_1\}\mathsf{Be}(\mathbf{v}_2 \mid 1,\theta)\delta_{\mathbf{v}_2}(\mathbf{v}_1) \\
&\qquad + (1-\mathrm{p})\mathsf{Be}(\mathbf{v}_2 \mid 1,\theta)\mathsf{Be}(\mathbf{v}_1 \mid 1,\theta)\mathbf{1}\{\mathbf{a}_1 < \mathbf{v}_1 < \mathbf{b}_1\}.
\end{aligned}$$

Hence to update $\mathbf{v}_1$ *we set* $\mathbf{v}_1 = \mathbf{v}_2$ *with probability proportional to*

$$\mathbf{q}_2 = \mathrm{p}\mathbf{1}\{\mathbf{a}_1 < \mathbf{v}_2 < \mathbf{b}_1\},$$

or with probability proportional to

$$\mathbf{q}_0 = (1-\mathrm{p})\left[(1-\mathbf{a}_1)^\theta - (1-\mathbf{b}_1)^\theta\right]$$

we sample $\mathbf{v}_1$ *from a truncated Beta distribution. Now, for* $j \geq 2$*, we get that* (5.27) *reduces to*

$$\begin{aligned}
\mathbb{p}(\mathbf{v}_j \mid \dots) &\propto \mathbb{p}(\mathbf{v}_j \mid \mathbf{v}_{j-1})\mathbb{p}(\mathbf{v}_{j+1} \mid \mathbf{v}_j)\mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\} \\
&\propto \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\}\left\{\mathrm{p}\, \mathbf{1}_{\{\mathbf{v}_j = \mathbf{v}_{j-1}\}} + (1-\mathrm{p})\mathsf{Be}(\mathbf{v}_j \mid 1,\theta)\right\} \times \\
&\qquad\qquad \left\{\mathrm{p}\, \mathbf{1}_{\{\mathbf{v}_j = \mathbf{v}_{j+1}\}} + (1-\mathrm{p})\mathsf{Be}(\mathbf{v}_{j+1} \mid 1,\theta)\right\} \\
&\propto \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j-1} = \mathbf{v}_{j+1} < \mathbf{b}_j\}\mathrm{p}^2\delta_{\mathbf{v}_{j-1}}(\mathbf{v}_j) + \\
&\qquad \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j-1} < \mathbf{b}_j\}\mathrm{p}(1-\mathrm{p})\mathsf{Be}(\mathbf{v}_{j+1} \mid 1,\theta)\delta_{\mathbf{v}_{j-1}}(\mathbf{v}_j) + \\
&\qquad \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j+1} < \mathbf{b}_j\}\mathrm{p}(1-\mathrm{p})\mathsf{Be}(\mathbf{v}_{j+1} \mid 1,\theta)\delta_{\mathbf{v}_{j+1}}(\mathbf{v}_j) + \\
&\qquad (1-\mathrm{p})^2\mathsf{Be}(\mathbf{v}_{j+1} \mid 1,\theta)\mathsf{Be}(\mathbf{v}_j \mid 1,\theta)\mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\}.
\end{aligned}$$

Thus to update $\mathbf{v}_j$*, with probability proportional to*

$$\mathbf{q}_{j-1} = \frac{\mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j-1} = \mathbf{v}_{j+1} < \mathbf{b}_j\}\mathrm{p}^2}{\mathsf{Be}(\mathbf{v}_{j+1} \mid 1, \theta)} + \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j-1} < \mathbf{b}_j\}\mathrm{p}(1-\mathrm{p})$$

we set $\mathbf{v}_j = \mathbf{v}_{j-1}$*, with probability proportional to*

$$\mathbf{q}_{j+1} = \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j+1} < \mathbf{b}_j\}\mathrm{p}(1-\mathrm{p})$$

we set $\mathbf{v}_j = \mathbf{v}_{j+1}$*, or with probability proportional to*

$$\mathbf{q}_0 = (1-\mathrm{p})^2 \left[(1-\mathbf{a}_1)^\theta - (1-\mathbf{b}_1)^\theta\right]$$

we sample $\mathbf{v}_j$ *from a truncated Beta distribution. If* $\mathbf{v}_{j+1}$ *is not given when updating* $\mathbf{v}_j$*, the full conditional of the random variable in question is slightly different. In this case we have that*

$$\begin{aligned}\mathbb{p}(\mathbf{v}_j \mid \ldots) &\propto \mathbb{p}(\mathbf{v}_j \mid \mathbf{v}_{j-1})\mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\} \\ &\propto \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j-1} < \mathbf{b}_j\}\mathrm{p}\,\delta_{\mathbf{v}_{j-1}}(\mathbf{v}_j) + (1-\mathrm{p})\mathbf{1}\{\mathbf{a}_j < \mathbf{v}_j < \mathbf{b}_j\}\mathsf{Be}(\mathbf{v}_j \mid 1, \theta).\end{aligned}$$

This means that with probability proportional to

$$\mathbf{q}_{j-1} = \mathbf{1}\{\mathbf{a}_j < \mathbf{v}_{j-1} < \mathbf{b}_j\}\mathrm{p}$$

we fix $\mathbf{v}_j = \mathbf{v}_{j-1}$*, or with probability proportional to*

$$\mathbf{q}_0 = (1-\mathrm{p}) \left[(1-\mathbf{a}_1)^\theta - (1-\mathbf{b}_1)^\theta\right]$$

we sample $\mathbf{v}_j$ *from the corresponding truncated Beta distribution.*

As is the case of DSBs, when implementing SSB priors it can be interesting to estimate the parameter p*, this due to Corollary 4.22. Indeed, by considering* $\mathbf{p}$ *random and choosing a prior for it, we allow the model to choose between SSB priors that are similar to a Dirichlet process, to a Geometric process or some SSP in between. If we decide to regard* $\mathbf{p}$ *as an extra random variable, it is easy to see that the full conditionals of* $\boldsymbol{\xi}_j$*,* $\mathbf{d}_i$ *and* $\mathbf{u}_i$ *remain identical, and the full conditional of* $\mathbf{v}_j$ *also remains the same by conditioning on* $\mathbf{p}$*. As to the full conditional of* $\mathbf{p}$ *we have that*

$$\begin{aligned}\mathbb{p}(\mathbf{p} \mid \ldots) &\propto \mathbb{p}(\mathbf{p})\mathbb{p}(\mathbf{v}_1)\prod_{j=2}^{\mathbf{m}} \mathbb{p}(\mathbf{v}_j \mid \mathbf{v}_{j-1}) \\ &\propto \mathbb{p}(\mathbf{p})\prod_{j=2}^{\mathbf{m}} \mathbf{p}\,\delta_{\mathbf{v}_{j-1}}(\mathbf{v}_j) + (1-\mathbf{p})\mathsf{Be}(\mathbf{v}_j \mid 1, \theta)\end{aligned}$$

where $(\mathbf{v}_j)_{j=1}^{\mathbf{m}}$ *are the updated* $\mathbf{v}_j$*'s. Note that since the Beta distribution if diffuse,* $\mathsf{Be}(\mathbf{v}_j \mid 1, \theta)\mathbf{1}_{\{\mathbf{v}_j \neq \mathbf{v}_{j-1}\}}$ *remains the density of a Beta distribution which means we can re-express*

$$\begin{aligned}\mathbb{p}(\mathbf{p} \mid \ldots) &\propto \mathbb{p}(\mathbf{p})\prod_{j=2}^{\mathbf{m}} \mathbf{p}\mathbf{1}_{\{\mathbf{v}_j = \mathbf{v}_{j-1}\}} + (1-\mathbf{p})\mathbf{1}_{\{\mathbf{v}_j \neq \mathbf{v}_{j-1}\}}\mathsf{Be}(\mathbf{v}_j \mid 1, \theta) \\ &\propto \mathbb{p}(\mathbf{p})\prod_{j=2}^{\mathbf{m}} \mathbf{p}^{\mathbf{1}_{\{\mathbf{v}_j = \mathbf{v}_{j-1}\}}}\left[(1-\mathbf{p})\mathsf{Be}(\mathbf{v}_j \mid 1, \theta)\right]^{\mathbf{1}_{\{\mathbf{v}_j \neq \mathbf{v}_{j-1}\}}} \\ &\propto \mathbb{p}(\mathbf{p})\mathbf{p}^{\sum_{j=2}^{\mathbf{m}} \mathbf{1}_{\{\mathbf{v}_j = \mathbf{v}_{j-1}\}}}(1-\mathbf{p})^{\sum_{j=2}^{\mathbf{m}} \mathbf{1}_{\{\mathbf{v}_j \neq \mathbf{v}_{j-1}\}}}.\end{aligned}$$

Thus if we pick $\mathbb{p}(\mathbf{p}) = \mathsf{Be}(\mathbf{p} \mid \alpha, \beta)$*, then to sample* $\mathbf{p}$ *from its full conditional we simply have to sample its from a* $\mathsf{Be}\left(\alpha + \sum_{j=2}^{\mathbf{m}} \mathbf{1}_{\{\mathbf{v}_j = \mathbf{v}_{j-1}\}}, \beta + \sum_{j=2}^{\mathbf{m}} \mathbf{1}_{\{\mathbf{v}_j \neq \mathbf{v}_{j-1}\}}\right)$ *distribution.*

For some stick-breaking priors the slice sampler can be modified as suggested by Kalli et al. (2011), so instead of updating $(\mathbf{u}_i)_{i=1}^n$ separately from $(\mathbf{v}_j)_{j\geq 1}$ we update them as block.

Updating $(\mathbf{u}_i)_{i=1}^n$ **and** $\mathbf{V} = (\mathbf{v}_j)_{j\geq 1}$ **as a block:**

$$\begin{aligned}
\mathbb{p}(\mathbf{u}_1, \ldots, \mathbf{u}_n, \mathbf{V} \mid \ldots) &\propto \mathbb{p}(\mathbf{V}) \prod_{i=1}^{n} \mathbf{1}_{\{\mathbf{u}_i < \mathbf{w}_{\mathbf{d}_i}\}} \\
&\propto \mathbb{p}(\mathbf{V}) \prod_{i=1}^{n} \mathbf{w}_{\mathbf{d}_i} \mathbf{1}_{\{\mathbf{u}_i < \mathbf{w}_{\mathbf{d}_i}\}} \mathbf{w}_{\mathbf{d}_i}^{-1} \\
&\propto \left\{ \mathbb{p}(\mathbf{V}) \prod_{j=1}^{\mathbf{k}} \mathbf{v}_j^{\mathbf{s}_j} (1 - \mathbf{v}_j)^{\mathbf{r}_j} \right\} \left\{ \prod_{i=1}^{n} \mathsf{Unif}(\mathbf{u}_i \mid 0, \mathbf{w}_{\mathbf{d}_i}) \right\},
\end{aligned}$$

where $\mathbf{s}_j = \sum_{i=1}^n \mathbf{1}_{\{\mathbf{d}_i = j\}}$, $\mathbf{r}_j = \sum_{i=1}^n \mathbf{1}_{\{\mathbf{d}_i > j\}}$ and $\mathbf{k} = \max\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$. This way we can first update the length variables from

$$\mathbb{p}(\mathbf{V} \mid \ldots (\text{exclude } (\mathbf{u}_i)_{i=1}^n) \ldots) \propto \mathbb{p}(\mathbf{V}) \prod_{j=1}^{\mathbf{k}} \mathbf{v}_j^{\mathbf{s}_j} (1 - \mathbf{v}_j)^{\mathbf{r}_j}, \tag{5.32}$$

and then given $\mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_l)$ for every $j \leq \mathbf{k}$, sample $(\mathbf{u}_i)_{i\geq 1}$ from

$$\mathbb{p}(\mathbf{u}_1, \ldots, \mathbf{u}_n \mid \ldots) = \prod_{i=1}^{n} \mathsf{Unif}(\mathbf{u}_i \mid 0, \mathbf{w}_{\mathbf{d}_i}). \tag{5.33}$$

That is we sample $\mathbf{u}_i \sim \mathsf{Unif}(\mathbf{u}_i \mid 0, \mathbf{w}_{\mathbf{d}_i})$ independently for $i \in \{1, \ldots, n\}$.

This modified Gibbs sampler is particularly useful when there exist a latent random element, $\mathbf{Z}$, such that given $\mathbf{Z}$ (a priori), the length variables are independent and Beta distributed, this is for every $m \in \mathbb{N}$

$$\mathbb{p}(\mathbf{v}_1, \ldots, \mathbf{v}_m \mid \mathbf{Z}) \propto \prod_{j=1}^{m} \mathsf{Be}(\mathbf{v}_j \mid \alpha_j(\mathbf{Z}), \beta_j(\mathbf{Z}))$$

where α_j and β_j are measurable functions of $\mathbf{Z}$, for every $j \geq 1$. In this case conditional conjugacy is attained for the length variables, hence, taking into account $\mathbf{Z}$, we can update the first $m \geq \mathbf{k}$ length variables from

$$\begin{aligned}
\mathbb{p}(\mathbf{v}_1, \ldots, \mathbf{v}_m \mid \ldots (\text{exclude } (\mathbf{u}_i)_{i=1}^n) \ldots) &\propto \mathbb{p}(\mathbf{v}_1, \ldots, \mathbf{v}_m \mid \mathbf{Z}) \prod_{j=1}^{\mathbf{k}} \mathbf{v}_j^{\mathbf{s}_j} (1 - \mathbf{v}_j)^{\mathbf{r}_j} \\
&\propto \prod_{j=1}^{m} \mathsf{Be}(\mathbf{v}_j \mid \alpha_j(\mathbf{Z}), \beta_j(\mathbf{Z})) \prod_{j=1}^{\mathbf{k}} \mathbf{v}_j^{\mathbf{s}_j} (1 - \mathbf{v}_j)^{\mathbf{r}_j} \\
&\propto \prod_{j=1}^{m} \mathbf{v}_j^{\alpha_j(\mathbf{Z}) + \mathbf{s}_j - 1} (1 - \mathbf{v}_j)^{\beta_j(\mathbf{Z}) + \mathbf{r}_j - 1},
\end{aligned} \tag{5.34}$$

where, $\mathbf{s}_j = 0 = \mathbf{r}_j$ for $j > \mathbf{k}$. This means that to update the length variables, we can sample independently for $j \leq m$, $\mathbf{v}_j \sim \mathsf{Be}(\alpha_j(\mathbf{Z}) + \mathbf{s}_j, \beta_j(\mathbf{Z}) + \mathbf{r}_j)$. Now, if we are expanding to model to include $\mathbf{Z}$, clearly we must update this random element at each iteration of the Gibbs sampler. Whenever $\mathbf{Z}$ is conditionally independent of the rest of the random elements involved, given $\mathbf{V}$, its full conditional distribution is given by

$$\mathbb{p}(\mathbf{Z} \mid \ldots) \propto \mathbb{p}(\mathbf{V} \mid \mathbf{Z})\mathbb{p}(\mathbf{Z}).$$

If the mentioned conditional independence holds then we also have that the remaining full conditionals are not affected by including $\mathbf{Z}$. Let us consider some examples.

Example 5.5 (Updating independent Beta distributed length variables)**.** *Say that $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$ is a proper species sampling process, with stick-breaking weights $\mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_j)$ for some independent length variables $\mathbf{v}_j \sim \mathsf{Be}(\alpha_j, \beta_j)$. For this species sampling prior, to update $(\mathbf{u}_i)_{i=1}^n$ and $(\mathbf{v}_j)_{j\geq 1}$ as a block, we can first sample $\mathbf{v}_j \sim \mathsf{Be}(\alpha_j + \mathbf{s}_j, \beta_j + \mathbf{r}_j)$ for $j \leq m$, where $m > \mathbf{k} = \max\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$ and latter sample $\mathbf{u}_i \sim \mathsf{Unif}(\mathbf{u}_i \mid 0, \mathbf{w}_{\mathbf{d}_i})$ independently for $i \in \{1, \ldots, n\}$. This method can be applied for SSP such as the Dirichlet and the Pitman-Yor process.*

Example 5.6 (Updating the length variables of a Beta-Binomial stick breaking prior (BBSB))**.** *If $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$ is a BBSB process with parameters $(\alpha, \beta, \kappa, \mu_0)$ (see Definition 4.5) then $\mathbf{w}_j = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_j)$ for every $j \geq 1$, where $(\mathbf{v}_j)_{j\geq 1}$ is Markov chain with initial and stationary distribution $\mathsf{Be}(\alpha, \theta)$, and Beta-Binomial transition as introduced in Definition 4.4. As explained below this definition, we can define a latent Markov chain $\mathbf{Z} = (\mathbf{z}_j)_{j\geq 1}$ such that for every $j \geq 1$ $\{\mathbf{z}_j \mid \mathbf{v}_j\} \sim \mathsf{Bin}(\kappa, \mathbf{v}_j)$ and $\{\mathbf{v}_{j+1} \mid \mathbf{z}_j\} \sim \mathsf{Be}(\alpha + \mathbf{z}_j, \theta + \kappa - \mathbf{z}_j)$, where $\mathbf{z}_j$ is conditionally independent of $(\mathbf{v}_l, \mathbf{z}_j)_{l=1}^{j-1}$ given $\mathbf{v}_j$, and $\mathbf{v}_{j+1}$ is conditionally independent of $(\mathbf{v}_l)_{l=1}^{j}$ and $(\mathbf{z}_l)_{l=1}^{j-1}$ given $\mathbf{z}_j$. In other words, $(\mathbf{v}_j, \mathbf{z}_j)_{j\geq 1}$ is a Markov chain with initial and stationary distribution*

$$\mathbb{p}(\mathbf{z}_1, \mathbf{x}_1) = \mathsf{Be}(\mathbf{v}_1 \mid \alpha, \theta)\mathsf{Bin}(\mathbf{z}_1 \mid \kappa, \mathbf{v}_1) \tag{5.35}$$

and one step ahead transition

$$\mathbb{p}(\mathbf{v}_{j+1}, \mathbf{z}_{j+1} \mid \mathbf{v}_j, \mathbf{z}_j) = \mathsf{Be}(\mathbf{v}_{j+1} \mid \alpha + \mathbf{z}_j, \theta + \kappa - \mathbf{z}_j)\mathsf{Bin}(\mathbf{z}_{j+1} \mid \kappa, \mathbf{v}_{j+1}). \tag{5.36}$$

Evidently, integrating over $\mathbf{Z}$ we recover the Markov chain $\mathbf{V} = (\mathbf{v}_j)_{j\geq 1}$, whilst integrating over $\mathbf{V}$ we obtain that $\mathbf{Z} = (\mathbf{z}_j)_{j\geq 1}$ is Markov chain with transition

$$\begin{aligned}\mathbb{p}(\mathbf{z}_{j+1} \mid \mathbf{z}_j) &= \int \mathsf{Bin}(\mathbf{z} \mid \kappa, \mathbf{v})\mathsf{Be}(\mathbf{v} \mid \alpha, \theta) d\mathbf{v} \\ &= \binom{\kappa}{\mathbf{z}_{j+1}} \frac{(\alpha + \mathbf{z}_j)_{\mathbf{z}_{j+1}} (\theta + \kappa - \mathbf{z}_j)_{\kappa - \mathbf{z}_{j+1}}}{(\alpha + \theta + \kappa)_\kappa} \mathbf{1}_{\{\mathbf{z}_{j+1} \in \{0,\ldots,\kappa\}\}}\end{aligned}$$

and initial and stationary distribution

$$\mathbb{p}(\mathbf{z}_1) = \binom{\kappa}{\mathbf{z}_1} \frac{(\alpha)_{\mathbf{z}_1} (\theta)_{\kappa - \mathbf{z}_1}}{(\alpha + \theta)_\kappa} \mathbf{1}_{\{\mathbf{z}_{j+1} \in \{0,\ldots,\kappa\}\}}.$$

Exploiting the reversibility of $(\mathbf{v}_j, \mathbf{z}_j)_{j\geq 1}$ it is easy to see that for every $m \in \mathbb{N}$

$$\mathbb{p}(\mathbf{v}_1, \ldots, \mathbf{v}_m \mid \mathbf{Z}) = \prod_{j=1}^{m} \mathsf{Be}(\mathbf{v}_j \mid \alpha_j(\mathbf{Z}), \beta_j(\mathbf{Z}))$$

where $\alpha_1(\mathbf{Z}) = \alpha + \mathbf{z}_1$, $\beta_1(\mathbf{Z}) = \theta + \kappa - \mathbf{z}_1$ and for $j \geq 2$, $\alpha_j(\mathbf{Z}) = \alpha + \mathbf{z}_{j-1} + \mathbf{z}_j$ and $\beta_j(\mathbf{Z}) = \theta + 2\kappa - \mathbf{z}_{j-1} - \mathbf{z}_j$. Hence, we have augmented the model through the inclusion of $\mathbf{Z}$ in such way that the following hold:

a) *The length variables are independent and Beta distributed given $\mathbf{Z}$.*

b) *The conditional distribution of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ given $\mathbf{Z}$ depends only on finitely many elements of $\mathbf{Z}$.*

c) *The rest of the random elements are conditionally independent of $\mathbf{Z}$ given $\mathbf{V}$.*

With these considerations taken into account, when implementing the slice sampler for this model we can update $(\mathbf{v}_j)_{j\geq 1}$ and $(\mathbf{u}_i)_{i=1}^n$ as a block by sampling them from (5.34) and (5.33), respectively, where the measurable functions α_j and β_j are as defined above in this example. Due to (c)*, the full conditionals of $(\boldsymbol{\xi}_j)_{j\geq 1}$ and $(\mathbf{d}_i)_{i=1}^n$ are not affected by including $\mathbf{Z}$. Also as a consequence of* (c) *and equations* (5.35) *and* (5.36)*, the full conditional distribution of the first m elements of $\mathbf{Z}$ is*

$$\begin{aligned}\mathbb{p}(\mathbf{z}_1, \ldots, \mathbf{z}_m \mid \ldots) &\propto \prod_{i=1}^{m} \mathsf{Bin}(\mathbf{z}_j \mid \kappa, \mathbf{v}_j)\mathsf{Be}(\mathbf{v}_{j+1} \mid \alpha + \mathbf{z}_j, \theta + \kappa - \mathbf{z}_j)\\ &\propto \prod_{i=1}^{m} \frac{(\mathbf{v}_j\mathbf{v}_{j+1})^{\mathbf{z}_j}[(1-\mathbf{v}_j)(1-\mathbf{v}_{j+1})]^{\kappa-\mathbf{z}_j}}{\mathbf{z}_j!(\kappa-\mathbf{z}_j)!(\alpha)_{\mathbf{z}_j}(\theta)_{\kappa-\mathbf{z}_j}}\mathbf{1}_{\{\mathbf{z}_j\in\{0,\ldots,\kappa\}\}}.\end{aligned}$$

This means that to update $\mathbf{z}_j$, we sample it independently from

$$\mathbb{p}(\mathbf{z}_j \mid \ldots) \propto \frac{(\mathbf{v}_j\mathbf{v}_{j+1})^{\mathbf{z}_j}[(1-\mathbf{v}_j)(1-\mathbf{v}_{j+1})]^{\kappa-\mathbf{z}_j}}{\mathbf{z}_j!(\kappa-\mathbf{z}_j)!(\alpha)_{\mathbf{z}_j}(\theta)_{\kappa-\mathbf{z}_j}}\mathbf{1}_{\{\mathbf{z}_j\in\{0,\ldots,\kappa\}\}},$$

which is a discrete distribution with finite support, therefore it is easy to draw samples from $\mathbb{p}(\mathbf{z}_j \mid \ldots)$. Summarizing, we know how to update each random element involved, moreover Remark 5.2 remains true for this modified slice sampler, this together with (b) *imply that only finitely many random elements need to be updated at each iteration of the Gibbs sampler, hence it is practically feasible to implement the algorithm for BBSB priors.*

As done for DSBs and SSBs (see Examples 5.3 and 5.4), motivated by Corollary 4.18, we can put a prior on the tuning parameter, which in this case is $\boldsymbol{\kappa} = \kappa$, to estimate it. If we do this, the full conditional of $\boldsymbol{\kappa}$ is

$$\mathbb{p}(\boldsymbol{\kappa} \mid \ldots) \propto \mathbb{p}(\boldsymbol{\kappa}) \prod_{j=1}^{m} \mathsf{Be}(\mathbf{v}_{j+1} \mid \alpha + \mathbf{z}_j, \theta + \kappa - \mathbf{z}_j)\mathsf{Bin}(\mathbf{z}_j \mid \kappa, \mathbf{z}_j).$$

where $(\mathbf{v}_j, \mathbf{z}_j)_{j=1}^m$ are updated $\mathbf{v}_j$'s and $\mathbf{z}_j$'s. Being that $\boldsymbol{\kappa}$ takes values in $\mathbb{N}$, sampling from its full conditional can be quiet involved. However, if the prior $\mathbb{p}(\boldsymbol{\kappa})$ has finite support, sampling from $\mathbb{p}(\boldsymbol{\kappa} \mid \ldots)$ becomes easy.

Regardless of the particular specifications of the stick-breaking prior, the general algorithm for this modified slice sampler can be described as follows:

Algorithm 5: First T iterations of the modified Gibbs sampler algorithm for stick-breaking priors using slices

```
Initialize $\mathbf{d}_i = \mathbf{d}_i^{(0)} \in \mathbb{N}$ for every $i \in \{1, \ldots, n\}$, the latent object $\mathbf{Z} = \mathbf{Z}^{(0)}$, and
  the hyper-parameter $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(0)}$, if the model has been augmented to include
  them.
for $t \in \{1, \ldots, T\}$ do
    Set $\mathbf{m} = \max\{\mathbf{d}_1, \ldots, \mathbf{d}_n\}$;
    for $j \leq \mathbf{m}$ do
        Sample $\mathbf{v}_j$ from (5.32);
        Set $\mathbf{w}_j = \mathbf{w}_j^{(t)} = \mathbf{v}_j \prod_{l<j}(1 - \mathbf{v}_l)$;
    for $i \in \{1, \ldots, n\}$ do
        Sample $\mathbf{u}_i$ from (5.33);
        Set $\mathbf{u}_i^{(t)} = \mathbf{u}_i$;
    for $j \leq \mathbf{m}$ do
        Sample $\boldsymbol{\xi}_j$ from (5.22);
        Set $\boldsymbol{\xi}_j^{(t)} = \boldsymbol{\xi}_j$;
    while $\sum_{j=1}^{\mathbf{m}} \mathbf{w}_j < \max_i(1 - \mathbf{u}_i)$ do
        Set $\mathbf{m} = \mathbf{m} + 1$
        Sample $\boldsymbol{\xi}_{\mathbf{m}}$ from $\mu_0$;
        Sample $\mathbf{v}_{\mathbf{m}}$ from $\mathbb{p}(\mathbf{v}_{\mathbf{m}} \mid \mathbf{V}_{-\mathbf{m}}, \mathbf{Z})$;
        Set $\mathbf{w}_{\mathbf{m}} = \mathbf{w}_{\mathbf{m}}^{(t)} = \mathbf{v}_{\mathbf{m}} \prod_{l<\mathbf{m}}(1 - \mathbf{v}_l)$ and $\boldsymbol{\xi}_{\mathbf{m}} = \boldsymbol{\xi}_{\mathbf{m}}^{(t)}$;
    Set $\mathbf{m}^{(t)} = \mathbf{m}$;
    if the latent random element $\mathbf{Z}$ is defined then
        Sample $\mathbf{Z}$ from $\mathbb{p}(\mathbf{Z} \mid \ldots)$;
        Set $\mathbf{Z}^{(t)}$;
    if a prior distribution has been assigned to an hyper-parameter $\boldsymbol{\lambda}$: then
        Sample $\boldsymbol{\lambda}$ from $\mathbb{p}(\boldsymbol{\lambda} \mid \ldots)$;
        Set $\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}$;
    for $i \in \{1, \ldots, n\}$ do
        Define $\boldsymbol{\Psi}_i = \{j : \mathbf{u}_i < \mathbf{w}_j\}$;
        Sample $\mathbf{d}_i$ from (5.23);
        Set $\mathbf{d}_i^{(t)} = \mathbf{d}_i$;
Result: The Markov chain
  $\left(\mathbf{d}_1^{(t)}, \ldots, \mathbf{d}_n^{(t)}, \mathbf{u}_1^{(t)}, \ldots, \mathbf{u}_n^{(t)}, \left(\mathbf{w}_j^{(t)}\right)_{j \leq \mathbf{m}^{(t)}}, \left(\boldsymbol{\xi}_j^{(t)}\right)_{j \leq \mathbf{m}^{(t)}}, \mathbf{Z}^{(t)}, \boldsymbol{\lambda}^{(t)}\right)_{t=1}^{T}$.
```

Of course if the latent object $\mathbf{Z}$ was not introduced or no hyper-parameter was assigned a prior, we omit the steps where they are initialized and updated.

Despite whether we implement Algorithm 4 or 5, posterior inference for the estimated density and clustering structures, can be performed analogously as for the algorithm that uses random sets, by defining

$$\boldsymbol{\Psi}_i^{(t)} = \left\{ j : \mathbf{u}_i^{(t)} < \mathbf{w}_j^{(t)} \right\},$$

for every $i \in \{1, \ldots, n\}$. Now, if we chose to assign a prior distribution to some hyper-parameter, $\boldsymbol{\lambda}$ (such as $\boldsymbol{\rho_\nu}$ for DSBs, $\mathbf{p}$ for SSBs or $\boldsymbol{\kappa}$ for BBSBs) we can use the MAP $\boldsymbol{\lambda}^{(\hat{t})}$, or alternative, exploit the fact that $\left(\boldsymbol{\lambda}^{(t)}\right)_{t=T_0+1}^{T}$ are samples from $\mathbb{p}(\boldsymbol{\lambda} \mid \mathbf{y}_1, \ldots, \mathbf{y}_n)$, to estimate relevant features of the posterior distribution.

5.3 Illustrations

For this section we designed four small experiments to test the performance of the new Bayesian non-parametric priors, introduced in Section 4, and compare the results with the ones provided by their limiting processes: the Dirichlet and the Geometric priors. In all cases by means of the slice sampler we will adjust mixtures of Gaussian distributions to data points $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ that present no repetitions. Explicitly, we model the data $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ as i.i.d. sampled from $\boldsymbol{\phi} = \sum_{j\geq 1} \mathbf{w}_j \mathcal{K}(\cdot \mid \boldsymbol{\xi}_j)$, where $\mathcal{K}$ denotes a Gaussian distribution and the species sampling process $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$ is a DSB, a SSB, a BBSB, or one of the limiting processes. We will call $\boldsymbol{\phi}$ a DSB mixture whenever $\boldsymbol{\mu}$ is a DSB and analogously for the other classes of stick-breaking process. For the first three experiments we will be working with univariate data and fix the corresponding tuning parameter (ρ_ν, p or κ) to distinct deterministic values. The main objective of these experiments is to analyse the posterior impact of Theorems 4.3 and 4.14, by means of DSBs, SSBs and BBSBs. For univariate data, we will assume a Normal kernel with random location and scale parameters, i.e. $\boldsymbol{\xi}_j = (\mathbf{m}_j, \boldsymbol{\tau}_j)$, and $\mathcal{K}(\cdot \mid \boldsymbol{\xi}_j) = \mathsf{N}\left(\mathbf{m}_j, \boldsymbol{\tau}_j^{-1}\right)$, to attain conjugacy for $\mathcal{K}$ and μ_0 we assume that a priori $\boldsymbol{\xi}_j$ follows a Normal-Gamma distribution, this is $\mu_0(\boldsymbol{\xi}_j) = \mathsf{N}\left(\mathbf{m}_j \mid \mu, (\lambda\boldsymbol{\tau}_j)^{-1}\right) \mathsf{Ga}(\boldsymbol{\tau}_j \mid a, b)$. For the fourth and last experiment, we will assign a prior distribution to the tuning parameters, so the models can perform posterior inference and altogether chose between a Dirichlet process, a Geometric process, or some stick-breaking prior with dependent length variables amidst. For this last experiment, we will adjust mixtures to bivariate data. Here we consider $\mathcal{K}(\cdot \mid \boldsymbol{\xi}_j) = \mathsf{N}_2(\mathbf{m}_j, \boldsymbol{\Sigma}_j)$ and to achieve conjugacy we assume a Normal-inverse-Wishart prior for $\boldsymbol{\xi}_j = (\mathbf{m}_j, \boldsymbol{\Sigma}_j)$, so that $\mu_0(\boldsymbol{\xi}_j) = \mathsf{N}_2(\mathbf{m}_j \mid \mu, \lambda^{-1}\boldsymbol{\Sigma}_j)\mathsf{W}^{-1}(\boldsymbol{\Sigma}_j \mid \mathrm{P}, \nu)$.

5.3.1 Results for DSB mixtures with fixed tuning

For this experiment we simulated 200 data points from a mixture of seven Normal distributions, and we will be adjusting six DSB mixtures with parameters (β, θ, μ_0). In all cases the we will fix $\theta = 1$, and for each distinct DSB mixture we chose a different value of β in the set $\{0, 1/3, 1, 4, 9, \infty\}$ so that the underlying tie probability $\rho_\nu = 1/(\beta + 1)$ varies in $\{1, 0.75, 0.5, 0.2, 0.1, 0\}$. The DSBs with $\beta = 0$ and $\beta = \infty$ refer to a Geometric and Dirichlet process respectively.

In Figure 32 we can observe that all the models do a good job estimating the density through the EAP, and recover each of the seven modes that the histogram of the data features. The Dirichlet process and the DSB with $\beta = 9$ struggle more than the other models to differentiate the second and third modes from left to right, this can be due to the initial election of the parameter θ and the fact that a priori the DSB with $\beta = 9$ behaves similarly to a Dirichlet process. In Figure 32 we can also observe that it is at the high density areas that the estimated density from model to model varies slightly.

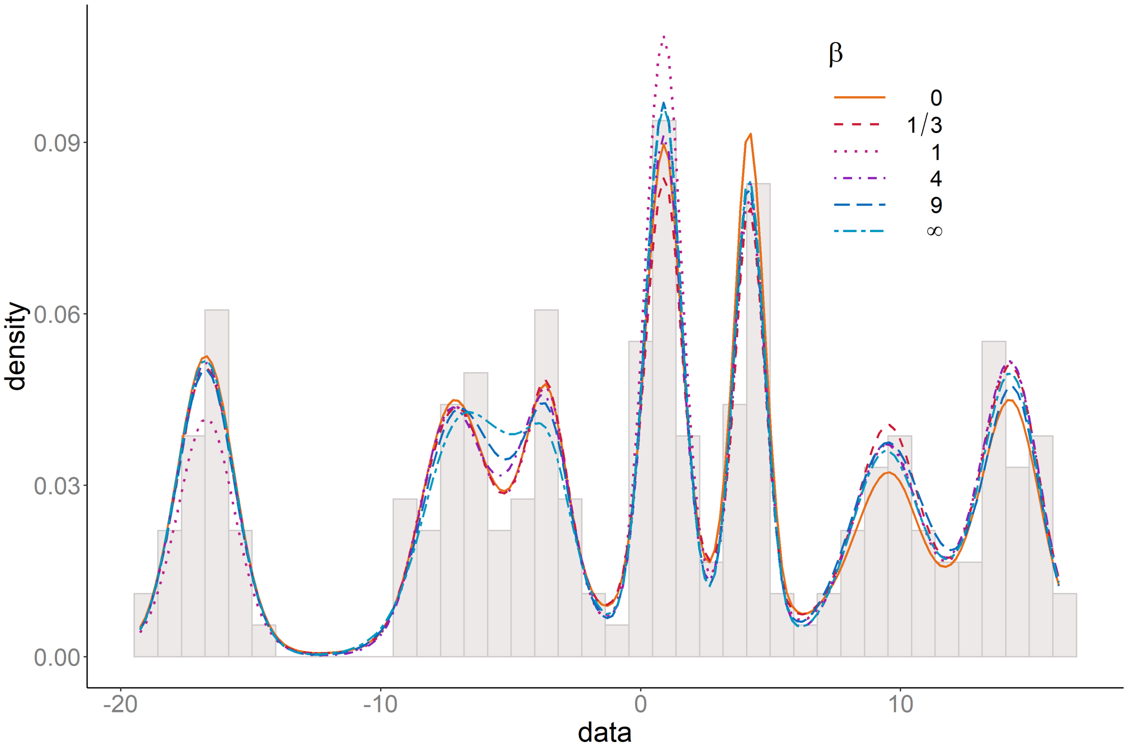


Figure 32: Estimated densities by means of the EaP, taking into account 4000 iterations of the Gibbs sampler, skipping 4 iterations, after a burn-in period of 7000, for a Geometric mixture ($\beta = 0$, $\rho_\nu = 1$), a Dirichlet mixture ($\beta = \infty$, $\rho_\nu = 0$) and four DSB mixtures with parameter $\beta \in \{1/3, 1, 4, 9\}$ ($\rho_\nu \in \{0.75, 0.5, 0.2, 0.1\}$, respectively). In all cases the parameter $\theta = 1$.

Figure 33 illustrates the posterior distribution of $\mathbf{K}_n$, with $n = 200$, for each of the six DSB mixtures implemented. Here we see that the Dirichlet process and the DSBs with a larger value of β, give high probability to numbers close to seven, which is the true number of components of the mixture from which the data was sampled. In contrast, as the parameter β approaches zero, we observe that the models tend to assign higher probability to larger values through the posterior distribution of $\mathbf{K}_n$. This means that these models use more components to provide the estimations illustrated in Figure 32. Indeed, since the Geometric weights decrease at a constant rate, in order to estimate the size and shape of some components, the model is forced to overlap many small components. If we were interested in clustering the data points, this can be a disadvantage of DSB mixtures with a small value of β, (a large value of the underlying tie probability ρ_ν) because it is likely that the number of clusters will be overestimated. However, if we are only interested in density estimation this feature actually makes the models that behave similar to Geometric processes more likely to capture subtle changes in the histogram of the data set.

Overall we see that the results are consistent with Corollary 4.9 put together with Remark 5.1, in the sense that as β grows, the results provided by the DSB mixtures are similar to those given by a Dirichlet prior, and at the other extreme, when $\beta \to 0$, the estimations are closer to those provided by a Geometric prior.

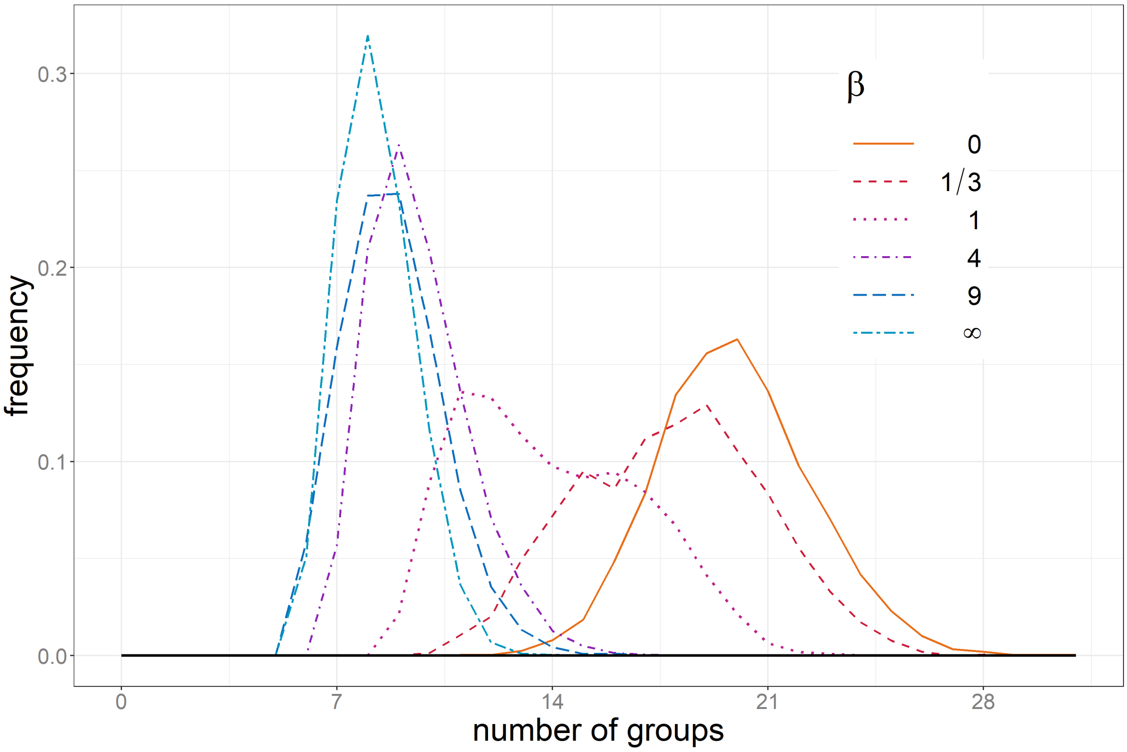


Figure 33: Frequency polygons corresponding to the posterior distribution of $\mathbf{K}_{200}$, for the Geometric mixture ($\beta = 0$, $\rho_\nu = 1$), the Dirichlet mixture ($\beta = \infty$, $\rho_\nu = 0$) and the four DSB mixtures with parameter $\beta \in \{1/3, 1, 4, 9\}$ ($\rho_\nu \in \{0.75, 0.5, 0.2, 0.1\}$, respectively). In all cases the parameter $\theta = 1$.

5.3.2 Results for SSB mixtures with tuning

In this experiment we will be testing SSB mixtures in an analogous way we analysed DSB mixtures in Section 5.3.1. For this study we simulated 220 observations from a different mixture of seven Gaussian distributions. To this database we will adjust six distinct SSB mixtures with parameters $(\mathrm{p}, \nu_0, \mu_0)$, where $\nu_0 = \mathsf{Be}(1, \theta)$. For the six models we fix $\theta = 1$ and for each SSB mixture we will chose a distinct value of p in the set $\{0, 0.2, 0.5, 0.8, 0.97, 1\}$. Here the SSB mixtures with $\mathrm{p} = 0$ and $\mathrm{p} = 1$ refer to a Dirichlet and a Geometric process, respectively.

Figure 34 exhibits the EAP estimator of the density, given by the the six SSB mixtures. In this figure we observe that in general all the models provide a fairly good estimation of the density and coincide in most points. It is at the high density areas were the estimation varies from model to model, most evidently the SSB mixtures with a higher value of p differ from the rest of the models at the size of the local modes. In particular we see that for the second mode of the histogram, from left to right, the Geometric process ($\mathrm{p} = 1$) estimates two distinct modes, whereas the rest of the models estimate one local mode. This is due to the fact that the Geometric priors tend to estimate the density using a larger number of smaller components which results in the model being able to capture subtle changes in the dataset. Although in this case the true mixture features seven modes and not eight as the Geometric process estimates, this should not be consider a mistake by the Geometric prior, as it is possible that we require more sampled data points for the dataset to be representative of fine details of the true density.

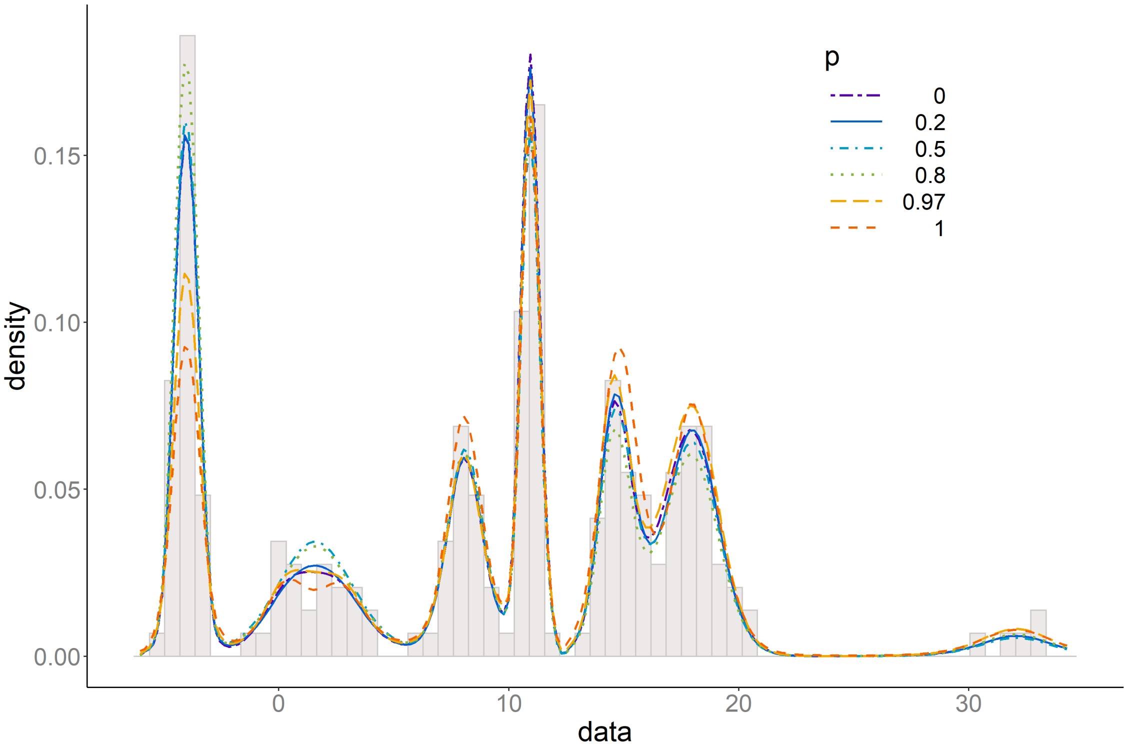


Figure 34: Estimated densities through the EaP, taking into account 4000 iterations of the Gibbs sampler, skipping 4 iterations, after a burn-in period of 5000, for a Geometric mixture, a Dirichlet mixture and four SSB mixtures with parameter p $= 0.2, 0.4, 0.6, 0.8$. In all cases the parameter $\theta = 1$

On the other side, in Figure 35 we can observe the posterior distribution of $\mathbf{K}_n$ for $n = 220$, corresponding to each SSB prior. Here we see that SSB mixtures with p ≤ 0.8 accumulate more mass at smaller values of $\mathbf{K}_n$ than the SSB models with p ≈ 1. Notice that for the mixtures with p $\in \{0.2, 0.5, 0.8\}$ the posterior distribution of $\mathbf{K}_n$ even gives higher probability to smaller values than the Dirichlet model. In particular, the SSB with p $= 0.5$ recovers the true number of mixture components through the posterior mode of $\mathbf{K}_n$. This phenomena can be explained by analysing the prior distribution of $\mathbf{K}_n$ for each SSB. As illustrated in Figure 29, if p $\not\approx 1$, the prior distribution of $\mathbf{K}_n$ is similar to that of a Dirichlet process, with the difference that the prior variance is slightly bigger for SSBs, with p > 0, without significantly affecting the prior mean. When implementing the models this translates to a bigger flexibility of SSBs with p $\in \{0.2, 0.5, 0.8\}$, in terms of the number of components the model requires to provide the estimates, without favouring significantly larger values. Now, for the case where p ≈ 1, we observe that generally the posterior distribution of $\mathbf{K}_n$ favours larger values. Despite, note that even for p $= 0.97$ which is close to one, the posterior distribution of $\mathbf{K}_n$ still accumulates mass at smaller values than in the case of a Geometric prior. This is a consequence of the fact that the convergence rates of SSBs to Geometric processes, as p $\to 1$ is very slow. Hence to approximate the results provided by a Geometric prior we require a value of p closer to one, than p $= 0.97$.

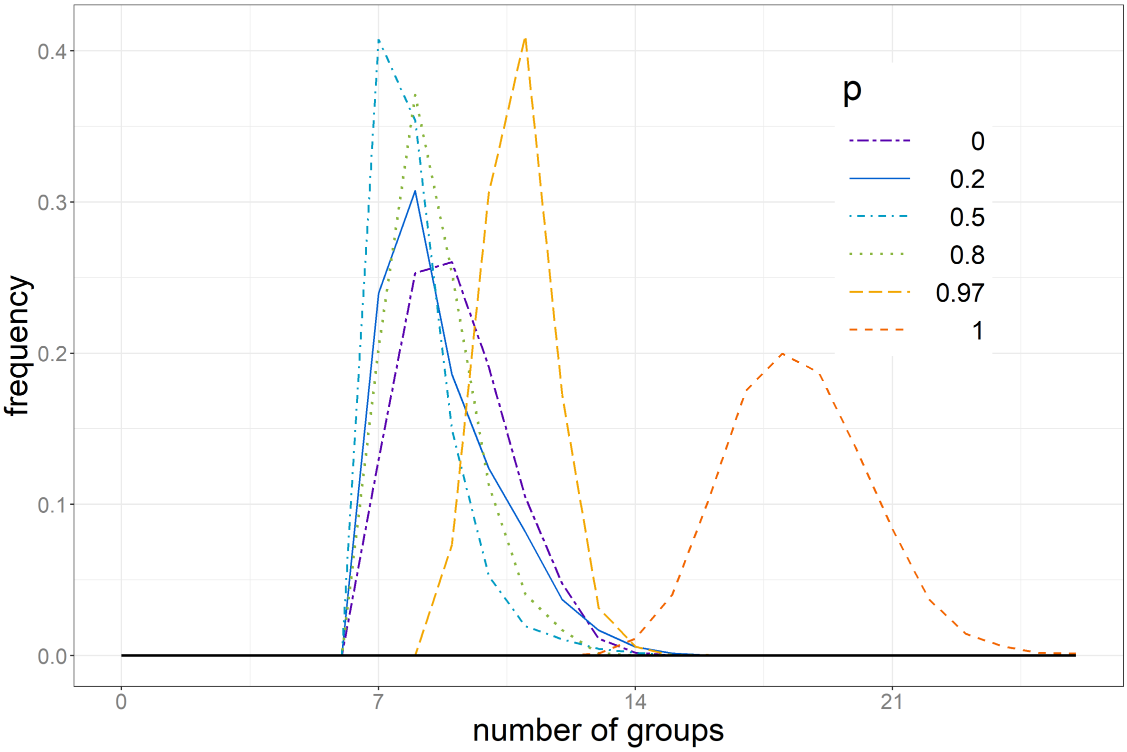


Figure 35: Frequency polygons corresponding to the posterior distribution of $\mathbf{K}_{220}$, for the Geometric mixture, the Dirichlet mixture and the four SSB mixtures with parameter $\mathrm{p} = 0.2, 0.4, 0.6, 0.8$. In all cases the parameter $\theta = 1$.

5.3.3 Results for BBSB mixtures with fixed tuning parameter

This third experiment consists in contrasting BBSB mixtures against Dirichlet and Geometric mixtures. For this study we simulated 200 data points from a mixture of eleven Gaussian kernels. In order to estimate the density of the data we will adjust five BBSB mixtures with parameters $(\alpha, \theta, \kappa, \mu_0)$. For all mixtures we fix $\alpha = \theta = 1$, and for each BBSB model we choose a distinct value of κ in the set $\{0, 10, 100, 200, \infty\}$. Supported by Corollary 4.18, the BBSB mixtures with $\kappa = 0$ and $\kappa = \infty$ refer to Dirichlet and Geometric processes, respectively.

Figure 36 shows the estimated densities through the EAP estimator for the five BBSB mixtures, we see that all models recover each of the eleven modes, and at most points the estimated densities are similar from model to model. Now, if we look thoroughly at the right part of the histogram, we see that the Geometric process ($\kappa = \infty$) differs from other models at the size of high density regions, and if we focus on the left side of the histogram we can appreciate that, in contrast to the other BBSB models, the Dirichlet mixture ($\kappa = 0$) struggles to separate the three modes on the left.

Looking at Figure 37, we observe analogous results to those observed for DSB mixtures in Section 5.3.1. For smaller values of the parameter κ, the posterior distribution of $\mathbf{K}_n$ concentrates mass at values nearer to the true number of components. Alternatively, for larger values of κ, the posterior distribution of $\mathbf{K}_n$ exhibits a larger mean and variance, this is consistent with the prior analysis of BBSB process in Section 4.2.3.

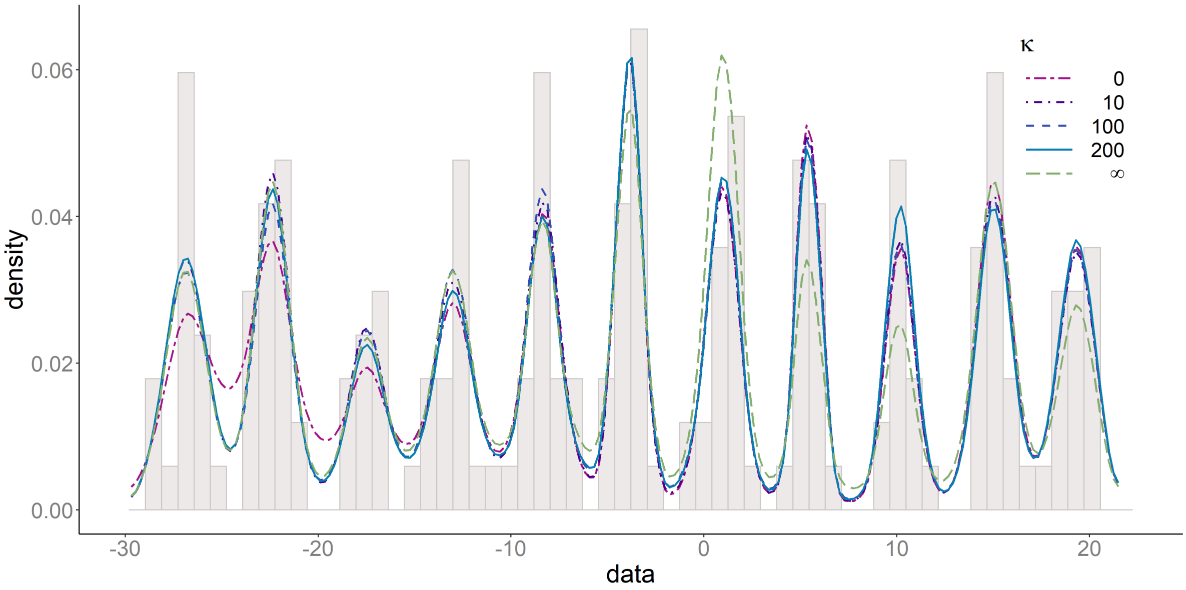


Figure 36: Estimated densities through the EAP, taking into account 4000 iterations of the Gibbs sampler, skipping 4 iterations, after a burn-in period of 2000, for a Geometric mixture, a Dirichlet mixture and three BBSB mixtures with parameter $\kappa \in \{10, 100, 200\}$. In all cases we fixed $\theta = 1$.

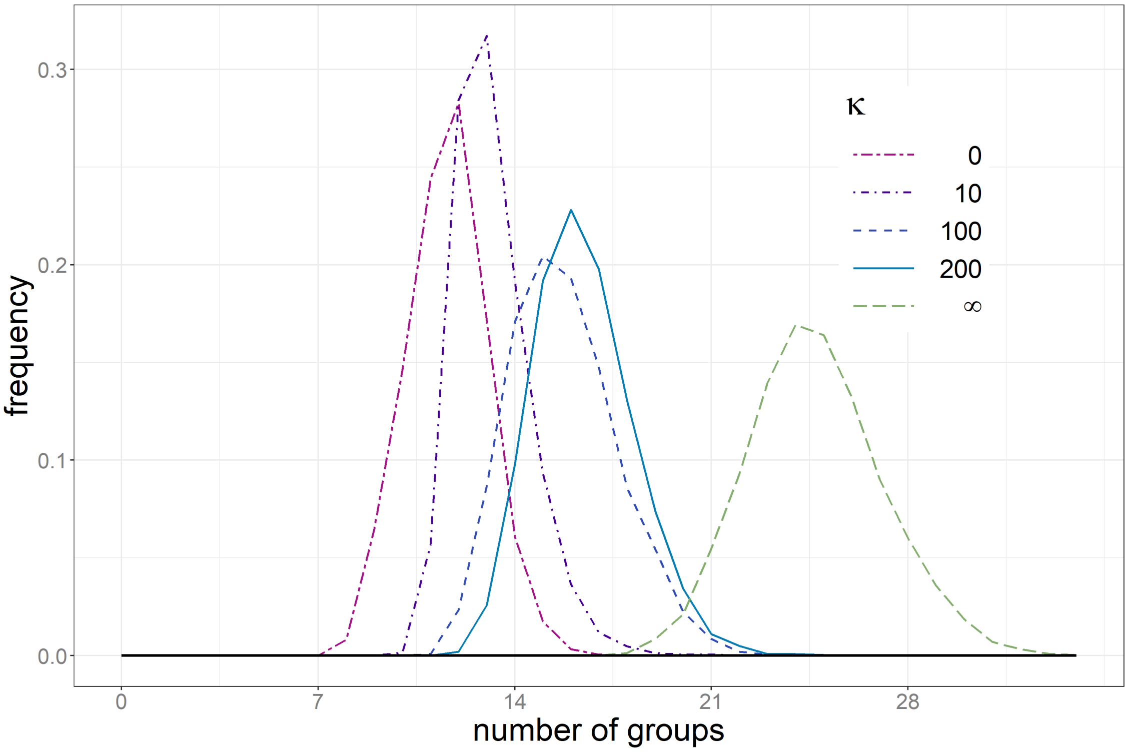


Figure 37: Frequency polygons corresponding to the posterior distribution of $\mathbf{K}_{200}$, for the Geometric mixture, the Dirichlet mixture and the three BBSB mixtures parameter $\kappa \in \{10, 100, 200\}$. In all cases we fixed $\theta = 1$.

Overall we see that DSBs, SSBs and BBSBs propose three distinct ways to approximate Dirichlet and Geometric processes. In general we observed that models similar to Dirichlet processes tend to use fewer components to estimate the density, in contrast, model close to the Geometric process use a larger number of components and are more sensible to subtle features in the histogram. We can conclude that by fixing the tuning parameters to distinct values, the posterior inference for these classes of stick-breaking priors with non independent length variables, is consistent with the analysis of the models developed in Section 4.

5.3.4 Results for DSB, SSB and BBSB mixtures with random tuning parameters

For this last study we simulated 510 data points from a paw-shaped mixture of seven Gaussian distributions. The objective of this experiment is to analyse the results provided by DSBs, SSBs and BBSBs when the corresponding tuning parameter is considered random and assigned a prior. In principle this allows the model to chose the best value of the tuning parameter for the dataset. In this simulation study we will concentrate in estimating the following quantities of interest: (a) the clusters of the data points using the MAP, (b) the density of the data using both the MAP and the EAP estimators, (c) the posterior distribution of the number of significant components, $\mathbf{K}_n$, provided by the EAP, and (d) the posterior distribution of the tuning parameters using the EAP estimator.

In order to infer about the quantities (a)–(c) we will adjust five models a Dirichlet, a DSB, a SSB, a BBSB and a Geometric mixture. In all cases we will consider that the marginal distribution of the length variables is a $\mathsf{Unif}(0, 1)$ distribution, so the parameter $\theta = 1$. For the DSB mixture we will assume that the underlying tie probability $\boldsymbol{\rho_\nu} \sim \mathsf{Unif}(0, 1)$, for the SSB we will also assign a $\mathsf{Unif}(0, 1)$ prior to the tuning parameter, $\mathbf{p}$, as to the BBSB we will consider that a priori $\boldsymbol{\kappa} \sim \mathsf{Unif}(\{1, \dots, 100\})$.

In Figure 38 we see that the DSB, the SSB and the BBSB estimate better the clusters of the data points than the limiting processes. In particular, the DSB recovers exactly seven clusters as there are according to the true model. Whilst the SSB and the BBSB estimate eight clusters, the eighth one consist of a very small group of data points, so the error is relatively small. As to shape of the clusters, the DSB and the SSB are the ones that perform best. In the same figure we see that the Dirichlet model underestimates the number of clusters by one when compared with the true model. Indeed, the estimation provided by the Dirichlet process merges two clusters of the model true. On the other extreme, the Geometric model overestimates the number of clusters by seven, and under estimates the sizes of the real clusters. As Figure 38 illustrates, the main advantage of DSBs, SBS's and BBSBs with random tuning parameter, over the limiting processes, is that they are more general, thus allowing the model to chose between a model similar to the Dirichlet process, to the Geometric process or some model in between that combines characteristics of both limits. This behaviour of the models is also reflected through the MAP estimators of the density, presented in Figures 39 and 40. Here we see that the SSB is the one that provides the best estimation of the density through the MAP, followed by the DSB. While the Dirichlet process and the BBSB do recover the paw shape of the density these models do not perform this as neatly as the DSB and specially as the SSB. As to the Geometric process we observe that through the MAP it does not even recover the paw shape of the true model.

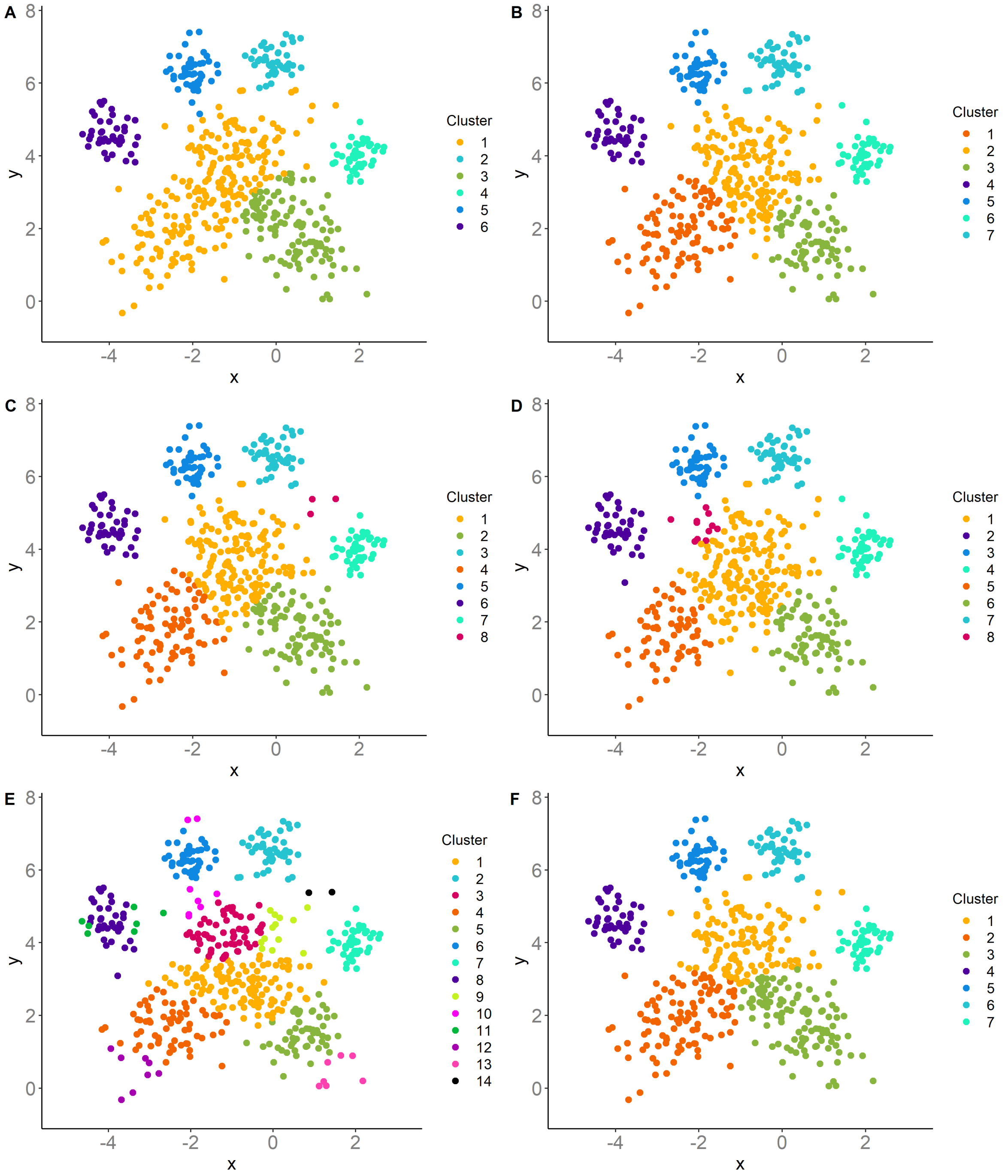


Figure 38: Estimated clusters of the data points using the MAP, taking into account 8000 iterations of the Gibbs sampler after a burn-in period of 2000 iterations, according to the Dirichlet prior (A) a DSB prior (B), a SSB prior (C), a BBSB prior (D) and a Geometric prior (E). F shows the clustering of the data points according which component of the true mixture are the data points more likely to come from.

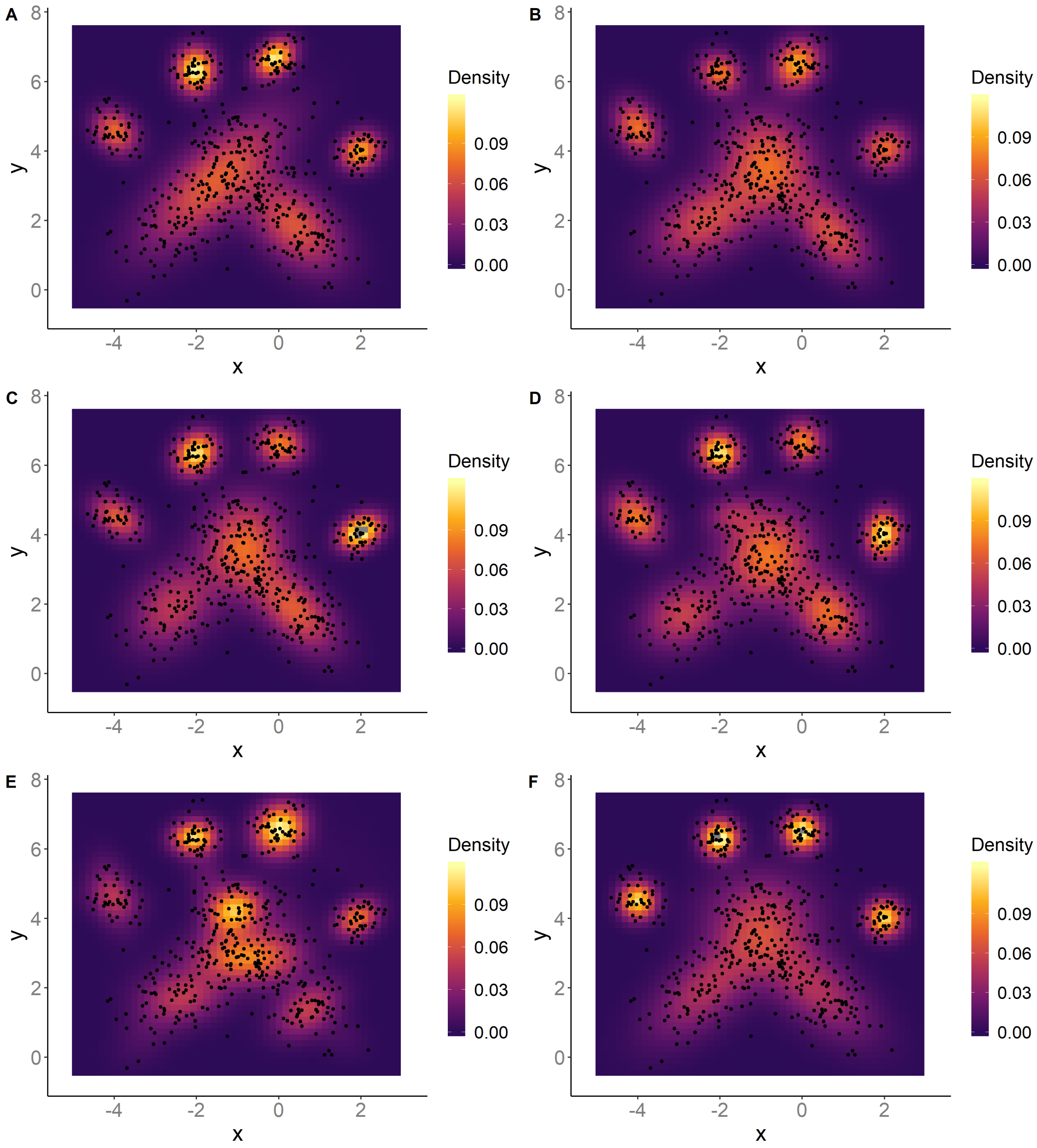


Figure 39: Estimated densities of the data points using the MAP, taking into account 8000 iterations of the Gibbs sampler after a burn-in period of 2000 iterations, according to the Dirichlet prior (A) a DSB prior (B), a SSB prior (C), a BBSB prior (D) and a Geometric prior (E). F shows the true density from which the data points were i.i.d. sampled.

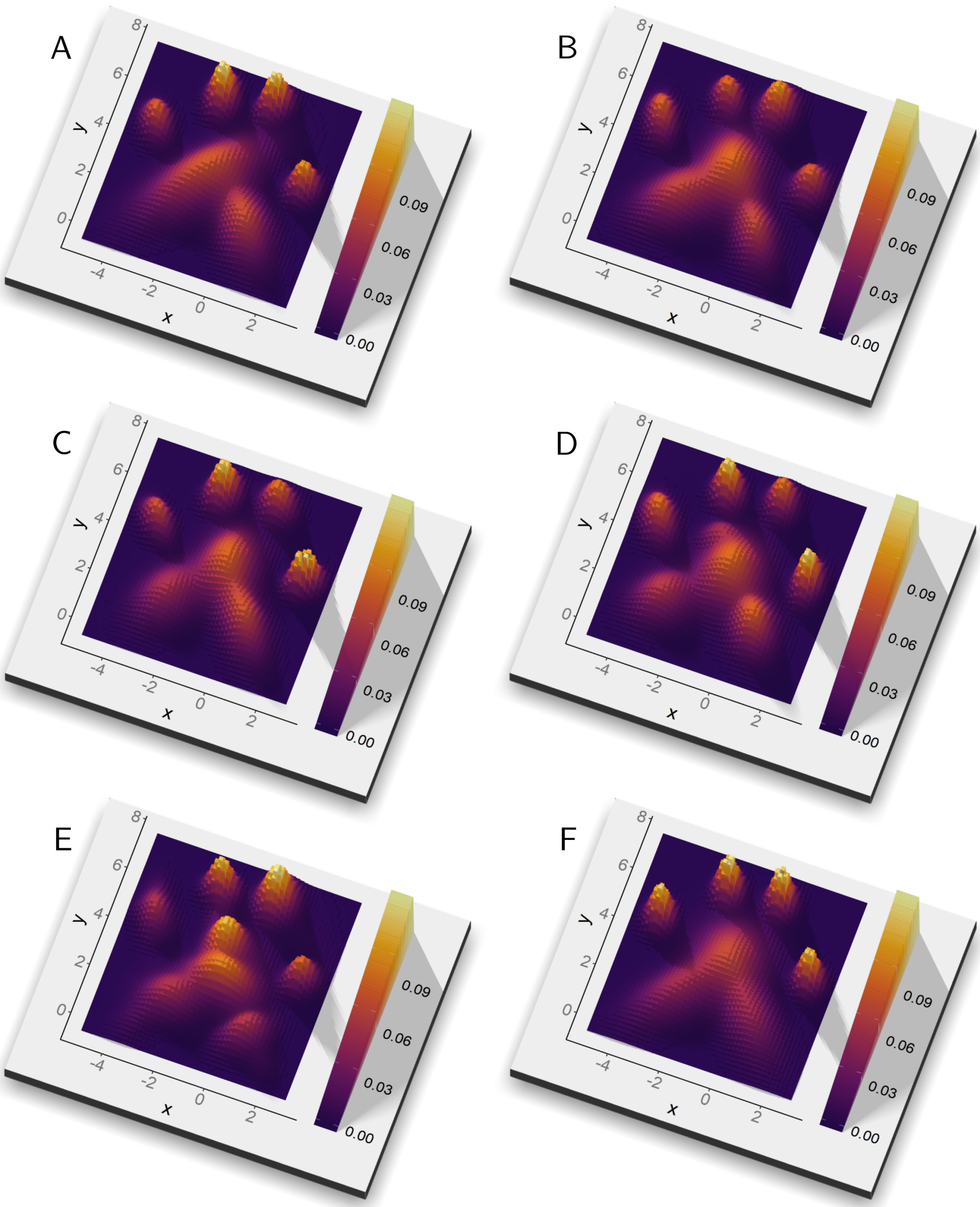


Figure 40: 3D view of the estimated densities in Figure 39

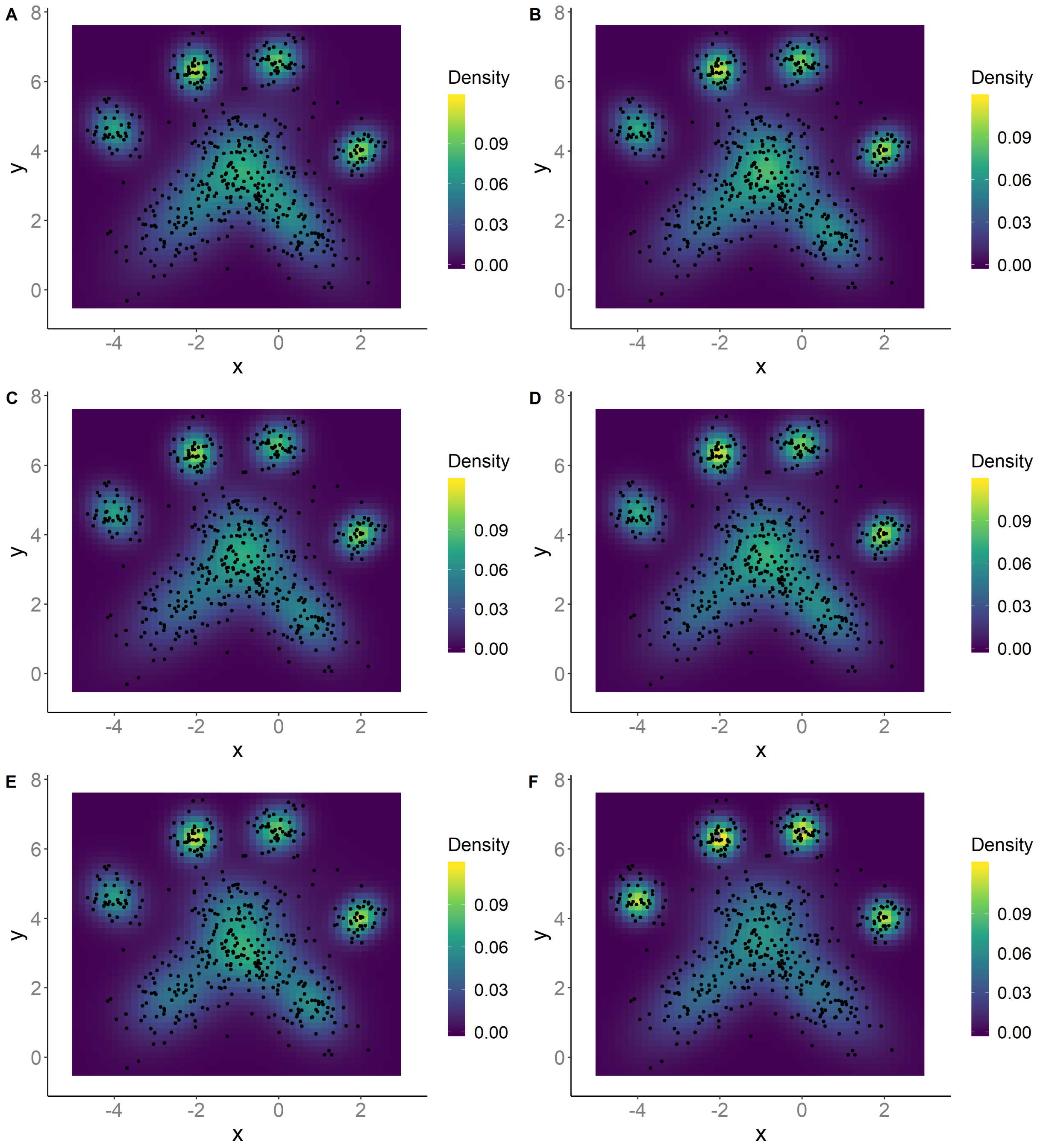


Figure 41: Estimated densities of the data points using the EAP, taking into account each fourth iteration among 8000 iterations of the Gibbs sampler after a burn-in period of 2000, according to the Dirichlet prior (A) a DSB prior (B), a SSB prior (C), a BBSB prior (D) and a Geometric prior (E). F shows the true density from which the data points were i.i.d. sampled.

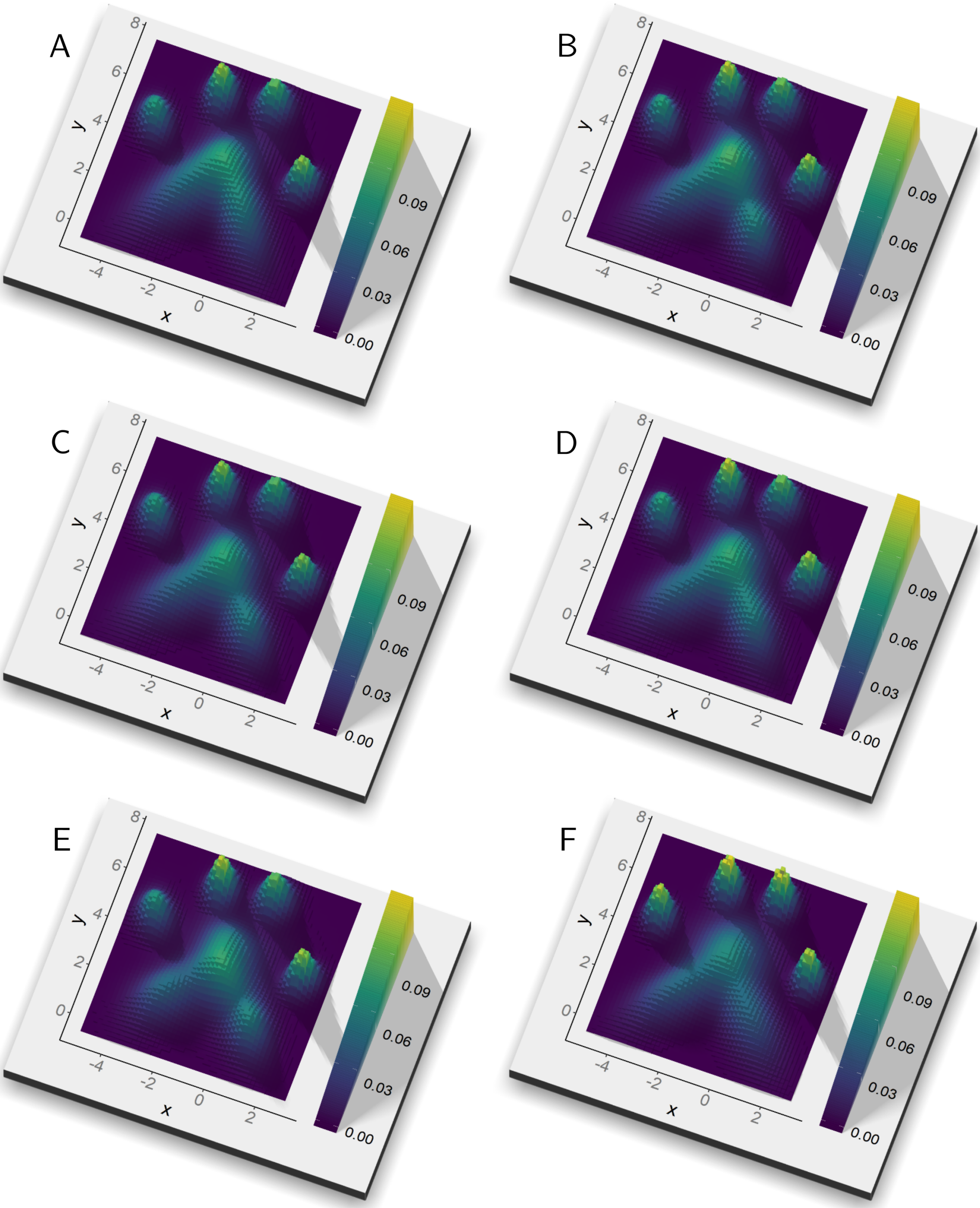


Figure 42: 3D view of the estimated densities in Figure 41

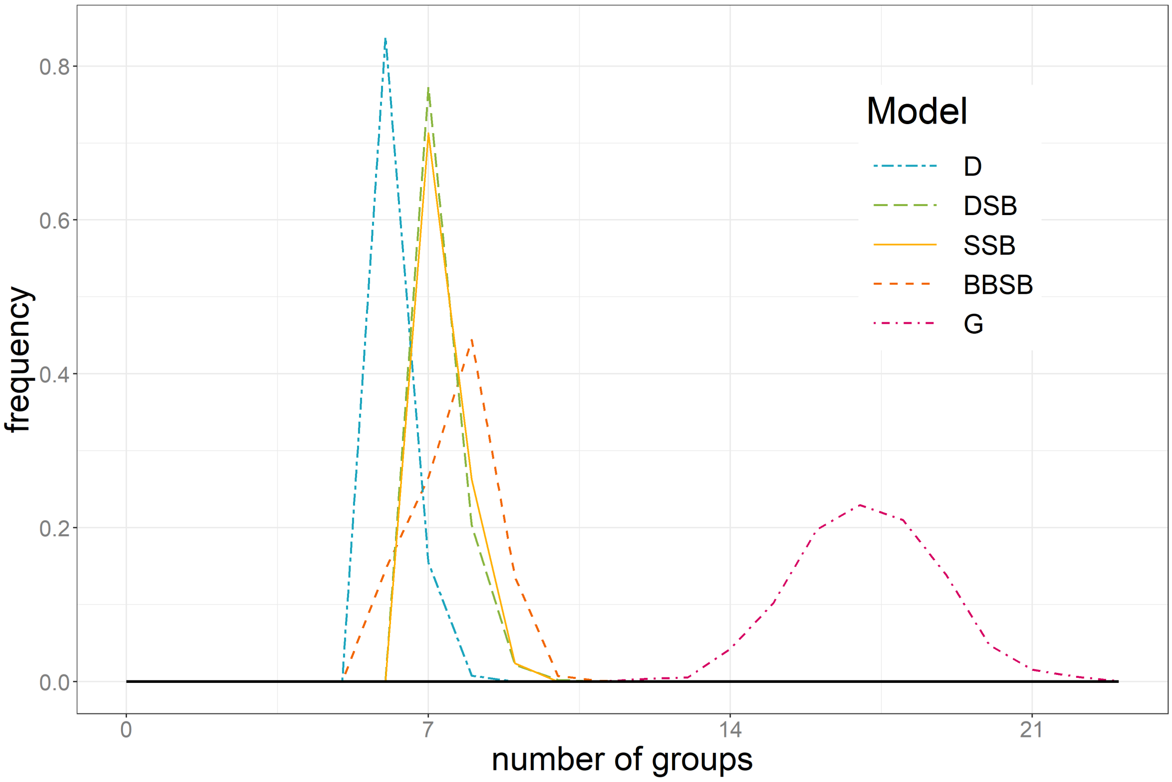


Figure 43: Frequency polygons corresponding to the posterior distribution of $\mathbf{K}_{510}$, for the Dirichlet (D), DSB, SSB, BBSB and the Geometric (G) mixtures.

Figures 41 and 42 exhibit the EAP estimators of the density provided by each model. In comparison to the MAP estimators we see that these ones are much smoother. Looking at the EAP estimations of the densities, we appreciate that the differences from one model to another are much more subtle and overall all models estimate the density quite nicely. However, if we focus in the palm of the paw we see that the DSB, the SSB and the BBSB recover the shape slightly better than the Dirichlet and the Geometric processes, perhaps the one that performs best at this task is the BBSB or the SSB. Before we move on, let us make a small parenthesis to comment that the fact that all models provide good estimations of the densities through the EAP, is closely related to the fact that Gaussian mixtures are in general very flexible models and that all of the species sampling priors considered here have full support. In effect, species sampling priors with full support lead to very flexible mixtures, so one would expect that, after enough valid iterations of the Gibbs sampler, all models analysed here provide decent estimations of the densities.

In Figure 43 we observe the posterior distribution of $\mathbf{K}_n$, (for $n = 510$) for each of the models. Here we see that through the posterior mode of $\mathbf{K}_n$ the DSB and the SSB recover the true number of mixtures components. The Dirichlet and the BBSB models also assing a probability larger than zero to the true number of components. Despite this, the posterior mode of $\mathbf{K}_n$ for the Dirichlet process is one unit smaller than the true number of components, and posterior mode of $\mathbf{K}_n$ for the BBSB model is one unit bigger. As to the Geometric process, similarly as for other datasets, the posterior distribution of $\mathbf{K}_n$ concentrates in significantly larger values than the real number of mixture components, which in this particular case is seven.

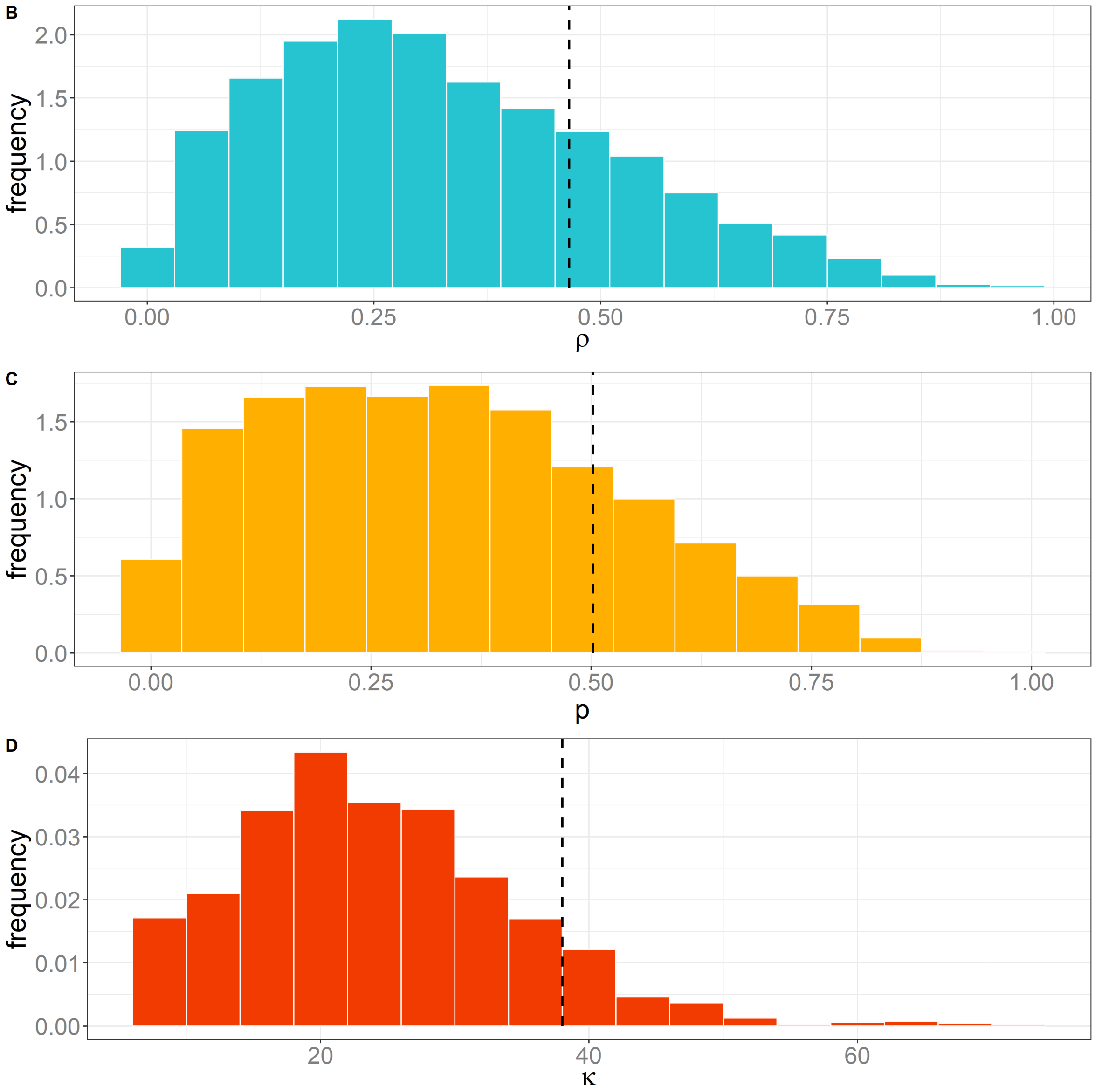


Figure 44: Posterior distributions of the tuning parameters $\boldsymbol{\rho_{\nu}}$, $\mathbf{p}$ and $\boldsymbol{\kappa}$ (B, C and D, respectively. The dashed lines indicate the MAP estimator of each random parameter.

The last figure we will analyse here is Figure 44, which presents the posterior distribution of the tuning parameters and the MAP estimator for each of these. For the underlying tie probability of the DSB, $\boldsymbol{\rho_{\nu}}$, we see in B that the posterior mode is close to 0.25, suggesting the for this database a model closer to the Dirchlet process than the Geometric process is preferred. In C a similar pattern is exhibited in the posteriori distribution of the tuning parameter, $\mathbf{p}$, of the SSB. As to the posterior distribution $\boldsymbol{\kappa}$ for the BBSB, we observe that the mode is at 20, so it coincides with the DSB and SSB models that a model more similar to a Dirichlet process fits better the dataset.

6 Conclusions and final comments

To the best of our knowledge Section 4 represents the first general treatment of stick-breaking processes with non-independent length variables. Since the class of random sequences with dependent items remains very hard to manage, we decided to restrict ourselves to the case where the length variables, $(\mathbf{v}_i)_{i\geq 1}$, are either exchangeable or Makorvian. Exploiting the conditional independence of these sequences, we were able to prove important properties regarding the validity of the models. Namely, under minor conditions, we proved that the stick-breaking process in question are proper and have full support. We also derived limiting properties that give a new interpretation to the well-known Bayesian non-parametric priors, the Dirichlet and the Geometric processes, as the extrema of stick-breaking processes with stationary (exchangeable or Markovian) length variables. Additionally, we implemented mixtures of Gaussian distributions where the mixing prior generalizes Dirichlet and Geometric processes, and by means of specific examples we showed the advantages of our models. When we put a prior on the tuning parameter, DSB, SSB and BBSB mixtures seem to efficiently combine the best features of both limiting mixing priors. In addition, for DSBs and SSBs we computed the probability that consecutive weights are decreasingly ordered.

Whilst the present work was mainly motivated by Bayesian non-parametric theory, and was conducted in this setting, stick-breaking processes continue to be widely used in other probabilistic frameworks, and the results provided here can easily emigrate to such contexts. As an example, the ordering of the weights is an appealing result for other areas of probability.

Outside the principal contribution of the thesis, smaller contributions were made in other sections. For example, the construction of SSPs by means of latent random sets, has already been used before, but to the best of our knowledge, the formalization of the method and the algorithm in Section 5.2.2, have not been described in the general setting presented here. Another small contribution, or rather remark, that seemingly is not a widely known fact, is that the classes of SSPs, stick-breaking processes, and exchangeable random probability measures are exactly the same class. On a similar line of thought, while I am completely aware that Theorem 3.4 is well-known for Dirichlet processes, I personally have not seen it written, in terms of the tie probability for the general class of SSPs. Another remark here is that Corollary 2.24 together with Theorem 3.4 characterize the extreme points of the class of exchangeable random probability measures. Overall, I personally hope that the compendium of results presented here, contributes to whoever reads the thesis.

To conclude the document I want to propose a couple of research lines for further work:

- The transitions and symmetries we defined in Section 4 and used to defined new Bayesian non-parametric priors, can also be used to model dependence between two or more distinct SSPs and latter adjust the corresponding mixtures to partially exchangeable data.

- It may be interesting to consider stick-breaking process where the length variables form a martingale. My conjecture is that, it can be proven that most of these stick-breaking processes, at least lead to feasible Bayesian non-parametric priors, meaning that they have full support and are proper.

A Proofs of Section 1

A.1 Proof of Lemma 1.2

Let Σ_1, Σ_2 and Σ_3 be the σ-algebras generated by I, II and III respectively.

$\Sigma_1 \subseteq \Sigma_3$: For any $B \in \hat{\mathcal{S}}$ and any $\mu \in \mathcal{P}(S)$, we have that $\int \mathbf{1}_B d\mu = \mu(B)$, hence the projection map π_B equals the integration map $\pi_{\mathbf{1}_B}$. This proves π_B is measurable with respect to Σ_3, for every $B \in \hat{\mathcal{S}}$, and by definition of Σ_1 and Σ_3, it follows $\Sigma_1 \subseteq \Sigma_3$.

$\Sigma_3 \subseteq \Sigma_2$: If f is a simple measurable function, $f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$, for some $A_1, \dots, A_n \in \mathscr{B}_S$(with A_i and A_j disjoint, for $i \neq j$). By linearity of the integral we have that

$$\pi_f(\mu) = \mu(f) = \sum_{i=1}^n a_i \mu(A_i) = \sum_{i=1}^n a_i \, \pi_{A_i}(\mu), \quad \mu \in \mathcal{M}(S),$$

that is $\pi_f = \sum_{i=1}^n a_i \, \pi_{A_i}$, from which we obtain that π_f is Σ_2-measurable. Now, if f is an arbitrary measurable and non-negative function then f and be approximated by simple functions $f_n \nearrow f$, and by monotone convergence theorem we obtain $\pi_{f_n}(\mu) \nearrow \pi_f(\mu)$ for every $\mu \in \mathcal{M}(S)$. This is $\pi_{f_n} \nearrow \pi_f$, hence π_f is Σ_2-measurable, and we get $\Sigma_3 \subseteq \Sigma_2$.

$\Sigma_2 \subseteq \Sigma_1$: Fix a localizing sequence, $(S_n)_{n=1}^\infty$, and let $B \in \hat{\mathcal{S}}$. Then $B_n = B \cap S_n \nearrow B$, with $B_n \in \hat{\mathcal{S}}$, for all $n \geq 1$. Evidently, $\pi_{B_n}(\mu) = \mu(B_n) \nearrow \mu(B) = \pi_B(\mu)$, for every $\mu \in \mathcal{M}(S)$. Thus, π_B is Σ_1-measurable. □

A.2 Proof of Theorem 1.4

I $\Leftrightarrow$ II: Let $\boldsymbol{\mu}$ be a kernel from S into T. If $f = \mathbf{1}_B$, for some $B \in \mathscr{B}_T$, we get $\mathcal{A}^{\boldsymbol{\mu}}(f) = \boldsymbol{\mu}(B) \in S_+$. By linearity we see that for the a simple function, $f = \sum_{i=1}^n b_i \mathbf{1}_{Bi}$, with $B_1, \dots, B_n \in \mathscr{B}_T$, $\mathcal{A}^{\boldsymbol{\mu}}(f) = \sum_{i=1}^n b_i \boldsymbol{\mu}(B_i) \in S_+$. Finally for arbitrary measurable $f : S \to \overline{\mathbb{R}}_+$, there exist a collection of simple functions $f_n \nearrow f$, hence by monotone convergence theorem $\mathcal{A}^{\boldsymbol{\mu}}(f) = \lim_n \mathcal{A}^{\boldsymbol{\mu}}(f_n) \in S_+$. Conversely, say $\mathcal{A}^{\boldsymbol{\mu}}(f) \in S_+$, for every $f \in T_+$. The choice $f = \mathbf{1}_B$, for $B \in \mathscr{B}_T$, shows that $\boldsymbol{\mu}(\cdot, B) = \boldsymbol{\mu}(B) = \mathcal{A}^{\boldsymbol{\mu}}(f)$ is a measurable function from S into $\overline{\mathbb{R}}_+$. The fact that $\boldsymbol{\mu}_s(\cdot) = \boldsymbol{\mu}(s, \cdot)$ is a measure is implicit in the definition of the mapping $\mathcal{A}^{\boldsymbol{\mu}} : T_+ \to S_+$.

I,II $\Leftrightarrow$ III, follows directly from the definition of the Borel σ-algebra, $\mathscr{B}_{\mathcal{M}(T)}$, of $\mathcal{M}(T)$. □

A.3 Proof of Theorem 1.5

I $\Leftrightarrow$ II is a straight-forward consequence of Kolomogorov's consistency theorem. This is the finite dimensional distributions, characterize the law of $\boldsymbol{\mu}$. It is also obvious that II $\Rightarrow$ III. Now say that III holds and fix $B_1, \dots, B_n \in \mathscr{B}_S$. It is easy to see that there exist $m \in \mathbb{N}$ and a disjoint collection $\{A_i\}_{i=1}^m$, such that for every $j \leq n$ there exist $K_j \subseteq \{1, \dots, m\}$ that satisfies $B_j = \bigcup_{i \in K_j} A_i$ (for example for $B_1, B_2 \in \mathscr{B}_S$ we might choose $A_1 = B_1 \setminus (B_1 \cap B_2)$, $A_2 = B_1 \cap B_2$ and $A_3 = B_2 \setminus (B_1 \cap B_2)$). Then, we have

$$(\boldsymbol{\mu}(B_1), \dots, \boldsymbol{\mu}(B_n)) = \left(\sum_{i \in K_1} \boldsymbol{\mu}(A_1), \dots, \sum_{i \in K_n} \boldsymbol{\mu}(A_i) \right).$$

Since $(\boldsymbol{\mu}(A_1),\dots,\boldsymbol{\mu}(A_m)) \overset{d}{=} (\boldsymbol{\nu}(A_1),\dots,\boldsymbol{\nu}(A_m))$, II follows. We have shown so far I $\Leftrightarrow$ II $\Leftrightarrow$ III. Now we prove III $\Leftrightarrow$ IV. If III holds, for any simple function $f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ we have that

$$\boldsymbol{\mu}(f) = \sum_{i=1}^n a_i \boldsymbol{\mu}(A_i) \overset{d}{=} \sum_{i=1}^n a_i \boldsymbol{\nu}(A_i) = \boldsymbol{\nu}(f).$$

For arbitrary $f \in S_+$ there exist a collection of simple functions with $f_n \nearrow f$, hence by monotone convergence theorem, we obtain

$$\boldsymbol{\mu}(f) = \lim_{n\to\infty} \boldsymbol{\mu}(f_n) \overset{d}{=} \lim_{n\to\infty} \boldsymbol{\nu}(f_n) = \boldsymbol{\nu}(f),$$

which proves IV. Conversely if IV holds, for any disjoint collection of measurable sets $\{A_i\}_{i=1}^n$ we have that

$$\sum_{i=1}^n a_i \boldsymbol{\mu}(A_i) \overset{d}{=} \sum_{i=1}^n a_i \boldsymbol{\nu}(A_i).$$

for every $(a_1,\dots,a_n) \in \mathbb{R}_+^n$, and by the Cramér-Wold theorem (see Corollary 5.5 in Kallenberg (2002)) III follows. To finish the proof we will show IV $\Leftrightarrow$ V. Since $\boldsymbol{\mu}(f)$ and $\boldsymbol{\nu}(f)$ are positive random variables, we know that IV is equivalent to $\mathbb{E}\left[e^{-t\boldsymbol{\mu}(f)}\right] = \mathbb{E}\left[e^{-t\boldsymbol{\nu}(f)}\right]$ for all $t \in \mathbb{R}_+$ and $f \in S_+$ (see Theorem 5.3 in Kallenberg (2002)), the choice $t = 1$ gives V. Finally, assume V holds and fix $f \in S_+$ and $t \in \mathbb{R}_+$. Then $tf \in S_+$ and we obtain $\mathbb{E}\left[e^{-t\boldsymbol{\mu}(f)}\right] = \mathbb{E}\left[e^{-\boldsymbol{\mu}(tf)}\right] = \mathbb{E}\left[e^{-\boldsymbol{\nu}(tf)}\right] = \mathbb{E}\left[e^{-t\boldsymbol{\nu}(f)}\right]$. □

A.4 Proof of Proposition 1.6

Say that $S = \mathbb{R}$ and $\boldsymbol{\mu}(B) \overset{a.s.}{=} \boldsymbol{\nu}(B)$ for every $B \in B_{\mathbb{R}}$. For $a, b \in \mathbb{Q}$ with $a < b$, define

$$A_{a,b} = \{\omega \in \Omega : \boldsymbol{\mu}(\omega,(a,b)) = \boldsymbol{\nu}(\omega,(a,b))\} \quad \text{and} \quad A = \bigcap_{\substack{a,b\in\mathbb{Q}\\ a<b}} A_{a,b},$$

evidently $A \in \mathcal{F}$ is measurable, and $\mathbb{P}[A] = 1$. Now, fix $\omega \in A$ and $c < d \in \mathbb{R}$. There exist two sequences of rational numbers $(a_i)_{i\in\mathbb{N}}$ and $(b_i)_{i\in\mathbb{N}}$ such that $a_1 < b_1$, $a_n \nearrow c$, and $b_n \searrow d$, thus $(a_n, b_n) \nearrow (c,d)$. By continuity of measures we obtain

$$\boldsymbol{\mu}(\omega,(c,d)) = \lim_{n\to\infty} \boldsymbol{\mu}(\omega,(a_n,b_n)) = \lim_{n\to\infty} \boldsymbol{\nu}(\omega,(a_n,b_n)) = \boldsymbol{\nu}(\omega,(c,d)).$$

This shows that the π-system $\mathcal{C} = \{(c,d) : c \le d, c, d \in \mathbb{R}\}$ is contained in the λ-system $\mathcal{D} = \{B \in \mathscr{B}_{\mathbb{R}} : \boldsymbol{\mu}(\omega,B) = \boldsymbol{\nu}(\omega,B)\}$, with the convention $(c,d) = \emptyset$ for $c = d$. Hence, a monotone class argument shows that $\boldsymbol{\nu}(\omega,B) = \boldsymbol{\mu}(\omega,B)$ for every $B \in \mathscr{B}_{\mathbb{R}}$. That is $\boldsymbol{\mu} \overset{a.s.}{=} \boldsymbol{\nu}$. The same is true for random measures over an arbitrary Borel space $(S, \mathscr{B}_S)$, and the result for this case follows by picking a suitable Borel bijection between $\mathbb{R}$ and S. □

A.5 Proof of Theorem 1.7

Fix a localizing sequence $S_n \nearrow S$ in $\hat{\mathcal{S}}$, and set $B_n = S_n \setminus S_{n-1}$ with $S_0 = \emptyset$. Clearly $\mathbf{1}_{B_n}\boldsymbol{\mu}$ is finite and $\boldsymbol{\mu} = \sum_{n\ge1} \mathbf{1}_{B_n}\boldsymbol{\mu}$, hence, it suffices to derive the desired representation for finite measures. This said we may assume without loss of generality that $\boldsymbol{\mu}$ is finite.

Furthermore, using the Borel property of S we can further reduce the analysis to the case $S=[0,1)$. The decomposition follows from the jump structure of $t \mapsto \boldsymbol{\mu}([0,t])$, for $t \in [0,1)$.

To attain a measurable representation, define the array of sets $\mathcal{I}=\{I_{n,j}\}_{n,j}$ with

$$I_{n,j}=[2^{-n}(j-1),2^{-n}j), \quad n \in \mathbb{N},\, j \in \{1,\dots 2^n\}.$$

Note that for all $n \geq 1$, $\mathcal{I}_n=\{I_{n,j}\}_{j=1}^{2^n}$ is a partition of $[0,1)$ and that $\mathcal{I}_{n+1}$ is a refinement of $\mathcal{I}_n$. Moreover for $s<t$ in $[0,1)$ there exist $n, j_1, j_2 \geq 1$ such that $s \in I_{n,j_1}$, $t \in I_{n,j_2}$ and $j_1 \neq j_2$. In this sense, $\mathcal{I}$ separates points. Fix $\epsilon>0$ and define

$$\boldsymbol{\kappa}_n^\epsilon=\sum_{j=1}^{2^n}\mathbf{1}_{\{\boldsymbol{\mu}(I_{n,j})>\epsilon\}}, \quad \boldsymbol{\sigma}_{n,j}^\epsilon=2^{-n}j\mathbf{1}_{\{\boldsymbol{\mu}(I_{n,j})>\epsilon\}}.$$

Set $\boldsymbol{\xi}_{n,1}^\epsilon=\inf\{\boldsymbol{\sigma}_{n,j}^\epsilon : \boldsymbol{\sigma}_{n,j}^\epsilon \neq 0\}$, and $\boldsymbol{\alpha}_{n,1}^\epsilon=\boldsymbol{\mu}\left(I_{n,(2^n\boldsymbol{\xi}_{n,1}^\epsilon)}\right)$, recursively for $2 \leq k \leq \boldsymbol{\kappa}_n^\epsilon$, define

$$\boldsymbol{\xi}_{n,k}^\epsilon=\inf\{\boldsymbol{\sigma}_{n,j}^\epsilon>\boldsymbol{\xi}_{n,k-1}^\epsilon : \boldsymbol{\sigma}_{n,j}^\epsilon \neq 0\}, \quad \text{and} \quad \boldsymbol{\alpha}_{n,k}^\epsilon=\boldsymbol{\mu}\left(I_{n,(2^n\boldsymbol{\xi}_{n,k}^\epsilon)}\right)$$

By construction the above are measurable and the following limits exist

$$\boldsymbol{\kappa}^\epsilon=\lim_{n\to\infty}\boldsymbol{\kappa}_n^\epsilon, \quad \boldsymbol{\xi}_k^\epsilon=\lim_{n\to\infty}\boldsymbol{\xi}_{n,k}^\epsilon, \quad \text{and} \quad \boldsymbol{\alpha}_k^\epsilon=\lim_{n\to\infty}\boldsymbol{\alpha}_{n,k}^\epsilon$$

for every $k \leq \boldsymbol{\kappa}^\epsilon$. Note that $\boldsymbol{\kappa}^\epsilon$ is the number of atoms of $\boldsymbol{\mu}$ with size bigger that ϵ, $(\boldsymbol{\xi}_k^\epsilon)_{k=1}^{\boldsymbol{\kappa}^\epsilon}$ are such atoms in increasing order and $(\boldsymbol{\alpha}_k^\epsilon)_{k=1}^{\boldsymbol{\kappa}^\epsilon}$ are the corresponding sizes. This way

$$\boldsymbol{\eta}^\epsilon=\sum_{k=1}^{\boldsymbol{\kappa}^\epsilon}\boldsymbol{\alpha}_k^\epsilon\delta_{\boldsymbol{\xi}_k^\epsilon}$$

represents the atomic part of $\boldsymbol{\mu}$ with atom sizes bigger that ϵ. Since $\boldsymbol{\mu}$ is finite, $\boldsymbol{\eta}^\epsilon$ is also finite and there is no subtlety in defining $\boldsymbol{\nu}^\epsilon=\boldsymbol{\mu}-\boldsymbol{\eta}^\epsilon$, which has no atoms with sizes bigger than ϵ. Now, for $\epsilon'<\epsilon$ we can do the same procedure for $\boldsymbol{\nu}^\epsilon$, to attain a measure $\boldsymbol{\eta}^{\epsilon'}$ that encodes the atomic part of $\boldsymbol{\nu}^\epsilon$ with atom sizes bigger than ϵ', and a measure $\boldsymbol{\nu}^{\epsilon'}$ with no atoms with sizes bigger than ϵ', that satisfy $\boldsymbol{\nu}^\epsilon=\boldsymbol{\eta}^{\epsilon'}+\boldsymbol{\nu}^{\epsilon'}$. Notice that $\boldsymbol{\mu}=\boldsymbol{\eta}^\epsilon+\boldsymbol{\eta}^{\epsilon'}+\boldsymbol{\nu}^{\epsilon'}$. Continuing recursively into countably many steps, we obtain a measurable representation of all the atoms, and a remainder, $\boldsymbol{\nu}$, such that $\boldsymbol{\nu}(\{s\})=0$, for all $s \in [0,1)$. The uniqueness assertion follows easily from the construction.

□

A.6 Proof of Lemma 1.8

If $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$ we clearly have $\mathbb{P}[\boldsymbol{\mu}(B)=0]=\mathbb{P}[\boldsymbol{\nu}(B)=0]$ for every $B \in \hat{\mathcal{S}}$. To derive the converse result we will need require the following Lemma whose proof can be found in Kallenberg (2017)

Lemma A.1. *Every Borel space $(S,\mathscr{B}_S)$, with localizing ring $\hat{\mathcal{S}}$, contains a dissection system, that is an array of subsets $\mathcal{I}=\{I_{n,j}\}_{n,j}$ such that*

a) *For every $n \geq 1$, $\mathcal{I}_n=\{I_{n,j}\}_j$ is a countable partition of S.*

b) *For every $m > n$, $\mathcal{I}_m$ is a refinement of $\mathcal{I}_n$.*

c) *For every $n \in \mathbb{N}$ and $B \in \hat{\mathcal{S}}$, B is covered by finitely many $I_{n,j}$'s.*

d) *The σ-field generated by $\mathcal{I}$ equals $\mathscr{B}_S$.*

Now, note that as a consequence of Theorem 1.7, the class of point processes over $(S, \mathscr{B}_S)$, here denoted by $\mathcal{N}(S)$ is a Borel subset of $\mathcal{M}(S)$, hence it is Borel itself, and to its σ-algebra we denote $\mathscr{B}_{\mathcal{N}(S)}$. Define the subsets of $\mathscr{B}_{\mathcal{N}(S)}$

$$\mathcal{C}_B = \{\mu \in \mathcal{N}(S) : \mu(B) = 0\} = \pi_B^{-1}[\{0\}] \in \mathscr{B}_{\mathcal{N}(S)},$$

recalling that π_B denotes the projection map $\mu \mapsto \mu(B)$. Set $\mathcal{C} = \{\mathcal{C}_B : B \in \hat{\mathcal{S}}\}$, since $\hat{\mathcal{S}}$ is a ring, $\mathcal{C}$ is a π-system. Indeed, for $B, A \in \hat{\mathcal{S}}$, $\mu(A) = 0$ and $\mu(B) = 0$ if and only if $\mu(A \cup B) = 0$, that is $\mathcal{C}_A \cap \mathcal{C}_B = \mathcal{C}_{A\cup B}$. By hypothesis, for every $B \in \hat{\mathcal{S}}$

$$\mathbb{P}[\boldsymbol{\mu} \in \mathcal{C}_B] = \mathbb{P}[\boldsymbol{\mu}(B) = 0] = \mathbb{P}[\boldsymbol{\nu}(B) = 0] = \mathbb{P}[\boldsymbol{\nu} \in \mathcal{C}_B].$$

Thus the π-system $\mathcal{C}$ is contained in the λ-system $\mathcal{D} = \{M \in \mathscr{B}_{\mathcal{N}(S)} : \mathbb{P}[\boldsymbol{\mu} \in M] = \mathbb{P}[\boldsymbol{\nu} \in M]\}$, and a monotone class argument shows $\mathbb{P}[\boldsymbol{\mu} \in M] = \mathbb{P}[\boldsymbol{\nu} \in M]$ for every $M \in \sigma(\mathcal{C})$. To finish the proof it suffices to see $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are $\sigma(\mathcal{C})$- measurable. To this aim consider a dissection system $\mathcal{I} = \{I_{n,j}\}_{n,j}$ and define define the mapping $\mu \mapsto \mu^*$ from $\mathcal{N}(S)$ into $\mathcal{N}(S)$ given by

$$\mu^*(B) = \lim_{n\to\infty} \sum_j \mu(I_{n,j} \cap B) \wedge 1$$

for all $B \in \hat{\mathcal{S}}$. Note that by Lemma 1.2 the mapping is measurable with respect to $\mathscr{B}_{\mathcal{N}(S)} = \mathscr{B}_{\mathcal{M}(S)} \cap \mathcal{N}(S)$. Moreover, $\mu^*(B) = m$ if and only if there exist $n \geq 1$, and indexes $j_1, \dots, j_m$ such that $\mu(I_{n,j_i} \cap B) > 0$ for every $i \in \{1, \dots, m\}$ and $\mu(I_{n,j} \cap B) = 0$ for all $j \notin \{j_1, \dots, j_m\}$. Hence $\mu \mapsto \mu^*$ is even $\sigma(\mathcal{C})$-measurable. Finally since $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are simple, we get $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ and $\boldsymbol{\nu} = \boldsymbol{\nu}^*$, which shows $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are $\sigma(\mathcal{C})$-measurable. □

A.7 Proof of Proposition 1.9

First we prove (i), if $\boldsymbol{\kappa}$ is non-random, so that $\boldsymbol{\kappa} = n$ for some $n \in \mathbb{N}$, we get $\boldsymbol{\mu} = \sum_{j=1}^n \delta_{\boldsymbol{\xi}_j}$, for some $(\boldsymbol{\xi}_j)_{j=1}^n \overset{\text{iid}}{\sim} \mu_0$. Thus

$$\boldsymbol{\mu}(f) = \int f(s)\boldsymbol{\mu}(ds) = \sum_{j=1}^n f(\boldsymbol{\xi}_j),$$

and

$$\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right] = \prod_{j=1}^n \mathbb{E}\left[e^{-f(\boldsymbol{\xi}_j)}\right] = \left(\int e^{-f(s)}\mu_0(ds)\right)^n = \left(\mu_0\left(e^{-f}\right)\right)^n.$$

if $\boldsymbol{\kappa}$ is random, (i) follows by conditioning. To prove (ii) first assume $\boldsymbol{\nu} = \nu$ for some deterministic random measure $\nu \in \mathcal{M}(S)$. Recall that for $\mathbf{x} \sim \mathsf{Poi}(\theta)$, its probability generating function is $\mathbb{E}[a^{\mathbf{x}}] = e^{-\theta(1-a)}$. Then, for a positive simple function $f = \sum_{i=1}^n b_i \mathbf{1}_{B_i}$, (where $\{B_i\}_{i=1}^n$ are disjoint),

$$\begin{aligned}
\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right] &= \mathbb{E}\left[\exp\left\{-\sum_{i=1}^n b_i\boldsymbol{\mu}(B_i)\right\}\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[\left(e^{-b_i}\right)^{\boldsymbol{\mu}(B_i)}\right] \\
&= \prod_{i=1}^n \exp\left\{-\nu(B_i)\left(1-e^{-b_i}\right)\right\} \\
&= \exp\left\{-\sum_{i=1}^n \left(1-e^{-b_i}\right)\nu(B_i)\right\} \\
&= \exp\left\{-\sum_{i=1}^n \int \left(1-e^{-b_i\mathbf{1}_{B_i}(s)}\right)\nu(ds)\right\} \\
&= \exp\left\{-\int \left(1-e^{-\sum_{i=1}^n b_i\mathbf{1}_{B_i}(s)}\right)\nu(ds)\right\} \\
&= \exp\left\{-\int \left(1-e^{-f(s)}\right)\nu(ds)\right\} \\
&= \exp\left\{-\nu\left(1-e^{-f}\right)\right\}
\end{aligned}$$

Any $f \in S_+$ can be approximated by (positive) simple functions $f_n \nearrow f$, and using monotone convergence and Lebesgue dominated convergence theorem, the result extends for arbitrary $f \in S_+$. Finally the case where $\boldsymbol{\mu}$ is directed by the random measure, $\boldsymbol{\nu}$, follows by conditioning. □

A.8 Proof of Proposition 1.10

First we prove (i), by conditioning we can reduce to that case where $\boldsymbol{\mu}$ is a Poisson process, so that $\boldsymbol{\nu} = \nu$ is non-random. Note that both properties are local, that is $\boldsymbol{\mu}$ is simple if and only if the restriction, $\mathbf{1}_B\boldsymbol{\mu}$, is simple for every bounded set $B \in \hat{\mathcal{S}}$ and analogously ν is diffuse if and only if $\mathbf{1}_B\nu$ is diffuse. This said, since $\boldsymbol{\mu}$ and ν are locally finite, we may assume without loss of generality that $\boldsymbol{\mu}$ and ν are finite. Now, if $\nu(S) = 0$, $\boldsymbol{\mu}(S) = 0$ almost surely, and we fall in degenerate non-interesting scenario. If $\nu(S) > 0$ we can normalize this measure to attain a probability measure $\mu = \nu/\nu(S)$. Let $\boldsymbol{\eta}$ be a mixed Binomial process based on $(\mathbf{n}, \mu)$, for some $\mathbf{n} \sim \mathsf{Poi}(\nu(S))$. By Proposition 1.9

$$\mathbb{E}\left[e^{-\boldsymbol{\eta}(f)}\right] = \mathbb{E}\left[\left(\mu\left(e^{-f}\right)\right)^{\mathbf{n}}\right] = \exp\left\{-\nu(S)\mu\left(1-e^{-f}\right)\right\} = \exp\left\{-\nu\left(1-e^{-f}\right)\right\},$$

that is $\mathbb{E}\left[e^{-\boldsymbol{\eta}(f)}\right] = \mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right]$ for any $f \in S_+$, and by Theorem 1.5, we get $\boldsymbol{\eta} \overset{d}{=} \boldsymbol{\mu}$. This means that $\boldsymbol{\mu}$ is a mixed Binomial process, and we may write $\boldsymbol{\mu} = \sum_{j=1}^{\boldsymbol{\kappa}} \delta_{\boldsymbol{\xi}_j}$ for a collection $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \mu$ and some independent Poisson random variable $\boldsymbol{\kappa}$. Let $\mathcal{A} = \{s \in S : \mu(\{s\}) > 0\}$ be the set of atoms of μ, which is at most countable. Then, By Fubini's theorem we have that, for every $i \neq j$

$$\mathbb{P}\left[\boldsymbol{\xi}_i = \boldsymbol{\xi}_j\right] = \int\int \mathbf{1}_{\{s\}}(t)\,\mu(dt)\mu(ds) = \int \mu(\{s\})\mu(ds) = \sum_{s\in A} \mu(\{s\})^2.$$

Hence $\mathbb{P}\left[\boldsymbol{\xi}_i = \boldsymbol{\xi}_j\right] = 0$ if and only if $\mathcal{A} = \emptyset$, that is $\boldsymbol{\mu}$ is simple if and only if μ is diffuse. Finally note that the set of atoms of μ is precisely that of ν.

To prove (ii) assume one more time that $\boldsymbol{\mu}$ is a Poisson random measure directed by ν. By Proposition 1.9, we get that for every $f \in S_+$ and each $t > 0$,

$$\exp\left\{-\nu\left(1 - e^{-tf}\right)\right\} = \mathbb{E}\left[e^{-\boldsymbol{\mu}(tf)}\right] = \mathbb{E}\left[e^{-t\boldsymbol{\mu}(f)}\mathbf{1}_{\{\boldsymbol{\mu}(f)<\infty\}}\right].$$

Thus, by Lebesgue dominated convergence theorem, as $t \to 0$,

$$\exp\left\{-\nu\left(1 - e^{-tf}\right)\right\} \to \mathbb{E}\left[\mathbf{1}_{\{\boldsymbol{\mu}(f)<\infty\}}\right] = \mathbb{P}\left[\boldsymbol{\mu}(f) < \infty\right]. \tag{A.1}$$

Now, if $\nu(f \wedge 1) = \infty$ we get $\nu\left(1 - e^{-tf}\right) = \infty$, so by equation A.1, $\mathbb{P}\left[\boldsymbol{\mu}(f) < \infty\right] = 0$. Conversely if $\nu(f \wedge 1) < \infty$ we have that $\nu\left(1 - e^{-tf}\right) \to 0$ as $t \to 0$, thus $\mathbb{P}\left[\boldsymbol{\mu}(f) < \infty\right] = 1$. This proves the statement for Poisson process, the case where $\boldsymbol{\mu}$ is a Cox process, follows by conditioning. □

A.9 Proof of Lemma 1.11

To prove (i), first assume $\boldsymbol{\mu} = \sum_k \delta_{s_k}$, is not random. Choosing $(\boldsymbol{\tau}_k)_{k\geq 1}$ some independent random elements taking values in T such that $\boldsymbol{\tau}_k$ has distribution $\boldsymbol{\nu}_{s_k}$ we obtain that for $f \in S_+$

$$\begin{aligned}
\mathbb{E}\left[e^{-\boldsymbol{\eta}(f)}\right] &= \mathbb{E}\left[\exp\left\{-\sum_k f(\tau_k)\right\}\right] \\
&= \prod_k \mathbb{E}\left[e^{-f(\tau_k)}\right] \\
&= \prod_k \boldsymbol{\nu}_{s_k}\left(e^{-f}\right) \\
&= \exp\left\{\sum_k \log\left(\boldsymbol{\nu}_{s_k}\left(e^{-f}\right)\right)\right\}
\end{aligned}$$

Now, set $g(s) = \log\left(\boldsymbol{\nu}_s\left(e^{-f}\right)\right)$ for each $s \in S$, clearly $g : S \to \mathbb{R}_-$, it is a measurable function and

$$\exp\left\{\sum_k \log\left(\boldsymbol{\nu}_{s_k}\left(e^{-f}\right)\right)\right\} = \exp\left\{\sum_k g(s_k)\right\} = \exp\left\{\boldsymbol{\mu}(g)\right\} = \exp\left\{\boldsymbol{\mu}(\log\left(\boldsymbol{\nu}\left(e^{-f}\right)\right)\right\}$$

The scenario where $\boldsymbol{\mu}$ is not deterministic follows by conditioning on $\boldsymbol{\mu}$.

To prove (ii), again assume $\boldsymbol{\mu} = \sum_k \delta_{s_k}$ is not random, note that we can write $\tilde{\boldsymbol{\eta}} = \sum_k \boldsymbol{\alpha}_k \delta_{s_k}$, where $(\boldsymbol{\alpha}_k)_{k\geq 1}$ are independent random variables such that $\boldsymbol{\alpha}_k \sim \mathsf{Ber}(p(s_k))$, then $\mathbb{E}\left[e^{-t\boldsymbol{\alpha}_k}\right] = e^{-t}p(s_k) + (1 - p(s_k)) = 1 - p(s_k)(1 - e^{-t})$ for every $t \in \mathbb{R}$. This implies that for every $f \in S_+$

$$\begin{aligned}
\mathbb{E}\left[e^{-\tilde{\boldsymbol{\eta}}(f)}\right] &= \mathbb{E}\left[\exp\left\{-\sum_k \boldsymbol{\alpha}_k f(s_k)\right\}\right] \\
&= \prod_k \mathbb{E}\left[\exp\left\{-\boldsymbol{\alpha}_k f(s_k)\right\}\right] \\
&= \prod_k \left\{1 - p(s_k)\left(1 - e^{-f(s_k)}\right)\right\} \\
&= \exp\left\{\sum_k \log\left(\left\{1 - p(s_k)\left(1 - e^{-f(s_k)}\right)\right\}\right)\right\}
\end{aligned}$$

setting $g(s) = \log\left\{1 - p(s)\left(1 - e^{f(s)}\right)\right\}$ get that $g : S \to \mathbb{R}_-$ is a measurable and

$$\exp\left\{\sum_k \log\left(\left\{1 - p(s_k)(1 - e^{-f(s_k)})\right\}\right)\right\} = \exp\left\{\sum_k g(s_k)\right\} = \exp\left\{\boldsymbol{\mu}(g)\right\}.$$

Whenever $\boldsymbol{\mu}$ is not deterministic, we simply condition on $\boldsymbol{\mu}$ to obtain the desired statement. □

A.10 Proof of Theorem 1.12

To provide the desired result we will need some preliminary lemmas.

Lemma A.2. *Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be either simple or diffuse locally finite random measures over $(S, \mathscr{B}_S)$, and consider a constant $c > 0$, then $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$ if and only if $\mathbb{E}\left[e^{-c\boldsymbol{\mu}(B)}\right] = \mathbb{E}\left[e^{-c\boldsymbol{\nu}(B)}\right]$ and every $B \in \hat{\mathcal{S}}$.*

Proof: Clearly $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$ implies $\mathbb{E}\left[e^{-c\boldsymbol{\mu}(B)}\right] = \mathbb{E}\left[e^{-c\boldsymbol{\nu}(B)}\right]$ for every $B \in \mathscr{B}_S$. To prove the converse result, first assume $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are both diffuse. Define the Cox processes $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ directed by $c\boldsymbol{\mu}$ and $c\boldsymbol{\nu}$ respectively. Note that Proposition 1.9 remains true for $0 \leq f \leq \infty$. Let $B \in \hat{\mathcal{S}}$, set $f = \infty \mathbf{1}_B$ and note that

$$e^{-\boldsymbol{\mu}^*(f)} = \begin{cases} 1 & \text{if } \boldsymbol{\mu}^*(B) = 0 \\ 0 & \text{if } \boldsymbol{\mu}^*(B) > 0 \end{cases}$$

with the convention $\infty * 0 = 0$. That is $e^{-\boldsymbol{\mu}^*(f)} \sim \mathsf{Ber}(\mathbb{P}[\boldsymbol{\mu}^*(B) = 0])$, and analogously for $\boldsymbol{\nu}^*$. So by hypothesis and Proposition 1.9 we get

$$\mathbb{P}[\boldsymbol{\mu}^*(B) = 0] = \mathbb{E}\left[e^{-\boldsymbol{\mu}^*(f)}\right] = \mathbb{E}\left[e^{-c\boldsymbol{\mu}(B)}\right] = \mathbb{E}\left[e^{-c\boldsymbol{\nu}(B)}\right] = \mathbb{E}\left[e^{-\boldsymbol{\nu}^*(f)}\right] = \mathbb{P}[\boldsymbol{\nu}^*(B) = 0].$$

Since $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are diffuse we must have $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ are simple, and by Lemma 1.8 we get $\boldsymbol{\mu}^* \overset{d}{=} \boldsymbol{\nu}^*$. Finally, since the laws of a Cox process and its intensity measure characterize each other we get $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$.

Now assume $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are simple. Fix $B \in \hat{\mathcal{S}}$ and define $p = (1 - e^{-c})$. Let $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ be p-thinnings of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ respectively. Let $f = \infty \mathbf{1}_B$ and note that Lemma 1.11 remains true for $0 \leq f \leq \infty$. Thus

$$\mathbb{P}[\boldsymbol{\mu}^*(B) = 0] = \mathbb{E}\left[e^{-\boldsymbol{\mu}^*(f)}\right] = \mathbb{E}\left[e^{-\boldsymbol{\mu}(\log(1 - p\mathbf{1}_B))}\right] = \mathbb{E}\left[e^{-\boldsymbol{\mu}(B)\log(1-p)}\right] = \mathbb{E}\left[e^{-c\boldsymbol{\mu}(B)}\right],$$

and analogously for $\boldsymbol{\nu}^*$ and $\boldsymbol{\nu}$. Since $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are simple, we must have that the same holds for $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$, so by Lemma 1.8 we obtain $\boldsymbol{\nu}^* \overset{d}{=} \boldsymbol{\mu}^*$, and by construction it is obvious this implies $\boldsymbol{\nu} \overset{d}{=} \boldsymbol{\mu}$ as desired. □

Lemma A.3. *Let $(S, \mathscr{B}_S)$ be a Borel space and $\boldsymbol{\mu}$ a locally finite random measure.*

i) *If $\boldsymbol{\mu}$ is simple and $\mathbb{E}[\boldsymbol{\mu}(\{s\})] = 0$ for all $s \in S$, then $\boldsymbol{\mu}$ is completely random if and only if it is a Poisson random measure.*

ii) *If $\boldsymbol{\mu}$ is diffuse, then it is completely random if and only if it is non-random.*

Proof: Clearly if $\boldsymbol{\mu}$ is Poisson or non-random then it is completely random, so we will prove the converse results. Let $\boldsymbol{\mu}$ be a locally finite completely random measure and define the set function $\rho : \hat{\mathcal{S}} \to \mathbb{R}_+$ by

$$\rho(B) = -\log\left(\mathbb{E}\left[e^{-\boldsymbol{\mu}(B)}\right]\right) \tag{A.2}$$

Note that $\rho(\emptyset) = 0$ and that for disjoint $A, B \in \hat{\mathcal{S}}$

$$\rho(A \cup B) = -\log\left(\mathbb{E}\left[e^{-\boldsymbol{\mu}(A)}\right]\mathbb{E}\left[e^{-\boldsymbol{\mu}(B)}\right]\right) = \rho(A) + \rho(B),$$

which shows ρ is finitely additive. In fact, ρ is even countably additive since $B_n \nearrow B$ in $\hat{\mathcal{S}}$ implies $\rho(B_n) \nearrow \rho(B)$ by monotone and Lebesgue dominated convergence theorems. Further $\rho(B) < \infty$ for all $B \in \hat{\mathcal{S}}$, $\hat{\mathcal{S}}$ is a ring and $(S, \mathscr{B}_S)$ is Borel, thus by Carathéodory's extesion theorem ρ can be uniquely extended to a measure over $(S, \mathscr{B}_S)$. Finally note that $\mathbb{E}[\boldsymbol{\mu}(\{s\})] = 0$ in either (i) or (ii), thus $\boldsymbol{\mu}(\{s\}) = 0$ almost surely, and we get $\rho(\{s\}) = -\log\left(\mathbb{E}\left[e^{-\boldsymbol{\mu}(\{s\})}\right]\right) = 0$, that is ρ is diffuse.

Now, say (i) holds, and let $\boldsymbol{\nu}$ be a Poisson random measure with intensity $c\rho$, where $c = (1 - e^{-1})^{-1}$, by Proposition 1.9 we know

$$\mathbb{E}\left[e^{-\boldsymbol{\nu}(B)}\right] = \exp\left\{-c\rho\left(1 - e^{-\mathbf{1}_B}\right)\right\} = \exp\left\{-\rho(B)\right\} = \mathbb{E}\left[e^{-\boldsymbol{\mu}(B)}\right].$$

$\boldsymbol{\mu}$ is simple by hypothesis and since ρ is diffuse, $\boldsymbol{\nu}$ is simple by Proposition 1.10. Hence, Lemma A.2 yields $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu}$, which proves $\boldsymbol{\mu}$ is Poisson.

If $\boldsymbol{\mu}$ is diffuse, set $\boldsymbol{\nu} = \rho$, then $\mathbb{E}\left[e^{-\boldsymbol{\nu}(B)}\right] = \exp\left\{-\rho(B)\right\} = \mathbb{E}\left[e^{-\boldsymbol{\mu}(B)}\right]$. By Lemma A.2 the result follows. □

For a locally finite completely random measure, $\boldsymbol{\mu}$, we define its the set of fixed atoms as $\mathcal{D}_{\boldsymbol{\mu}} = \{s \in S : \rho(s) > 0\}$ where ρ is as in (A.2). It is easy to see that $\mathcal{D}_{\boldsymbol{\mu}} = \{s \in S : \mathbb{E}\left[\boldsymbol{\mu}(\{s\})\right] > 0\} = \{s \in S : \mathbb{P}[\boldsymbol{\mu}(\{s\}) > 0] > 0\}$.

Lemma A.4. *Let $\boldsymbol{\mu} = \sum_{j \geq 1} \boldsymbol{\alpha}_j \delta_{\boldsymbol{\xi}_j}$ be a locally finite completely random measure over $(S, \mathscr{B}_S)$ with no fixed atoms. Then $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j \geq 1}$ defines a Poisson random measure over $S \times \mathbb{R}_+$ whose diffuse intensity, ν, satisfies*

$$\int_{\mathbb{R}_+} (x \wedge 1)\, \nu(B, dx) < \infty,$$

$B \in \hat{\mathcal{S}}$. In particular, we get that $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j \geq 1}$ are i.i.d.

Proof: Define the random measure over $S \times \mathbb{R}_+$,

$$\boldsymbol{\eta} = \sum_{j\geq 1} \delta_{(\boldsymbol{\xi}_j,\boldsymbol{\alpha}_j)} = \sum_{j\geq 1} \delta_{\boldsymbol{\xi}_j} \otimes \delta_{\boldsymbol{\alpha}_j}.$$

Notice that since $\boldsymbol{\mu}$ has no fixed atoms, then the atoms $\boldsymbol{\xi}_j$'s of $\boldsymbol{\mu}$ must be pairwise distinct almost surely, hence $\boldsymbol{\eta}$ must be simple. Let $\hat{\mathcal{T}}$ be a localizing ring of $S \times \mathbb{R}_+$, fix any $B \in \hat{\mathcal{T}}$ and consider the random measure over S, $\boldsymbol{\eta}_B = \mathbf{1}_B\boldsymbol{\eta}(\cdot \times \mathbb{R}_+)$. Note that $\boldsymbol{\eta}(\cdot \times \mathbb{R}_+) = \sum_{j\geq 1} \delta_{\boldsymbol{\xi}_j}$ inherits the independent increments from $\boldsymbol{\mu}$ and so does $\boldsymbol{\eta}_B$. Further $\boldsymbol{\eta}_B$ is simple and $\mathbb{E}[\boldsymbol{\eta}_B(\{s\})] = 0$ for every $s \in S$. Then, by Lemma A.3, we get $\boldsymbol{\eta}_B$ is Poisson, in particular $\boldsymbol{\eta}(B) = \boldsymbol{\eta}_B(S)$ is Poisson distributed. Since $B \in \hat{\mathcal{T}}$ was chosen arbitrarily, we have shown that $\boldsymbol{\eta}(B)$ follows a Poisson distribution for all $B \in \hat{\mathcal{T}}$, and Lemma A.2 yields $\boldsymbol{\eta}$ is a Poisson random measure. Now, as explained in the proof of the first part of Proposition 1.10, $\mathbf{1}_B\boldsymbol{\eta}$ is a mixed Poisson Binomial process for all $B \in \hat{\mathcal{T}}$, which implies $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ are i.i.d. Finally, given that $\boldsymbol{\eta}$ is simple, by Proposition 1.10 its intensity measure, ν, must be diffuse, further as $\boldsymbol{\mu}$ is locally finite, we get that for every $A \in \hat{\mathcal{S}}$, $t\boldsymbol{\mu}(A) = \sum_{j\geq 1} t\boldsymbol{\alpha}_j\delta_{\boldsymbol{\xi}_j}(A) = \boldsymbol{\eta}(f) < \infty$, where $f \in (S \times \mathbb{R}_+)_+$ is given by $f(s, x) = x\mathbf{1}_A(s)$. Thus, by the second part of Proposition 1.10 we get,

$$\int_{\mathbb{R}_+} (x \wedge 1)\, \nu(A, dx) = \nu\,(f \wedge 1) < \infty,$$

□

Proof of Theorem 1.12: The necessity is obvious, so we prove the sufficiency. As $\boldsymbol{\mu}$ is locally finite, its set of fixed atoms is at most countable, let $s_1, s_2, \ldots$, be such atoms. The independent increments of $\boldsymbol{\mu}$ assure $\boldsymbol{\gamma}_j = \boldsymbol{\mu}(\{s_j\})$ is a non-negative random variable independent of the rest of the process. We can subtract from $\boldsymbol{\mu}$, the component representing the fixed atoms, $\sum_{j\geq 1} \boldsymbol{\gamma}_j\delta_{s_j}$. The remaining discontinuities can be measurably encoded by

$$\boldsymbol{\mu}^* = \sum_{j\geq 1} \boldsymbol{\alpha}_j\delta_{\boldsymbol{\xi}_j},$$

as explained in Theorem 1.7. $\boldsymbol{\mu}^*$ inherits from $\boldsymbol{\mu}$ the independent increments, and has no fixed atoms, so Lemma A.4 yields $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ are as in (b). Finally, subtracting even this part, we end up with a diffuse measure, β, with independent increments, and by Lemma A.3 we get it is non-random. □

A.11 Proof of Theorem 1.13

I ⇒ II: Let A be a closed set. Define the function $f_\varepsilon(s) = (1 - d(s, A)/\varepsilon)^+$, with the shorthand $z^+ = \max\{0, z\}$. So that f_ε is bounded and continuous and

$$f_\varepsilon(s) \in \begin{cases} \{1\} & \text{if } s \in A \\ \{0\} & \text{if } d(s, A) \geq \varepsilon \\ (0, 1) & \text{if } 0 < d(s, A) < \varepsilon, \end{cases}$$

hence, $\mathbf{1}_A \leq f_\varepsilon \leq \mathbf{1}_{A^\varepsilon}$, where $A^\varepsilon = \{s \in S : d(s, A) < \varepsilon\}$. By I. we obtain

$$\limsup_n \mu_n(A) = \limsup_n \mu_n(\mathbf{1}_A) \leq \limsup_n \mu_n(f_\varepsilon) = \mu(f_\varepsilon) \leq \mu(A^\varepsilon). \tag{A.3}$$

Since A is closed, we have $A = \bigcap_\varepsilon A^\varepsilon$, so by making $\varepsilon \to 0$, we get $\mu(A^\varepsilon) \to \mu(A)$, which together with equation (A.3) gives the result.

II $\Leftrightarrow$ III: To show II $\Leftrightarrow$ III, it suffices to consider complements of the corresponding sets.

II,III $\Rightarrow$ IV: For any set A, the boundary of A can be written as $\partial A = \overline{A} \setminus A^\circ$, where $\overline{A}$ denotes the closure of A, and A° its interior. As $\overline{A}$ is closed and A° is open, by II and III,

$$\mu\left(\overline{A}\right) \geq \limsup_n \mu_n\left(\overline{A}\right) \geq \liminf_n \mu_n(A^\circ) \geq \mu(A^\circ).$$

Since A is a μ-continuity set we have that $\mu\left(\overline{A}\right) = \mu(A) = \mu(A^\circ)$, thus the above inequalities are in fact equalities, which gives $\lim_n \mu(A) = \mu(A)$.

IV $\Rightarrow$ I: Fix a continuous and bounded function f. As f is bounded, there exist $x, y \in \mathbb{R}$ such that $x \leq f \leq y$. Now,

$$\mu(f) = \int f d\mu = \int_x^y \mu(\{f > t\})dt, \tag{A.4}$$

where $\{f > t\} = \{s \in S : f(s) > t\}$, and the same holds for μ_n. For any $t \in (x, y)$, and $s \in \partial\{f > t\}$, we may take $a_1, a_2, \ldots \in \{f > t\}$ and $b_1, b_2, \ldots \in \{f \leq t\}$ with $a_n \to s$ and $b_n \to s$. As f is continuous we have that $f(s) = \lim_n f(a_n) \geq t$ and $f(s) = \lim_n f(b_n) \leq t$, hence $f(s) = t$. That is, $\partial\{f > t\} \subseteq \{f = t\}$, which means that, $\mu(\{f = t\}) = 0$ implies $\{f > t\}$ is a μ-continuity set. As μ is a probability measure, there are at most countably many t's for which $\mu(\{f = t\}) > 0$, hence by IV,

$$\mu_n(\{f > t\}) \to \mu(\{f > t\})$$

almost everywhere on (x, y). This together with (A.4), show that $\mu_n(f) \to \mu(f)$. □

A.12 Proof of Theorem 1.14

Evidently $d_{\mathcal{P}}(\mu, \nu) = 0$ if and only if $\mu = nu$, and $d_{\mathcal{P}}(\mu, \nu) = d_{\mathcal{P}}(\nu, \mu)$ for every $\mu, \nu \in \mathcal{P}(S)$. So to prove that $d_{\mathcal{P}}$ is a metric it suffices to check the triangle inequality. Let $\mu, \nu, \xi \in \mathcal{P}(S)$, and say $d_{\mathcal{P}}(\mu, \nu) < \varepsilon$, $d_{\mathcal{P}}(\nu, \xi) < \lambda$. Then, for every $B \in \mathscr{B}_S$,

$$\mu(B) \leq \nu(B^\varepsilon) + \varepsilon \leq \xi\left([B^\varepsilon]^\lambda\right) + \lambda + \varepsilon \leq \xi\left(B^{\varepsilon+\lambda}\right) + \lambda + \varepsilon$$

and analogously $\xi(B) \leq \mu\left(B^{\varepsilon+\lambda}\right) + \lambda + \varepsilon$. Thus $d_{\mathcal{P}}(\mu, \xi) \leq \lambda + \varepsilon$, for all such ε and λ. By considering the corresponding infimum over ε and λ we obtain $d_{\mathcal{P}}(\mu, \xi) \leq d_{\mathcal{P}}(\mu, \nu) + d_{\mathcal{P}}(\nu, \xi)$.

To prove the remaining part of the statement, first say that $d_{\mathcal{P}}(\mu_n, \mu) \to 0$, for some $\mu, \mu_1, \mu_2, \ldots \in \mathcal{P}(S)$. Then, we may take $\varepsilon_1, \varepsilon_2, \ldots$, with $d_{\mathcal{P}}(\mu_n, \mu) < \varepsilon_n$ and $\varepsilon_n \to 0$. This way for every closed set A,

$$\limsup_n \mu_n(A) \leq \limsup_n \mu\left(A^{\varepsilon_n}\right) + \varepsilon_n = \mu\left(\bigcap_n A^{\varepsilon_n}\right) = \mu(A).$$

and we obtain $\mu_n \xrightarrow{w} \mu$. To show the converse result we first prove the following lemma

Lemma A.5. *If S is separable. Then, for every $\mu \in \mathcal{P}(S)$ and $\delta > 0$, we may find a countable collection of open balls, $B_1, B_2, \ldots$, whose radius is smaller than δ, $\mu(\partial B_n) = 0$ for all n, and $\bigcup_n B_n = S$.*

Proof of Lemma A.5: Let D be a countable dense set in S. For $x \in D$ define $S(x,r) = \{s \in S : d(x,s) = r\}$ and $B(x,r) = \{s \in S : d(x,s) < r\}$. Observe that $\partial B(x,r) \subseteq S(x,r)$. The collection $\mathcal{S} = \{S(x,r) : \delta/2 < r < \delta\}$ is disjoint, hence at most countably many of its members satisfy $\mu(S(x,r)) > 0$. As, $\mathcal{S}$ is uncountable, there exist $r_x \in (\delta/2, \delta)$ with $\mu(S(x,r_x)) = 0$, thus $\mu(\partial B(x,r_x)) = 0$. Finally as D is dense $\bigcup_{x\in D} B(x,r_x) = S$, and as D is countable $\{B(x,r_x)\}_{x\in D}$ is also countable. $\square$

Returning to the proof of Theorem 1.14, say that $\mu_n \xrightarrow{w} \mu$ for some $\mu, \mu_1, \mu_2, \ldots, \in \mathcal{P}(S)$. Let $\varepsilon > 0$ and take $0 < \delta < \varepsilon/3$. By Lemma A.5 we may find a countable collection of open balls, $B_1, B_2, \ldots$, whose radius is smaller than $\delta/2$, $\mu(\partial B_j) = 0$ for all j, and $\bigcup_j B_j = S$. Fix k such that $\mu\left(\bigcup_{j=1}^k B_j\right) \geq 1 - \delta$, and consider the collection

$$\mathcal{A} = \left\{ \bigcup_{j\in K} B_j : K \subseteq \{1, \ldots, k\} \right\}$$

For every $A \in \mathcal{A}$, $\partial A \subseteq \bigcup_{j=1}^k \partial B_n$, thus $\mu(\partial A) \leq \sum_{j=1}^k \mu(\partial B_n) = 0$. By the Portmanteau theorem (IV) we get $\mu_n(A) \to \mu(A)$ for every $A \in \mathcal{A}$. As $\mathcal{A}$ is finite we may find N such that, $|\mu_n(A) - \mu(A)| < \delta$, for every $n \geq N$ and every $A \in \mathcal{A}$. Particularly, $\mu_n\left(\bigcup_{j=1}^k B_j\right) \geq \mu_n\left(\bigcup_{j=1}^k B_j\right) - \delta \geq 1 - 2\delta$, for $n \geq N$. Now fix $B \in \mathscr{B}_S$, and set

$$A = \bigcup \{B_j : j \in \{1, \ldots, k\}, B_j \cap B \neq \emptyset\} \in \mathcal{A}.$$

We find that

- $A \subseteq B^\delta \subseteq B^\varepsilon$, as the diameter of B_j is smaller than δ.
- $B \subseteq A \cup \left(\bigcup_{j=1}^k B_j\right)^c$
- $\mu(A) \leq \mu_n(A) + \delta$ and $\mu_n(A) \leq \mu(A) + \delta$, for all $n \geq N$.
- $\mu\left(\left[\bigcup_{j=1}^k B_j\right]^c\right) \leq \delta$ and $\mu_n\left(\left[\bigcup_{j=1}^k B_j\right]^c\right) \leq 2\delta$, for every $n \geq N$.

Putting this together we obtain

$$\mu(B) \leq \mu(A) + \mu\left(\left[\bigcup_{j=1}^k B_j\right]^c\right) \leq \mu_n(A) + 2\delta \leq \mu_n\left(B^\varepsilon\right) + \varepsilon$$

and

$$\mu_n(B) \leq \mu_n(A) + \mu_n\left(\left[\bigcup_{j=1}^k B_j\right]^c\right) \leq \mu(A) + 3\delta \leq \mu\left(B^\varepsilon\right) + \varepsilon,$$

that is $d_{\mathcal{P}}(\mu_n, \mu) \leq \varepsilon$, for every $n \geq N$. As ε was arbitrary, this shows $d_{\mathcal{P}}(\mu_n, \mu) \to 0$. $\square$

A.13 Proof of Theorem 1.16

Let $\Sigma_1, \Sigma_2, \Sigma_3$ and Σ_4 denote the σ-algebras generated by I,II,III and IV respectively.

$\Sigma_1 = \Sigma_2$: For any continuous and bounded function $f : S \to \mathbb{R}$, we have that the mapping $\pi_f : \mu \mapsto \mu(f)$ is continuous with respect to the topology of weak converge (this is a consequence of Theorems 1.13 and 1.14). Hence π_f is measurable with respect to Σ_1, and we get $\Sigma_2 \subseteq \Sigma_1$. Conversely, by Theorem 1.15 we have that $(\mathcal{P}(S), d_{\mathcal{P}})$ is separable, hence every (weakly) open set is a countable union of basis sets

$$\{\mu : |\mu(f) - r| < \varepsilon, f \text{ is continuous and bounded}, r, \varepsilon > 0\} \in \Sigma_2.$$

Thus $\Sigma_1 \subseteq \Sigma_2$, and we obtain $\Sigma_1 = \Sigma_2$.

$\Sigma_3 = \Sigma_4$: For any $B \in \mathscr{B}_S$ and any $\mu \in \mathcal{P}(S)$, we have that $\int \mathbf{1}_B d\mu = \mu(B)$, hence the projection map π_B equals the integration map $\pi_{\mathbf{1}_B}$ and $\Sigma_4 \subseteq \Sigma_3$ follows. Conversely, if f is a simple measurable function, $f = \sum_{i=1}^n a_1 \mathbf{1}_{A_i}$, (with A_i and A_j disjoint, for $i \neq j$). By linearity of the integral we have that $\pi_f = \sum_{i=1}^n a_i \pi_{A_i}$, from which we obtain that π_f is Σ_4-measurable. Now, if f is an arbitrary measurable and non-negative function then f and be approximated by simple functions $f_n \nearrow f$ and by monotone convergence theorem we obtain $\pi_{f_n} \nearrow \pi_f$, hence π_f is also Σ_4-measurable, and we get $\Sigma_3 = \Sigma_4$.

$\Sigma_4 \subseteq \Sigma_2$: Let $B \in \mathscr{B}_S$ and define $f_\varepsilon : S \to \mathbb{R}$ by $f_\varepsilon = (1 - d(s,B)/\varepsilon)^+$. Then $0 \leq f_\varepsilon \leq 1$ is continuous and bounded, and $f_\varepsilon \to \mathbf{1}_B$ as $\varepsilon \to 0$. By Lebesgue dominated convergence theorem, we obtain $\pi_{f_\varepsilon} \to \pi_{\mathbf{1}_B} = \pi_B$. Since π_{f_ε} is Σ_2-measurable for each $\varepsilon > 0$, this shows the projection map π_B is Σ_2-measurable. As B was chosen arbitrarily, this shows $\Sigma_4 \subseteq \Sigma_2$.

$\Sigma_2 \subseteq \Sigma_3$: Let $f : S \to \mathbb{R}$ be a continuous and bounded function. As it is continuous, then it is measurable and we may decompose it as the sum of two measurable non-negative functions, its positive and negative parts. That is $f = f^+ - f^-$, by linearity of the integral $\pi_f = \pi_{f^+} - \pi_{f^-}$, from which is easy to see that π_f is Σ_3 measurable. Thus $\Sigma_2 \subseteq \Sigma_3$. □

A.14 Proof of Lemma 1.17

First note that for every bounded function $f : S \to \mathbb{R}$ and each random probability measure $\boldsymbol{\mu}$ over $(S, \mathscr{B}_S)$, the random variable $\boldsymbol{\mu}(f)$ is bounded, hence $\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right]$ exists. This said it is trivial that II implies III. To see the converse simply note that for every $t \in \mathbb{R}$, tf is also a continuous and bounded function from which it is clear that III implies II. Now we prove III implies I, fix $n \in \mathbb{N}$ and let $B_1, \ldots, B_n \in \mathscr{B}_S$ be mutually disjoint. For $\varepsilon \in (0,1)$ and $i \in \{1, \ldots, n\}$ define $f_i^{(\varepsilon)} : S \to \mathbb{R}$ by

$$f_i^{(\varepsilon)}(s) = \begin{cases} 1 - d(s, B_i)/\varepsilon & \text{if } s \in B_i^\varepsilon \\ 0 & \text{if } s \notin B_i^\varepsilon \end{cases}$$

where $B_i^\varepsilon = \{s \in S : d(s, B_i) < \varepsilon\}$. Then $f_i^{(\varepsilon)}$ is continuous and bounded and $f_i^{(\varepsilon)} \to \mathbf{1}_{B_i}$ as $\varepsilon \to 0$. Let $b_1, \ldots, b_n \in \mathbb{R}$ and note that $f^{(\varepsilon)} = \sum_{i=1}^n b_i f_i^{(\varepsilon)}$ is also continuous and bounded for all $\varepsilon \in (0,1)$, and $f^{(\varepsilon)} \to \sum_{i=1}^n b_i \mathbf{1}_{B_i}$ as $\varepsilon \to 0$. By III we know that $\mathbb{E}\left[e^{-\boldsymbol{\mu}\left(f^{(\varepsilon)}\right)}\right] = \mathbb{E}\left[e^{-\boldsymbol{\nu}\left(f^{(\varepsilon)}\right)}\right]$, and by Lebesgue dominated convergence theorem we obtain $\mathbb{E}\left[e^{-\sum_{i=1}^n b_i \boldsymbol{\mu}(B_i)}\right] = \mathbb{E}\left[e^{-\sum_{i=1}^n b_i \boldsymbol{\nu}(B_i)}\right]$. This shows

$$(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n)) \overset{d}{=} (\boldsymbol{\nu}(B_1), \ldots, \boldsymbol{\nu}(B_n)),$$

and by Theorem 1.5, this proves I. Conversely, say I holds and fix a continuous and bounded function $f : S \to \mathbb{R}_+$, consider its positive and negative parts, $f^+(s) = \max\{0, f(s)\}$ and $f^-(s) = -\min\{0, f(s)\}$, so that $f = f^+ - f^-$ with $f^+, f^- \in S_+$. For every $a, b \in \mathbb{R}_+$, $af^+ + bf^- \in S_+$, thus, by Theorem 1.5, we get $\mathbb{E}\left[e^{-a\boldsymbol{\mu}(f^+)-b\boldsymbol{\mu}(f^-)}\right] = \mathbb{E}\left[e^{-a\boldsymbol{\nu}(f^+)-b\boldsymbol{\nu}(f^-)}\right]$. This means $(\boldsymbol{\mu}(f^+), \boldsymbol{\mu}(f^-)) \overset{d}{=} (\boldsymbol{\nu}(f^+), \boldsymbol{\nu}(f^-))$, from which is evident that $\boldsymbol{\mu}(f) = \boldsymbol{\mu}(f^+) - \boldsymbol{\mu}(f^-) \overset{d}{=} \boldsymbol{\nu}(f^+) - \boldsymbol{\nu}(f^-) = \boldsymbol{\nu}(f)$, and the result follows. □

A.15 Proof of Lemma 1.18

(i): Let $w = (w_1, w_2, \ldots)$, $w^{(n)} = \left(w_1^{(n)}, w_2^{(n)}, \ldots\right)_{n\geq 1}$ be elements of Δ_∞, and $\mu = (\mu_1, \mu_2, \ldots)$, $\mu^{(n)} = \left(\mu_1^{(n)}, \mu_2^{(n)}, \ldots\right)_{n\geq 1}$, be elements of $\mathcal{P}(S)^\infty$, such that $w_j^{(n)} \to w_j$ and $\mu_j^{(n)} \overset{w}{\to} \mu_j$, for every $j \geq 1$. Define $\nu^{(n)} = \sum_{j\geq 1} w_j^{(n)} \mu_j^{(n)}$ and $\nu = \sum_{j\geq 1} w_j \mu_j$. Fix a continuous and bounded function $f : S \to \mathbb{R}$. Then, for $j \geq 1$, $w_j^{(n)} \mu_j^{(n)}(f) \to w_j \mu_j(f)$. Since f is bounded, there exist M such that $|f| \leq M$, hence $|w_j^{(n)} \mu_j^{(n)}(f)| \leq w_j^{(n)} \mu_j^{(n)}(|f|) \leq w_j^{(n)} M$, for every $n \geq 1$, and $j \geq 1$. Evidently, $M w_j^{(n)} \to M w_j$, and $\sum_{j\geq 1} M w_j^{(n)} = M = \sum_{j\geq 1} M w_j$. Hence, by general Lebesgue dominated convergence theorem, we obtain

$$\nu^{(n)}(f) = \sum_{j\geq 1} w_j^{(n)} \mu_j^{(n)}(f) \to \sum_{j\geq 1} w_j \mu_j(f) = \nu(f)$$

That is $\nu^{(n)} \overset{w}{\to} \nu$.

(ii): To prove the second part, using (i) it suffices to see that the mapping $s \to \delta_s$ from S into $\mathcal{P}(S)$ is continuous. So fix $s_n \to s$ in S and let $f : S \to \mathbb{R}$ be a continuous and bounded function. Then $\delta_{s_n}(f) = f(s_n) \to f(s) = \delta_s$, which shows $\delta_{s_n} \overset{w}{\to} \delta_s$ as desired.

(iii): Consider some discrete probability measures $\left(\mu^{(n)} = \sum_{j\geq 1} w_j^{(n)} \delta_{s_j^{(n)}}\right)_{n\geq 1}$ over $(S, \mathscr{B}_S)$, such that $\mu^{(n)} \overset{w}{\to} \sum_{j\geq 1} w_j \delta_{s_j} = \mu$, and set $\Phi^{(n)} = \int \nu_s \, \mu^{(n)}(ds) = \sum_{j\geq 1} w_j^{(n)} \nu_{s_j^{(n)}}$ and $\Phi = \int \nu_s \, \mu(ds) = \sum_{j\geq 1} w_j \nu_{s_j}$. Let $f : T \to \mathbb{R}$ be a continuous and bounded function and define the function $h : S \to \mathbb{R}$, by

$$h(s) = \int f(t) \nu_s(dt).$$

Evidently h is bounded because f is bounded and ν_s is a probability measure. Furthermore, as $\nu_{s_n} \overset{w}{\to} \nu_s$, for every $s_n \to s$ in S, h is also continuous. Thus,

$$\Phi^{(n)}(f) = \sum_{j\geq 1} w_j^{(n)} h\left(s_j^{(n)}\right) = \mu^{(n)}(h) \to \mu(h) = \sum_{j\geq 1} w_j h(s_j) = \Phi(f).$$

That is $\Phi^{(n)} \overset{w}{\to} \Phi$.

(iv): The fourth and last part follows easily by composing the mappings in (ii) and (iii), or alternatively directly from (i). □

A.16 Proof of Proposition 1.19

Since $\mathbb{WS}(\boldsymbol{\mu})$ coincides with the intersection of all closed sets $C \in \mathscr{B}_{\mathcal{P}(S)}$ such that $\mathsf{Q}(C) = \mathbb{P}[\boldsymbol{\mu} \in C] = 1$, to prove this proposition it suffices to show that $\mathcal{C} = \{\varphi \in \mathcal{P}(S) : \mathbb{S}(\varphi) \subseteq \mathbb{S}(\mu_0)\}$ is closed and that $\boldsymbol{\mu}$ belongs to $\mathcal{C}$ almost surely. To prove closeness let $\left(\varphi^{(n)}\right)_{n\geq 1}$ be elements of $\mathcal{C}$ such that $\varphi^{(n)} \overset{w}{\to} \varphi^{(\infty)}$ for some $\varphi^{(\infty)} \in \mathcal{P}(S)$. Note that, as $\mathbb{S}(\mu_0)$ is a closed set, $\mathbb{S}\left(\varphi^{(n)}\right) \subseteq \mathbb{S}(\mu_0)$ implies $\varphi^{(n)}\left(\mathbb{S}(\mu_0)\right) = 1$, hence by the Portmanteau theorem

$$\varphi^{(\infty)}\left(\mathbb{S}(\mu_0)\right) \geq \limsup_n \varphi^{(n)}\left(\mathbb{S}(\mu_0)\right) = 1.$$

This means $\mathbb{S}\left(\varphi^{(\infty)}\right) \subseteq \mathbb{S}(\mu_0)$, and we get $\varphi^{(\infty)} \in \mathcal{C}$, that is $\mathcal{C}$ is closed. To prove $\boldsymbol{\mu} \in \mathcal{C}$ almost surely, realize that $\boldsymbol{\mu}\left(\mathbb{S}(\mu_0)^c\right) \geq 0$ almost surely and by definition $0 = \mu_0\left(\mathbb{S}(\mu_0)^c\right) = \mathbb{E}\left[\boldsymbol{\mu}\left(\mathbb{S}(\mu_0)^c\right)\right]$. Thus, we must have $\boldsymbol{\mu}\left(\mathbb{S}(\mu_0)^c\right) = 0$ almost surely, that is $\mathbb{S}(\boldsymbol{\mu}) \subseteq \mathbb{S}(\mu_0)$ almost surely, which shows $\boldsymbol{\mu} \in \mathcal{C}$ almost surely. □

B Proofs of Section 2

B.1 Proof of Theorem 2.1

1. $\Rightarrow$ 2: There exist $\mathcal{G}$ sub-σ-algebra such that $(\mathbf{x}_i)_{i\geq 1}$ is conditionally i.i.d. given $\mathcal{G}$, let $\boldsymbol{\mu}$ be a version of $\mathbb{P}[\mathbf{x}_1 \in \cdot \mid \mathcal{G}]$, so it is, of course, a version of $\mathbb{P}[\mathbf{x}_k \in \cdot \mid \mathcal{G}]$ for every $k \in \mathbb{N}$, and we have

$$\mathbb{P}[\mathbf{X} \in \cdot \mid \mathcal{G}] = \boldsymbol{\mu}^\infty$$

a.s. which implies

$$\mathbb{P}[\mathbf{X} \in \cdot\,] = \mathbb{E}[\boldsymbol{\mu}^\infty] = \int_{\mathcal{P}(S)} \mu^\infty \mathsf{Q}(d\mu).$$

where Q denotes the law of $\boldsymbol{\mu}$.

2. $\Rightarrow$ 3: Let $n \geq 1$, $\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$ and σ permutation of $[n]$, then

$$\mathbb{P}\left[\bigcap_{i=1}^n (\mathbf{x}_i \in B_i)\right] = \int_{\mathcal{P}(S)} \prod_{i=1}^n \mu(B_i)\mathsf{Q}(d\mu) = \mathbb{P}\left[\bigcap_{i=1}^n (\mathbf{x}_{\sigma(i)} \in B_i)\right].$$

3. $\Rightarrow$ 4: Let $n \geq 1$ and $0 < k_1 < k_2 < \cdots < k_n$, construct a permutation $\rho : [k_n] \to [k_n]$ such that for every $i \in [n]$, $\rho(i) = k_i$, and apply the definition of exchangeability.

4. $\Rightarrow$ 1: Let us denote by θ_n to the shift operator, so that $\theta_n(X) = (x_{n+1}, x_{n+2}, \ldots)$, for a sequence $X = (x_1, x_2, \ldots)$. If $\mathbf{X}$ is contractable then for every $k \leq m \leq n$

$$(\mathbf{x}_m, \theta_m(\mathbf{X})) \overset{d}{=} (\mathbf{x}_k, \theta_m(\mathbf{X})) \overset{d}{=} (\mathbf{x}_k, \theta_n(\mathbf{X})). \tag{B.1}$$

Let $\mathcal{G}_n := \sigma(\theta_n(\mathbf{X})) = \sigma(\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots)$ and define $\tau := \bigcap_{n\geq 1} \mathcal{G}_n$ (the tail σ-algebra of the sequence), also fix $B \in \mathscr{B}_s$. By equation (B.1) we have that

$$\mathbb{P}[\mathbf{x}_m \in B \mid \mathcal{G}_m] = \mathbb{P}[\mathbf{x}_k \in B \mid \mathcal{G}_m] = \mathbb{P}[\mathbf{x}_k \in B \mid \mathcal{G}_n],$$

a.s. Moreover, by the tower property of conditional expectation

$$\mathbb{E}\left[\mathbb{P}[\mathbf{x}_k \in B \mid \mathcal{G}_n] \,|\, \mathcal{G}_{n+1}\right] = \mathbb{P}[\mathbf{x}_k \in B \mid \mathcal{G}_{n+1}],$$

a.s for every $n \geq k$, hence $(\mathbb{P}[\mathbf{x}_k \in B \mid \mathcal{G}_n])_{n\geq k}$ is a reverse martingale, and by the reverse martingale convergence theorem, as $n \to \infty$,

$$\mathbb{P}[\mathbf{x}_m \in B \mid \mathcal{G}_m] = \mathbb{P}[\mathbf{x}_k \in B \mid \mathcal{G}_m] = \mathbb{P}[\mathbf{x}_k \in B \mid \mathcal{G}_n] \to \mathbb{P}[\mathbf{x}_k \in B \mid \tau], \tag{B.2}$$

almost surely. Particularly, by choosing $k = m$ and latter $k = 1$ we get

$$\mathbb{P}[\mathbf{x}_m \in B \mid \mathcal{G}_m] = \mathbb{P}[\mathbf{x}_m \in B \mid \tau] = \mathbb{P}[\mathbf{x}_1 \in B \mid \tau]. \tag{B.3}$$

The first equality shows that $\mathbf{x}_m$ is independent of $\theta_m(\mathbf{X})$ given τ, i.e. for every $n \in \mathbb{N}$ and $m < k_1 < k_2 < \cdots < k_n$, $\mathbf{x}_m$ is independent of $(\mathbf{x}_{k_i})_{i=1}^n$ given τ. Thus, for every

$\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$,

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{i=1}^n (\mathbf{x}_i \in B_i) \middle| \tau\right] &= \mathbb{P}[\mathbf{x}_1 \in B_1 \mid \tau]\mathbb{P}\left[\bigcap_{i=2}^n (\mathbf{x}_i \in B_i) \middle| \tau\right] \\
&= \mathbb{P}[\mathbf{x}_1 \in B_1 \mid \tau]\left[\mathbb{P}[\mathbf{x}_2 \in B_2 \mid \tau]\mathbb{P}\left[\bigcap_{i=3}^n (\mathbf{x}_i \in B_i) \middle| \tau\right]\right] \\
&\ \vdots \\
&= \prod_{i=1}^n \mathbb{P}[\mathbf{x}_i \in B_i \mid \tau]
\end{aligned}$$

The second equality of equation (B.3) yields $\mathbb{P}[\mathbf{x}_1 \in B \mid \tau]$ is a version of $\mathbb{P}[\mathbf{x}_m \in B \mid \tau]$, so if we let $\boldsymbol{\mu}$ be any regular version of the conditional distribution of $\mathbf{x}_1$ given τ, then $\mathbb{P}[\mathbf{x}_m \in B \mid \tau] = \boldsymbol{\mu}(B)$ a.s. for every $m \geq 1$, and we can conclude that

$$\mathbb{P}\left[\bigcap_{i=1}^n (\mathbf{x}_i \in B_i) \middle| \tau\right] = \prod_{i=1}^n \mathbb{P}[\mathbf{x}_i \in B_i \mid \tau] = \prod_{i=1}^n \boldsymbol{\mu}(A_i)$$

a.s. for every $n \in \mathbb{N}$ and $\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$. □

B.2 Proof of Theorem 2.2

a) Clearly $\boldsymbol{\mu}$ is a $\mathcal{G}$-measurable random probability measure, so by the tower property of the conditional expectation we have that for every $n \geq 1$ and every $\{B_i\}_{i=1}^n \subseteq \mathscr{B}_S$

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{i=1}^n (\mathbf{x}_i \in B_i) \middle| \boldsymbol{\mu}\right] &= \mathbb{E}\left(\mathbb{P}\left[\bigcap_{i=1}^n (\mathbf{x}_i \in B_i) \middle| \mathcal{G}\right] \middle| \boldsymbol{\mu}\right) \\
&= \mathbb{E}\left[\prod_{i=1}^n \boldsymbol{\mu}(B_i) \middle| \boldsymbol{\mu}\right] \\
&= \prod_{i=1}^n \boldsymbol{\mu}(B_i),
\end{aligned}$$

almost surely. Or equivalently $\mathbb{P}[\mathbf{X} \in \cdot \mid \boldsymbol{\mu}] \overset{a.s.}{=} \boldsymbol{\mu}^\infty$. This implies $\mathbb{P}[\mathbf{X} \in \cdot \mid \mathcal{G}] = \mathbb{P}[\mathbf{X} \in \cdot \mid \boldsymbol{\mu}]$, and as $\boldsymbol{\mu}$ is $\mathcal{G}$-measurable we obtain that $\mathbf{X}$ is conditionally independent of $\mathcal{G}$ given $\boldsymbol{\mu}$.

b) This is obvious by the proof of a).

c) Let $B \in \mathscr{B}_S$ and let $\mathbf{Y} = \boldsymbol{\mu}(B)$. Define $g : S^\infty \times [0,1] \to \mathbb{R}$ by

$$g(x,z) = \mathbf{1}_{\left\{n^{-1}\sum_{i\leq n} \mathbf{1}_B(x_i) \to z\right\}}, \text{ where } x = (x_1, x_2, \ldots)$$

Also define the event

$$A = \left\{n^{-1}\sum_{i\leq n} \mathbf{1}_B(\mathbf{x}_i) \to \boldsymbol{\mu}(B)\right\}$$

and note $\mathbf{1}_A = g(\mathbf{X}, \mathbf{Y})$ is a measurable function and $\mathbf{Y}$ is $\mathcal{G}$-measurable, hence (by the disintegration theorem)

$$\begin{aligned}\mathbb{P}[A] &= \mathbb{E}[\mathbb{E}[g(\mathbf{X}, \mathbf{Y})|\mathcal{G}]] \\ &= \mathbb{E}\left[\int g(x, \mathbf{Y})\boldsymbol{\mu}^\infty(dx)\right] \\ &= \mathbb{E}\left[\boldsymbol{\mu}^\infty\left\{x : n^{-1}\sum_{i\leq n}\mathbf{1}_B(x_i) \to \mathbf{Y}\right\}\right],\end{aligned}$$

by b) and the strong law of large numbers we also have that conditionally given $\boldsymbol{\mu}$

$$\boldsymbol{\mu}^\infty\left\{x : n^{-1}\sum_{i\leq n}\mathbf{1}_B(x_i) \to \mathbf{Y}\right\} = 1$$

and we conclude $\mathbb{P}[A] = 1$. The $\mathbf{X}$-measurability of $\boldsymbol{\mu}$ is obvious.

d) In the proof of Theorem 2.1 we already showed

$$\mathbb{P}[\mathbf{X} \in \cdot\,] = \int_{\mathcal{P}(S)} \mu^\infty \mathsf{Q}(d\mu),$$

where Q is the law of $\boldsymbol{\mu}$. Now let $\tilde{\mathsf{Q}}$ be any probability measure over $(\mathcal{P}(S), \mathscr{B}_{\mathcal{P}(S)})$ such that the above equation holds, then there exist a random measure $\tilde{\boldsymbol{\mu}}$ such that $\tilde{\boldsymbol{\mu}} \sim \tilde{\mathsf{Q}}$ and it is possible to construct a random sequence $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_i)_{i\geq 1}$ such that $\mathbb{P}[\tilde{\mathbf{X}} \in \cdot \mid \tilde{\boldsymbol{\mu}}] = \tilde{\boldsymbol{\mu}}^\infty$. By the proof of Theorem 2.1 we get

$$\mathbb{P}[\mathbf{X} \in \cdot\,] = \int_{\mathcal{P}(S)} \mu^\infty \tilde{\mathsf{Q}}(d\mu) = \mathbb{P}[\tilde{\mathbf{X}} \in \cdot\,].$$

Thus $\mathbf{X} \overset{d}{=} \tilde{\mathbf{X}}$, which together with $\mathbb{P}[\tilde{\mathbf{X}} \in \cdot \mid \tilde{\boldsymbol{\mu}}] = \tilde{\boldsymbol{\mu}}^\infty$ and $\mathbb{P}[\mathbf{X} \in \cdot \mid \boldsymbol{\mu}] = \boldsymbol{\mu}^\infty$ yield $\boldsymbol{\mu} \overset{d}{=} \tilde{\boldsymbol{\mu}}$, this is $\mathsf{Q} = \tilde{\mathsf{Q}}$.

□

B.3 Proof of Proposition 2.6

Let σ_m be a permutation of $[m]$ and define the permutation of $[n]$, σ by $\sigma(j) = \sigma_m(j)$ if $j \leq m$ and $\sigma(j) = j$ for $j > m$. Then for every partition $A = \{A_1, \ldots, A_k\}$ of $[m]$,

$$\mathbb{P}[\mathbf{\Pi}_m = A] = \sum_{B\in\mathcal{P}_{[n]}(A)} \mathbb{P}[\mathbf{\Pi}_n = B] = \sum_{B\in\mathcal{P}_{[n]}(A)} \mathbb{P}[\mathbf{\Pi}_n = \sigma(B)] = \mathbb{P}[\mathbf{\Pi}_m = \sigma_m(A)],$$

which shows $\mathbf{\Pi}_m$ is exchangeable. □

B.4 Proof of Proposition 2.7

Each unordered composition of n into k parts, $\{n_j\}_{j=1}^k$ is uniquely identified with an element in $\{(n_1,\dots,n_k)\in\mathcal{C}_n^k : n_1\ge n_2\ge\cdots\ge n_k\}$ and it is also uniquely identified with an element $(m_1,\dots,m_n)\in\mathcal{M}_n^k$, where $m_i=|\{j\in[k]: n_j=i\}|$. The number of partitions of $[n]$ having exactly m_i blocks containing i elements, for every $i\in[n]$, is $n!/(\prod_{i=1}^n (i!)^{m_i}(m_j!))$, which shows

$$\mathbb{P}[\mathbf{M}_n=(m_1,\dots,m_n)]=\mathbb{P}[\mathbf{N}_n=\{n_1,\dots,n_k\}]=\frac{n!}{\prod_{i=1}^n (i!)^{m_i}(m_i!)}\pi_n(n_1,\dots,n_k),$$

and

$$\mathbb{P}[\mathbf{N}_n^{\downarrow}=(n_1,\dots,n_k)]=\frac{n!}{\prod_{i=1}^n (i!)^{m_i}(m_i!)}\pi_n(n_1,\dots,n_k)\mathbf{1}_{\{n_1\ge n_2\ge\cdots\ge n_k\}},$$

Now, the number of ordered partitions of $[n]$ having k blocks and such that the jth block contains n_j elements for every $j\in[k]$ is $n!/\prod_{j=1}^k n_j!$ and for any such partition $A=(A_1,\dots,A_k)$,

$$\mathbb{P}\left[\mathbf{\Pi}_n^{\mathrm{ex}}=A\right]=\frac{1}{k!}\pi_n(n_1,\dots,n_k),$$

where $\mathbf{\Pi}_n^{\mathrm{ex}}$ is as in Definition 2.14. Hence

$$\mathbb{P}[\mathbf{N}_n^{\mathrm{ex}}=(n_1,\dots,n_k)]=\frac{n!}{k!\prod_{j=1}^k n_j!}\pi_n(n_1,\dots,n_k).$$

Finally, note that the number of partitions of $[n]$ such that the block containing the element 1 has n_1 elements is $N_1=(n-1)!/((n_1-1)!(n-n_1)!)$, among those, the ones whose second block (according to the order of the least element) contains n_2 elements, are

$$N_2=N_1\frac{((n-n_1)-1)!}{(n_2-1)!((n-n_1)-n_2)!}=\frac{(n)!}{(n_1-1)!(n_2-1)!n(n-n_1)(n-n_1-n_2)!}.$$

partitions. From those N_2, the numbers of partitions whose third block, contains n_3 elements is

$$\begin{aligned}N_3&=N_2\frac{((n-n_1-n_2)-1)!}{(n_3-1)!((n-n_1-n_2)-n_3)!}\\&=\frac{n!}{\left\{\prod_{j=1}^3(n_j-1)!\left(n-\sum_{i=1}^{j-1}n_i\right)\right\}((n-n_1-n_2)-n_3)!}.\end{aligned}$$

with the convention that the empty sum equals 0. Inductively, we find that the number of partitions of $[n]$ whose jth block in the least element order contains n_j elements, is

$$\frac{n!}{\prod_{j=1}^k\left(\sum_{i\ge j}n_i\right)(n_j-1)!},$$

from which is easy to see that

$$\mathbb{P}\left[\tilde{\mathbf{N}}_n=(n_1,\dots,n_k)\right]=\frac{n!}{\prod_{j=1}^k\left(\sum_{i\ge j}n_i\right)(n_j-1)!}\pi_n(n_1,\dots,n_k),$$

□

B.5 Proof of Proposition 2.8

Let $\mathbf{N}^{\text{ex}} = (\mathbf{n}_1^{\text{ex}}, \dots, \mathbf{n}_{\mathbf{K}_n}^{\text{ex}})$ be as in Proposition 2.7. Then, for every permutation σ of $[k]$

$$\mathbb{P}[\mathbf{N}^{\text{ex}} = (n_1, \dots, n_k), \mathbf{N} = (n_{\sigma(1)}, \dots, n_{\sigma(k)})] = \frac{1}{k!}\pi_n^*(n_{\sigma(1)}, \dots, n_{\sigma(k)})$$

Summing over all $k!$ permutations of $[k]$, we obtain

$$\mathbb{P}[\mathbf{N}^{\text{ex}} = (n_1, \dots, n_k)] = \frac{1}{k!}\sum_{\sigma}\pi_n^*(n_{\sigma(1)}, \dots, n_{\sigma(k)})$$

By Proposition 2.7 we also know that

$$\mathbb{P}[\mathbf{N}^{\text{ex}} = (n_1, \dots, n_k)] = \frac{n!}{k!\prod_{j=1}^k n_j!}\pi_n(n_1, \dots, n_k).$$

Putting together the last couple of equations we conclude

$$\pi_n(n_1, \dots, n_k) = \frac{\prod_{j=1}^k n_j!}{n!}\sum_{\sigma}\pi_n^*(n_{\sigma(1)}, \dots, n_{\sigma(k)}).$$

□

B.6 Proof of Corollary 2.9

Let $n_1 \geq n_2 \geq \dots n_k$ be a ranked composition of n and consider the collection $X = \{x_1, \dots, x_n\}$ exhibiting n_j values equal j. Fix $m_i = \sum_{j=1}^k \mathbf{1}_{\{n_j=i\}}$, and for every $i \in [n]$ let $C_{i,1} \dots, C_{i,m_i}$ be the m_i distinct sub-collections of X having i identical elements. Let $\mathbf{z}_1, \dots, \mathbf{z}_n$ be sampled without replacement from X and consider, the random partition of $[n]$, $\mathbf{\Pi}(\mathbf{z}_{1:n})$, generated by the equivalence relation $i \sim j$ if and only if $\mathbf{z}_i = \mathbf{z}_j$. Then, for a fix partition A of $[n]$ having exactly m_i blocks, $A_{i,1}, \dots, A_{i,m_i}$, containing exactly i elements we get

$$\begin{aligned}\mathbb{P}[\mathbf{\Pi}(\mathbf{z}_{1:n}) = A] &= \sum_{(\sigma_i)_i}\mathbb{P}\left[\bigcap_{\{i:m_i>0\}}\bigcap_{j=1}^{m_i}\bigcap_{l\in A_{i,j}}(\mathbf{z}_l \in C_{i,\sigma_i(j)})\right] \\ &= \sum_{(\sigma_i)_i}\frac{\prod_{i=1}^n (i!)^{m_i}}{n!} = \frac{\prod_{i=1}^n (i!)^{m_i}(m_i!)}{n!}\end{aligned}$$

where the sum ranges over all collections $(\sigma_i)_{\{i:m_i>0\}}$, such that σ_i is a permutation of $[m_i]$. This shows that for any partition of $[n]$,

$$\mathbb{P}[\mathbf{\Pi}(\mathbf{z}_{1:n}) = A \mid \mathbf{N}^{\downarrow} = (n_1, \dots, n_k)] = \frac{\prod_{i=1}^n (i!)^{m_i}(m_i!)}{n!}$$

if A has exactly m_i blocks containing i elements, and the above probability equals 0 otherwise. Now, note that since $\mathbf{N}^{\downarrow}$ is completely determined by $\mathbf{\Pi}_n$,

$$\mathbb{P}[\mathbf{\Pi}_n = A, \mathbf{N}^{\downarrow} = (n_1, \dots, n_k)] = \mathbb{P}[\mathbf{\Pi}_n = A] = \pi(n_1, \dots, n_k),$$

if A has exactly m_i blocks containing i elements, and $\mathbb{P}[\mathbf{\Pi}_n = A, \mathbf{N}^{\downarrow} = (n_1, \dots, n_k)] = 0$, otherwise. Hence, by Proposition 2.7 we obtain

$$\mathbb{P}[\mathbf{\Pi}_n = A \mid \mathbf{N}^{\downarrow} = (n_1, \dots, n_k)] = \frac{\pi(n_1, \dots, n_k)}{\mathbb{P}[\mathbf{N}^{\downarrow} = (n_1, \dots, n_k)]} = \frac{\prod_{i=1}^{n} (i!)^{m_i} (m_i!)}{n!}$$

if A has exactly m_i blocks containing i elements, and the above probability equals 0 otherwise. Insomuch as $\mathbf{N}^{\downarrow}$ is a discrete random element, this shows that for every partition A of $[n]$,

$$\mathbb{P}[\mathbf{\Pi}(\mathbf{z}_{1:n}) = A \mid \mathbf{N}^{\downarrow}] = \mathbb{P}[\mathbf{\Pi}_n = A \mid \mathbf{N}^{\downarrow}]$$

□

B.7 Proof of Theorem 2.10

Let $\tilde{\mathbf{\Pi}}_n = \left(\tilde{\mathbf{\Pi}}_{n,1}, \dots, \tilde{\mathbf{\Pi}}_{n,\mathbf{K}_n}\right)$ be the ordering of the blocks of $\mathbf{\Pi}_n$ according to the least element, so that $\tilde{\mathbf{\Pi}}_{n,1}$ contains the element 1, $\tilde{\mathbf{\Pi}}_{n,2}$ contains the smallest element that is not in $\tilde{\mathbf{\Pi}}_{n,1}$, and so on. Suppose without loss of generality that on the same probability space where $\mathbf{\Pi}$ is defined, there exist a sequence $(\mathbf{u}_i)_{i \geq 1}$ of i.i.d. $\mathsf{Unif}(0,1)$ random variables and define

$$\mathbf{x}_m = \sum_{i \geq 1} \mathbf{u}_i \mathbf{1}_{\left\{m \in \tilde{\mathbf{\Pi}}_{n,i}\right\}}, \quad m \in \mathbb{N},$$

where n is any natural number bigger that m, in other words $\mathbf{x}_m = \mathbf{u}_i$ if and only if $m \in \tilde{\mathbf{\Pi}}_{n,i}$. Note that if $m \in \tilde{\mathbf{\Pi}}_{n,i}$, then $m \in \tilde{\mathbf{\Pi}}_{n+k,i}$ for every $k \geq 1$, so the above is well defined. Since $\mathbf{\Pi}$ is exchangeable, so is $\mathbf{X} = (\mathbf{x}_m)_{m \in \mathbb{N}}$ (this can be easily corroborated by fixing a permutation of $\mathbb{N}$ and applying the corresponding definitions). Now, let us rename the uniform random variables in the following way: If $(\mathbf{\Pi}^{\downarrow}_{n,1}, \dots, \mathbf{\Pi}^{\downarrow}_{n,\mathbf{K}_n})$ is a rearrangement of the blocks in such way that $|\mathbf{\Pi}^{\downarrow}_{n,j}| = \mathbf{n}^{\downarrow}_{n,j}$, then $\hat{\mathbf{u}}_{n,j} = \mathbf{u}_i$ if and only if there exist $k \in \mathbf{\Pi}^{\downarrow}_{n,j}$ such that $\mathbf{x}_k = \mathbf{u}_i$, formally

$$\hat{\mathbf{u}}_{n,j} = \sum_{i \geq 1} \mathbf{u}_i \left(\sum_{k \in \mathbf{\Pi}^{\downarrow}_{n,j}} \mathbf{1}_{\{\mathbf{x}_k = \mathbf{u}_i\}} \right).$$

Note that for every $B \in \mathscr{B}_{[0,1]}$, $\hat{\mathbf{u}}_{n,j} \in B$ if and only if $\mathbf{x}_k \in B$ for every $k \in \mathbf{\Pi}^{\downarrow}_{n,j}$, thus

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{x}_i}(B) = \sum_{j \geq 1} \frac{\mathbf{n}^{\downarrow}_{n,j}}{n} \delta_{\hat{\mathbf{u}}_{n,j}}(B), \quad n \in \mathbb{N},$$

the exchangeability of $\mathbf{X}$ implies that, as $n \to \infty$, the above converges a.s. to the random variable $\boldsymbol{\mu}(B)$, where $\boldsymbol{\mu}$ is the directing random measure of $\mathbf{X}$. Particularly, we must have that $\mathbf{n}^{\downarrow}_{n,j}/n$ converges almost surely to a random variable $\mathbf{w}^{\downarrow}_j$ which is the jth largest size of the atoms of $\boldsymbol{\mu}$ (with the convention that if $\boldsymbol{\mu}$ has fewer than j atoms, the size of jth largest atoms is 0). By construction it is also obvious that $\mathbf{\Pi}(\mathbf{x}_{1:\infty}) = \mathbf{\Pi}$, that is, we could consider $\mathbf{\Pi}$ as if it was generated by sequentially sampling from $\boldsymbol{\mu}$, conditionally given $\boldsymbol{\mu}$.

□

B.8 Proof of Proposition 2.11

By Kingman's representation theorem, without loss of generality we may consider $\mathbf{\Pi} = \mathbf{\Pi}(\mathbf{x}_{1:\infty})$, for some exchangeable sequence $\mathbf{X} = (\mathbf{x}_i)_{i\geq 1}$, taking values in a Borel space $(S, \mathscr{B}_S)$, and such that $(\mathbf{w}_1^{\downarrow}, \mathbf{w}_2^{\downarrow}, \ldots)$ are the ranked sizes of the atoms of the directing random measure, $\boldsymbol{\mu}$, of $\mathbf{X}$ (with the convention if $\boldsymbol{\mu}$ has fewer than j atoms, the size of jth largest atoms is 0). Let $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j\geq 1}$ be the (almost surely distinct) atoms of $\boldsymbol{\mu}$, so that

$$\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j^{\downarrow} \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j^{\downarrow}\right) \mu_0$$

for some diffuse probability measure over, μ_0, over $(S, \mathscr{B}_S)$. Define $\mathbf{k}_1 = 1$ and for $j \geq 1$, $\mathbf{k}_{j+1} = \min\{i \geq 1 : \mathbf{x}_i \notin \{\mathbf{x}_{\mathbf{k}_1}, \ldots, \mathbf{x}_{\mathbf{k}_j}\}\}$, with the convention $\min \emptyset = \infty$, and where $\mathbf{k}_{i+1} = \infty$ if $\mathbf{k}_i = \infty$. Set $\tilde{\boldsymbol{\xi}}_j = \mathbf{x}_{\mathbf{k}_j}$ if $\mathbf{k}_j < \infty$ or sample $\tilde{\boldsymbol{\xi}}_j$ from μ_0 independently if $\mathbf{k}_j = \infty$, for every $j \geq 1$. In other words, if $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$ exhibits at least j distinct values, $\tilde{\boldsymbol{\xi}}_j$ is the jth value to appear in the sample $(\mathbf{x}_1, \mathbf{x}_2, \ldots)$. By construction, for every $i \leq n$, $\mathbf{x}_i = \tilde{\boldsymbol{\xi}}_j$ if and only if $i \in \tilde{\mathbf{\Pi}}_{n,j}$, from which is clear that

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}(B) = \sum_{j\geq 1} \frac{\tilde{\mathbf{n}}_{n,j}}{n} \delta_{\tilde{\boldsymbol{\xi}}_j}(B) = \sum_{j=1}^{\mathbf{K}_n} \frac{\tilde{\mathbf{n}}_{n,j}}{n} \delta_{\tilde{\boldsymbol{\xi}}_j}(B),$$

for every $B \in \mathscr{B}_S$. By taking limits as $n \to \infty$, and using the fact that the directing random measure of $\mathbf{X}$ is unique almost surely, we get that $\tilde{\mathbf{n}}_{n,j}/n \to \tilde{\mathbf{w}}_j$ almost surely, for every $j \geq 1$, where $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots)$ is some permutation of $\mathbf{W}^{\downarrow}$. Note for any atom of $\boldsymbol{\mu}$, $\boldsymbol{\xi}_j$, and every $i \geq 1$, we have that $\tilde{\boldsymbol{\xi}}_i = \boldsymbol{\xi}_j$ if and only if $\tilde{\mathbf{w}}_i = \mathbf{w}_j^{\downarrow}$, moreover $\tilde{\mathbf{w}}_i = 0$ if and only if $\tilde{\boldsymbol{\xi}}_i$ is not an atom of $\boldsymbol{\mu}$. Hence,

$$\mathbb{P}\left[\tilde{\boldsymbol{\xi}}_1 \in \cdot \,\middle|\, \mathbf{W}^{\downarrow}, \boldsymbol{\Xi}\right] = \mathbb{P}\left[\mathbf{x}_1 \in \cdot \,\middle|\, \mathbf{W}^{\downarrow}, \boldsymbol{\Xi}\right] = \mathbb{P}\left[\mathbf{x}_1 \in \cdot \,|\, \boldsymbol{\mu}\right] = \sum_{j\geq 1} \mathbf{w}_j^{\downarrow} \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j^{\downarrow}\right) \mu_0$$

implies

$$\mathbb{P}\left[\tilde{\mathbf{w}}_1 \in \cdot \,\middle|\, \mathbf{W}^{\downarrow}\right] = \sum_{j\geq 1} \mathbf{w}_j^{\downarrow} \delta_{\mathbf{w}_j^{\downarrow}} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j^{\downarrow}\right) \delta_0.$$

For $i \geq 1$, let $\boldsymbol{\Xi}_{(i)}$ be the set of atoms of $\boldsymbol{\mu}$ that have not appeared in $\{\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_i\}$ and $\tilde{\boldsymbol{\Xi}}_{(i)}$ be the set of atoms that already appeared in $\{\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_i\}$ also set $\boldsymbol{\Phi}_{(i)} = \{j : \boldsymbol{\xi}_j \in \boldsymbol{\Xi}_{(i)}\}$ and $\tilde{\boldsymbol{\Phi}}_{(i)} = \{j : \boldsymbol{\xi}_j \in \tilde{\boldsymbol{\Xi}}_{(i)}\}$. Conditionally given $\mathbf{W}, \boldsymbol{\Xi}$ and $\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_i$, we know that $\tilde{\boldsymbol{\xi}}_{i+1}$ can not be equal to any element of $\tilde{\boldsymbol{\Xi}}_{(i)}$, hence it is either sampled from μ_0 with probability proportional to $1 - \sum_{j\geq 1} \mathbf{w}_j$ or equals $\boldsymbol{\xi}_j \in \boldsymbol{\Xi}_{(i)}$ with probability proportional to $\mathbf{w}_j^{\downarrow}$. This is

$$\mathbb{P}\left[\tilde{\boldsymbol{\xi}}_{i+1} \in \cdot \,\middle|\, \mathbf{W}, \boldsymbol{\Xi}\right] = \frac{\sum_{j\in \boldsymbol{\Phi}_{(i)}} \mathbf{w}_j^{\downarrow} \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j^{\downarrow}\right) \mu_0}{1 - \sum_{j\in \tilde{\boldsymbol{\Phi}}_{(i)}} \mathbf{w}_j^{\downarrow}}$$

if $\left(1-\sum_{j\in\tilde{\mathbf{\Phi}}_{(i)}}\mathbf{w}_j^{\downarrow}\right) > 0$ and $\mathbb{P}\left[\tilde{\boldsymbol{\xi}}_{i+1}\in\cdot\,\middle|\,\mathbf{W},\boldsymbol{\Xi},\tilde{\boldsymbol{\xi}}_1,\ldots,\tilde{\boldsymbol{\xi}}_i\right] = \mu_0$ otherwise, with the convention that the empty sum equals 0. As to $\tilde{\mathbf{w}}_i$ we have that

$$\begin{aligned}\mathbb{P}\left[\tilde{\mathbf{w}}_{i+1}\in\cdot\,|\,\mathbf{W},\tilde{\mathbf{w}}_1,\ldots,\tilde{\mathbf{w}}_i\right] &= \frac{\sum_{j\in\mathbf{\Phi}_{(i)}}\mathbf{w}_j^{\downarrow}\delta_{\mathbf{w}_j^{\downarrow}}+\left(1-\sum_{j\geq 1}\mathbf{w}_j^{\downarrow}\right)\delta_0}{1-\sum_{j\in\tilde{\mathbf{\Phi}}_{(i)}}\mathbf{w}_j^{\downarrow}}\\ &= \frac{\sum_{j\geq 1}\mathbf{w}_j^{\downarrow}\delta_{\mathbf{w}_j^{\downarrow}}-\sum_{j=1}^{i}\tilde{\mathbf{w}}_j\delta_{\tilde{\mathbf{w}}_j}+\left(1-\sum_{j\geq 1}\mathbf{w}_j\right)\delta_0}{1-\sum_{j=1}^{i}\tilde{\mathbf{w}}_j},\end{aligned}$$

if $\sum_{j=1}^{i}\tilde{\mathbf{w}}_i=\sum_{j\in\tilde{\mathbf{\Phi}}_{(i)}}\mathbf{w}_j^{\downarrow}<1$, and $\mathbb{P}\left[\tilde{\mathbf{w}}_{i+1}\in\cdot\,\middle|\,\mathbf{W},\boldsymbol{\Xi},\tilde{\boldsymbol{\xi}}_1,\ldots,\tilde{\boldsymbol{\xi}}_i\right]=\delta_0$, otherwise. This shows that $(\tilde{\mathbf{w}}_1,\tilde{\mathbf{w}}_2,\ldots)$ is a size-biased pseudo-permutation of $\mathbf{W}^{\downarrow}$. In particular, if $\sum_{j\geq 1}\mathbf{w}_j=1$ almost surely, $(\tilde{\mathbf{w}}_1,\tilde{\mathbf{w}}_2,\ldots)$ is also size-biased permutation of $\mathbf{W}^{\downarrow}$. □

B.9 Proof of Proposition 2.12

Define the event A_n = the $(n+1)$th customers sits at table 1 and set $\mathbf{y}_n=\mathbf{1}_{A_n}$ then $\mathbb{P}[\mathbf{y}_1=1]=(1-\sigma)/(\theta+1)$ and for $n\geq 1$

$$\begin{aligned}\mathbb{P}[\mathbf{y}_{n+1}=1\mid\mathbf{y}_1\ldots,\mathbf{y}_n] &= \frac{\sum_{i=1}^{n}\mathbf{y}_i+1-\sigma}{n+1+\theta}\\ &= \frac{(1-\sigma)+\sum_{i=1}^{n}\mathbf{z}_i}{(1-\sigma+\sum_{i=1}^{n}\mathbf{z}_i)+(\theta+\sigma+n-\sum_{i=1}^{n}\mathbf{z}_i)}.\end{aligned}$$

Thus for $(y_1,\ldots,y_n)\in\{0,1\}^n$

$$\mathbb{P}\left(\bigcap_{i=1}^{n}(\mathbf{y}_i=y_i)\right)=\frac{(1-\sigma)_y(\theta+\sigma)_{n-y}}{(\theta+1)_n}$$

where $y=\sum_{i=1}^{n}y_i$, comparing the above equation with equation (2.3), and using the representation theorem for exchangeable sequences we obtain that the long-run proportion of customers sitting at table 1 is $\mathbf{w}_1=\mathbf{v}_1\sim\mathsf{Be}(1-\sigma,\theta+\sigma)$.

Now fix $j\geq 1$ and assume that for $m\leq j$ we know that $\mathbf{w}_m=\mathbf{v}_m\prod_{i=1}^{m-1}(1-\mathbf{v}_i)$ for some independent random variables $\mathbf{v}_i\sim\mathsf{Be}(1-\sigma,\theta+i\sigma)$. Note that in the Chinese restaurant model, after n customers have arrived and are currently occupying $k>j$ tables with corresponding frequencies $n_1,\ldots,n_k$, when a new customer arrives he/she sits at one the first j tables with probability $(n'-j\sigma)/(n+\theta)$, or at table with a number equal or greater than $j+1$, with probability

$$1-\frac{n'-\sigma}{n+\theta}=\frac{n''+\theta+j\sigma}{n+\theta}$$

where $n'=\sum_{i=1}^{j}n_i$ and $n''=n-n'$. Since the probability that the new customer sits at table $j+1$ is $(n_{j+1}-\sigma)/(n+\theta)$, we easily compute that $(n_{j+1}-\sigma)/(n''+\theta+j\sigma)$ is the conditional probability that he/she at table $j+1$ given that he/she does not sits at one of the first j tables. This said, conditioning on $\mathbf{w}_1,\ldots,\mathbf{w}_j$, imagine that all the $\mathbf{W}_{(j)}=\sum_{m=1}^{j}\mathbf{w}_m$ customers that will end up sitting at one of the first j tables have made a reservation, so when they arrive, they will just pass through. Among the

$(1-\mathbf{W}_{(j)})$ remaining customers, the first one to arrive will sit at table $j+1$, and after n customers without reservation have arrived and there are n_{j+1} customers sitting at table $j+1$, the $(n+1)$th customer will sit at table $j+1$ with probability $(n_{j+1}-\sigma)/(n+\theta+j\sigma)$ or at a table with number bigger than $j+1$ with probability $1-(n_{j+1}-\sigma)/(n+\theta+j\sigma)$. Note that by construction, the long-run proportion of customers that will end up sitting at table $j+1$, in the original chinese restaurant and in the modified chinese restaurant with reservation, coincide (in distribution). In the modified version, let us focus only in the customers without reservation, define the event $B_n =$ the $(n+1)$th customer sits at table $j+1$ and set $\mathbf{z}_n = \mathbf{1}_{B_n}$. Then similarly to table 1 (without reservation), for $(z_1, \ldots, z_n) \in \{0,1\}^n$

$$\mathbb{P}\left(\bigcap_{i=1}^{n}(\mathbf{z}_i = z_i)\right) = \frac{(1-\sigma)_z(\theta+(j+1)\sigma)_{n-z}}{(1+\theta+j\sigma)_z}$$

comparing the above equation with equation (2.3), and using the representation theorem for exchangeable sequences we get that among the customers that do not sit at the first j tables, the long-run proportion of them that will sit a table $j+1$ is $\mathbf{v}_{j+1} \sim \mathsf{Be}(1-\sigma, \theta+(j+1)\sigma)$, independently of $\mathbf{w}_1, \ldots, \mathbf{w}_j$. Hence the overall proportion of customers that will sit at table $j+1$ must be $\mathbf{v}_i(1-\mathbf{W}_{(j)})$. Using the induction hypothesis that $\mathbf{w}_m = \mathbf{v}_m \prod_{i=1}^{m-1}(1-\mathbf{v}_i)$ for every $m \leq j$. It is easy to see that $1-\mathbf{W}_{(j)} = \prod_{i=1}^{j}(1-\mathbf{v}_i)$ which shows that $\mathbf{w}_{j+1} = \mathbf{v}_{j+1}\prod_{i=1}^{j}(1-\mathbf{v}_i)$ for some independent $\mathbf{v}_{j+1} \sim \mathsf{Be}(1-\sigma, \theta+(j+1)\sigma)$, and by induction we have prove (i).

To prove (ii) let $\mathbf{w}_1, \mathbf{w}_2, \ldots$ be as in (i). It suffices to show that $\sum_{j\geq 1}\mathbf{w}_j = 1$ almost surely. This is equivalent to prove that $1-\sum_{i=1}^{j}\mathbf{w}_i = \prod_{i=1}^{j}(1-\mathbf{v}_i)$ goes to 0 almost surely as $j \to \infty$. Since $(\prod_{i=1}^{j}(1-\mathbf{v}_i))_{j\geq 1}$ are almost surely decreasing positive random variables and bounded by 1, we even get that it suffices to show

$$\mathbb{E}\left[\prod_{i=1}^{j}(1-\mathbf{v}_i)\right] = \prod_{i=1}^{j}\mathbb{E}\left[(1-\mathbf{v}_i)\right] = \prod_{i=1}^{j}(1-\mathbb{E}[\mathbf{v}_i]) \to 0$$

as $j \to \infty$. A famous calculus result for divergent series states that for a sequence of numbers $0 < a_i < 1$. $\prod_{i\geq 1}(1-a_i)$ diverges to 0 if and only if $\sum_{i\geq 1} a_i$ diverges to ∞. In our case we know that $\mathbf{v}_i \sim \mathsf{Be}(1-\sigma, \theta+i\sigma)$, so

$$\sum_{i\geq 1}\mathbb{E}[\mathbf{v}_i] = \sum_{i\geq 1}\frac{1-\sigma}{1+\theta+(i-1)\sigma}$$

which clearly diverges as the Harmonic series does. and we have proven (ii).

To prove (iii) let $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n\geq 1}$ be the exchangeable partition of $\mathbb{N}$ generated by the two-parameter chinese restaurant process. For each $n \in \mathbb{N}$, let $\left(\tilde{\mathbf{\Pi}}_{n,1}, \ldots, \tilde{\mathbf{\Pi}}_{n,\mathbf{K}_n}\right)$ be the ordering of the blocks of $\mathbf{\Pi}_n$ according to the least element, with corresponding block sizes $(\tilde{\mathbf{n}}_{n,1}, \ldots, \tilde{\mathbf{n}}_{n,\mathbf{K}_n})$. By (ii) we know that $\mathbf{\Pi}$ is proper and by Proposition 2.11 we now that $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots)$ are in size-biased random order, where $\tilde{\mathbf{w}}_j = \lim_{n\to\infty}\tilde{\mathbf{n}}_{n,j}/n$, almost surely. This said, simply note that the long-run proportion of customers that will end up sitting at table j, $\mathbf{w}_j$, is precisely $\lim_{n\to\infty}\tilde{\mathbf{n}}_{n,j}/n$, and (iii) follows. □

B.10 Proof of Proposition 2.13

Let $A = \{A_1, \dots, A_k\}$ be a partition of $[n]$, whose blocks are ordered according to the least element. For $B \in \mathcal{P}_{[n+1]}(A)$ with $|B|$ blocks in order of appearance, $\{B_1, \dots, B_{|B|}\}$, we necessarily have that either $|B| = k$ and ther exist $j \in [k]$ such that $B_j = A_j \cup \{n+1\}$ and for $i \neq j$, $B_i = A_i$, or $|B| = k+1$ and for every $i \in [k]$, $B_i = A_i$ and $B_{|B|} = \{n+1\}$. This said we easily compute

$$\begin{aligned}\mathbb{P}[\mathbf{\Pi}_n = A] = \sum_{B \in \mathcal{P}_{[n]}(A)} \mathbb{P}[\mathbf{\Pi}_{n+1} = B] &= \sum_{B \in \mathcal{P}_{[n]}(A)} \pi'_{n+1}(|B_1|, \dots, |B_{|B|}|) \\ &= \pi'_{n+1}(|A_1|, \dots, |A_k|, 1) + \sum_{j=1}^{k} \pi'_{n+1}(|A_1|, \dots, |A_{j-1}|, |A_j| + 1, |A_{j+1}|, \dots, |A_k|).\end{aligned}$$

□

B.11 Proof of Theorem 2.15

Assume (i) holds. Let $\mathbf{\Pi}' = (\mathbf{\Pi}'_n)_{n \geq 1}$ be the partition of $\mathbb{N}$ generated by the chinese restaurant process with random seating plan determined by $(\mathbf{w}_1, \mathbf{w}_2, \dots)$. By Proposition 2.14 we know $\mathbf{\Pi}'$ is partially exchangeable and its pEPPF is given by

$$\pi'(n_1, \dots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{w}_j^{\mathbf{n}_j - 1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^{j} \mathbf{w}_j\right)\right].$$

Moreover by construction the long-run proportion of elements that belong to the jth block of $\mathbf{\Pi}'$ is $\mathbf{w}_j$, where the blocks are ordered according to the least element. By Kingman's representation theorem together with Proposition 2.11, we know that there is a one to one correspondence between EPPF's and distributions of size-biased pseudo-permutations sequences of weights, so let π be the EPPF corresponding to $(\mathbf{w}_1, \mathbf{w}_2, \dots)$. Consider an exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi} = (\mathbf{\Pi}_n)_{n \geq 1}$, with EPPF π, and say that $\tilde{\mathbf{w}}_j$ is the long-run proportion of elements that belong to the jth block of $\mathbf{\Pi}$, where the blocks are ordered according to the least element. By construction and Proposition 2.11 it is clear that $(\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \dots)$ is equal in distribution to a size-biased pseudo-permutation of $(\mathbf{w}_1, \mathbf{w}_2, \dots)$. Since both sequences are size-biased pseudo-permutations we must have $(\mathbf{w}_1, \mathbf{w}_2, \dots) \stackrel{d}{=} (\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \dots)$. Since the distribution of the limiting frequencies of the blocks in order of appearance characterize completely the law of corresponding partially exchangeable partition this implies $\mathbf{\Pi} \stackrel{d}{=} \mathbf{\Pi}'$. Hence, $\mathbf{\Pi}'$ is exchangeable and its EPPF must be given by $\pi' = \pi$. That is π' is a symmetric function of its arguments.

Say (ii) holds. By Proposition 2.14 we know

$$\pi'(n_1, \dots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \mathbf{w}_j^{\mathbf{n}_j - 1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^{j} \mathbf{w}_j\right)\right].$$

describes the pEPPF of the partially exchangeable partition of $\mathbb{N}$, $\mathbf{\Pi}'$, generated by the chinese restaurant process with random seating plan determined by $(\mathbf{w}_1, \mathbf{w}_2, \dots)$. Since π' is symmetric we further have that is must be an EPPF and that $\mathbf{\Pi}'$ is exchangeable. By construction the long-run proportion of elements that belong to the jth block of $\mathbf{\Pi}'$, in the least element order, is $\mathbf{w}_j$. Since $\mathbf{\Pi}'$ is exchangeable, by Proposition 2.11 we obtain that $(\mathbf{w}_1, \mathbf{w}_2, \dots)$ is a size-biased pseudo-permutation of some weights sequence. □

B.12 Proof of Corollary 2.16

By Theorem 2.15 we know π' is a symmetric function of its argument if and only if $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ is a size-biased pseudo-permutation. Let $(\mathbf{w}_1^{\downarrow}, \mathbf{w}_2^{\downarrow}, \ldots)$ be the decreasing rearrangement of $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$. If $\sum_{j\geq 1} \mathbf{w}_j = 1$, must have $\mathbf{w}_1^{\downarrow} > 0$ almost surely and that $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ is invariant under size-biased permutations. Hence $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ is a size-biased permutation of $(\mathbf{w}_1^{\downarrow}, \mathbf{w}_2^{\downarrow}, \ldots)$, and by the definition of size-biased permutation we get $\mathbf{w}_1^{\downarrow} > 0$ almost surely implies $\mathbf{w}_1 > 0$ almost surely. Now if $\sum_{j\geq 1} \mathbf{w}_j < 1$ occurs with positive probability, since $(\mathbf{w}_1, \mathbf{w}_2, \ldots)$ is a size-biased pseudo-permutation of $(\mathbf{w}_1^{\downarrow}, \mathbf{w}_2^{\downarrow}, \ldots)$ we obtain that

$$\mathbb{P}[\mathbf{w}_1 = 0] = \mathbb{E}\left[1 - \sum_{j\geq 1} \mathbf{w}_j^{\downarrow}\right] = 1 - \mathbb{E}\left[\sum_{j\geq 1} \mathbf{w}_j\right] > 0$$

hence $\mathbf{w}_1 > 0$ almost surely can not hold. □

B.13 Proof of Corollary 2.17

Let π' be the pEPPF of $\mathbf{\Pi}'$ and π be the EPPF of $\mathbf{\Pi}$. By Theorem 2.15 we know that π' is symmetric, hence $\mathbf{\Pi}'$ is exchangeable. Kingman's correspondence together with Proposition 2.11 set up a one to one correspondence between the distribution of size-biased pseudo-permutation and EPPF's, furthermore this correspondence is given by considering the long-run proportion of elements in the blocks in order of appearance. By construction the long-run proportion in the jth block of $\mathbf{\Pi}'$, in order of appearance is $\tilde{\mathbf{w}}_j$ almost surely. And by Proposition 2.11 the long-run proportion of elements in the jth block of $\mathbf{\Pi}$, in order of appearance is $\mathbf{w}_j^*$, almost surely, where $(\mathbf{w}_j^*)_{j\geq 1}$ is some size-biasde pseudo-permutation of $(\mathbf{w}_j)_{j\geq 1}$. This means that $(\mathbf{w}_j^*)_{j\geq 1}$ is equal in distribution to $(\tilde{\mathbf{w}}_j)_{j\geq 1}$, hence both partitions that must have the same EPPF, that is (by equation (2.17))

$$\pi(n_1, \ldots, n_k) = \pi'(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} (\tilde{\mathbf{w}}_j)^{\mathbf{n}_j - 1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^{j} \tilde{\mathbf{w}}_j\right)\right],$$

or in other words $\mathbf{\Pi}' \overset{d}{=} \mathbf{\Pi}$. □

B.14 Proof of Proposition 2.18

Since all the contraction maps preserve λ under inverse images we trivially have (i) implies (iii). Now we see that (iii) implies (ii), so let $B_1, \ldots, B_n \in \mathscr{B}_S$ be disjoint measurable sets with $\lambda(B_i) = \lambda(B_j)$ for all $i \neq j$ and let σ be a permutation of $[n]$. Define $g, h : \bigcup_{i=1}^n B_i \to S$ by

$$g(s) = \sum_{i=1}^{n} [(i-1)b + f_{B_i}(s)] \mathbf{1}_{B_i}(s),$$

and

$$h(s) = \sum_{i=1}^{n} [(i-1)b + f_{B_{\sigma(i)}}(s)] \mathbf{1}_{B_{\sigma(i)}}(s),$$

where $b = \lambda(B_1) = \lambda(B_i)$, as illustrated in Figure 45.

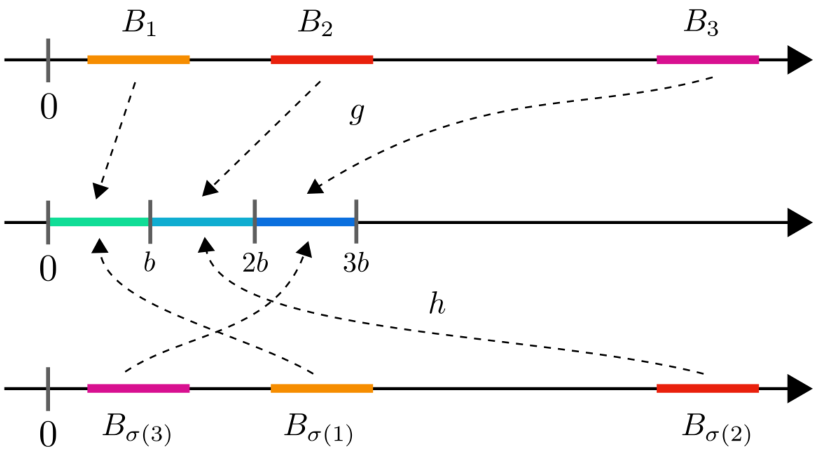


Figure 45: Illustration of g and h.

The contractability of $\boldsymbol{\mu}$ yields

$$\begin{aligned}(\boldsymbol{\mu}(B_1),\dots,\boldsymbol{\mu}(B_n)) &= \left(\boldsymbol{\mu}(g^{-1}[0,b]),\dots,\boldsymbol{\mu}(g^{-1}[(n-1)b,nb])\right)\\ &\overset{d}{=} (\boldsymbol{\mu}([0,b]),\dots,\boldsymbol{\mu}([(n-1)b,nb]))\,,\end{aligned}$$

and analogously for $\left(\boldsymbol{\mu}\left(B_{\sigma(1)}\right),\dots,\boldsymbol{\mu}\left(B_{\sigma(n)}\right)\right)$ using h. This shows $\boldsymbol{\mu}$ is λ-exchangeable.

Finally we prove (ii) implies (i). Fix $f : S \to S$ with $\lambda = \lambda(f^{-1}[\cdot])$ and define $\boldsymbol{\nu} = \boldsymbol{\mu}(f^{-1}[\cdot])$. Let $g = \sum_{i=1}^n b_i \mathbf{1}_{B_i}$ be a simple function where $B_1,\dots,B_n \in \hat{\mathcal{S}}$ are disjoint and $\lambda(B_i) = \lambda(B_j)$. Let us denote $B = \bigcup_{i=1}^n B_i$, and $A_i = f^{-1}[B_i]$.

a) If $f^{-1}[B] \cap B = \emptyset$, then $B_1,\dots,B_n,A_1,\dots,A_n$ are disjoint sets with the same λ-measure. Thus, the λ-exchangeability of $\boldsymbol{\mu}$ implies

$$(\boldsymbol{\mu}(B_1),\dots,\boldsymbol{\mu}(B_n),\boldsymbol{\mu}(A_1),\dots,\boldsymbol{\mu}(A_n)) \overset{d}{=} (\boldsymbol{\mu}(A_1),\dots,\boldsymbol{\mu}(A_n),\boldsymbol{\mu}(B_1),\dots,\boldsymbol{\mu}(B_n))$$

which in turn gives $(\boldsymbol{\mu}(B_1),\dots,\boldsymbol{\mu}(B_n)) \overset{d}{=} (\boldsymbol{\mu}(f^{-1}[B_1]),\dots,\boldsymbol{\mu}(f^{-1}[B_n]))$, and we get $\boldsymbol{\mu}(g) = \sum_{i=1}^n b_i\boldsymbol{\mu}(B_i) \overset{d}{=} \sum_{i=1}^n b_i\boldsymbol{\mu}(A_i) = \boldsymbol{\nu}(g)$.

b) If $f^{-1}[B]\cap B \neq \emptyset$ and $S = \mathbb{R}_+$, as B and $f^{-1}[B]$ are also bounded sets, we may take $C_1,\dots,C_n$ such that $C = \bigcup_{i=1}^n C_i \cap B = \emptyset$ and $C \cap f^{-1}[B] = \emptyset$. Define $\hat{f} : S \to S$, by

$$\hat{f} = \sum_{i=1}^n f_{C_i}^{-1} \circ f_{B_i}(s)\mathbf{1}_{B_i}(s) + \sum_{i=1}^n f_{B_i}^{-1} \circ f_{C_i}(s)\mathbf{1}_{C_i}(s) + s\mathbf{1}_{(B\cup C)^c}(s),$$

where f_D denotes the contraction map on D. Then $\hat{f}$ maps B_i into C_i, C_i into B_i, and preserves λ under inverse images. Also define $\tilde{f} : S \to S$, by

$$\tilde{f} = \sum_{i=1}^n f_{C_i}^{-1} \circ f_{A_i}(s)\mathbf{1}_{A_i}(s) + \sum_{i=1}^n f_{A_i}^{-1} \circ f_{C_i}(s)\mathbf{1}_{C_i}(s) + s\mathbf{1}_{(f^{-1}[B]\cup C)^c}(s),$$

so that $\tilde{f}$ maps A_i into C_i, C_i into A_i, and preserves λ under inverse images. As shown in (a) we get

$$(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n)) \overset{d}{=} \left(\boldsymbol{\mu}\left(\hat{f}^{-1}[B_1]\right), \ldots, \boldsymbol{\mu}\left(\hat{f}^{-1}[B_n]\right)\right)$$

and

$$(\boldsymbol{\mu}(A_1), \ldots, \boldsymbol{\mu}(A_n)) \overset{d}{=} \left(\boldsymbol{\mu}\left(\tilde{f}^{-1}[A_1]\right), \ldots, \boldsymbol{\mu}\left(\tilde{f}^{-1}[A_n]\right)\right).$$

And by construction $\hat{f}^{-1}[B_i] = C_i = \tilde{f}^{-1}[A_i]$, which yields $(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n)) \overset{d}{=} (\boldsymbol{\mu}(A_1), \ldots, \boldsymbol{\mu}(A_n))$, similarly as in (a) we conclude $\boldsymbol{\mu}(g) \overset{d}{=} \boldsymbol{\nu}(g)$.

c) The case where $f^{-1}[B] \cap B \neq \emptyset$ and $S = [0,1]$, follows by (b) by restriction of $\mathbb{R}_+$ into $[0,1]$.

In either case we get $\boldsymbol{\mu}(g) \overset{d}{=} \boldsymbol{\nu}(g)$. Note that if $g : S \to \mathbb{R}_+$ is an arbitrary measurable function, we can construct a sequence of simple functions $g_n \nearrow g$ where $g_n = \sum_{i=1}^{m_n} b_{n,i} \mathbf{1}_{B_{n,i}}$ for some disjoint sets $B_{n,1}, \ldots, B_{n,m_n} \in \hat{\mathcal{S}}$, with $\lambda(B_{n,i}) = \lambda(B_{n,j})$. As shown in the proof of Theorem 1.5, this means that $\boldsymbol{\mu}(g) \overset{d}{=} \boldsymbol{\nu}(g)$ also holds for arbitrary positive measurable functions, and by the same theorem we obtain $\boldsymbol{\mu} \overset{d}{=} \boldsymbol{\nu} = \boldsymbol{\mu}(f^{-1}[\cdot])$. □

B.15 Proof of Proposition 2.20

Since S is Borel we may assume without loss of generality that $S = \mathbb{R}_+$ or $S = [0,1]$ and λ stands for the Lebesgue measure. Note that if $\boldsymbol{\mu} = 0$ almost surely, we get $\mu = 0$ and this uninteresting case follows trivially, so assume this is not the case. To prove the statement it suffices to show $\lambda(B) = \lambda(A)$ if and only if $\mu(B) = \mu(A)$ for every $A, B \in \hat{\mathcal{S}}$. For $A, B \in \hat{\mathcal{S}}$ with $\lambda(A) = \lambda(B)$ we get by the symmetry of $\boldsymbol{\mu}$ that $\boldsymbol{\mu}(A) \overset{d}{=} \boldsymbol{\mu}(B)$, hence $\mu(A) = \mu(B)$. Now if $\lambda(A) < \lambda(B)$ we might take $C \subsetneq B$ with $\lambda(C) = \lambda(A)$, so that $\mu(C) = \mu(A)$. If $\mu(A) = \mu(B)$ this means that $\mu(D) = 0$, with $D = B \setminus C$, thus $\boldsymbol{\mu}(D) = 0$ almost surely. As $\lambda(D) = \lambda(B) - \lambda(A) > 0$ and $\boldsymbol{\mu}$ is λ-symmetric we get $\boldsymbol{\mu}(E) = 0$ almost surely for all $E \in \hat{\mathcal{S}}$, so $\boldsymbol{\mu} = 0$ almost surely, which contradicts the assumption that $\boldsymbol{\mu}$ does not lay in the uninteresting case. Alternatively, we must have $\mu(A) = \mu(C) < \mu(B)$ and the proof is complete.

□

B.16 Proof of Lemma 2.21

We have already shown the necessity of the statement so we turn to prove the sufficiency. First note that if $\lambda(S) < \infty$ and $\boldsymbol{\mu}$ is λ-symmetric, then $\boldsymbol{\mu}$ is also symmetric with respect to $\lambda/\lambda(S)$, so we may assume without loss of generality that $\lambda(S) = 1$ and through a suitable Borel bijection further reduce to the case where $S = [0,1]$ and λ the Lebesgue measure. If otherwise $\lambda(S) = \infty$, under similar arguments we might reduce to the case where $S = \mathbb{R}_+$ and λ stands for the Lebesgue measure.

i.a) Say that $\boldsymbol{\mu}$ is a λ-symmetric simple point process, $S = [0,1]$ and λ is the Lebesgue measure over $([0,1], \mathscr{B}_{[0,1]})$. Note that $\boldsymbol{\mu}(S) = \boldsymbol{\mu}(f^{-1}[S])$ for every $f : S \to S$

such that $\lambda = \lambda(f^{-1}[S])$, so conditionally given $\boldsymbol{\mu}(S)$, $\boldsymbol{\mu}$ remains exchangeable. This way, by conditioning on $\boldsymbol{\mu}(S)$ we may further reduce to the case $\boldsymbol{\mu}(S) = n$, for some $n \in \mathbb{N}$, so that $\boldsymbol{\mu} = \sum_{i=1}^{n} \delta_{\boldsymbol{\xi}_j}$ for some random elements $(\boldsymbol{\xi}_j)_{j=1}^{n}$ taking values in S. Let $\boldsymbol{\xi}_1^{\downarrow}, < \cdots < \boldsymbol{\xi}_n^{\downarrow}$ be the increasing rearrangement of the atoms of $\boldsymbol{\mu}$. Consider some intervals $I_1 < \cdots < I_n$ of $[0,1]$ (where $I < J$ means that $x < y$, for every $x \in I$ and $y \in J$), and a shift $(J_i)_{i=1}^{n}$ of $(I_i)_{i=1}^{n}$. That is $(J_i)_{i=1}^{n}$ is another collection of intervals of $[0,1]$ with $J_1 < \cdots < J_n$ and $\lambda(J_k) = \lambda(I_k)$ for every $k \in [n]$. The symmetry of $\boldsymbol{\mu}$ yields

$$\mathbb{P}\left[\bigcap_{k=1}^{n}\{\boldsymbol{\xi}_k^{\downarrow} \in I_k\}\right] = \mathbb{P}\left[\bigcap_{k=1}^{n}\{\boldsymbol{\mu}(I_k) > 0\}\right] = \mathbb{P}\left[\bigcap_{k=1}^{n}\{\boldsymbol{\mu}(J_k) > 0\}\right] = \mathbb{P}\left[\bigcap_{k=1}^{n}\{\boldsymbol{\xi}_k^{\downarrow} \in J_k\}\right],$$

and by making $\lambda(I_k) \to 0$ for all k, we obtain

$$\mathbb{P}\left[\bigcap_{k=1}^{n}\{\boldsymbol{\xi}_k^{\downarrow} \in dx_k\}\right] = \mathbb{P}\left[\bigcap_{k=1}^{n}\{\boldsymbol{\xi}_k^{\downarrow} \in dy_k\}\right],$$

for every elements of S, $x_1 < \cdots < x_n$ and $y_1 < \cdots < y_n$. That is $\boldsymbol{\xi}_1^{\downarrow}, \ldots, \boldsymbol{\xi}_n^{\downarrow}$ distribute as the ordered statistics of a collection of independent $\mathsf{Unif}(0,1)$ random variables, in other words $\mathbb{P}\left[(\boldsymbol{\xi}_1^{\downarrow}, \ldots, \boldsymbol{\xi}_n^{\downarrow}) \in \cdot\right] = n!\lambda^n$, so we can conclude that given $\boldsymbol{\mu}(S) = n$, $(\boldsymbol{\xi}_j)_{j=1}^{n} \overset{\text{iid}}{\sim} \lambda$, that is $\boldsymbol{\mu}$ is a mixed binomial process based on $\boldsymbol{\kappa} = \boldsymbol{\mu}(S)$ and λ.

i.b) Say that $\boldsymbol{\mu}$ is a λ-symmetric simple point process, $S = \mathbb{R}_+$ and λ is the Lebesgue measure over $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$. Fix $\epsilon > 0$ and $A_\epsilon = [0, \epsilon]$, by a suitable transformation and (i.a) we know that $\mathbf{1}_{A_\epsilon}\boldsymbol{\mu}$ is a mixed binomial based on $(\mathbf{1}_{A_\epsilon}\lambda)/\lambda(A_\epsilon)$. Further, since $\boldsymbol{\mu}$ is contractable, for every bounded set $B \in \hat{\mathcal{S}}$ we know that $\boldsymbol{\mu}(f_B^{-1}[\cdot]) \overset{d}{=} \mathbf{1}_{[0,\lambda(B)]}\boldsymbol{\mu}$, hence $\mathbf{1}_B\boldsymbol{\mu}$ is also a mixed binomial process based on $(\mathbf{1}_B\lambda)/\lambda(B)$. Now let $B_1, B_2, \ldots \in \hat{\mathcal{S}}$ such that $B_n \nearrow S$, set $\lambda_n = \mathbf{1}_{B_n}\lambda/\lambda(B_n)$ and $\boldsymbol{\kappa_n} = \boldsymbol{\mu}(B_n)$ so that $\mathbf{1}_{B_n}\boldsymbol{\mu}$ is a mixed Binomial process based on $\boldsymbol{\kappa}_n$ and λ_n. For $f : S \to \mathbb{R}_+$ supported on B_m we get for every $n > m$,

$$\lambda_n(e^{-f}) = 1 - \frac{\lambda(1 - e^{-f})}{\lambda(B_n)}.$$

and $\mathbf{1}_{B_n}\boldsymbol{\mu}(f) = \boldsymbol{\mu}(f)$, so by Proposition 1.9

$$\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right] = \mathbb{E}\left[\lambda_n(e^{-f})^{\boldsymbol{\kappa}_n}\right] = \mathbb{E}\left[\left\{1 - \frac{\lambda(1 - e^{-f})}{\lambda(B_n)}\right\}^{\lambda(B_n)\boldsymbol{\gamma}_n}\right] \tag{B.4}$$

where $\boldsymbol{\gamma}_n = \boldsymbol{\kappa}_n/\lambda(B_n)$. By Helly's selection theorem, we have that $\boldsymbol{\gamma}_n \xrightarrow{d} \boldsymbol{\gamma}$ in $[0,\infty]$, along a sub-sequence. Moreover, as $x_n \to x$ in $[0,\infty]$ and $m_n \to \infty$, we get $(1 - \{a/m_n\}^{m_n x_n}) \to e^{-ax}$, thus by taking limits as $n \to \infty$ in (B.4), along the acquired subsequence we get

$$\mathbb{E}\left[e^{-\boldsymbol{\mu}(f)}\right] = \mathbb{E}\left[\exp\left\{-\boldsymbol{\gamma}\lambda(1 - e^{-f})\right\}\right]. \tag{B.5}$$

The choice $f = \varepsilon \mathbf{1}_{B_m}$, yields $\mathbb{E}\left[e^{-\varepsilon \boldsymbol{\mu}(B_m)}\right] = \mathbb{E}\left[\exp\left\{-\boldsymbol{\gamma}\lambda(B_m)(1-e^{-\varepsilon})\right\}\right]$, and by letting $\varepsilon \to 0$ we find $\boldsymbol{\gamma} < \infty$, almost surely. Hence by a monotone convergence argument we can extend (B.5) to arbitrary measurable function $f : S \to \mathbb{R}_+$, therefore, by Proposition 1.9 and Theorem 1.5, we obtain $\boldsymbol{\mu}$ is a mixed Poisson process based on λ and $\boldsymbol{\gamma}$.

ii) Since $\boldsymbol{\mu} = \boldsymbol{\beta}\lambda$ if and only if $\boldsymbol{\mu}(B) = \boldsymbol{\beta}\lambda(B)$ for all $B \in \hat{\mathcal{S}}$, it suffices to consider the scenario where $S = [0,1]$ and λ is the Lebesgue measure. If $\boldsymbol{\beta} = \boldsymbol{\mu}(S) > 0$, $\boldsymbol{\nu} = \boldsymbol{\mu}/\boldsymbol{\beta}$ is a well defined random probability measure that remains λ-symmetric. By Proposition 2.20 we know that $\boldsymbol{\nu}$ is $\mathbb{E}[\boldsymbol{\nu}]$-symmetric, so we may assume without loss of generality that $\mathbb{E}[\boldsymbol{\nu}] = \lambda$. Consider the random measure $\boldsymbol{\nu}^2 = \boldsymbol{\nu} \otimes \boldsymbol{\nu}$ on $S \times S$ and note that the exchangeability of $\boldsymbol{\mu}$ yields

$$\mathbb{E}\left[\boldsymbol{\nu}^2\left(I_1 \times I_2\right)\right] = \mathbb{E}\left[\boldsymbol{\nu}\left(I_1\right)\boldsymbol{\nu}\left(I_2\right)\right] = \mathbb{E}\left[\boldsymbol{\nu}\left(J_1\right)\boldsymbol{\nu}\left(J_2\right)\right] = \mathbb{E}\left[\boldsymbol{\nu}^2\left(J_1 \times J_2\right)\right] \quad \text{(B.6)}$$

for every intervals $I_1 < I_2$, or $I_2 < I_1$, and $J_1 < J_2$, or $J_2 < J_1$, with $\lambda(I_k) = \lambda(J_k)$. Further, since $\boldsymbol{\nu}$ is diffuse we get that $\mathbb{E}\left[\boldsymbol{\nu}^2\left\{(s_1, s_2) : s_1 = s_2\right\}\right] = 0$. Hence by a suitable approximation (see Figure 46 for an illustration) we can extend (B.6) to any measurable rectangles $I_1 \times I_2$ and $J_1 \times J_2$ such that $\lambda(I_k) = \lambda(J_k)$.

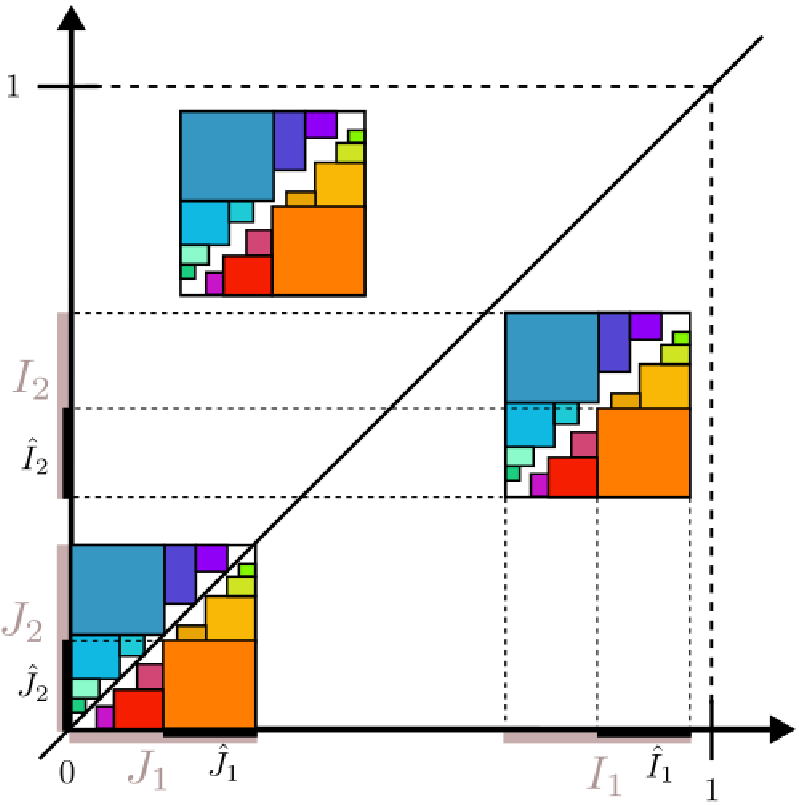


Figure 46: Approximation of $J_1 \times J_2$ by means of measurable rectangles $\tilde{J}_1 \times \tilde{J}_2$ with $\tilde{J}_1 < \tilde{J}_2$. In this Figure, sets with the same color have the same λ^2-measure.

This means that $\mathbb{E}\left[\boldsymbol{\nu}^2\right]$ is invariant under arbitrary shifts of measurable rectangles, which in turn, implies that $\mathbb{E}[\boldsymbol{\nu}^2]$ must be proportional to λ^2. Now, $\boldsymbol{\nu}^2(S \times S) =$

$\boldsymbol{\nu}(S)\boldsymbol{\nu}(S) = 1 = \lambda^2(S \times S)$, thus we even get $\boldsymbol{\nu}^2 = \lambda^2$. With this in mind we can compute

$$\mathsf{Var}(\boldsymbol{\nu}(B)) = \mathbb{E}\left[\boldsymbol{\nu}(B)^2\right] - \mathbb{E}[\boldsymbol{\nu}(B)]^2 = \lambda(B)^2 - \lambda(B)^2 = 0,$$

which proves $\boldsymbol{\nu}(B) = \mathbb{E}[\boldsymbol{\nu}(B)] = \lambda(B)$, almost surely, for all $B \in \mathscr{B}_S$. Therefore, $\boldsymbol{\mu} = \boldsymbol{\beta}\boldsymbol{\nu} = \boldsymbol{\beta}\lambda$, almost surely. Now, if there exist a measurable set, A, on the probability space where $\boldsymbol{\mu}$ is defined, such that $\boldsymbol{\mu} = 0$ over A, we simply fix $\boldsymbol{\beta}(\omega) = 0$ for every $\omega \in A$ and derive the desired representation on A^c, as above.

□

B.17 Proof of Lemma 2.22

Since $(S, \mathscr{B}_S)$ is Borel, by means of a suitable Borel bijection we may reduce to that case where $S = [0,1]$, if $\lambda(S) < \infty$, or $S = \mathbb{R}_+$, for $\lambda(S) = \infty$, and λ stands for the Lebesgue measure.

i) Say that $S = [0,1]$ and λ is the Lebesgue measure. Let $\boldsymbol{\nu} = \sum_{j \geq 1} \delta_{(\boldsymbol{\alpha}_j, \boldsymbol{\xi}_j)}$, and consider the point process of $\mathbb{R}_+$, $\boldsymbol{\nu}' = \boldsymbol{\nu}(\cdot \times S) = \sum_{j \geq 1} \boldsymbol{\kappa}_j \boldsymbol{\alpha}_j$, where clearly $\boldsymbol{\kappa}_j = |\{\boldsymbol{\xi}_j : \boldsymbol{\mu}(\{\boldsymbol{\xi}_j\}) = \boldsymbol{\alpha}_j\}|$. As $\boldsymbol{\nu}'$ remains invariant under λ-preserving transformation of S, we may reduce to the case where this one is non-random. Further, the local finiteness of $\boldsymbol{\mu}$ yields $\boldsymbol{\nu}'((a, \infty)) < \infty$, almost surely, for every $0 < a$ (otherwise we would have infinitely many atoms, $\boldsymbol{\xi}_j$, with corresponding jumps $\boldsymbol{\alpha}_j > a$). Since it suffices to derive the stated representation on sets of the form (a, ∞), we may even assume $\boldsymbol{\nu}'(\mathbb{R}_+) < \infty$, without loss of generality. With this considerations in mind, we now have $\boldsymbol{\nu}' = \sum_{i \leq n} k_i \delta_{\tau_i}$, for some constants $n, k_1, \ldots, k_n \in \mathbb{N}$ and $\tau_1, \ldots, \tau_n \in \mathbb{R}_+$. This means that $\boldsymbol{\nu}$ has got to take the form

$$\boldsymbol{\nu} = \sum_{i \leq n} \sum_{j \leq k_i} \delta_{(\tau_j, \boldsymbol{\sigma}_{i,j})}$$

for some S-valued random variables $(\boldsymbol{\sigma}_{i,j})_{i \leq n, j \leq k_i}$. For $i \leq n$, let $\boldsymbol{\sigma}^{\downarrow}_{i,1} < \cdots \boldsymbol{\sigma}^{\downarrow}_{i,k_i}$ be the increasing rearrangement of $(\boldsymbol{\sigma}_{i,j})_{j \leq k_i}$, consider a collection of disjoint interval $(I_{i,j})_{j \leq k_i}$ and shift $(J_{i,j})_{j \leq k_i}$, so that $I_{i,1} < \cdots < I_{i,k_i}$, $J_{i,1} < \cdots < J_{i,k_i}$ and $\lambda(I_{i,j}) = \lambda(J_{i,j})$. The λ-symmetry of $\boldsymbol{\mu}$ yields

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{i=1}^{n} \bigcap_{j=1}^{k_i} \{\boldsymbol{\sigma}^{\downarrow}_{i,j} \in I_{i,j}\}\right] &= \mathbb{P}\left[\bigcap_{i=1}^{n} \bigcap_{j=1}^{k_i} \{\boldsymbol{\nu}(\{\tau_i\} \times I_{i,j}) > 0\}\right] \\
&= \mathbb{P}\left[\bigcap_{i=1}^{n} \bigcap_{j=1}^{k_i} \{\boldsymbol{\nu}(\{\tau_i\} \times J_{i,j}) > 0\}\right] \\
&= \mathbb{P}\left[\bigcap_{i=1}^{n} \bigcap_{j=1}^{k_i} \{\boldsymbol{\sigma}^{\downarrow}_{i,j} \in J_{i,j}\}\right]
\end{aligned}$$

by making $\lambda(I_{i,j}) \to 0$ we get

$$\mathbb{P}\left[\bigcap_{i=1}^{n} \bigcap_{j=1}^{k_i} \{\boldsymbol{\sigma}^{\downarrow}_{i,j} \in dx_{i,j}\}\right] = \mathbb{P}\left[\bigcap_{i=1}^{n} \bigcap_{j=1}^{k_i} \{\boldsymbol{\sigma}^{\downarrow}_{i,j} \in dy_{i,j}\}\right]$$

for any elements of S, $x_{i,1} < \cdots < x_{i,k_i}$ and $y_{i,1} < \cdots < y_{i,k_i}$. Thus, the law of $\left(\boldsymbol{\sigma}^{\downarrow}_{i,j}\right)_{i\leq n, j\leq k_i}$ is proportional to the Lebesgue measure on the non-diagonal part of $\Delta_{k_1} \times \cdots \times \Delta_{k_n}$, where $\Delta_k = \{(x_1, \ldots, x_n) \in S^k : x_1 < \cdots < x_n\}$. Furthermore, from the proof of Proposition 2.20, we get $\mathbb{E}[\boldsymbol{\mu}(\{s\})] = 0$ for every $s \in S$, so $\boldsymbol{\mu}$ has no fixed atoms, and we obtain that the elements in $(\boldsymbol{\sigma}_{i,j})_{i\leq n, j\leq k_i}$ are distinct almost surely. Hence the law of $\left(\boldsymbol{\sigma}^{\downarrow}_{i,j}\right)_{i\leq n, j\leq k_i}$ vanishes on diagonal spaces. This shows that

$$\mathbb{P}\left[\left(\boldsymbol{\sigma}^{\downarrow}_{i,j}\right)_{i\leq n, j\leq k_i} \in \cdot\right] \propto \lambda^{\sum_{i\leq n} k_i}.$$

Thus, the elements in $\{\boldsymbol{\nu}(\{\tau_i\} \times \cdot)\}_{i\leq n}$ are independent binomial process as illustrated in Figure 47.

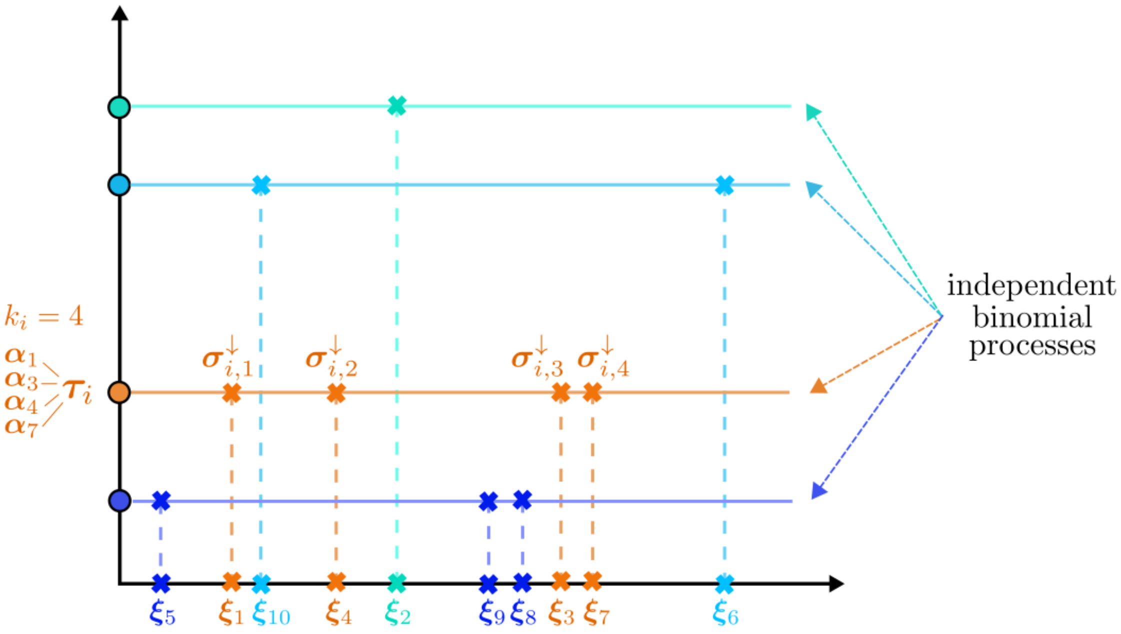


Figure 47: Illustration of the underlying independent binomial processes in $\{(\boldsymbol{\alpha}_j, \boldsymbol{\xi}_j)\}_{j\geq 1}$.

In terms of $\{(\boldsymbol{\alpha}_j, \boldsymbol{\xi}_j)\}_{j\geq 1}$ this shows that $\boldsymbol{\nu}$ is a λ-randomization of the point process $\sum_{i\geq 1} \delta_{\boldsymbol{\alpha}_j}$ (see Definition of randomization in Section 1.3.2). Finally, noting that $\lambda : \mathbb{R}_+ \to S$, regarded as a kernel is constant on $\mathbb{R}_+$, we even get $\boldsymbol{\xi}_j$ is independent of $\boldsymbol{\alpha}_j$, and we can conclude $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \lambda$ is independent of $(\boldsymbol{\alpha}_j)_{j\geq 1}$.

ii) The second part of the statement will not be required for subsequent developments, so we will sketch this proof avoiding technical details. If $S = \mathbb{R}_+$ and λ represents the Lebesgue measure, then for every $\epsilon > 0$ we may construct a measurable partition of S, $B_{\epsilon,1}, B_{\epsilon,2}, \ldots$, where $\lambda(B_{\epsilon,i}) = \epsilon$ for every $i \geq 1$. This means that $\{\boldsymbol{\mu}(B_{\epsilon,i})\}_{i\geq 1}$ is an exchangeable sequence, and by Theorem 2.1 it is conditionally i.i.d. The conditional independence also holds for disjoint bounded sets $B_1, \ldots, B_n$, (despite whether $\lambda(B_i) = \lambda(B_j)$ or not). Indeed, if we denote $\mathbf{x}_t = \boldsymbol{\mu}([0, t])$, $\mathbf{x}^{(t)} = (\mathbf{x}_s)_{s>t}$, and consider the tail σ-algebra $\tau = \sigma(\mathbf{x}^{(t)} - \mathbf{x}_t)$, similarly as in Theorem 2.1, and using the λ-symmetry of $\boldsymbol{\mu}$, it can be shown that

$$\mathbb{P}[\mathbf{x}_t - \mathbf{x}_s \in \cdot \mid \mathbf{x}^{(t)} - \mathbf{x}_t] = \mathbb{P}[\mathbf{x}_t - \mathbf{x}_s \in \cdot \mid \tau] = \mathbb{P}[\mathbf{x}_{t-s} \in \cdot \mid \tau],$$

almost surely. This means that $\mathbf{x}_t - \mathbf{x}_s$ is independent of $\mathbf{x}^{(t)} - \mathbf{x}_t$ given τ and that the distribution of this increment only depends on $t - s$. Translating this into $\boldsymbol{\mu}$ we get that given τ, $\boldsymbol{\mu}$ has conditionally independent increments, that is Theorem 1.12 holds for $\boldsymbol{\mu}$ conditionally given τ. By the proof of Proposition 2.20 we see $\mathbb{E}[\boldsymbol{\mu}(\{s\})] = 0$ for every $s \in S$, so $\boldsymbol{\mu}$ has no fixed atoms. Thus, there must exist τ-measurable locally finite random measures $\boldsymbol{\nu}$ over $S \times \mathbb{R}_+$ satisfying

$$\int_{\mathbb{R}_+} (x \wedge 1) \boldsymbol{\nu}(B, dx) < \infty$$

almost surely, for every $B \in \hat{\mathcal{S}}$, such that conditionally given $\boldsymbol{\nu}$, $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ defines a Poisson process over $S \times \mathbb{R}_+$ directed by $\boldsymbol{\nu}$. This shows that $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j\geq 1}$ forms a Cox. To attain the required decomposition of $\boldsymbol{\nu}$, consider the Cox process $\boldsymbol{\gamma} = \sum_{j\geq 1} \delta_{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)}$, so that $\mathbb{E}[\boldsymbol{\gamma} \mid \tau] = \boldsymbol{\nu}$. Note that $\boldsymbol{\gamma}$ inherits the λ-symmetry of $\boldsymbol{\mu}$, that is, for every $f : S \to S$ with $\lambda = \lambda(f^{-1}[\cdot])$ and bounded sets B and A, $\boldsymbol{\gamma}(B \times A) \overset{d}{=} \boldsymbol{\gamma}(f^{-1}[B] \times A)$. Hence, $\boldsymbol{\nu}$ is invariant under (inverse images of) λ-preserving transformations of S, which in turn yields $\boldsymbol{\nu} = \lambda \otimes \boldsymbol{\varrho}$ for some random measure, $\boldsymbol{\varrho}$, over $(\mathbb{R}_+, \mathscr{B}_{\mathbb{R}_+})$ satisfying $\int_{\mathbb{R}_+} (x \wedge 1) \boldsymbol{\varrho}(dx) < \infty$. The proof of this last assertion appears in Kallenberg (2002), Theorem 2.6. For a thorough proof of the result in question see for instance Kallenberg (2005) or Kallenberg (2017).

□

C Proofs of Section 3

C.1 Proof of Proposition 3.1

Let $\boldsymbol{\sigma}^{-1}$ be the inverse function of $\boldsymbol{\sigma}$, which evidently exists and is independent of the atoms, as $\boldsymbol{\sigma}$ is a permutation that is independent of the atoms. Since the collection of atoms $\Xi = (\boldsymbol{\xi}_j)_{j\geq 1}$ is i.i.d. we clearly have that $\Xi \overset{d}{=} (\boldsymbol{\xi}_{\boldsymbol{\sigma}^{-1}(j)})_{j\geq 1} = \boldsymbol{\sigma}^{-1}(\Xi)$ and by theorem 1.7 a random probability measure is a measurable function of its weights and atoms, this implies

$$\boldsymbol{\mu} \overset{d}{=} \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_{\boldsymbol{\sigma}^{-1}(j)}} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \mu_0.$$

As the sum of the weights is bounded by 1, we have that $\sum_{j\geq 1} \mathbf{w}_j = \sum_{j\geq 1} \mathbf{w}_{\boldsymbol{\sigma}(j)}$ almost surely and that $\sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_{\boldsymbol{\sigma}^{-1}(j)}} = \sum_{j\geq 1} \mathbf{w}_{\boldsymbol{\sigma}(j)} \delta_{\boldsymbol{\xi}_{\boldsymbol{\sigma}(\boldsymbol{\sigma}^{-1}(j))}} = \sum_{j\geq 1} \mathbf{w}_{\boldsymbol{\sigma}(j)} \delta_{\boldsymbol{\xi}_j}$ almost surely. Putting this together with the last equation we obtain

$$\boldsymbol{\mu} \overset{d}{=} \sum_{j\geq 1} \mathbf{w}_{\boldsymbol{\sigma}(j)} \delta_{\boldsymbol{\xi}_j} + \left(1 - \sum_{j\geq 1} \mathbf{w}_{\boldsymbol{\sigma}(j)}\right) \mu_0.$$

□

C.2 Proof of Lemma 3.2

Let us denote $\mathbf{w}_0 = 1 - \sum_{j\geq 1} \mathbf{w}_j$, so that $\sum_{j\geq 0} \mathbf{w}_j = 1$ almost surely, and by a monotone convergence argument we also obtain $\sum_{j\geq 0} \mathbb{E}\left[\mathbf{w}_j\right] = 1$. Note that

$$\boldsymbol{\mu}(f) = \sum_{j\geq 1} \mathbf{w}_j f(\boldsymbol{\xi}_j) + \mathbf{w}_0 \mu_0(f),$$

and that if M is a bound of f, we have that $|\sum_{j=1}^n \mathbf{w}_j f(\boldsymbol{\xi}_j)| < M$ almost surely for every $n \geq 1$. Hence, by linearity of the expectation, Lebesgue dominated convergence theorem, and since $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0$ independently of the weights, we get

$$\mathbb{E}[\boldsymbol{\mu}(f)] = \sum_{j\geq 1} \mathbb{E}[\mathbf{w}_j]\mathbb{E}[f(\boldsymbol{\xi}_j)] + \mathbb{E}\left[\mathbf{w}_0\right] \mu_0(f) = \left(\sum_{j\geq 0} \mathbb{E}\left[\mathbf{w}_j\right]\right) \mu_0(f) = \mu_0(f).$$

This proves the first part. To prove the second and thirds parts, first realize that

$$1 = \mathbb{E}\left[\left(\sum_{j\geq 0} \mathbf{w}_j\right)^2\right] = \mathbb{E}\left[\sum_{j\geq 1} \mathbf{w}_j^2\right] + \mathbb{E}\left[\sum_{i\neq j} \mathbf{w}_i \mathbf{w}_j\right] + \mathbb{E}\left[\mathbf{w}_0^2\right],$$

and by a monotone convergence argument we get, $1 - \rho = \sum_{i\neq j} \mathbb{E}\left[\mathbf{w}_i \mathbf{w}_j\right] + \mathbb{E}\left[\mathbf{w}_0^2\right]$, where $\sum_{i\neq j} a_i a_j$ denotes $\sum_{i\geq 0}\sum_{j\geq 0} a_i a_j \mathbf{1}_{\{i\neq j\}}$. Secondly, since f and g are bounded and $\boldsymbol{\mu}$ is a random probability measure we have that $\boldsymbol{\mu}(|f|), \boldsymbol{\mu}(|g|) < \infty$ almost surely. Then,

$$\begin{aligned}
\boldsymbol{\mu}(f)\boldsymbol{\mu}(g) &= \left(\sum_{j\geq 1} \mathbf{w}_j f(\boldsymbol{\xi}_j) + \mathbf{w}_0\mu_0(f)\right)\left(\sum_{j\geq 1} \mathbf{w}_j g(\boldsymbol{\xi}_j) + \mathbf{w}_0\mu_0(g)\right) \\
&= \sum_{j\geq 1} \mathbf{w}_j^2 f(\boldsymbol{\xi}_j) g(\boldsymbol{\xi}_j) + \sum_{\substack{i,j\geq 1 \\ i\neq j}} \mathbf{w}_i\mathbf{w}_j f(\boldsymbol{\xi}_i) g(\boldsymbol{\xi}_j) + \left(\sum_{j\geq 1} \mathbf{w}_0\mathbf{w}_j g(\boldsymbol{\xi}_j)\right)\mu_0(f) \\
&\qquad + \left(\sum_{j\geq 1} \mathbf{w}_0\mathbf{w}_j f(\boldsymbol{\xi}_j)\right)\mu_0(g) + \mathbf{w}_0^2\mu_0(f)\mu_0(g).
\end{aligned}$$

Now, if M is a bound for f, and N is a bound of g we have that for every $n \geq 1$, $|\sum_{j=1}^n \mathbf{w}_j\mathbf{w}_0 f(\boldsymbol{\xi}_j))| \leq M$, $|\sum_{j=1}^n \mathbf{w}_j\mathbf{w}_0 g(\boldsymbol{\xi}_j))| \leq N$, $|\sum_{j=1}^n \mathbf{w}_j^2 f(\boldsymbol{\xi}_j) g(\boldsymbol{\xi}_j)| \leq MN$, and $|\sum_{i=1}^n \sum_{j=1}^n \mathbf{w}_i\mathbf{w}_j f(\boldsymbol{\xi}_j) g(\boldsymbol{\xi}_i)\mathbf{1}_{\{i\neq j\}}| \leq MN$. Thus, by linearity of the expectation, Lebesgue dominated convergence theorem, and since $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0$ independently of the weights, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{\mu}(f)\boldsymbol{\mu}(g)\right] \\
&= \sum_{j\geq 1} \mathbb{E}\left[\mathbf{w}_j^2\right]\mathbb{E}\left[f(\boldsymbol{\xi}_j)g(\boldsymbol{\xi}_j)\right] + \sum_{\substack{i,j\geq 1 \\ i\neq j}} \mathbb{E}\left[\mathbf{w}_i\mathbf{w}_j\right]\mathbb{E}\left[f(\boldsymbol{\xi}_i)\right]\mathbb{E}\left[g(\boldsymbol{\xi}_j)\right] + \mathbb{E}\left[\mathbf{w}_0^2\right]\mu_0(f)\mu_0(g) \\
&\qquad + \left(\sum_{j\geq 1} \mathbb{E}\left[\mathbf{w}_0\mathbf{w}_j\right]\mathbb{E}\left[g(\boldsymbol{\xi}_j)\right]\right)\mu_0(f) + \left(\sum_{j\geq 1} \mathbb{E}\left[\mathbf{w}_0\mathbf{w}_j\right]\mathbb{E}\left[f(\boldsymbol{\xi}_j)\right]\right)\mu_0(g). \\
&= \sum_{j\geq 1} \mathbb{E}\left[\mathbf{w}_j^2\right]\mu_0(fg) + \sum_{i\neq j} \mathbb{E}\left[\mathbf{w}_i\mathbf{w}_j\right]\mu_0(f)\mu_0(g) + \mathbb{E}\left[\mathbf{w}_0^2\right]\mu_0(f)\mu_0(g) \\
&= \rho\mu_0(fg) + (1-\rho)\mu_0(f)\mu_0(g).
\end{aligned}$$

This proves the third part of the lemma, and the choice $g = f$ gives the second part. □

C.3 Proof of Theorem 3.4

We may assume without loss of generality that all the species sampling processes are defined on the same probability space. First we prove (i), let $f : S \to \mathbb{R}$ be a continuous and bounded function. Since f is continuous it is also measurable, and by Lemma 3.2 we have that

$$\begin{aligned}
&\mathbb{E}\left[\left\{\boldsymbol{\mu}^{(n)}(f) - \mu_0(f)\right\}^2\right] \\
&\quad = \mathbb{E}\left[\left\{\boldsymbol{\mu}^{(n)}(f)\right\}^2\right] - 2\mathbb{E}\left[\boldsymbol{\mu}^{(n)}(f)\right]\mu_0(f) + \{\mu_0(f)\}^2 \\
&\quad = \rho^{(n)}\mu_0^{(n)}\left(f^2\right) + \left(1-\rho^{(n)}\right)\left\{\mu_0^{(n)}(f)\right\}^2 - 2\mu_0^{(n)}(f)\mu_0(f) + \{\mu_0(f)\}^2.
\end{aligned} \tag{C.1}$$

By hypothesis we know that $\mu_0^{(n)} \overset{w}{\to} \mu_0$ and $\rho_n \to 0$, as $n \to \infty$, by taking limits in (C.1), we found that

$$\mathbb{E}\left[\left\{\boldsymbol{\mu}^{(n)}(f) - \mu_0(f)\right\}^2\right] \to 0,$$

as $n \to \infty$. That is, $\boldsymbol{\mu}^{(n)}(f)$ converges to $\mu_0(f)$ in $\mathcal{L}_2$, which implies $\boldsymbol{\mu}^{(n)}(f) \xrightarrow{d} \mu_0(f)$. Since f was an arbitrary continuous and bounded function, this proves (i).

To prove (ii) let $\mathbf{w}_1^{(n)} \geq \mathbf{w}_2^{(n)} \geq \cdots$ be the decreasingly ordered weights of $\boldsymbol{\mu}^{(n)}$ and let $\boldsymbol{\xi}_j^{(n)}$ be the atom corresponding to $\mathbf{w}_j^{(n)}$. Let us denote $\mathbf{w}_0^{(n)} = 1 - \sum_{j\geq 1} \mathbf{w}_j^{(n)}$. Recall that

$$\rho^{(n)} = \sum_{j\geq 1} \mathbb{E}\left[\left(\mathbf{w}_j^{(n)}\right)^2\right] \tag{C.2}$$

and by the proof of Lemma 3.2 we also know

$$1 - \rho^{(n)} = \mathbb{E}\left[\left(\mathbf{w}_0^{(n)}\right)^2\right] + \sum_{i\neq j} \mathbb{E}\left[\mathbf{w}_j^{(n)}\mathbf{w}_i^{(n)}\right] \tag{C.3}$$

for $n \geq 1$. Since the weights are decreasing, we must have that for every $i \geq j \geq 2$, $\mathbb{E}\left[\mathbf{w}_i^{(n)}\mathbf{w}_j^{(n)}\right] \leq \mathbb{E}\left[\mathbf{w}_i^{(n)}\mathbf{w}_{j-1}^{(n)}\right]$, hence

$$\sum_{i\neq j} \mathbb{E}\left[\mathbf{w}_i^{(n)}\mathbf{w}_j^{(n)}\right] \geq \sum_{j\geq 1}\sum_{i\geq j+1} \mathbb{E}\left[\mathbf{w}_i^{(n)}\mathbf{w}_j^{(n)}\right] \geq \sum_{j\geq 2}\sum_{i\geq j} \mathbb{E}\left[\mathbf{w}_i^{(n)}\mathbf{w}_j^{(n)}\right] \geq \sum_{j\geq 2} \mathbb{E}\left[\left(\mathbf{w}_j^{(n)}\right)^2\right] \geq 0.$$

for $n \geq 1$. By taking limits, as $n \to \infty$, in (C.3) we found $\sum_{j\geq 2} \mathbb{E}\left[\left(\mathbf{w}_j^{(n)}\right)^2\right] \to 0$, which together with (C.2) proves that $\mathbb{E}\left[\left(\mathbf{w}_1^{(n)}\right)^2\right] \to 1$. Since $0 \leq \mathbf{w}_1^{(n)} \leq 1$, and $\sum_{j\geq 0} \mathbb{E}\left[\mathbf{w}_j^{(n)}\right] = 1$, we obtain

$$\mathbb{E}\left[\mathbf{w}_1^{(n)}\right] \to 1 \quad \text{and} \quad \sum_{j\neq 1} \mathbb{E}\left[\mathbf{w}_j^{(n)}\right] \to 0, \tag{C.4}$$

as $n \to \infty$. Given that all the corresponding spaces are Polish, and $\mu_0^{(n)} \xrightarrow{w} \mu_0$, we might construct on a probability space $\left(\hat{\Omega}, \hat{\mathbf{F}}, \hat{\mathbb{P}}\right)$, some independent sequences, $\left(\hat{\boldsymbol{\xi}}_j^{(n)}\right)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0^{(n)}$, and $\left(\hat{\mathbf{w}}_j^{(n)}\right)_{j\geq 1} \overset{d}{=} \left(\mathbf{w}_j^{(n)}\right)_{j\geq 1}$, such that $\hat{\boldsymbol{\xi}}_j^{(n)} \to \hat{\boldsymbol{\xi}}_j \sim \mu_0$, almost surely, as $n \to \infty$, independently for $j \geq 1$. Define $\hat{\boldsymbol{\mu}}^{(n)} = \sum_{j\geq 1} \hat{\mathbf{w}}_j^{(n)}\delta_{\hat{\boldsymbol{\xi}}_j^{(n)}} + \hat{\mathbf{w}}_0^{(n)}\mu_0^{(n)}$, where $\hat{\mathbf{w}}_0^{(n)} = 1 - \sum_{j\geq 1} \hat{\mathbf{w}}_j^{(n)}$. Then for any continuous and bounded function, f, by equations Lemma 3.2

$$\begin{aligned}
&\mathbb{E}\left[\left\{\hat{\boldsymbol{\mu}}^{(n)}(f) - \delta_{\hat{\boldsymbol{\xi}}_1}(f)\right\}^2\right] \\
&= \mathbb{E}\left[\left\{\hat{\boldsymbol{\mu}}^{(n)}(f)\right\}^2\right] - 2\mathbb{E}\left[\hat{\boldsymbol{\mu}}^{(n)}(f)\, f\left(\hat{\boldsymbol{\xi}}_1\right)\right] + \mathbb{E}\left[\left\{f\left(\hat{\boldsymbol{\xi}}_1\right)\right\}^2\right] \\
&= \rho^{(n)}\,\mu_0^{(n)}\left(f^2\right) + \left(1-\rho^{(n)}\right)\left\{\mu_0^{(n)}(f)\right\}^2 - 2\mathbb{E}\left[\hat{\boldsymbol{\mu}}^{(n)}(f)\, f\left(\hat{\boldsymbol{\xi}}_1\right)\right] + \mu_0(f^2).
\end{aligned} \tag{C.5}$$

As f is bounded, we can write

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathbf{w}}_1^{(n)}\right]\mathbb{E}\left[f\left(\hat{\boldsymbol{\xi}}_1^{(n)}\right) f\left(\hat{\boldsymbol{\xi}}_1\right)\right] - M^2\sum_{j\neq 1}\mathbb{E}\left[\hat{\mathbf{w}}_j^{(n)}\right] &\leq \mathbb{E}\left[\hat{\boldsymbol{\mu}}^{(n)}(f)\, f(\hat{\boldsymbol{\xi}}_1)\right] \\
&\leq \mathbb{E}\left[\hat{\mathbf{w}}_1^{(n)}\right]\mathbb{E}\left[f\left(\hat{\boldsymbol{\xi}}_1^{(n)}\right) f\left(\hat{\boldsymbol{\xi}}_1\right)\right] + M^2\sum_{j\neq 1}\mathbb{E}\left[\hat{\mathbf{w}}_j^{(n)}\right]
\end{aligned}$$

where M is a bound of f. By taking limits as $n \to \infty$ in the last equation and by (C.4), we get

$$\mathbb{E}\left[\hat{\boldsymbol{\mu}}^{(n)}(f)\, f\left(\hat{\boldsymbol{\xi}}_1\right)\right] \to \mathbb{E}\left[\left\{f\left(\hat{\boldsymbol{\xi}}_1\right)\right\}^2\right] = \mu_0(f^2).$$

From which is evident that

$$\mathbb{E}\left[\left\{\hat{\boldsymbol{\mu}}^{(n)}(f) - \delta_{\hat{\boldsymbol{\xi}}_1}(f)\right\}^2\right] \to 0$$

as $n \to \infty$. That is $\hat{\boldsymbol{\mu}}^{(n)}(f) \to \delta_{\hat{\boldsymbol{\xi}}_1}(f)$ in $\mathcal{L}_2$, which implies $\boldsymbol{\mu}^{(n)}(f) \overset{d}{=} \hat{\boldsymbol{\mu}}^{(n)}(f) \overset{d}{\to} \delta_{\hat{\boldsymbol{\xi}}_1}(f)$. As this holds for every continuous and bounded function, f, we obtain $\boldsymbol{\mu}^{(n)} \overset{dw}{\to} \delta_{\hat{\boldsymbol{\xi}}_1}$, as $n \to \infty$. □

C.4 Proof of Theorem 3.6

In this proof for any random (or deterministic) sequence $\mathbf{z}_1, \mathbf{z}_2, \ldots$ we write $\boldsymbol{\Pi}(\mathbf{z}_{1:n})$ to the partition of $[n]$ generated by the equivalence relation $i \sim j$ if and only if $\mathbf{z}_i = \mathbf{z}_j$, and for any random (or deterministic) vector of positive numbers $\mathbf{n} = (\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}})$ set $\mathbf{n}^{(j)} = (\mathbf{n}_1, \ldots, \mathbf{n}_{j-1}, \mathbf{n}_j + 1, \mathbf{n}_{j+1}, \ldots, \mathbf{n}_{\mathbf{K}})$ for $1 \leq j \leq \mathbf{K}$ and $\mathbf{n}^{(\mathbf{K}+1)} = (\mathbf{n}_1, \ldots, \mathbf{n}_{\mathbf{K}}, 1)$. To prove this Theorem we will show that $(\mathbf{x}_i)_{i\geq 1}$ as in I–V satisfy:

1) The distinct values that $(\mathbf{x}_i)_{i\geq 1}$ exhibits are i.i.d. and have distribution μ_0. This should be intuitive from each of the statements, because in all cases if $\mathbf{x}_n$ contributes with a new value not yet observed in $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}\}$, this new value is chosen independently of the previous values.

2) The EPPF of $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$ is π. For this we will use Kingman representation theorem, the Chinese restaurant construction with a given prediction rule and the Chinese restaurant construction with random seating plan (see Sections 2.2.5–2.2.7).

To provide a formal proof we start with a small Lemma.

Lemma C.1. *Let $\Xi = (\boldsymbol{\xi}_j)_{j\geq 1} \overset{iid}{\sim} \mu_0$, where μ_0 is diffuse and $(\mathbf{d}_i)_{i\geq 1}$ be a sequence such that its elements take values in $\mathbb{N} \cup \{0\}$. Set $\mathbf{z}_i = \boldsymbol{\xi}_{\mathbf{d}_i}$ if $\mathbf{d}_i \in \mathbb{N}$, and sample $\mathbf{z}_i \sim \mu_0$ independently if $\mathbf{d}_i = 0$. Then for every $n \geq 1$*

$$\mathbb{P}[\mathbf{z}_1 \in B_1, \ldots, \mathbf{z}_n \in B_n \mid \boldsymbol{\Pi}(\mathbf{z}_{1:n})] = \prod_{j=1}^{\mathbf{K}_n} \mu_0\left(\bigcap_{i\in\boldsymbol{\Pi}_j} B_i\right)$$

where $\boldsymbol{\Pi}(\mathbf{z}_{1:n}) = \{\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_{\mathbf{K}_n}\}$.

Proof of Lemma C.1: Fix $n \geq 1$ and set $\mathbf{D}_j = \{i \leq n : \mathbf{d}_i = j\}$, for $j \in \mathbb{N}$, and

$\mathbf{D}_0 = \{i \leq n : \mathbf{d}_i = 0\}$, then

$$\begin{aligned}
\mathbb{P}[\mathbf{z}_1 \in B_1, \ldots, \mathbf{z}_n \in B_n \mid \mathbf{d}_1, \ldots, \mathbf{d}_n, \boldsymbol{\Xi}] &= \prod_{i=1}^{n} \mathbb{P}[\mathbf{z}_i \in B_i \mid \mathbf{d}_i, \boldsymbol{\Xi}] \\
&= \prod_{i=1}^{n} \left\{ \sum_{j\geq 1} \delta_{\boldsymbol{\xi}_j}(B_i) \mathbf{1}_{\{\mathbf{d}_i = j\}} + \mu_0(B_i) \mathbf{1}_{\{\mathbf{d}_i = 0\}} \right\} \\
&= \left\{ \prod_{j\geq 1} \prod_{i \in \mathbf{D}_j} \delta_{\boldsymbol{\xi}_j}(B_i) \right\} \left\{ \prod_{i \in \mathbf{D}_0} \mu_0(B_i) \right\} \\
&= \left\{ \prod_{\{j : \mathbf{D}_j \neq \emptyset\}} \delta_{\boldsymbol{\xi}_j} \left(\bigcap_{i \in \mathbf{D}_j} B_i \right) \right\} \left\{ \prod_{i \in \mathbf{D}_0} \mu_0(B_i) \right\},
\end{aligned} \tag{C.6}$$

for measurable sets $B_1, \ldots, B_n$ and using the convention that the empty product equals 1. By the tower property of conditional expectation and the fact that $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0$, we get

$$\begin{aligned}
\mathbb{P}[\mathbf{z}_1 \in B_1, \ldots, \mathbf{z}_n \in B_n \mid \mathbf{d}_{1:n}] &= \mathbb{E}\left[\left\{ \prod_{\{j : \mathbf{D}_j \neq \emptyset\}} \delta_{\boldsymbol{\xi}_j} \left(\bigcap_{i \in \mathbf{D}_j} B_i \right) \right\} \left\{ \prod_{i \in \mathbf{D}_0} \mu_0(B_i) \right\} \middle| \mathbf{d}_{1:n} \right] \\
&= \left\{ \prod_{\{j : \mathbf{D}_j \neq \emptyset\}} \mu_0 \left(\bigcap_{i \in \mathbf{D}_j} B_i \right) \right\} \left\{ \prod_{i \in \mathbf{D}_0} \mu_0(B_i) \right\}
\end{aligned}$$

where $\mathbf{d}_{1:n} = (\mathbf{d}_1, \ldots, \mathbf{d}_n)$. Since μ_0 is diffuse we have that outside a $\mathbb{P}$-null set, $\mathbf{z}_i = \mathbf{z}_k$ if and only if there exist $j \geq 1$ such that $i, k \in \mathbf{D}_j$. Thus if we denote $\boldsymbol{\Pi}(\mathbf{z}_{1:n}) = \{\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_{\mathbf{K}_n}\}$, we may rewrite

$$\mathbb{P}[\mathbf{z}_1 \in B_1, \ldots, \mathbf{z}_n \in B_n \mid \mathbf{d}_{1:n}] = \prod_{i=1}^{\mathbf{K}_n} \mu_0 \left(\bigcap_{i \in \boldsymbol{\Pi}_j} B_i \right)$$

almost surely. As the right side of the above equation is $\boldsymbol{\Pi}(\mathbf{z}_{1:n})$-measurable, this yields

$$\mathbb{P}[\mathbf{z}_1 \in B_1, \ldots, \mathbf{z}_n \in B_n \mid \boldsymbol{\Pi}(\mathbf{z}_{1:n})] = \prod_{i=1}^{\mathbf{K}_n} \mu_0 \left(\bigcap_{i \in \boldsymbol{\Pi}_j} B_i \right).$$

□

Proof of Theorem 3.6: (I. ⇒ IV.): Without loss of generality (after possibly enlarging the original probability space), we may define a sequence $(\mathbf{d}_i)_{i\geq 1}$ such that conditionally given $\mathbf{W}$, $\mathbf{d}_1, \mathbf{d}_2, \ldots$ are independent of the atoms $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j\geq 1}$, and with

$$\{\mathbf{d}_1, \mathbf{d}_2, \ldots \mid \mathbf{W}\} \overset{\text{iid}}{\sim} \sum_{j\geq 1} \mathbf{w}_j \delta_j + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \delta_0$$

That is $\mathbb{P}[\mathbf{d}_i = j \mid \mathbf{W}] = \mathbf{w}_j$ for every $j \in \mathbb{N}$ and with probability $1 - \sum_{j\geq 1} \mathbf{w}_j$, $\mathbf{d}_i = 0$, independently for $i \geq 1$. Conditioning on $\mathbf{d}_i$ and $\boldsymbol{\Xi}$, set $\mathbf{z}_i = \boldsymbol{\xi}_{\mathbf{d}_i}$ if and only if $\mathbf{d}_i \in \mathbb{N}$, or sample $\mathbf{z}_i$ independently from μ_0 if and only if $\mathbf{d}_i = 0$, independently for $i \geq 1$. Again, without loss of generality the sequence $(\mathbf{z}_i)_{i\geq 1}$ may be defined and we may further assume that $(\mathbf{z}_i)_{i\geq 1}$ is conditionally independent of $\mathbf{W} = (\mathbf{w}_j)_{j\geq 1}$ given $(\mathbf{d}_i)_{i\geq 1}$. Formally we have that for every $n \geq 1$ and any measurable sets $B_1, \ldots, B_n$,

$$\begin{aligned}\mathbb{P}[\mathbf{z}_1 \in B_1, \ldots, \mathbf{z}_n \in B_n \mid \mathbf{d}_1, \ldots, \mathbf{d}_n, \boldsymbol{\Xi}] &= \prod_{i=1}^{n} \mathbb{P}[\mathbf{z}_i \in B_i \mid \mathbf{d}_i, \boldsymbol{\Xi}] \\ &= \prod_{i=1}^{n} \left\{ \sum_{j\geq 1} \delta_{\boldsymbol{\xi}_j}(B_i) \mathbf{1}_{\{\mathbf{d}_i = j\}} + \mu_0(B_i) \mathbf{1}_{\{\mathbf{d}_i = 0\}} \right\}. \end{aligned} \tag{C.7}$$

By the tower property of conditional expectation, monotone convergence theorem and by the assumed conditional independences.

$$\begin{aligned}\mathbb{P}[\mathbf{z}_1 \in B_1, \ldots, \mathbf{z}_n \in B_n \mid \mathbf{W}, \boldsymbol{\Xi}] &= \mathbb{E}\left[\prod_{i=1}^{n} \left\{ \sum_{j\geq 1} \delta_{\boldsymbol{\xi}_j}(B_i) \mathbf{1}_{\{\mathbf{d}_i = j\}} + \mu(B_i) \mathbf{1}_{\{\mathbf{d}_i = 0\}} \right\} \middle| \mathbf{W}, \boldsymbol{\Xi} \right] \\ &= \prod_{i=1}^{n} \left\{ \sum_{j\geq 1} \delta_{\boldsymbol{\xi}_j} \mathbb{P}[\mathbf{d}_i = j \mid \mathbf{W}] + \mu(B_i) \mathbb{P}[\mathbf{d}_i = 0 \mid \mathbf{W}] \right\} \\ &= \prod_{i=1}^{n} \left\{ \sum_{j\geq 1} \delta_{\boldsymbol{\xi}_j} \mathbf{w}_j + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \mu_0(B_i) \right\} \\ &= \prod_{i=1}^{n} \boldsymbol{\mu}(B_i) \end{aligned}$$

Since $\boldsymbol{\mu}$ is $(\mathbf{W}, \boldsymbol{\Xi})$-measurable, this clearly shows that $\{\mathbf{z}_1, \mathbf{z}_2, \ldots \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$ so that $(\mathbf{x}_i)_{i\geq 1}$ as in I is identically distributed as $(\mathbf{z}_i)_{i\geq 1}$, this together with Lemma C.1 yields

$$\mathbb{P}[\mathbf{x}_1 \in B_1, \ldots, \mathbf{x}_n \in B_n \mid \boldsymbol{\Pi}(\mathbf{x}_{1:n})] = \prod_{j=1}^{\mathbf{K}_n} \mu_0 \left(\bigcap_{i \in \boldsymbol{\Pi}_j} B_i \right)$$

where $\boldsymbol{\Pi}(\mathbf{x}_{1:n}) = \{\boldsymbol{\Pi}_1, \ldots, \boldsymbol{\Pi}_{\mathbf{K}_n}\}$. Finally from Corollary 2.17 we know that the EPPF of $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$, is given by

$$\pi(n_1, \ldots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \tilde{\mathbf{w}}_j^{n_j - 1} \prod_{j=1}^{k-1} \left(1 - \sum_{i=1}^{j} \tilde{\mathbf{w}}_j\right)\right],$$

where $(\tilde{\mathbf{w}}_j)_{j\geq 1}$ is a size-biased pseudo-permutation of $(\mathbf{w}_j)_{j\geq 1}$. Thus IV holds for $(\mathbf{x}_i)_{i\geq 1}$ as in I.

(II. ⇒ IV.): Without loss of generality, we may define the following random elements. Let $\boldsymbol{\Xi} = (\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0$, and independently, set $\mathbf{d}_1 = 1$, and for $n \geq 1$,

$$\mathbb{P}[\mathbf{d}_{n+1} = i \mid \tilde{\mathbf{W}}, \mathbf{d}_1, \ldots, \mathbf{d}_n] = \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j \delta_j(i) + \left(1 - \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j\right) \delta_{\mathbf{K}_n + 1}(i) \tag{C.8}$$

where $\mathbf{K}_n = \max\{\mathbf{d}_1, \dots, \mathbf{d}_n\}$. Define $(\mathbf{z}_i)_{i\geq 1}$ by $\mathbf{z}_i = \boldsymbol{\xi}_{\mathbf{d}_i}$, so that $\mathbb{P}[\mathbf{z}_1 \in \cdot \mid \Xi, \mathbf{d}_1] = \delta_{\boldsymbol{\xi}_{\mathbf{d}_1}}$, implying that $\mathbf{z}_1 \sim \mu_0$, and for $n \geq 1$

$$\mathbb{P}[\mathbf{z}_{n+1} \in \cdot \mid \Xi, \tilde{\mathbf{W}}, \mathbf{d}_1, \dots, \mathbf{d}_n] = \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j \delta_{\boldsymbol{\xi}_{\mathbf{d}_j}} + \left(1 - \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j\right) \delta_{\boldsymbol{\xi}_{\mathbf{K}_n+1}}$$

Note that $\mathbf{z}_1^* = \boldsymbol{\xi}_{\mathbf{d}_1}, \dots, \mathbf{z}_{\mathbf{K}_n}^* = \boldsymbol{\xi}_{\mathbf{d}_{\mathbf{K}_n}}$ are precisely the distinct values that $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ exhibits in order of appearance, so the last equation implies

$$\mathbb{P}[\mathbf{z}_{n+1} \in \cdot \mid \tilde{\mathbf{W}}, \mathbf{z}_1, \dots, \mathbf{z}_n, \boldsymbol{\xi}_{\mathbf{K}_n+1}] = \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j \delta_{\mathbf{z}_j^*} + \left(1 - \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j\right) \delta_{\boldsymbol{\xi}_{\mathbf{K}_n+1}},$$

and by the tower property of conditional expectation we obtain

$$\mathbb{P}[\mathbf{z}_{n+1} \in \cdot \mid \tilde{\mathbf{W}}, \mathbf{z}_1, \dots, \mathbf{z}_n] = \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j \delta_{\mathbf{z}_j^*} + \left(1 - \sum_{j=1}^{\mathbf{K}_n} \tilde{\mathbf{w}}_j\right) \mu_0.$$

This proves that $(\mathbf{x}_i)_{i\geq 1}$ as in II and $(\mathbf{z}_i)_{i\geq 1}$ as constructed here are identically distributed, which together with Lemma C.1 show that

$$\mathbb{P}[\mathbf{x}_1 \in B_1, \dots, \mathbf{x}_n \in B_n \mid \boldsymbol{\Pi}(\mathbf{x}_{1:n})] = \prod_{j=1}^{\mathbf{K}_n} \mu_0\left(\bigcap_{i \in \boldsymbol{\Pi}_j} B_i\right)$$

where $\boldsymbol{\Pi}(\mathbf{x}_{1:n}) = \{\boldsymbol{\Pi}_1, \dots, \boldsymbol{\Pi}_{\mathbf{K}_n}\}$. The fact that $\boldsymbol{\Pi}(\mathbf{x}_{1:n})$ is exchangeable and its EPPF is π, follows by putting together the hypothesis

$$\pi(n_1, \dots, n_k) = \mathbb{E}\left[\prod_{j=1}^{k} \tilde{\mathbf{w}}_j^{n_j - 1} \prod_{j=1}^{k-1}\left(1 - \sum_{i=1}^{j} \tilde{\mathbf{w}}_j\right)\right],$$

the fact that μ_0 is diffuse, the Chinese restaurant construction with random seating plan, Theorem 2.15 and Corollary 2.17.

(III. ⇒ IV.): Once more, without loss of generality, we may define the following random elements. Let $(\mathbf{z}_j^*)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0$, and independently, set $\mathbf{d}_1 = 1$, and for $n \geq 1$,

$$\mathbb{P}[\mathbf{d}_{n+1} = i \mid \mathbf{d}_1, \dots, \mathbf{d}_n] = \sum_{j=1}^{\mathbf{K}_n+1} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})} \delta_j(i) \tag{C.9}$$

where $\mathbf{K}_n = \max\{\mathbf{d}_1, \dots, \mathbf{d}_n\}$ and $\mathbf{n} = (\mathbf{n}_1, \dots \mathbf{n}_{\mathbf{K}_n})$ is given by $\mathbf{n}_j = |\{i : \mathbf{d}_i = j\}|$. Also define $(\mathbf{z}_i)_{i\geq 1}$ by $\mathbf{z}_i = \mathbf{z}_{\mathbf{d}_i}^*$, and let us denote $\boldsymbol{Z}^* = (\mathbf{z}_j^*)_{j\geq 1}$, $\mathbf{d}_{1:n} = (\mathbf{d}_1, \dots, \mathbf{d}_n)$ and $\mathbf{z}_{1:n} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$. Clearly $\mathbf{z}_1 \sim \mu_0$, and from (C.9), we get that

$$\mathbb{P}[\mathbf{z}_{n+1} \in \cdot \mid \mathbf{z}_{1:n}, \boldsymbol{Z}^*] = \sum_{j=1}^{\mathbf{K}_n+1} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})} \delta_{\mathbf{z}_j^*}(B),$$

where $\mathbf{n} = (\mathbf{n}_1, \dots, \mathbf{n}_{\mathbf{K}_n})$ are the frequencies of the $\mathbf{K}_n$ distinct values, $\mathbf{z}_1^*, \dots, \mathbf{z}_{\mathbf{K}_n}^*$, that $\mathbf{z}_1, \dots, \mathbf{z}_n$ exhibits. By the tower property of conditional expectation, we obtain

$$\begin{aligned}\mathbb{P}[\mathbf{z}_{n+1} \in \cdot \mid \mathbf{z}_{1:n}] &= \mathbb{E}\left[\sum_{j=1}^{\mathbf{K}_n} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})}\delta_{\mathbf{z}_j^*}(B) + \frac{\pi\left(\mathbf{n}^{(\mathbf{K}_n+1)}\right)}{\pi(\mathbf{n})}\delta_{\mathbf{z}^*_{\mathbf{K}_n+1}}(B) \,\middle|\, \mathbf{z}_{1:n}\right]\\ &= \sum_{j=1}^{\mathbf{K}_n} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})}\delta_{\mathbf{z}_j^*}(B) + \frac{\pi\left(\mathbf{n}^{(\mathbf{K}_n+1)}\right)}{\pi(\mathbf{n})}\mathbb{E}[\delta_{\mathbf{z}^*_{\mathbf{K}_n+1}}(B)]\\ &= \sum_{j=1}^{\mathbf{K}_n} \frac{\pi\left(\mathbf{n}^{(j)}\right)}{\pi(\mathbf{n})}\delta_{\mathbf{z}_j^*}(B) + \frac{\pi\left(\mathbf{n}^{(\mathbf{K}_n+1)}\right)}{\pi(\mathbf{n})}\mu_0(B)\end{aligned}$$

This shows that $(\mathbf{z}_i)_{i\geq 1}$ is equal in distribution to $(\mathbf{x}_i)_{i\geq 1}$ as in III. Thus Lemma C.1 proves

$$\mathbb{P}[\mathbf{x}_1 \in B_1, \dots, \mathbf{x}_n \in B_n \mid \mathbf{\Pi}(\mathbf{x}_{1:n})] = \prod_{j=1}^{\mathbf{K}_n} \mu_0\left(\bigcap_{i\in\mathbf{\Pi}_j} B_i\right).$$

The fact that the EPPF of $\mathbf{\Pi}(\mathbf{x}_{1:n})$ is π is immediate from construction (see the Chinese restaurant construction with a given prediction rule in Section 2.2.7).

(IV $\Rightarrow$ V): Fix $n \geq 1$, Let $A = \{A_1, \dots, A_k\}$ be a partition of $[n]$ with $|A_i| = n_i$, and consider some measurable sets $B_1, \dots, B_n$. By IV we have that if $\mathbb{P}[\mathbf{\Pi}(\mathbf{x}_{1:n}) = A] > 0$,

$$\begin{aligned}\mathbb{P}[\mathbf{x}_1 \in B_1, \dots, &\mathbf{x}_n \in B_n, \mathbf{\Pi}(\mathbf{x}_{1:n}) = A]\\ &= \mathbb{P}[\mathbf{x}_1 \in B_1, \dots \mathbf{x}_n \in B_n \mid \mathbf{\Pi}(\mathbf{x}_{1:n}) = A]\mathbb{P}[\mathbf{\Pi}(\mathbf{x}_{1:n}) = A]\\ &= \prod_{j=1}^{k} \mu_0\left(\bigcap_{i\in A_j} B_i\right)\pi(n_1, \dots, n_k),\end{aligned}$$

and this probability equals zero otherwise. Now, fix $x_1, \dots, x_n \in S$, then

$$\mathbb{P}[\mathbf{x}_1 \in dx_1, \dots, \mathbf{x}_n \in dx_n] = \sum_{A\in\mathcal{P}_{[n]}} \mathbb{P}[\mathbf{x}_1 \in dx_1, \dots, \mathbf{x}_n \in dx_n, \mathbf{\Pi}(\mathbf{x}_{1:n}) = A],$$

where the sum ranges over all partitions of $[n]$. Now, for $i \neq j$ we have that if $\mathbf{x}_i \in dx_i$ and $\mathbf{x}_j \in dx_j$, then i and j belong to the same block of $\mathbf{\Pi}(\mathbf{x}_{1:n})$ if and only if $x_i = x_j$. Hence,

$$\mathbb{P}[\mathbf{x}_1 \in dx_1, \dots, \mathbf{x}_n \in dx_n, \mathbf{\Pi}(\mathbf{x}_{1:n}) = A] \neq 0,$$

if and only if $A = \mathbf{\Pi}(x_{1:n})$. In which case.

$$\mathbb{P}[\mathbf{x}_1 \in dx_1, \dots, \mathbf{x}_n \in dx_n, \mathbf{\Pi}(\mathbf{x}_{1:n}) = A] = \pi(n_1, \dots, n_k)\prod_{i=1}^{k} \mu_0(dx_j^*),$$

where $x_1^*, \dots, x_k^*$ are the distinct values in $\{x_1, \dots, x_n\}$, and $n_j = |\{i : x_i = x_j^*\}|$. Since this is the only positive term in the sum we conclude

$$\mathbb{P}[\mathbf{x}_1 \in dx_1, \dots, \mathbf{x}_n \in dx_n] = \pi(n_1, \dots, n_k)\prod_{i=1}^{k} \mu_0(dx_j^*).$$

We have proven that I,II,III $\Rightarrow$ IV $\Rightarrow$ V. Since the prediction rule, the finite dimensional distributions and the law of directing random measure characterize completely the law of an exchangeable sequence, we must also have V $\Rightarrow$ I, II, III. □

C.5 Proof of Theorem 3.7

Let $\mathbf{\Pi}_1, \dots, \mathbf{\Pi}_{\mathbf{K}_n}$ denote the blocks of $\mathbf{\Pi}(\mathbf{x}_{1:n})$, and consider $(\mathbf{x}_j^*)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0$ independently. By Theorem 3.6 IV, and the tower property of conditional expectation

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_1, \dots, \mathbf{x}_n)\right] &= \mathbb{E}\left[\mathbb{E}\left[f(\mathbf{x}_1, \dots, \mathbf{x}_n) \mid \mathbf{\Pi}(\mathbf{x}_{1:n})\right]\right] \\ &= \mathbb{E}\left[f(\mathbf{x}_{l_1}^*, \dots, \mathbf{x}_{l_n}^*) \prod_{j=1}^{\mathbf{K}_n} \prod_{r\in \mathbf{\Pi}_j} \mathbf{1}_{\{l_r = j\}}\right] \\ &= \sum_{A\in\mathcal{P}_{[n]}} \left\{\int f(x_{l_1}, \dots, x_{l_n}) \prod_{j=1}^{k} \prod_{r\in A_j} \mathbf{1}_{\{l_r=j\}}\, \mu_0(dx_1)\dots\mu_0(dx_k)\right\} \pi(|A_1|, \dots, |A_k|),\end{aligned}$$

whenever the integral in the right side exist, and where, $k = |A|$ and $A_1, \dots, A_k$ stand for the blocks of A, for $A \in \mathcal{P}_{[n]}$ fixed. This proves (3.2). Note that we may rewrite the such equation as

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_1, \dots, \mathbf{x}_n)\right] = \sum_{k=1}^{n} \sum_{(m_1,\dots,m_n)\in\mathcal{M}_n^k} \sum_{\{A_1,\dots,A_k\}} \pi(|A_1|, \dots, |A_k|)\times \\ \times\left\{\int f(x_{l_1}, \dots, x_{l_n}) \prod_{j=1}^{k} \prod_{r\in A_j} \mathbf{1}_{\{l_r=j\}}\, \mu_0(dx_1)\dots\mu_0(dx_k)\right\},\end{aligned}$$

where the inner sum ranges over all partitions of $[n]$ with exactly m_i blocks containing i elements. If f is symmetric, for every partition $A = \{A_1, \dots, A_k\}$ of $[n]$ and any $(x_1, \dots, x_k) \in S^k$ we have that

$$f(x_{l_1}, \dots, x_{l_n}) \prod_{j=1}^{k} \prod_{r\in A_j} \mathbf{1}_{\{l_r=j\}} = f(x_{[n_1,\dots,n_k]})$$

where $n_1, \dots, n_k$ are the ranked sizes of the blocks of A, and $x_{[n_1,\dots,n_k]}$ denotes the vector of size n with the first n_1 entries equal to x_1, the next n_2 entries equal to x_2, and so on (see Figure 48). This together with the symmetric of π imply

$$\pi(|A_1|, \dots, |A_k|) \int f(x_{l_1}, \dots, x_{l_n}) \prod_{j=1}^{k} \prod_{r\in A_j} \mathbf{1}_{\{l_r=j\}}\, \mu_0(dx_1)\dots\mu_0(dx_k)$$

is identical to

$$\pi(n_1, \dots, n_k) \int f(x_{[n_1,\dots,n_k]})\, \mu_0(dx_1)\dots\mu_0(dx_k).$$

Finally recall that for every $(m_1, \dots, m_n) \in \mathcal{M}_n^k$ there exist a unique ranked composition of n into k parts, $(n_1, \dots, n_k)$ such that $m_i = \sum_{j=1}^{k} \mathbf{1}_{\{n_j=i\}}$, and that the total number of partitions of $[n]$, having exactly m_i blocks that contain i elements, is $n!/(\prod_{i=1}^{n}(i!)^{m_i}(m_i!))$. Hence, we conclude

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}_1, \dots, \mathbf{x}_n)\right] = \sum_{k=1}^{n} \sum_{(m_1,\dots,m_n)\in\mathcal{M}_n^k} \frac{n!}{\prod_{i=1}^{n}(i!)^{m_i}(m_i!)} \pi(n_1, \dots, n_k)\times \\ \times \int f\left(x_{[n_1,\dots,n_k]}\right)\, \mu_0(dx_1)\dots\mu_0(dx_k),\end{aligned}$$

□

$$x_{[n_1,\ldots,n_k]} = (\underbrace{x_1,\ldots,x_1}_{n_1 \text{ times}}, \underbrace{x_2,\ldots,x_2}_{n_2 \text{ times}}, \ldots, \underbrace{x_k,\ldots,x_k}_{n_k \text{ times}})$$

Figure 48: Graphical explanation of $x_{[n_1,\ldots,n_k]}$.

C.6 Proof of Corollary 3.9

Recalling that $\mathbf{x}_i \sim \mu_0$, and using Corollary 3.8 (a), we found that for any measurable function $f : [0,1] \to \mathbb{R}_+$, and for every $i \neq j$,

$$\mathbb{E}\left[f(\mathbf{x}_j) \,|\, \mathbf{x}_i\right] = \rho\, f(\mathbf{x}_i) + (1-\rho) \int f(x)\mu_0(dx) = \rho\, f(\mathbf{x}_i) + (1-\rho)\mathbb{E}\left[f(\mathbf{x}_i)\right]$$

The choice $f(x) = x$ proves (a). To prove (b) note that for $f(x) = x^2$, we obtain $\mathbb{E}\left[\mathbf{x}_j^2 \,\middle|\, \mathbf{x}_i\right] = \rho\, \mathbf{x}_i^2 + (1-\rho)\mathbb{E}\left[\mathbf{x}_i^2\right]$, this together with (a) show that

$$\begin{aligned}\mathsf{Var}(\mathbf{x}_j \mid \mathbf{x}_i) &= \rho\, \mathbf{x}_i^2 + (1-\rho)\mathbb{E}[\mathbf{x}_i^2] - \left(\rho\, \mathbf{x}_i + (1-\rho)\mathbb{E}[\mathbf{x}_i]\right)^2 \\ &= (1-\rho)\left\{\rho\left(\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i]\right)^2 + \mathsf{Var}(\mathbf{x}_i)\right\}.\end{aligned}$$

To prove (c) we first compute, using (a), $\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_j\right] = \mathbb{E}[\mathbf{x}_i\mathbb{E}\left[\mathbf{x}_j \,|\, \mathbf{x}_i\right]] = \rho\mathbb{E}\left[\mathbf{x}_i^2\right] + (1-\rho)\mathbb{E}[\mathbf{x}_i]^2$. Thus

$$\mathsf{Cov}(\mathbf{x}_i, \mathbf{x}_j) = \rho\mathbb{E}\left[\mathbf{x}_i^2\right] + (1-\rho)\mathbb{E}[\mathbf{x}_i]^2 - \mathbb{E}[\mathbf{x}_i]^2 = \rho\,\mathsf{Var}(\mathbf{x}_i).$$

Finally, (d) follows by diving the last equation by $\mathsf{Var}(\mathbf{x}_i) = \sqrt{\mathsf{Var}(\mathbf{x}_i)\mathsf{Var}(\mathbf{x}_j)}$. □

C.7 Proof of Theorem 3.10

Say that $\rho^{(n)} \to 0$ as $n \to \infty$. Being that S and $\mathcal{P}(S)$ are Polish, and by Theorem 3.4 we might construct on some probability space $\left(\hat{\Omega}, \hat{\mathbf{F}}, \hat{\mathbb{P}}\right)$ some exchangeable sequences $\left\{\hat{\mathbf{X}}^{(n)} = \left(\hat{\mathbf{x}}_i^{(n)}\right)_{i\geq 1}\right\}_{n\geq 1}$, such that $\mathbf{X}^{(n)}$ is directed by a SSP, $\hat{\boldsymbol{\mu}}^{(n)}$, with base measure $\mu_0^{(n)}$ and tie probability $\rho^{(n)}$, and where $\hat{\boldsymbol{\mu}}^{(n)}$ converges weakly almost surely to μ_0, as $n \to \infty$. Fix $m \geq 1$ and $B_1, \ldots, B_m \in \mathscr{B}_S$. Since μ_0 is diffuse, B_i is a μ_0-continuity set, and by the Portmanteau theorem we know $\hat{\boldsymbol{\mu}}^{(n)}(B_i) \to \mu_0(B_i)$ almost surely, as $n \to \infty$. This together with the representation theorem for exchangeable sequences imply

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{x}}_i^{(n)} \in B_i\right) \,\middle|\, \hat{\boldsymbol{\mu}}^{(n)}\right] = \prod_{i=1}^{m} \hat{\boldsymbol{\mu}}^{(n)}(B_i) \to \prod_{i=1}^{m} \mu_0(B_i),$$

almost surely, as $n \to \infty$, and by taking expectations we obtain

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{x}}_i^{(n)} \in B_i\right)\right] \to \hat{\mathbb{E}}\left[\prod_{i=1}^{m} \mu_0(B_i)\right] = \prod_{i=1}^{m} \mu_0(B_i) = \hat{\mathbb{P}}\left[\bigcap_{i=1}^{m} (\hat{\mathbf{x}}_i \in B_i)\right],$$

where $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \mu_0$. Thus $\mathbf{X}^{(n)} \overset{d}{=} \hat{\mathbf{X}}^{(n)} \xrightarrow{d} \hat{\mathbf{X}}$ as $n \to \infty$, and we have proven (i).

If $\rho^{(n)} \to 1$, analogously as in (i) we may construct sequences $\left\{\hat{\mathbf{X}}^{(n)} = \left(\hat{\mathbf{x}}_i^{(n)}\right)_{i\geq 1}\right\}_{n\geq 1}$, and SSPs $\{\hat{\boldsymbol{\mu}}^{(n)}\}_{n\geq 1}$ as above, but where $\boldsymbol{\mu}^{(n)}$ converges weakly almost surely to $\delta_{\hat{\mathbf{x}}}$, as $n \to \infty$, for some $\hat{\mathbf{x}} \sim \mu_0$. Fix $m \geq 1$ and $B_1, \ldots, B_m \in \mathscr{B}_S$. The diffuseness of μ_0 implies that $\hat{\mathbf{x}} \notin \partial B_i$ almost surely, so that outside a $\hat{\mathbb{P}}$-null set, B_i is a $\delta_{\hat{\mathbf{x}}}$-continuity set, and using the Portmanteau theorem we obtain $\boldsymbol{\mu}^{(n)}(B_i) \to \delta_{\hat{\mathbf{x}}}(B_i)$, almost surely, as $n \to \infty$. The representation theorem for exchangeable sequences assures

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{x}}_i^{(n)} \in B_i\right) \middle| \hat{\boldsymbol{\mu}}^{(n)}\right] = \prod_{i=1}^{m} \hat{\boldsymbol{\mu}}^{(n)}(B_i) \to \prod_{i=1}^{m} \delta_{\hat{\mathbf{x}}}(B_i),$$

almost surely, as $n \to \infty$, and by taking expectations we get

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{x}}_i^{(n)} \in B_i\right)\right] \to \hat{\mathbb{E}}\left[\prod_{i=1}^{m} \delta_{\hat{\mathbf{x}}}(B_i)\right] = \hat{\mathbb{P}}\left[\hat{\mathbf{x}} \in B_1, \ldots, \hat{\mathbf{x}} \in B_m\right].$$

Hence $\mathbf{X}^{(n)} \overset{d}{=} \hat{\mathbf{X}}^{(n)} \xrightarrow{d} (\hat{\mathbf{x}}, \hat{\mathbf{x}}, \ldots)$ as $n \to \infty$. □

C.8 Proof of Theorem 3.11

Lemma C.2. *Let $\boldsymbol{\mu}$ be a species sampling process over $(S, \mathscr{B}_S)$ as in (3.1) with $\sum_{j\geq 1} \mathbf{w}_j > 0$. Set $\boldsymbol{\nu} = \sum_{j\geq 1} \overline{\mathbf{w}}_j \delta_{\boldsymbol{\xi}_j}$ where $\overline{\mathbf{w}}_j = \mathbf{w}_j / \sum_{j\geq 1} \mathbf{w}_j$, for every $j \geq 1$. If $\mu \in \mathcal{P}(S)$ is such that*

$$\mathbb{P}\left[\boldsymbol{\nu} \in \mathcal{U}_{\epsilon_1,\ldots,\epsilon_k}(\mu; B_1, \ldots, B_k), \sum_{j\geq 1} \mathbf{w}_j > \epsilon\right] > 0,$$

for each positive integer k, every k-tuple $(B_1, \ldots, B_k)$ of μ-continuity sets and all $(\epsilon_1, \ldots, \epsilon_k, \epsilon) \in (0,1)^{k+1}$, then $\mu \in \mathbb{WS}(\boldsymbol{\mu})$.

Proof: Fix a k-tuple of μ-continuity sets $(B_1, \ldots, B_k)$, and $\varepsilon_1, \ldots, \varepsilon_k \in (0,1)^k$. Choose

$$1 > \epsilon > \max\{0, 1 - \varepsilon_i / |\mu_0(B_i) - \mu(B_i)| : i \in \{1, \ldots, k\}\} \tag{C.10}$$

and for every $1 \leq i \leq k$, define

$$\epsilon_i = \varepsilon_i/\epsilon - |\mu_0(B_i) - \mu(B_i)|(1/\epsilon - 1).$$

Note that by (C.10), $-|\mu_0(B_i) - \mu(B_i)| > -\varepsilon_i/(1-\epsilon)$ and $1/\epsilon > 1$, thus $\epsilon_i > 0$. Now, under the event

$$\bigcap_{i=1}^{k}\{|\boldsymbol{\nu}(B_i) - \mu(B_i)| < \epsilon_i\} \cap \left\{\sum_{j\geq 1} \mathbf{w}_j > \epsilon\right\},$$

and given that $\boldsymbol{\mu} = \left(\sum_{j\geq 1} \mathbf{w}_j\right) \boldsymbol{\nu} + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \mu_0$, we get, for every $1 \leq i \leq k$,

$$\begin{aligned}
|\boldsymbol{\mu}(B_i) - \mu(B_i)| &= \left|\sum_{j\geq 1} \mathbf{w}_j \left(\boldsymbol{\nu}(B_i) - \mu_0(B_i)\right) + \left(\mu_0(B_i) - \mu(B_i)\right)\right| \\
&\leq \sum_{j\geq 1} \mathbf{w}_j \left|\boldsymbol{\nu}(B_i) - \mu(B_i)\right| + \left(1 - \sum_{j\geq 1} \mathbf{w}_j\right) \left|\mu_0(B_i) - \mu(B_i)\right| \\
&< \sum_{j\geq 1} \mathbf{w}_j \left(\frac{\varepsilon_i - |\mu_0(B_i) - \mu(B_i)|(1-\epsilon)}{\epsilon}\right) + (1-\epsilon)\left|\mu_0(B_i) - \mu(B_i)\right| \\
&= (1-\epsilon)\left|\mu_0(B_i) - \mu(B_i)\right| \left(1 - \frac{\sum_{j\geq 1} \mathbf{w}_j}{\epsilon}\right) + \frac{\varepsilon_i \sum_{j\geq 1} \mathbf{w}_j}{\epsilon}.
\end{aligned}$$

From (C.10) we obtain $|\mu_0(B_i) - \mu(B_i)| < \varepsilon_i/(1-\epsilon)$, hence

$$|\boldsymbol{\mu}(B_i) - \mu(B_i)| < \varepsilon_i \left(1 - \frac{\sum_{j\geq 1} \mathbf{w}_j}{\epsilon}\right) + \frac{\varepsilon_i \sum_{j\geq 1} \mathbf{w}_j}{\epsilon} = \varepsilon_i.$$

This means that

$$\mathbb{P}\left[\boldsymbol{\mu} \in \mathcal{U}_{\varepsilon_1,\ldots,\varepsilon_k}(\mu; B_1, \ldots, B_k)\right] \geq \mathbb{P}\left[\boldsymbol{\nu} \in \mathcal{U}_{\epsilon_1,\ldots,\epsilon_k}(\mu; B_1, \ldots, B_k), \sum_{j\geq 1} \mathbf{w}_j > \epsilon\right],$$

and the result follows. □

Lemma C.3. *Let $(\mathbf{w}_j)_{j\geq 1}$ be a weights sequence and consider a k-tuple of non-negative real numbers, $(p_1, \ldots, p_k)$ that sum up to one, with $k \geq 2$. For every $\epsilon^* > 0$, there exist $\varepsilon > 0$ such that*

$$\bigcap_{j\geq 1} \{\mathbf{w}_j < \varepsilon\} \cap \left\{\sum_{j\geq 1} \mathbf{w}_j = 1\right\} \subseteq \bigcup_{(n_1,\ldots,n_k)\in\mathbb{N}^k} \bigcap_{i=1}^{k} \left\{\left|p_i - \sum_{j=1}^{n_i} \mathbf{w}_{m_i+j}\right| < \frac{\epsilon^*}{k+1}\right\}$$

where $m_1 = 0$ and $m_i = \sum_{j=1}^{i-1} n_j$, for $2 \leq i \leq k+1$.

Proof: Without loss of generality we may assume

$$0 < \epsilon^* < \min\{p_i : 1 \leq i \leq k\} \tag{C.11}$$

Choose, $0 < \varepsilon \leq \min\{\epsilon^*/[(k-1)(k+1)], p_i - \epsilon^*/(k+1) : 1 \leq i \leq k\}$. Fix $\omega \in \Omega$ such that $\mathbf{w}_j(\omega) < \varepsilon$ for every $j \geq 1$ and $\sum_{j\geq 1} \mathbf{w}_j(\omega) = 1$. In the rest of the proof let us denote $W_j = \mathbf{w}_j(\omega)$, for $j \geq 1$. To complete the proof we must show that there exists a k-tuple, $(n_1, \ldots, n_k)$, of positive integers such that

$$p_i - \epsilon^*/(k+1) < \sum_{j=1}^{n_i} W_{m_i+j} < p_i + \epsilon^*/(k+1), \tag{C.12}$$

for every $1 \leq i \leq k$. To this aim we will first prove that for some positive integers, $n_1, \ldots, n_k$,

$$p_i - \lambda < \sum_{j=1}^{n_i} W_{m_i+j} < p_i + \lambda \tag{C.13}$$

for every $1 \le i \le k-1$ and where $\lambda = \epsilon^*/[2(k-1)(k+1)]$. We start with the case $i = 1$. As $\sum_{j\ge 1} W_j = 1$, there exists an integer $l \ge 1$ such that $\sum_{j=1}^{l} W_j > p_1 - \lambda > 0$. Define $n_1 = \min\{l \ge 1 : \sum_{j=1}^{l} W_j > p_1 - \lambda\}$. Being that $W_1 < \varepsilon \le p_1 - \epsilon^*/(k+1) < p_1 - \lambda$ we get $n_1 \ge 2$. Clearly $\sum_{j=1}^{n_1-1} W_j \le p_1 - \lambda$, and since $W_{n_1} < \varepsilon < \epsilon^*/[(k-1)(k+1)] = 2\lambda$, we obtain $\sum_{j=1}^{n_1} W_j < p_1 + \lambda$. Thus (C.13) holds for $i = 1$.

Now, assume that for some $1 \le i < k-1$, (C.13) holds for every $1 \le h \le i$. By summation, the induction hypothesis yields

$$\sum_{j=1}^{m_{i+1}} W_j = \sum_{h=1}^{i} \sum_{j=1}^{n_h} W_{m_h+j} < \sum_{h=1}^{i} p_h + i\lambda \tag{C.14}$$

Given that $\sum_{j\ge 1} W_j = 1$ and $\sum_{j=1}^{i+1} p_j + (i-1)\lambda < 1$ (by (C.11)), there exist an integer $l > 1$ such that

$$\sum_{j=1}^{l} W_j > \sum_{j=1}^{i+1} p_j + (i-1)\lambda. \tag{C.15}$$

For every such integer l, (C.14) and (C.15) imply $\sum_{j=1}^{l-m_{i+1}} W_{m_{i+1}+j} = \sum_{j=1}^{l} W_j - \sum_{j=1}^{m_{i+1}} W_j > p_{i+1} - \lambda$. So, we can set

$$n_{i+1} = \min\left\{ h \ge 1 : \sum_{j=1}^{h} W_{m_{i+1}+j} > p_{i+1} - \lambda \right\}.$$

Since $W_{m_{i+1}+1} < \varepsilon \le p_{i+1} - \epsilon^*/(k+1) < p_{i+1} - \lambda$ we get $n_{i+1} \ge 2$. Clearly, $\sum_{j=1}^{n_{i+1}-1} W_{m_{i+1}+j} \le p_{i+1} - \lambda$, and being that $W_{m_{i+1}+n_{i+1}} < \varepsilon < \epsilon^*/[(k-1)(k+1)] = 2\lambda$, we obtain $\sum_{j=1}^{n_{i+1}} W_{m_{i+1}+j} < p_{i+1} + \lambda$, which shows (C.13) for $i+1$.

Thus we have shown by induction that (C.13) is true for every $1 \le i \le k-1$. Given that $\lambda < \epsilon^*/(k+1)$, this yields (C.12) for all such i. It remains to prove (C.12) also holds for $i = k$. Summing up (C.13) for $1 \le i \le k-1$ we get

$$\sum_{j=1}^{m_k} W_j = \sum_{i=1}^{k-1} \sum_{j=1}^{n_i} W_{m_i+j} < 1 - p_k + \epsilon^*/[2(k+1)] \tag{C.16}$$

Evidently, for some integer $l \ge 1$,

$$\sum_{j=1}^{l} W_j > 1 - \epsilon^*/[2(k+1)], \tag{C.17}$$

and for every such number l, subtracting (C.16) from (C.17), we get $\sum_{j=1}^{l-m_k} W_{m_k+j} = \sum_{j=1}^{l} W_j - \sum_{j=1}^{m_k} W_j > p_k - \epsilon^*/(k+1)$. Define

$$n_k = \min\left\{ h \ge 1 : \sum_{j=1}^{h} W_{m_k+j} > p_k - \epsilon^*/(k+1) \right\}.$$

As $W_{m_k+1} < \varepsilon \le p_k - \epsilon^*/(k+1)$ we get $n_k \ge 2$. By definition of n_k, $\sum_{j=1}^{n_k-1} W_{m_k+j} \le p_k - \epsilon^*/(k+1)$, and being that $W_{m_k+n_k} < \varepsilon < \epsilon^*/[(k-1)(k+1)] < 2\epsilon^*/(k+1)$, we obtain $\sum_{j=1}^{n_k} W_{m_k+j} < p_k + \epsilon^*/(k+1)$, which shows (C.12) holds for $i = k$. □

Lemma C.4. *Let $\boldsymbol{\mu} = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}$ be a proper species sampling process with base measure μ_0, such that for every $\varepsilon > 0$,*

$$\mathbb{P}\left[A \cap \left(\bigcap_{j\geq 1}\{\mathbf{w}_j < \varepsilon\}\right)\right] > 0$$

for some $A \in \sigma(\mathbf{w}_j : j \geq 1)$. Then, for every collection of disjoint sets $\{B_1, \ldots, B_k\} \subseteq \mathscr{B}_S$ with $\mu_0(B_i) > 0$ and $\sum_{i=1}^{k} \mu_0(B_i) = 1$, every k-tuple of positive real numbers $(p_1, \ldots, p_k)$ that sum up to one, and every $\epsilon^* > 0$,

$$\mathbb{P}\left[A \cap \left(\bigcap_{i=1}^{k}\{|\boldsymbol{\mu}(B_i) - p_i| < \epsilon^*\}\right)\right] > 0.$$

Proof: Let ϵ^*, $\{B_1, \ldots, B_k\}$ and $(p_1, \ldots, p_k)$ be as in the statement of the Lemma in question. Let us denote $\mathbb{P}_A[E] = \mathbb{P}[A \cap E]$, for every event $E \in \mathcal{F}$. By hypothesis and Lemma C.3 we have that

$$\mathbb{P}_A\left[\bigcup_{(n_1,\ldots,n_k)\in\mathbb{N}^k} \bigcap_{i=1}^{k}\left\{\left|p_i - \sum_{j=1}^{n_i} \mathbf{w}_{m_i+j}\right| < \frac{\epsilon^*}{k+1}\right\}\right] > 0.$$

Hence, for some $(n_1, \ldots, n_k) \in \mathbb{N}^k$

$$\mathbb{P}_A\left[\bigcap_{i=1}^{k}\left\{\left|p_i - \sum_{j=1}^{n_i} \mathbf{w}_{m_i+j}\right| < \frac{\epsilon^*}{k+1}\right\}\right] > 0.$$

Given that $(\mathbf{w}_j)_{j\geq 1}$ is independent of $(\boldsymbol{\xi}_j)_{j\geq 1} \overset{\text{iid}}{\sim} \mu_0$, and $\mu_0(B_i) > 0$, this gives

$$\mathbb{P}_A\left[\bigcap_{i=1}^{k}\left(\left\{\left|p_i - \sum_{j=1}^{n_i} \mathbf{w}_{m_i+j}\right| < \frac{\epsilon^*}{k+1}\right\} \cap \left\{\bigcap_{j=1}^{n_i}[\boldsymbol{\xi}_{m_i+j} \in B_i]\right\}\right)\right] > 0. \tag{C.18}$$

for some $(n_1, \ldots, n_k) \in \mathbb{N}^k$. Now, as $B_1, \ldots, B_k$ are disjoint, under the event in (C.18) we get

$$p_i - \epsilon^*/(k+1) < \sum_{j=l}^{n} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}(B_i) < p_i + \epsilon^*/(k+1) \tag{C.19}$$

for every $1 \leq i \leq k$ and some integer n. Summing over $i = 1, \ldots, k$, the first inequality in (C.19) shows $1 - k\epsilon^*/(k+1) < \sum_{j=1}^{n} \mathbf{w}_j$, hence $\sum_{j>n} \mathbf{w}_j \leq k\epsilon^*/(k+1)$. Being that $\boldsymbol{\mu}(B_i) = \sum_{j\geq 1} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}(B_i)$, (C.19) yields

$$p_i - \epsilon^*/(k+1) \leq \sum_{j=1}^{n} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}(B_i) \leq \boldsymbol{\mu}(B_i) \leq \sum_{j=l}^{n} \mathbf{w}_j \delta_{\boldsymbol{\xi}_j}(B_i) + \sum_{j>n} \mathbf{w}_j \leq p_i + \epsilon^*,$$

Thus

$$\bigcap_{i=1}^{k}\left(\left\{\left|p_i - \sum_{j=1}^{n_i} \mathbf{w}_{m_i+j}\right| < \frac{\epsilon^*}{k+1}\right\} \cap \left\{\bigcap_{j=1}^{n_i}[\boldsymbol{\xi}_{m_i+j} \in B_i]\right\}\right) \subseteq \bigcap_{i=1}^{k}\{|\boldsymbol{\mu}(B_i) - p_i| < \epsilon^*\},$$

and by (C.18) we conclude $\mathbb{P}\left[\bigcap_{i=1}^{k}\{|\boldsymbol{\mu}(B_i) - p_i| < \epsilon^*\}\right] > 0$. □

Proof of Theorem 3.11:

I ⇒ II: Fix $\varepsilon > 0$ and $0 < \gamma < 1$. Pick an integer $k > \max\{(1+\varepsilon)/\varepsilon, 1/(1-\gamma)\}$ and choose a measurable partition of S, $\{B_1, \ldots, B_k\}$, such that $0 < \mu_0(B_i) < 1$, B_i is a μ_0-continuity set, for all $1 \leq i \leq k$, and

$$\mu^* = \max_{i \in \{1,\ldots,k\}} \mu_0(B_i) > \max\{1/[(1-\gamma)k], \varepsilon/(k\varepsilon - 1)\}. \tag{C.20}$$

Notice that we may take such a partition, as $(S, \mathscr{B}_S)$ is Borel, μ_0 is diffuse and by the choice of k we get $1 > \max\{1/[(1-\gamma)k], \varepsilon/(k\varepsilon - 1)\}$. Define the probability measure, $\mu^{(k)}$, through

$$\mu^{(k)}(B) = \frac{1}{k}\sum_{i=1}^{k}\frac{\mu_0(B \cap B_i)}{\mu_0(B_i)}$$

for every $B \in \mathscr{B}_S$. Clearly $\mathbb{S}\left(\mu^{(k)}\right) = \mathbb{S}(\mu_0)$ and by I we get that $\mu^{(k)}$ belongs to the support of $\boldsymbol{\mu}$. Now set

$$1/k < \epsilon < \min\{\varepsilon\mu^*/(\mu^* + \varepsilon), (1-\gamma)\mu^*\}, \tag{C.21}$$

and note that by (C.20), $1/k < \min\{\varepsilon\mu^*/(\mu^* + \varepsilon), (1-\gamma)\mu^*\}$, hence ϵ is well defined, and $\epsilon - 1/k > 0$. With this in mind, assumption I yields

$$\begin{aligned}
0 &< \mathbb{P}\left[\boldsymbol{\mu} \in \mathcal{U}_{\epsilon-1/k,\ldots,\epsilon-1/k}\left(\mu^{(k)}, B_1, \ldots, B_k\right)\right] \\
&= \mathbb{P}\left[\bigcap_{i=1}^{k}\{|\boldsymbol{\mu}(B_i) - \mu^{(k)}(B_i)| < \epsilon - 1/k\}\right] \\
&= \mathbb{P}\left[\bigcap_{i=1}^{k}\{|\boldsymbol{\mu}(B_i) - 1/k| < \epsilon - 1/k\}\right] \\
&\leq \mathbb{P}\left[\bigcap_{i=1}^{k}\left\{\sum_{\{j:\boldsymbol{\xi}_j \in B_i\}}\mathbf{w}_j + \left(1 - \sum_{j\geq 1}\mathbf{w}_j\right)\mu_0(B_i) < \epsilon\right\}\right] \\
&\leq \mathbb{P}\left[\bigcap_{j\geq 1}\{\mathbf{w}_j < \epsilon\} \cap \left\{\left(1 - \sum_{j\geq 1}\mathbf{w}_j\right)\mu^* < \epsilon\right\}\right] \\
&\leq \mathbb{P}\left[\bigcap_{j\geq 1}\{\overline{\mathbf{w}}_j < \epsilon/(1-\epsilon/\mu^*)\} \cap \left\{\sum_{j\geq 1}\mathbf{w}_j > 1 - \epsilon/\mu^*\right\}\right]
\end{aligned}$$

which by (C.21) yields

$$\mathbb{P}\left[\bigcap_{j\geq 1}\{\overline{\mathbf{w}}_j < \varepsilon\} \cap \left\{\sum_{j\geq 1}\mathbf{w}_j > \gamma\right\}\right] > 0,$$

this in turn implies II.

II ⇒ I: Fix $\mu \in \mathcal{P}(S)$ such that $\mathbb{S}(\mu) \subseteq \mathbb{S}(\mu_0)$. By virtue of Lemma C.2, to complete the proof it suffices to show

$$\mathbb{P}\left[\boldsymbol{\nu} \in \mathcal{U}_{\epsilon_1,\ldots,\epsilon_k}(\mu; B_1, \ldots, B_k), \sum_{j\geq 1} \mathbf{w}_j > \epsilon\right] > 0,$$

for each positive integer k, every k-tuple $(B_1, \ldots, B_k)$ of μ-continuity sets and all $(\epsilon_1, \ldots, \epsilon_k, \epsilon) \in (0,1)^{k+1}$, where $\boldsymbol{\nu} = \sum_{j\geq 1} \overline{\mathbf{w}}_j \delta_{\boldsymbol{\xi}_j}$. Let $\{C_1, \ldots, C_m\}$ be the partition of S generated by $\{B_1, \ldots, B_k\}$, the first thing we want to prove is that if $\mu_0(C_i) = 0$, then $\mu(C_i) = 0$ for all $i \in \{1, \ldots, m\}$. To this aim denote by C_i° to the interior of C_i, clearly if $\mu_0(C_i) = 0$, then C_i° is an open set with μ_0 measure equal to 0, thus $\mathbb{S}(\mu) \subseteq \mathbb{S}(\mu_0) \subseteq (C_i^\circ)^c$, which means $\mu(C_i^\circ) = 0$. Further as $\partial(A_1 \cap A_2) \subseteq \partial A_1 \cup \partial A_2$, for all $A_1, A_2 \in B_S$, and $B_1, \ldots, B_k$ are μ-continuity sets, we must have $\mu(\partial C_i) = 0$ for all $i \in \{1, \ldots, m\}$. Thus, for each i such that $\mu_0(C_i) = 0$, we obtain $\mu(C_i) = 0$. Note that $m \leq 2^k$, hence, by assumption II, Lemma C.4 with $A = \left\{\sum_{j\geq 1} \mathbf{w}_j > \epsilon\right\}$ and for $\epsilon^* = \min\{\epsilon_1, \ldots, \epsilon_k\}$, we can write

$$\begin{aligned}
0 &< \mathbb{P}\left[\bigcap_{\{i:\mu_0(C_i)>0\}} \left\{|\boldsymbol{\nu}(C_i) - \mu(C_i)| < 2^{-k}\epsilon^*\right\} \cap \left\{\sum_{j\geq 1} \mathbf{w}_j > \epsilon\right\}\right] \\
&\leq \mathbb{P}\left[\bigcap_{i=1}^{k} \{|\boldsymbol{\nu}(B_i) - \mu(B_i)| < \epsilon^*\} \cap \left\{\sum_{j\geq 1} \mathbf{w}_j > \epsilon\right\}\right] \\
&\leq \mathbb{P}\left[\boldsymbol{\nu} \in \mathcal{U}_{\epsilon_1,\ldots,\epsilon_k}(\mu; B_1, \ldots, B_k), \sum_{j\geq 1} \mathbf{w}_j > \epsilon\right],
\end{aligned}$$

as desired.

II ⇒ III: Let $\varepsilon > 0$, by definition $\sum_{j\geq 1} \mathbf{w}_j \leq 1$ almost surely. Hence

$$\mathbb{P}\left[\bigcup_{m\geq 1} \left\{\sum_{j>m} \mathbf{w}_j < \epsilon/2\right\}\right] = 1$$

This equation together with II for $\gamma = 1 - \varepsilon/2$, yield

$$\begin{aligned}
0 &< \mathbb{P}\left[\bigcap_{j\geq 1} \left\{\mathbf{w}_j < \varepsilon \sum_{i\geq 1} \mathbf{w}_i\right\} \cap \left\{\sum_{i\geq 1} \mathbf{w}_i > 1 - \varepsilon/2\right\}\right] \\
&= \mathbb{P}\left[\bigcap_{j\geq 1} \left\{\mathbf{w}_j < \varepsilon \sum_{i\geq 1} \mathbf{w}_i\right\} \cap \left\{\sum_{i\geq 1} \mathbf{w}_i > 1 - \varepsilon/2\right\} \cap \bigcup_{m\geq 1} \left\{\sum_{i>m} \mathbf{w}_i < \epsilon/2\right\}\right] \\
&\leq \sum_{m\geq 1} \mathbb{P}\left[\bigcap_{j\geq 1} \left\{\mathbf{w}_j < \varepsilon \sum_{i\geq 1} \mathbf{w}_i\right\} \cap \left\{\sum_{i\geq 1} \mathbf{w}_i > 1 - \varepsilon/2\right\} \cap \left\{\sum_{i>m} \mathbf{w}_i < \epsilon/2\right\}\right].
\end{aligned}$$

Hence, there exist $m \geq 1$ such that

$$\mathbb{P}\left[\bigcap_{j\geq 1} \left\{\mathbf{w}_j < \varepsilon \sum_{i\geq 1} \mathbf{w}_i\right\} \cap \left\{\sum_{i\geq 1} \mathbf{w}_i > 1 - \varepsilon/2\right\} \cap \left\{\sum_{i>m} \mathbf{w}_i < \epsilon/2\right\}\right] > 0. \tag{C.22}$$

Noting that for all $m \geq 1$, $\sum_{i=1}^m \mathbf{w}_i = \sum_{i\geq 1} \mathbf{w}_i - \sum_{i>m} \mathbf{w}_i$ and $\sum_{i\geq 1} \mathbf{w}_i \leq 1$ almost surely, (C.22) shows

$$\mathbb{P}\left[\bigcap_{j\geq 1} \{\mathbf{w}_j < \varepsilon\} \cap \left\{\sum_{i=1}^m \mathbf{w}_i > 1-\varepsilon\right\}\right] > 0,$$

which in turn implies III.

III ⇒ II: Fix $\varepsilon > 0$ and $0 < \gamma < 1$. Set $0 < \epsilon < \min\{\varepsilon/(1+\varepsilon), 1-\gamma\}$. Note that if $\mathbf{w}_1 < \epsilon, \ldots \mathbf{w}_m < \epsilon$ and $\sum_{j=1}^m \mathbf{w}_j > 1-\epsilon$ for some $m \geq 1$, then for all $j \in \{1,\ldots,m\}$

$$\mathbf{w}_j < \epsilon < \frac{\epsilon}{1-\epsilon}\sum_{j=1}^m \mathbf{w}_j < \frac{\epsilon}{1-\epsilon}\sum_{j\geq 1}\mathbf{w}_j,$$

and for $j > m$

$$\mathbf{w}_j \leq \sum_{j>m}\mathbf{w}_j = \sum_{j\geq 1}\mathbf{w}_j - \sum_{j=1}^m \mathbf{w}_j \leq 1 - \sum_{j=1}^m \mathbf{w}_j < \epsilon < \frac{\epsilon}{1-\epsilon}\sum_{j\geq 1}\mathbf{w}_j.$$

In either case we get $\overline{\mathbf{w}}_j < \varepsilon$, further $\sum_{j\geq 1}\mathbf{w}_j \geq \sum_{j=1}^m \mathbf{w}_j > 1-\epsilon > \gamma$. This together with III, show that

$$\begin{aligned} 0 &< \mathbb{P}\left[\mathbf{w}_1 < \epsilon, \ldots, \mathbf{w}_m < \epsilon, \sum_{j=1}^m \mathbf{w}_j > 1-\epsilon\right] \\ &\leq \mathbb{P}\left[\bigcap_{j\geq 1}\{\overline{\mathbf{w}}_j < \varepsilon\} \cap \left\{\sum_{j\geq 1}\mathbf{w}_i > \gamma\right\}\right] \end{aligned}$$

for some $m \geq 1$. In particular we obtain, $\mathbb{P}\left[\max_{j\geq 1}\mathbf{w}_j < \varepsilon, \sum_{j\geq 1}\mathbf{w}_i > \gamma\right] > 0$. □

C.9 Proof of Proposition 3.13

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space over which $(\mathbf{w}_j)_{j\geq 1}$ is defined. Given that we are interested in an almost surely decomposition, we may assume without loss of generality that $\sum_{j\geq 1}\mathbf{w}_j \leq 1$ and $0 \leq \mathbf{w}_j \leq 1$, for every $j \geq 1$, hold over Ω. Fix $\mathbf{v}_1 = \mathbf{w}_1$, and for $k \geq 2$, define the event $E_k = \left\{\omega \in \Omega : \sum_{j=1}^{k-1}\mathbf{w}_j(\omega) < 1\right\}$ and set

$$\mathbf{v}_k = \frac{\mathbf{w}_k}{1-\sum_{j=1}^{k-1}\mathbf{w}_j}\mathbf{1}_{E_k}.$$

Evidently $\mathbf{v}_k$ is measurable as $\mathbf{w}_1, \ldots, \mathbf{w}_k$ are. Also, since $\sum_{j\geq 1}\mathbf{w}_j \leq 1$, we have that $\mathbf{w}_k(\omega) \leq 1 - \sum_{j=1}^{k-1}\mathbf{w}_j(\omega)$, for every $\omega \in E_k$, which yields $0 \leq \mathbf{v}_k \leq 1$. This shows $(\mathbf{v}_i)_{i\geq 1}$ is a sequence of $[0,1]$-valued random variables. Now, for $k < k'$ we have that $E_{k'} \subseteq E_k$. Hence, for every $k \geq 2$,

$$\mathbf{w}_k(\omega) = \frac{\mathbf{w}_k(\omega)}{1-\sum_{j=1}^{k-1}\mathbf{w}_j(\omega)}\prod_{j=1}^{k-1}\left(\frac{1-\sum_{i=1}^{j}\mathbf{w}_i(\omega)}{1-\sum_{i=1}^{j-1}\mathbf{w}_i(\omega)}\right) = \mathbf{v}_k(\omega)\prod_{j=1}^{k-1}(1-\mathbf{v}_j(\omega)),$$

for all $\omega \in E_k$, and since $\sum_{j\geq 1}\mathbf{w}_j \leq 1$, $\mathbf{w}_k(\omega) = 0 = \mathbf{v}_k(\omega)$, for $\omega \in (E_k)^c$. This show that for every $k \geq 1$, $\mathbf{w}_k = \mathbf{v}_k\prod_{j=1}^{k-1}(1-\mathbf{v}_j)$, as desired. □

C.10 Proof of Lemma 3.14

By definition $\boldsymbol{\mu}$ is proper if and only if $\sum_{j\geq 1} \mathbf{w}_j = 1$, which can be rewritten as $\lim_{j\to\infty}(1 - \sum_{i=1}^{j} \mathbf{w}_i) = \lim_{j\to\infty} \prod_{i=1}^{j}(1-\mathbf{v}_i) = 0$. Since the length variables are non-negative and bounded by 1 we get, $\prod_{i=1}^{j}(1-\mathbf{v}_i) \to 0$ almost surely, is equivalent to

$$\lim_{j\to\infty} \mathbb{E}\left[\prod_{i=1}^{j}(1-\mathbf{v}_i)\right] = \mathbb{E}\left[\lim_{j\to\infty}\prod_{i=1}^{j}(1-\mathbf{v}_i)\right] = 0,$$

which yields (i.a). To prove (i.b) note that for any real numbers $(a_i)_{i\geq 1}$ such that $0 \leq a_i < 1$, $\prod_{i=1}^{\infty}(1-a_i) = 0$ if and only $\sum_{i\geq 1} a_i = \infty$. Now fix ω in the original probability space and note that if there exist $i \geq 1$ such that $\mathbf{v}_i(\omega) = 1$ then $\prod_{i=1}^{\infty}(1-\mathbf{v}_i(\omega)) = 0$ trivially. Alternatively, if $\mathbf{v}_i(\omega) < 1$, for every $i \geq 1$, and $\sum_{i\geq 1}\mathbf{v}_i(\omega) = \infty$, then we also have $\prod_{i=1}^{\infty}(1-\mathbf{v}_i(\omega)) = 0$ which implies (i.b).

It remains to prove (ii). Fix $0 < \gamma < \varepsilon$ such that for every $n \geq 1$, $\mathbb{P}\left[\bigcap_{i=1}^{n}(\gamma < \mathbf{v}_i < \varepsilon)\right] > 0$. Being that $1 - \sum_{j=1}^{n}\mathbf{w}_j = \prod_{i=1}^{n}(1-\mathbf{v}_i)$, we get

$$\begin{aligned}
0 &< \mathbb{P}\left[\bigcap_{i=1}^{n}(\gamma < \mathbf{v}_i < \varepsilon)\right] \\
&\leq \mathbb{P}\left[\bigcap_{j=1}^{n}\left\{\mathbf{w}_j < \varepsilon(1-\gamma)^{j-1}\right\} \cap \left\{\sum_{j=1}^{n}\mathbf{w}_j > 1-(1-\gamma)^n\right\}\right] \\
&\leq \mathbb{P}\left[\bigcap_{j=1}^{n}\left\{\mathbf{w}_j < \varepsilon\right\} \cap \left\{\sum_{j=1}^{n}\mathbf{w}_j > 1-(1-\gamma)^n\right\}\right]
\end{aligned}$$

At this stage choose $n > \log(\varepsilon)/\log(1-\gamma)$ so that $(1-\gamma)^n < \varepsilon$. This yields III of Theorem 3.11 and the result follows. □

C.11 Proof of Proposition 3.16

Fix $\boldsymbol{\tau} = (\mathbf{w}_j)_{j\geq 1}$, and define a random variable $\mathbf{u}$ such that, conditionally given $\boldsymbol{\tau}$, has density

$$\mathbb{p}_{\mathbf{u}}(u \mid \boldsymbol{\tau}) = \sum_{j\geq 1}\mathbf{1}_{\{u<\mathbf{w}_j\}} = |\{j \geq 1 : u < \mathbf{w}_j\}|,$$

as illustrated below in Figure 49. Now, define $\boldsymbol{\Psi}_u = \{j \geq 1 : \mathbf{w}_j > u\}$ For every $u \in [0,1]$. Clearly $\boldsymbol{\Psi}_{\mathbf{u}} \neq \emptyset$ almost surely, since $\mathbf{u} \leq \max_{j\geq 1}\mathbf{w}_j$ almost surely. Also, as $\sum_{j\geq 1}\mathbf{w}_j = 1$ almost surely, we must have $\boldsymbol{\Psi}_{\mathbf{u}}$ is finite almost surely, That is $\boldsymbol{\Psi}$ takes values in $\mathcal{F}_{\mathbb{N}}$ with probability one. Finally, note that

$$\mathbf{w}_j = \int_0^1 \mathbf{1}_{\{u<\mathbf{w}_j\}}du = \int_0^1 \frac{1}{|\boldsymbol{\Psi}_u|}\mathbf{1}_{\{u<\mathbf{w}_j\}}\mathbb{p}_{\mathbf{u}}(u \mid \boldsymbol{\tau})du = \mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}_{\mathbf{u}}|}\mathbf{1}_{\{j\in\boldsymbol{\Psi}_{\mathbf{u}}\}}\,\middle|\,\boldsymbol{\tau}\right].$$

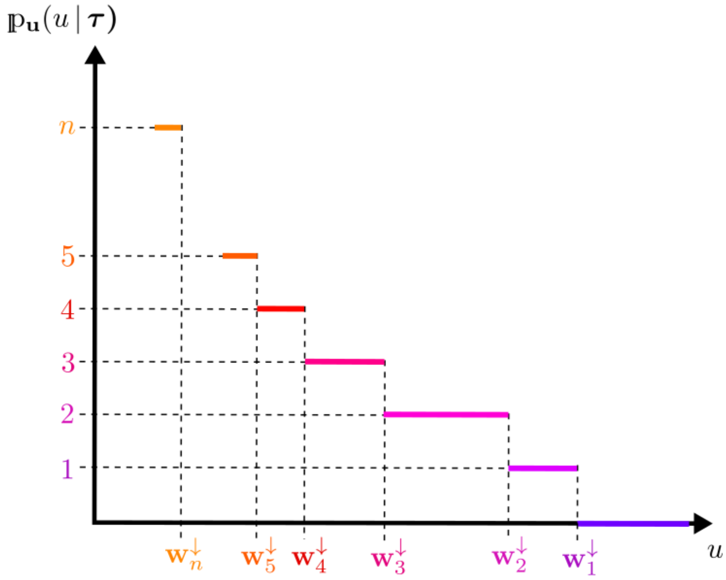


Figure 49: Illustration of $\mathbb{p}_{\mathbf{u}}(u \mid \boldsymbol{\tau})$, where $\left(\mathbf{w}_j^{\downarrow}\right)_{j\geq 1}$ stands for the decreasing rearrangement of $(\mathbf{w}_j)_{j\geq 1}$.

□

C.12 Proof of Proposition 3.17

Fix $B_1, \dots, B_n \in \mathscr{B}_S$. By the tower property of conditional expectation and conditional monotone convergence theorem we obtain

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{k=1}^{n}(\mathbf{x}_k \in B_k) \,\middle|\, \boldsymbol{\tau}, \boldsymbol{\Xi}\right] &= \mathbb{E}\left[\mathbb{P}\left[\bigcap_{k=1}^{n}(\mathbf{x}_k \in B_k) \,\middle|\, \boldsymbol{\tau}, \boldsymbol{\Xi}, \boldsymbol{\Psi}_1, \dots \boldsymbol{\Psi}_n\right] \,\middle|\, \boldsymbol{\tau}, \boldsymbol{\Xi}\right] \\
&= \mathbb{E}\left[\prod_{k=1}^{n}\left(\frac{1}{|\boldsymbol{\Psi}_k|}\sum_{j\in\boldsymbol{\Psi}_k}\delta_{\boldsymbol{\xi}_j}(B_k)\right) \,\middle|\, \boldsymbol{\tau}, \boldsymbol{\Xi}\right] \\
&= \prod_{k=1}^{n}\mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}_k|}\sum_{j\geq 1}\mathbf{1}_{\{j\in\boldsymbol{\Psi}_k\}}\delta_{\boldsymbol{\xi}_j}(B_k) \,\middle|\, \boldsymbol{\tau}, \boldsymbol{\Xi}\right] \\
&= \prod_{k=1}^{n}\left(\sum_{j\geq 1}\mathbb{E}\left[\frac{1}{|\boldsymbol{\Psi}_k|}\mathbf{1}_{\{j\in\boldsymbol{\Psi}_k\}} \,\middle|\, \boldsymbol{\tau}\right]\delta_{\boldsymbol{\xi}_j}(B_k)\right) \\
&= \prod_{k=1}^{n}\left(\sum_{j\geq 1}\mathbf{w}_j\delta_{\boldsymbol{\xi}_j}(B_k)\right)
\end{aligned}$$

Finally from (3.9) it is evident that $\mathbf{W}$ is $\boldsymbol{\tau}$-measurable, from which we conclude

$$\mathbb{P}\left[\bigcap_{k=1}^{n}(\mathbf{x}_k \in B_k) \,\middle|\, \mathbf{W}, \boldsymbol{\Xi}\right] = \prod_{k=1}^{n}\left(\sum_{j\geq 1}\mathbf{w}_j\delta_{\boldsymbol{\xi}_j}(B_k)\right).$$

□

C.13 Proof of Theorem 3.18

Recall that

a) If $\mathbf{z} \sim \mathsf{Gamma}(a, 1)$, then for every $t \in \mathbb{R}_+$,

$$\mathbb{E}\left[e^{-t\mathbf{z}}\right] = \frac{1}{(1+t)^a}.$$

b) If $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent random variables with $\mathbf{z}_i \sim \mathsf{Gamma}(a_i, 1)$, then $\mathbf{z} = \sum_{i=1}^n \mathbf{z}_i \sim \mathsf{Gamma}(a, 1)$, where $a = \sum_{i=1}^n a_i$. Further,

$$\left(\frac{\mathbf{z}_1}{\mathbf{z}}, \ldots, \frac{\mathbf{z}_n}{\mathbf{z}}\right) \sim \mathsf{Dir}\left(a_1, \ldots, a_n\right).$$

c) If $(\mathbf{z}_1, \ldots, \mathbf{z}_n) \sim \mathsf{Dir}\left(a_1, \ldots, a_n\right)$, then, for every $1 \leq k \leq n$

$$\mathbb{E}[\mathbf{z}_k] = \frac{a_k}{\sum_{k=1}^n a_k}.$$

First we see that Definition 3.8 implies Definition 3.6. Let $\boldsymbol{\lambda}^* = \sum_{j \geq 1} \delta_{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)}$ be the Poisson process over $S \times \mathbb{R}_+$ defined by $\{(\boldsymbol{\xi}_j, \boldsymbol{\alpha}_j)\}_{j \geq 1}$, and denote $\lambda^* = \lambda \otimes \varrho$. Fix $A \in \mathscr{B}_S$ and $t \in \mathbb{R}_+$, and define $f_t : S \times \mathbb{R}_+ \to \mathbb{R}$ by $f_t(s, x) = xt\mathbf{1}_A(s)$, so that $\boldsymbol{\lambda}^*(f_t) = t\boldsymbol{\lambda}(A)$. By Proposition 1.9 we get

$$\begin{aligned}
\mathbb{E}\left[e^{-\boldsymbol{\lambda}^*(f_t)}\right] &= \exp\left\{-\lambda^*\left(1 - e^{-f_t}\right)\right\} \\
&= \exp\left\{-\lambda(A)\int_0^\infty (1 - e^{-tx})\frac{e^{-x}}{x}dx\right\} \\
&= \frac{1}{(1+t)^{\lambda(A)}},
\end{aligned}$$

which yields $\mathbb{E}\left[e^{-t\boldsymbol{\lambda}(A)}\right] = (1+t)^{-\lambda(A)}$. That is $\boldsymbol{\lambda}(A) \sim \mathsf{Gamma}(\lambda(A), 1)$. Since $\boldsymbol{\lambda}$ has independent increments and is λ-homogeneous we even get that for every measurable partition $\{B_i\}_{i=1}^n$ of S, $\boldsymbol{\lambda}(B_1), \ldots, \boldsymbol{\lambda}(B_n)$ are independent random variables with $\boldsymbol{\lambda}(B_i) \sim \mathsf{Gamma}(\lambda(B_i), 1)$, hence by (b) above we get

$$(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n)) = \left(\frac{\boldsymbol{\lambda}(B_1)}{\boldsymbol{\lambda}(S)}, \ldots, \frac{\boldsymbol{\lambda}(B_n)}{\boldsymbol{\lambda}(S)}\right) \sim \mathsf{Dir}(\lambda(B_1), \ldots, \lambda(B_n)).$$

Definition 3.6, then follows trivially by fixing $\mu = \lambda$.

Secondly, we show Definition 3.6 yields Definition 3.7. So let $\boldsymbol{\mu}$ be as in the first definition of Dirichlet process and consider an exchangeable sequence $(\mathbf{x}_i)_{i \geq 1}$ directed by $\boldsymbol{\mu}$, that is $\{\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \boldsymbol{\mu}\} \overset{\text{iid}}{\sim} \boldsymbol{\mu}$. Let $B = \{B_k\}_{k=1}^n$ be a measurable partition of S and denote $\boldsymbol{\mu}_B = (\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$. Let us compute the conditional distribution of $\boldsymbol{\mu}_B$ given $\mathbf{x}_1, \ldots, \mathbf{x}_m$. To this aim define $\mathbf{y}_k = \sum_{i=1}^m \mathbf{1}_{\{\mathbf{x}_i \in B_k\}}$, for every $1 \leq k \leq n$, so that $\mathbf{y}_k$ counts the number of $\mathbf{x}_i$'s that fall into B_k, and set $\mathbf{Y} = (\mathbf{y}_k)_{k=1}^n$. Note that given $\boldsymbol{\mu}_B$, $\mathbf{Y} \sim \mathsf{Multinomial}(\boldsymbol{\mu}(B_1), \ldots, \boldsymbol{\mu}(B_n))$, that is

$$\mathbb{P}\left[\mathbf{Y} = (y_1, \ldots, y_n) \mid \boldsymbol{\mu}_B\right] = \frac{n!}{y_1! \cdots y_n!}\prod_{k=1}^n \boldsymbol{\mu}(B_k)^{y_k}.$$

As to the distribution of $\boldsymbol{\mu}_B$, by hypothesis know

$$\mathbb{P}\left[\bigcap_{k=1}^{n}\{\boldsymbol{\mu}(B_k)\in dx_k\}\right]=\frac{\Gamma\left(\sum_{k=1}^{n}\mu(B_k)\right)}{\prod_{k=1}^{n}\Gamma(\mu(B_k))}\prod_{k=1}^{n}x_k^{\mu(B_k)-1}dx_k.$$

From the last couple of equation it is easy to compute the conditional density of $\boldsymbol{\mu}_B$ given $\mathbf{Y}$, denoted by $\mathbb{p}_{\{\boldsymbol{\mu}_B|\mathbf{Y}\}}$. In effect,

$$\mathbb{p}_{\{\boldsymbol{\mu}_B|\mathbf{Y}\}}(x_1,\dots,x_n)\propto\left(\prod_{k=1}^{n}x_k^{\mathbf{y}_k}\right)\left(\prod_{k=1}^{n}x_k^{\mu(B_k)-1}\right)=\prod_{k=1}^{n}x_k^{\mathbf{y}_k+\mu(B_k)-1},$$

which yields

$$\mathbb{p}_{\{\boldsymbol{\mu}_B|\mathbf{Y}\}}(x_1,\dots,x_n)=\frac{\Gamma\left(\sum_{k=1}^{n}\mathbf{y}_k+\mu(B_k)\right)}{\prod_{k=1}^{n}\Gamma(\mathbf{y}_k+\mu(B_k))}\prod_{k=1}^{n}x_k^{\mathbf{y}_k+\mu(B_k)-1}.$$

Hence, we obtain $\{\boldsymbol{\mu}_B\mid\mathbf{Y}\}\sim\mathsf{Dir}\left(\mathbf{y}_1+\mu(B_1),\dots,\mathbf{y}_n+\mu(B_n)\right)$, which can be rewritten as

$$\{(\boldsymbol{\mu}(B_1),\dots,\boldsymbol{\mu}(B_n))\mid\mathbf{Y}\}\sim\mathsf{Dir}\left(\sum_{i=1}^{m}\delta_{\mathbf{x}_i}(B_1)+\theta\mu_0(B_1),\dots,\sum_{i=1}^{m}\delta_{\mathbf{x}_i}(B_n)+\theta\mu_0(B_n)\right).\tag{C.23}$$

Now, the exchangeability of $(\mathbf{x}_i)_{i\geq1}$ together with the definition of directing random measure imply that for $A\in\mathscr{B}_S$, $\boldsymbol{\mu}(A)$ depends on $\mathbf{x}_1,\dots,\mathbf{x}_m$ only through $\sum_{i=1}^{m}\delta_{\mathbf{x}_i}(A)$. This together with (C.23) and (c) in the initial remainder yield

$$\begin{aligned}\mathbb{P}\left[\mathbf{x}_{m+1}\in A\mid\mathbf{x}_1,\dots,\mathbf{x}_m\right]&=\mathbb{E}\left[\mathbb{P}\left[\mathbf{x}_{m+1}\in A\mid\boldsymbol{\mu},\mathbf{x}_1,\dots,\mathbf{x}_m\right]\mid\mathbf{x}_1,\dots,\mathbf{x}_m\right]\\&=\mathbb{E}\left[\boldsymbol{\mu}(A)\mid\mathbf{x}_1,\dots,\mathbf{x}_m\right]\\&=\frac{\sum_{i=1}^{m}\delta_{\mathbf{x}_i}(A)+\theta\mu_0(A)}{m+\theta}.\end{aligned}$$

To finish this part of the proof note the $\mathbb{P}[\mathbf{x}_1\in A]=\mathbb{E}[\boldsymbol{\mu}(A)]=\mu_0$. Thus Definition 3.7 holds.

Finally, the proof that Definition 3.7 implies Definition 3.9 follows from Proposition 2.12. Realize that the stick-breaking decomposition, the construction through normalization, the EPPF (prediction rule) and the finite dimensional distributions all characterize completely the law of a SSP, so we must have the four definitions are equivalent.

□

D Proofs of Section 4

D.1 Proof of Theorem 4.1

(i) By the second part of Lemma 3.14 it suffices to show that for every $0 < \varepsilon' < 1$, there exist $0 < \delta < \varepsilon'$ such that $\mathbb{P}\left[\bigcap_{i=1}^{n}(\delta < \mathbf{v}_i < \varepsilon')\right] > 0$, for every $n \geq 1$. So fix $0 < \varepsilon' < 1$ and consider $\varepsilon'' = \min\{\varepsilon, \varepsilon'\}$, where $\varepsilon > 0$ is such $(0, \varepsilon)$ is contained in the support of ν_0. Set $\delta = \varepsilon''/2$, by the representation theorem for exchangeable sequences, Jensen's inequality and the fact that $(\delta, \varepsilon'') \subseteq (0, \varepsilon)$ is contained in the support of ν_0,

$$\mathbb{P}\left[\bigcap_{i=1}^{n}(\delta < \mathbf{v}_i < \varepsilon'')\right] \geq \mathbb{E}\left[\prod_{i=1}^{n} \boldsymbol{\nu}(\delta, \varepsilon'')\right] = \mathbb{E}\left[\{\boldsymbol{\nu}(\delta, \varepsilon'')\}^n\right] \geq \{\nu_0(\delta, \varepsilon'')\}^n > 0,$$

for every $n \geq 1$. As $\varepsilon'' \leq \varepsilon'$, we conclude $\mathbb{P}\left[\bigcap_{i=1}^{n}(\delta < \mathbf{v}_i < \varepsilon')\right] \geq \mathbb{P}\left[\bigcap_{i=1}^{n}(\delta < \mathbf{v}_i < \varepsilon'')\right] > 0$, for $n \geq 1$.

(ii) By the first part of Lemma 3.14 we know $\boldsymbol{\mu}$ is proper if and only if $\mathbb{E}\left[\prod_{i=1}^{j}(1 - \mathbf{v}_i)\right] \to 0$, as $j \to \infty$. Since $(\mathbf{v}_i)_{i\geq 1}$ is exchangeable and directed by $\boldsymbol{\nu}$, we get $\mathbb{E}\left[\prod_{i=1}^{j}(1 - \mathbf{v}_i)\right] = \mathbb{E}\left[(1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^j\right]$. Now, if $\boldsymbol{\nu}(\{0\}) < 1$ almost surely, then $\mathbb{P}\left[\mathbf{v}_1 > 0 \mid \boldsymbol{\nu}\right] > 0$, almost surely. Since $\mathbf{v}_1$ is non-negative, this shows $\mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}] > 0$ almost surely, hence, as $j \to \infty$, $(1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^j \to 0$, almost surely. This yields $\mathbb{E}\left[\prod_{i=1}^{j}(1 - \mathbf{v}_i)\right] = \mathbb{E}\left[(1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^j\right] \to 0$ as $j \to \infty$. Alternatively, if $\mathbb{P}\left[\boldsymbol{\nu}(\{0\}) = 1\right] > 0$, then for every $j \geq 1$,

$$\mathbb{E}\left[(1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^j\right] = \mathbb{E}\left[(1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^j \mathbf{1}_{\{\boldsymbol{\nu}(\{0\})<1\}}\right] + \mathbb{P}\left[\boldsymbol{\nu}(\{0\}) = 1\right].$$

Which implies

$$\lim_{j\to\infty} \mathbb{E}\left[\prod_{i=1}^{j}(1 - \mathbf{v}_i)\right] = \lim_{j\to\infty} \mathbb{E}\left[(1 - \mathbb{E}[\mathbf{v}_1 \mid \boldsymbol{\nu}])^j\right] \geq \mathbb{P}\left[\boldsymbol{\nu}(\{0\}) = 1\right] > 0.$$

□

D.2 Proof of Theorem 4.3

(i) First we see that the sequences of length variables $\mathbf{V}^{(n)} = \left(\mathbf{v}_i^{(n)}\right)_{i\geq 1}$ converge in distribution to $\mathbf{V} = (\mathbf{v}_i)_{i\geq 1}$. Since $[0, 1]$ and $\mathcal{P}([0, 1])$ are Polish, we might construct on some probability space $\left(\hat{\Omega}, \hat{\mathbf{F}}, \hat{\mathbb{P}}\right)$ some exchangeable sequences $\left\{\hat{\mathbf{V}}^{(n)} = \left(\hat{\mathbf{v}}_i^{(n)}\right)_{i\geq 1}\right\}_{n\geq 1}$, such that $\hat{\mathbf{V}}^{(n)}$ is directed by a random probability measure $\hat{\boldsymbol{\nu}}^{(n)} \overset{d}{=} \boldsymbol{\nu}^{(n)}$, and where $\hat{\boldsymbol{\nu}}^{(n)}$ converges weakly almost surely to $\nu_0 \neq \delta_0$, as $n \to \infty$. Fix $m \geq 1$ and $B_1, \ldots, B_m \in \mathscr{B}_{[0,1]}$ such that B_i is a ν_0-continuity set for every $i \leq m$. By the Portmanteau theorem we know $\hat{\boldsymbol{\nu}}(B_i) \to \nu_0(B_i)$ almost surely as $n \to \infty$. This together with the representation theorem for exchangeable sequences imply

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{v}}_i^{(n)} \in B_i\right) \,\middle|\, \hat{\boldsymbol{\nu}}^{(n)}\right] = \prod_{i=1}^{m} \hat{\boldsymbol{\nu}}^{(n)}(B_i) \to \prod_{i=1}^{m} \nu_0(B_i),$$

almost surely, as $n \to \infty$, and by taking expectations we obtain

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{v}}_i^{(n)} \in B_i\right)\right] \to \prod_{i=1}^{m}\nu_0(B_i) = \mathbb{P}\left[\bigcap_{i=1}^{m}(\mathbf{v}_i \in B_i)\right].$$

Since each sequence of length variables is countable it is enough to prove the convergence of the finite dimensional distributions, and we get $\mathbf{V}^{(n)} \overset{d}{=} \hat{\mathbf{V}}^{(n)} \xrightarrow{d} \mathbf{V}$ as desired. Note that mapping

$$(v_1, v_2, \ldots, v_j) \mapsto \left(v_1, v_2(1-v_1), \ldots, v_j\prod_{i=1}^{j-1}(1-v_i)\right)$$

is continuous with respect to the product topology, thus the weights of $\boldsymbol{\mu}^{(n)}$, $\mathbf{W}^{(n)} = \mathsf{SB}\left[\mathbf{V}^{(n)}\right]$, converge in distribution to the weights of $\boldsymbol{\mu}$, $\mathbf{W} = \mathsf{SB}\left[\mathbf{V}\right]$. Further, the requirements of $\boldsymbol{\nu}^{(n)}$ and ν_0 assure $\mathbf{W}^{(n)}$ and $\mathbf{W}$ take values in the infinite dimensional simplex, Δ_∞. In addition, as the base measures $\mu_0^{(n)}$ converge weakly to μ_0, we also have that the atoms of $\boldsymbol{\mu}^{(n)}$, $\Xi^{(n)} = \left(\boldsymbol{\xi}_j^{(n)}\right)_{j\geq 1}$ (which are independent of $\mathbf{W}^{(n)}$) converge in distribution to the atoms of $\boldsymbol{\mu}$, $\Xi = (\boldsymbol{\xi}_j)_{j\geq 1}$ (which are independent of $\mathbf{W}$). Thus $\left(\mathbf{W}^{(n)}, \Xi^{(n)}\right) \xrightarrow{d} (\mathbf{W}, \Xi)$ in $\Delta_\infty \times S^\infty$, and Lemma 1.18 yields $\boldsymbol{\mu}^{(n)}$ converges weakly in distribution to $\boldsymbol{\mu}$. In particular, if ν_0 denotes a $\mathsf{Be}(1,\theta)$ distribution we get $\boldsymbol{\mu}$ is a Dirichlet process.

(ii) Analogously as in (i) we may construct sequences $\left\{\hat{\mathbf{V}}^{(n)} = \left(\hat{\mathbf{v}}_i^{(n)}\right)_{i\geq 1}\right\}_{n\geq 1}$, such that $\hat{\mathbf{V}}^{(n)}$ is directed by a random probability measure $\hat{\boldsymbol{\nu}}^{(n)} \overset{d}{=} \boldsymbol{\nu}^{(n)}$, and where $\hat{\boldsymbol{\nu}}^{(n)}$ converges weakly almost surely to $\delta_{\hat{\mathbf{v}}}$, with $\hat{\mathbf{v}} \sim \nu_0$, as $n \to \infty$. Fix $m \geq 1$ and $B_1, \ldots, B_m \in \mathscr{B}_{[0,1]}$ such that B_i is a ν_0-continuity set for every $i \leq m$. Then we get that $\hat{\mathbf{v}} \notin \partial B_i$ almost surely, so that outside a $\hat{\mathbb{P}}$-null set, B_i is a $\delta_{\hat{\mathbf{v}}}$-continuity set, and using the Portmanteau theorem we obtain $\boldsymbol{\nu}^{(n)}(B_i) \to \delta_{\hat{\mathbf{v}}}(B_i)$, almost surely, as $n \to \infty$. The representation theorem for exchangeable sequences assures

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{v}}_i^{(n)} \in B_i\right)\middle|\,\hat{\boldsymbol{\nu}}^{(n)}\right] = \prod_{i=1}^{m}\hat{\boldsymbol{\nu}}^{(n)}(B_i) \to \prod_{i=1}^{m}\delta_{\hat{\mathbf{v}}}(B_i),$$

almost surely, as $n \to \infty$, and by taking expectations we get

$$\hat{\mathbb{P}}\left[\bigcap_{i=1}^{m}\left(\hat{\mathbf{v}}_i^{(n)} \in B_i\right)\right] \to \hat{\mathbb{E}}\left[\prod_{i=1}^{m}\delta_{\hat{\mathbf{v}}}(B_i)\right] = \hat{\mathbb{P}}\left[\hat{\mathbf{v}} \in B_1, \ldots, \hat{\mathbf{v}} \in B_m\right].$$

Hence $\mathbf{V}^{(n)} \overset{d}{=} \hat{\mathbf{V}}^{(n)} \xrightarrow{d} (\hat{\mathbf{v}}, \hat{\mathbf{v}}, \ldots) \overset{d}{=} (\mathbf{v}, \mathbf{v}, \ldots)$ as $n \to \infty$. The rest of the proof of (ii) follows identically is in (i). □

D.3 Proof of Theorem 4.6

Fix $j \geq 1$. As ν_0 diffuse, $\left(1 - \mathbf{v}_j^{(\rho_\nu)}\right) > 0$ almost surely, for every $\rho_\nu \in (0,1)$ and $\left(1 - \mathbf{v}_j^{(\rho_\nu)}\right)^{-1}$ is well defined. Now, from the stick-breaking decomposition of the weights it follows that $\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}$ if and only if $\mathbf{v}_j^{(\rho_\nu)}\left(1-\mathbf{v}_j^{(\rho_\nu)}\right)^{-1} \geq \mathbf{v}_{j+1}^{(\rho_\nu)}$, or equivalently $\mathbf{v}_{j+1}^{(\rho_\nu)} \leq c\left(\mathbf{v}_j^{(\rho_\nu)}\right)$ where $c(v) = 1 \wedge v(1-v)^{-1}$. By Proposition 3.8 we know that under the event $\left\{\mathbf{v}_j^{(\rho_\nu)} \neq \mathbf{v}_{j+1}^{(\rho_\nu)}\right\}$, which occurs with probability $1-\rho_\nu$, the conditional distribution of $\left(\mathbf{v}_j^{(\rho_\nu)}, \mathbf{v}_{j+1}^{(\rho_\nu)}\right)$ is that of $(\mathbf{v}^*, \mathbf{v}) \overset{\text{iid}}{\sim} \nu_0$. Hence we can easily compute

$$\begin{aligned}\mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right] &= \mathbb{P}\left[\mathbf{v}_{j+1}^{(\rho_\nu)} \leq c\left(\mathbf{v}_j^{(\rho_\nu)}\right) \,\middle|\, \mathbf{v}_j^{(\rho_\nu)} = \mathbf{v}_{j+1}^{(\rho_\nu)}\right]\rho_\nu + \\ &\qquad \mathbb{P}\left[\mathbf{v}_{j+1}^{(\rho_\nu)} \leq c\left(\mathbf{v}_j^{(\rho_\nu)}\right) \,\middle|\, \mathbf{v}_j^{(\rho_\nu)} \neq \mathbf{v}_{j+1}^{(\rho_\nu)}\right](1-\rho_\nu) \\ &= \rho_\nu + (1-\rho_\nu)\mathbb{P}\left[\mathbf{v}^* \leq c(\mathbf{v})\right].\end{aligned} \tag{D.1}$$

As $0 \leq \mathbb{P}\left[\mathbf{v}^* \leq c(\mathbf{v})\right] \leq 1$, it is clear that $\rho_\nu \mapsto \mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right]$ is continuous and non-decreasing. Particularly, if there exist $\varepsilon > 0$ such that $(0, \varepsilon)$ is contained in the support of ν_0, then for $\varepsilon' = \min\{1/2, \varepsilon/2\}$ we have that

$$\mathbb{P}\left[\mathbf{v} < \varepsilon'/2,\, \varepsilon/2 < \mathbf{v}^* < \varepsilon\right] = \nu_0(0, \varepsilon'/2)\,\nu_0(\varepsilon/2, \varepsilon) > 0.$$

Since $\{\mathbf{v} < \varepsilon'/2,\ \varepsilon/2 < \mathbf{v}^*\} \subseteq \{\mathbf{v}^* > c(\mathbf{v})\}$, we get $\mathbb{P}\left[\mathbf{v}^* > c(\mathbf{v})\right] > 0$, that is $\mathbb{P}\left[\mathbf{v}^* \leq c(\mathbf{v})\right] < 1$. From (D.1) it is clear that, in this case, the mapping $\rho_\nu \mapsto \mathbb{P}\left[\mathbf{w}_j^{(\rho_\nu)} \geq \mathbf{w}_{j+1}^{(\rho_\nu)}\right]$ is even increasing, which proves (b).

The proof of (a) follows from (D.1) by noting that $\mathbb{P}\left[\mathbf{v}^* \leq c(\mathbf{v})\right] = \mathbb{E}\left[\overrightarrow{\nu_0}(c(\mathbf{v}))\right]$, where $\overrightarrow{\nu_0}$ is the distribution function of $\mathbf{v}^*$. Then, by taking limits, as $\rho_\nu \to 1$ and $\rho_\nu \to 0$, we obtain (c) and finally lower bound in (d) is immediate from (b) and (c). □

D.4 Proof of Corollary 4.10

For $\nu_0 = \mathsf{Be}(1, \theta)$, its distribution function is given by, $\nu_0([0, x]) = \overrightarrow{\nu_0}(x) = 1 - (1-x)^\theta$, hence by substituting the tie probability $\rho_\nu = 1/(\beta+1)$, in Theorem 4.3, we obtain

$$\mathbb{P}\left[\mathbf{w}_j^{(\beta)} \geq \mathbf{w}_{j+1}^{(\beta)}\right] = 1 - \frac{\beta}{\beta+1}\mathbb{E}\left[(1-c(\mathbf{v}))^\theta\right].$$

where $c(v) = 1 \wedge v(1-v)^{-1}$ and $\mathbf{v} \sim \mathsf{Be}(1, \theta)$. Now,

$$\mathbb{E}\left[(1-c(\mathbf{v}))^\theta\right] = \theta\int_0^{1/2}\left(1 - \frac{x}{1-x}\right)^\theta (1-x)^{\theta-1}dx = \theta\int_0^{1/2}\frac{(1-2x)^\theta}{(1-x)}dx,$$

and by the change of variables $y = 2x$,

$$\mathbb{E}\left[(1-c(\mathbf{v}))^\theta\right] = \frac{\theta}{2}\int_0^1 \frac{(1-y)^\theta}{(1-y/2)}dy = \frac{{}_2F_1(1,1;\theta+2,1/2)\theta}{2(\theta+1)}.$$

The rest of the proof follows easily by simple substitution and taking the corresponding limits.

□

D.5 Proof of Theorem 4.13

(i) Equation (4.5) proves $(\mathbf{v}_1, \dots, \mathbf{v}_n) \sim \nu_0 \circ \boldsymbol{\nu} \circ \dots \circ \boldsymbol{\nu}$, thus the first statement of (i) is straight forward from Lemma 3.14. To prove the second statement in (i), it suffices to show that for every $0 < \varepsilon' < 1$, there exist $0 < \delta < \varepsilon'$ such that $\mathbb{P}\left[\bigcap_{i=1}^{n}(\delta < \mathbf{v}_i < \varepsilon')\right] > 0$, for every $n \geq 1$. So fix $0 < \varepsilon' < 1$ and consider $\varepsilon'' = \min\{\varepsilon, \varepsilon'\}$, where $\varepsilon > 0$ is as in (i), also set $\delta = \varepsilon''/2$. As $(0, \varepsilon)$ is contained in the support of $\boldsymbol{\nu}(v; \cdot)$ for every $v \in (0, \varepsilon)$, we get that $(\delta, \varepsilon'') \subseteq (0, \varepsilon)$ is also contained in the support of $\boldsymbol{\nu}(v; \cdot)$ which yields

$$\int_{\delta}^{\varepsilon''} \boldsymbol{\nu}(v; dx) = \boldsymbol{\nu}(v; (\delta, \varepsilon'')) > 0$$

for every $v \in (\delta, \varepsilon'')$. Hence,

$$\begin{aligned}
\mathbb{P}\left[\bigcap_{i=1}^{n}(\delta < \mathbf{v}_i < \varepsilon')\right] &\geq \mathbb{P}\left[\bigcap_{i=1}^{n}(\delta < \mathbf{v}_i < \varepsilon'')\right] \\
&= \int_{(\delta,\varepsilon'')^n} \nu_0 \circ \boldsymbol{\nu} \circ \dots \circ \boldsymbol{\nu}(dv_1, dv_2 \dots, dv_n) \\
&= \int_{\delta}^{\varepsilon''} \cdots \int_{\delta}^{\varepsilon''} \boldsymbol{\nu}(v_{n-1}; dv_n) \cdots \boldsymbol{\nu}(v_1; dv_2)\nu_0(dv_1) > 0.
\end{aligned}$$

(ii) If ν_0 is $\boldsymbol{\nu}$-ergodic, the ergodic theorem for stationary Markov chains (Theorem 4.12) yields

$$\lim_{n \to \infty} \frac{\sum_{i \leq n} \mathbf{v}_i}{n} = \mathbb{E}[\mathbf{v}_1],$$

almost surely. Now, if $\nu_0 \neq \delta_0$, $\mathbb{E}[\mathbf{v}_1] > 0$, which implies $\sum_{i \geq 1} \mathbf{v}_i = \infty$ almost surely and the statement follows from Lemma 3.14. Conversely if $\nu_0 = \delta_0$, we get $\mathbb{E}[\mathbf{v}_1] = \mathbb{E}[\mathbf{v}_i] = 0$ for every $i \geq 1$, which implies $\mathbf{v}_i = 0$ almost surely for every $i \geq 1$ in which case $\boldsymbol{\mu}$ is diffuse almost surely.

□

D.6 Proof of Theorem 4.14

Lemma D.1 (Continuous mappings)**.** *Let S and T be Polish spaces. Let $\boldsymbol{\eta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ be random elements taking values in S, with $\boldsymbol{\eta}_n \xrightarrow{d} \boldsymbol{\eta}$, and consider some measurable mappings $f, f_1, f_2 \dots$ from S into T satisfying $f_n(s_n) \to f(s)$, for every $s_n \to s$ in S. Then $f_n(\boldsymbol{\eta}_n) \xrightarrow{d} f(\boldsymbol{\eta})$.*

Proof: Since S is Polish, it is Borel together with its Borel σ-algebra, hence we might construct on some probability space $\left(\hat{\Omega}, \hat{\mathbf{F}}, \hat{\mathbb{P}}\right)$ the random elements $\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2, \dots$ such that $\hat{\boldsymbol{\eta}} \overset{d}{=} \boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}_n \overset{d}{=} \boldsymbol{\eta}_n$ for every $n \geq 1$ and $\hat{\boldsymbol{\eta}}_n \to \hat{\boldsymbol{\eta}}$, almost surely as $n \to \infty$. By hypothesis this yields $f_n(\hat{\boldsymbol{\eta}}_n) \to f(\hat{\boldsymbol{\eta}})$ almost surely as $n \to \infty$, so in particular we obtain $f_n(\boldsymbol{\eta}_n) \overset{d}{=} f_n(\hat{\boldsymbol{\eta}}_n) \xrightarrow{d} f(\hat{\boldsymbol{\eta}}) \overset{d}{=} f(\boldsymbol{\eta})$ as desired. □

Lemma D.2. *Let S and T be Polish spaces. Consider some random elements $\boldsymbol{\gamma}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots$ and $\boldsymbol{\eta}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ taking values in S and T, respectively. Let ν be the distribution of $\boldsymbol{\gamma}$ and $\nu^{(n)}$ be the distribution of $\boldsymbol{\gamma}_n$, also consider some regular versions, $\boldsymbol{\nu}(\boldsymbol{\gamma}; \cdot)$ and $\boldsymbol{\nu}^{(n)}(\boldsymbol{\gamma}_n; \cdot)$,* *of* $\mathbb{P}[\boldsymbol{\eta} \in \cdot \mid \boldsymbol{\gamma}]$ *and* $\mathbb{P}[\boldsymbol{\eta}_n \in \cdot \mid \boldsymbol{\gamma}_n]$ *respectively. If* $\nu^{(n)} \xrightarrow{w} \nu$ *and for every* $s_n \to s$ *in* S *we have that* $\boldsymbol{\nu}^{(n)}(s_n; \cdot) \xrightarrow{w} \boldsymbol{\nu}(s; \cdot)$*, then* $(\boldsymbol{\gamma}_n, \boldsymbol{\eta}_n) \xrightarrow{d} (\boldsymbol{\gamma}, \boldsymbol{\eta})$*.*

Proof: Let $g : S \times T \to \mathbb{R}$ be a continuous and bounded function. Define $f, f_1, f_2, \ldots : S \to \mathbb{R}$ by

$$f_n(s) = \int g(s,t)\boldsymbol{\nu}^{(n)}(s; dt) \quad \text{and} \quad f(s) = \int g(s,t)\boldsymbol{\nu}(s; dt)$$

The first thing we will prove is that

$$f_n(s_n) \to f(s) \quad \text{as} \quad s_n \to s. \tag{D.2}$$

So let $s_n \to s$. Choose some random elements $\boldsymbol{\zeta}, \boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots$ with $\boldsymbol{\zeta}_n \sim \boldsymbol{\nu}^{(n)}(s_n; \cdot)$ and $\boldsymbol{\zeta} \sim \boldsymbol{\nu}(s; \cdot)$, this way, $\boldsymbol{\zeta}_n \xrightarrow{d} \boldsymbol{\zeta}$ by hypothesis. Define $h, h_1, h_2, \ldots : T \to \mathbb{R}$ by $h_n(t) = g(s_n, t)$ and $h(t) = g(s,t)$. As g is continuous, we have that $h_n(t_n) = g(s_n, t_n) \to g(s,t) = h(t)$, for every $t_n \to t$ in T. Hence, Lemma D.1 yields $h_n(\boldsymbol{\zeta}_n) \xrightarrow{d} h(\boldsymbol{\zeta})$, and in particular we obtain

$$\int g(s_n, t)\boldsymbol{\nu}^{(n)}(s_n; dt) = \mathbb{E}[h_n(\boldsymbol{\zeta}_n)] \to \mathbb{E}[h(\boldsymbol{\zeta})] = \int g(s,t)\boldsymbol{\nu}(s; dt).$$

Since $s_n \to s$ was arbitrary, this proves equation (D.2), which together with the hypothesis and Lemma D.1 show that $f_n(\boldsymbol{\gamma}_n) \xrightarrow{d} f(\boldsymbol{\gamma})$. Particularly,

$$\int\int g(s,t)\boldsymbol{\nu}^{(n)}(s; dt)\nu^{(n)}(ds) = \mathbb{E}[f_n(\boldsymbol{\gamma}_n)] \to \mathbb{E}[f(\boldsymbol{\gamma})] = \int\int g(s,t)\boldsymbol{\nu}(s; dt)\nu(ds). \tag{D.3}$$

Note that the double integral in the left side of equation (D.3) coincides with $\mathbb{E}[g(\boldsymbol{\gamma}_n, \boldsymbol{\eta_n})]$, whilst the one at the right side coincides with $\mathbb{E}[g(\boldsymbol{\gamma}, \boldsymbol{\eta})]$. That is, we have proven that $\mathbb{E}[g(\boldsymbol{\gamma}_n, \boldsymbol{\eta}_n)] \to \mathbb{E}[g(\boldsymbol{\gamma}, \boldsymbol{\eta})]$, for every continuous and bounded function $g : S \times T \to \mathbb{R}$, or equivalently $(\boldsymbol{\gamma}_n, \boldsymbol{\eta}_n) \xrightarrow{d} (\boldsymbol{\gamma}, \boldsymbol{\eta})$.

□

Proof of Theorem 4.14: (i) Let $(\mathbf{v}_i)_{i\geq 1} \overset{\text{iid}}{\sim} \nu_0$, and say that for some $m \geq 1$,

$$\left(\mathbf{v}_1^{(n)}, \ldots, \mathbf{v}_m^{(n)}\right) \xrightarrow{d} (\mathbf{v}_1, \ldots, \mathbf{v}_m) \tag{D.4}$$

as $n \to \infty$. Since $(\mathbf{v}_i)_{i\geq 1}$ and $\left(\mathbf{v}_i^{(n)}\right)_{i\geq 1}$ are Markov chains, we get $\mathbb{P}\left[\mathbf{v}_{m+1} \in \cdot \,|\, \mathbf{v}_1, \ldots, \mathbf{v}_m\right] = \nu_0$ and $\mathbb{P}\left[\mathbf{v}_{m+1}^{(n)} \in \cdot \,\middle|\, \mathbf{v}_1^{(n)}, \ldots, \mathbf{v}_m^{(n)}\right] = \boldsymbol{\nu}^{(n)}\left(\mathbf{v}_m^{(n)}; \cdot\right)$. By hypothesis we know that that for $v_n \to v \in [0,1]$, $\boldsymbol{\nu}^{(n)}(v_n; \cdot) \xrightarrow{w} \nu_0(v; \cdot) = \nu_0$, this together with (D.4) and Lemma D.2 yield

$$\left(\mathbf{v}_1^{(n)}, \ldots, \mathbf{v}_{m+1}^{(n)}\right) \xrightarrow{d} (\mathbf{v}_1, \ldots, \mathbf{v}_{m+1})$$

as $n \to \infty$. This induction argument together with the assumption $\mathbf{v}_1^{(n)} \xrightarrow{d} \mathbf{v}_1$ show that (D.4) holds for every $m \geq 1$ and we even obtain $\left(\mathbf{v}_i^{(n)}\right)_{i\geq 1} \xrightarrow{d} (\mathbf{v}_i)_{i\geq 1}$. The rest of the proof follows like that of Theorem 4.3, by noting that the mappings $(v_i)_{i\geq 1} \mapsto \mathsf{SB}[(v_i)_{i\geq 1}]$ and

$$[(w_1, w_2, \ldots), (s_1, s_2, \ldots)] \mapsto \sum_{j\geq 1} w_j \delta_{s_j}$$

are continuous with respect to the product and weak topologies. The proof of (ii) is completely analogous, with the difference that if $\boldsymbol{\nu}^{(n)}(v_n;\cdot) \xrightarrow{w} \delta_v$ for every $v_n \to v \in [0,1]$, then $\left(\mathbf{v}_i^{(n)}\right)_{i\geq 1} \xrightarrow{d} (\mathbf{v}, \mathbf{v}, \ldots)$ for some $\mathbf{v} \sim \nu_0$. □

D.7 Proof of Lemma 4.16

Lemma D.3. *For every $n \geq 1$ consider a Binomial random variable $\mathbf{x}_n \sim \mathsf{Bin}(n, p_n)$ where $p_n \to p$ in $[0,1]$. Then, as $n \to \infty$, $\mathbf{x}_n \xrightarrow{\mathcal{L}_2} p$.*

Proof: For $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\mathbf{x}_n}{n} - p\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\mathbf{x}_n^2\right] - \frac{2p}{n}\mathbb{E}[\mathbf{x}_n] + p^2 \\ &= \frac{p_n(1-p_n)}{n} + (p_n - p)^2.\end{aligned} \tag{D.5}$$

By taking limits as $n \to \infty$ in (D.5) we obtain

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{\mathbf{x}_n}{n} - p\right)^2\right] = 0.$$

□

Proof of Lemma 4.16: The first part of this result is straight forward. To prove (ii) fix $p_\kappa \to p \in [0,1]$. By Lemma D.3 we might construct on some probability space $\left(\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}}\right)$ some random variables $\hat{\mathbf{z}}^{(k)} \sim \mathsf{Bin}(\kappa, p_\kappa)$ such that $\hat{\mathbf{z}}^{(\kappa)}/\kappa \to p$ almost surely, as $\kappa \to \infty$. Also consider $\left\{\hat{\mathbf{v}}^{(k)} \,\middle|\, \hat{\mathbf{z}}^{(\kappa)}\right\} \sim \mathsf{Be}\left(\alpha^{(\kappa)} + \hat{\mathbf{z}}^{(\kappa)}, \theta^{(\kappa)} + \kappa - \hat{\mathbf{z}}^{(\kappa)}\right)$, so that marginally $\hat{\mathbf{v}}^{(k)} \sim \boldsymbol{\nu}^{(\kappa)}(p_\kappa;\cdot)$. Conditionally given $\hat{\mathbf{z}}^{(\kappa)}$, the moment generator function of $\hat{\mathbf{v}}^{(k)}$ is

$$\hat{\mathbb{E}}\left[e^{t\hat{\mathbf{v}}^{(\kappa)}} \,\middle|\, \hat{\mathbf{z}}^{(\kappa)}\right] = 1 + \sum_{n=1}^{\infty}\left(\prod_{r=0}^{n-1} \frac{\alpha^{(\kappa)} + \hat{\mathbf{z}}^{(\kappa)} + r}{\alpha^{(\kappa)} + \theta^{(\kappa)} + \kappa + r}\right)\frac{t^n}{n!}, \tag{D.6}$$

for every $t \in \mathbb{R}$. By construction we have that $\hat{\mathbf{z}}^{(\kappa)}/\kappa \to p$ almost surely, and by hypothesis $\alpha^{(\kappa)} \to \alpha$ and $\theta^{(\kappa)} \to \theta$ in $(0,\infty)$, which means that for every $r \geq 0$,

$$\frac{\alpha^{(\kappa)} + \hat{\mathbf{z}}^{(\kappa)} + r}{\alpha^{(\kappa)} + \theta^{(\kappa)} + \kappa + r} = \left(\frac{\alpha^{(\kappa)} + r}{\kappa} + \frac{\hat{\mathbf{z}}^{(\kappa)}}{\kappa}\right)\left(\frac{\alpha^{(\kappa)} + \theta^{(\kappa)} + r}{\kappa} + 1\right)^{-1} \to p, \tag{D.7}$$

almost surely, as $\kappa \to \infty$. Hence by the tower property of conditional expectation, equations (D.6) and (D.7), and Lebesgue dominated convergence theorem (the corresponding functions are dominated by e^t) we obtain

$$\begin{aligned}\lim_{\kappa\to\infty} \hat{\mathbb{E}}\left[e^{t\hat{\mathbf{v}}^{(\kappa)}}\right] &= \lim_{\kappa\to\infty} \hat{\mathbb{E}}\left[\hat{\mathbb{E}}\left[e^{t\hat{\mathbf{v}}^{(\kappa)}} \,\middle|\, \hat{\mathbf{z}}^{(\kappa)}\right]\right] \\ &= \hat{\mathbb{E}}\left[1 + \sum_{n=1}^{\infty}\left(\prod_{r=0}^{n-1} \lim_{\kappa\to\infty} \frac{\alpha^{(\kappa)} + \hat{\mathbf{z}}^{(\kappa)} + r}{\alpha^{(\kappa)} + \theta^{(\kappa)} + \kappa + r}\right)\frac{t^n}{n!}\right] \\ &= \hat{\mathbb{E}}\left[1 + \sum_{n=1}^{\infty}\frac{(pt)^n}{n!}\right] \\ &= e^{tp},\end{aligned}$$

which proves $\hat{\mathbf{v}}^{(\kappa)} \xrightarrow{d} p$, as $\kappa \to \infty$, or equivalently $\boldsymbol{\nu}^{(\kappa)}(p_\kappa;\cdot) \xrightarrow{w} \delta_p$. The fact that $\nu_0^{(\kappa)} \xrightarrow{w} \mathsf{Be}(\alpha, \theta)$ is obvious. □

D.8 Proof of Proposition 4.19

After possibly expanding the original probability space we might construct a chain $(\mathbf{z}_i)_{i\geq 1}$ such that for every $i \geq 1$, $\{\mathbf{z}_i \mid \mathbf{v}_i\} \sim \mathsf{Bin}(\kappa, \mathbf{v}_i)$, where $\mathbf{z}_i$ is conditionally independent of $(\mathbf{v}_1, \mathbf{z}_1, \dots, \mathbf{v}_{i-1}, \mathbf{z}_{i-1})$, given $\mathbf{v}_i$, and $\{\mathbf{v}_{i+1} \mid \mathbf{z}_i\} \sim \mathsf{Be}(\alpha + \mathbf{z}_i, \theta + \kappa - \mathbf{z}_i)$, where $\mathbf{v}_{i+1}$ is conditionally independent of $(\mathbf{v}_1, \mathbf{z}_1, \dots, \mathbf{v}_{i-1}, \mathbf{z}_{i-1}, \mathbf{v}_i)$ given $\mathbf{z}_i$.

a) Using elementary properties of conditional expectation, we obtain

$$\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{v}_i] = \mathbb{E}[\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{z}_i] \mid \mathbf{v}_i] = \mathbb{E}\left[\frac{\alpha + \mathbf{z}_i}{\alpha + \theta + \kappa} \,\middle|\, \mathbf{v}_i\right] = \frac{\alpha + \kappa \mathbf{v}_i}{\alpha + \theta + \kappa}.$$

b) Notice that

$$\mathsf{Var}(\mathbf{v}_{i+1} \mid \mathbf{v}_i) = \mathbb{E}[\mathsf{Var}(\mathbf{v}_{i+1} \mid \mathbf{z}_i) \mid \mathbf{v}_i] + \mathsf{Var}(\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{z}_i] \mid \mathbf{v}_i),$$

we first compute

$$\mathsf{Var}(\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{z}_i] \mid \mathbf{v}_i) = \mathsf{Var}\left(\frac{\alpha + \mathbf{z}_i}{\alpha + \theta + \kappa} \,\middle|\, \mathbf{v}_i\right) = \frac{\mathbf{v}_i(1 - \mathbf{v}_i)\kappa}{(\alpha + \theta + \kappa)^2},$$

secondly, we note that

$$\begin{aligned}
&\mathbb{E}[(\alpha + \mathbf{z}_i)(\theta + \kappa - \mathbf{z}_i) \mid \mathbf{v}_i] \\
&\qquad = \mathsf{Cov}(\alpha + \mathbf{z}_i, \theta + \kappa - \mathbf{z}_i \mid \mathbf{v}_i) + \mathbb{E}[\alpha + \mathbf{z}_i \mid \mathbf{v}_i]\mathbb{E}[\theta + \kappa - \mathbf{z}_i \mid \mathbf{v}_i] \\
&\qquad = -\mathsf{Var}(\mathbf{z}_i \mid \mathbf{v}_i) + (\alpha + \kappa \mathbf{v}_i)(\theta + \kappa - \kappa \mathbf{v}_i) \\
&\qquad = -\kappa \mathbf{v}_i(1 - \mathbf{v}_i) + (\alpha + \kappa \mathbf{v}_i)(\theta + \kappa(1 - \mathbf{v}_i)),
\end{aligned}$$

hence

$$\begin{aligned}
\mathbb{E}[\mathsf{Var}(\mathbf{v}_{i+1} \mid \mathbf{z}_i) \mid \mathbf{v}_i] &= \mathbb{E}\left[\frac{(\alpha + \mathbf{z}_i)(\theta + \kappa - \mathbf{z}_i)}{(\alpha + \theta + \kappa)^2(\alpha + \theta + \kappa + 1)} \,\middle|\, \mathbf{v}_i\right] \\
&= \frac{-\kappa \mathbf{v}_i(1 - \mathbf{v}_i) + (\alpha + \kappa \mathbf{v}_i)(\theta + \kappa(1 - \mathbf{v}_i))}{(\alpha + \theta + \kappa)^2(\alpha + \theta + \kappa + 1)},
\end{aligned}$$

and we can conclude the proof of (b),

$$\begin{aligned}
\mathsf{Var}(\mathbf{v}_{i+1} \mid \mathbf{v}_i) &= \frac{-\kappa \mathbf{v}_i(1 - \mathbf{v}_i) + (\alpha + \kappa \mathbf{v}_i)(\theta + \kappa(1 - \mathbf{v}_i)) + \mathbf{v}_i(1 - \mathbf{v}_i)\kappa(\alpha + \theta + \kappa + 1)}{(\alpha + \theta + \kappa)^2(\alpha + \theta + \kappa + 1)} \\
&= \frac{(\alpha + \kappa \mathbf{v}_i)(\theta + \kappa(1 - \mathbf{v}_i)) + \kappa \mathbf{v}_i(1 - \mathbf{v}_i)(\alpha + \theta + \kappa)}{(\alpha + \theta + \kappa)^2(\alpha + \theta + \kappa + 1)}.
\end{aligned}$$

c) We first note that from the Beta-Binomial conjugate model and since $\mathbf{v}_i \sim \mathsf{Be}(\alpha, \theta)$, $\{\mathbf{v}_i \mid \mathbf{z}_i\} \sim \mathsf{Be}(\alpha + \mathbf{z}_i, \theta + \kappa - \mathbf{z}_i)$, further by construction $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ are conditionally given $\mathbf{z}_i$, thus

$$\mathbb{E}[\mathbf{v}_i \mathbf{v}_{i+1}] = \mathbb{E}[\mathbb{E}[\mathbf{v}_i \mathbf{v}_{i+1} \mid \mathbf{z}_i]] = \mathbb{E}[\mathbb{E}[\mathbf{v}_i \mid \mathbf{z}_i]\mathbb{E}[\mathbf{v}_{i+1} \mid \mathbf{z}_i]] = \mathbb{E}\left[\left(\frac{\alpha + \mathbf{z}_i}{\alpha + \theta + \kappa}\right)^2\right],$$

conditioning on $\mathbf{v}_i$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\alpha+\mathbf{z}_i}{\alpha+\theta+\kappa}\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\alpha+\mathbf{x}_i}{\alpha+\theta+\kappa}\right)^2 \middle| \mathbf{v}_i\right]\right] \\
&= \mathbb{E}\left[\frac{\alpha^2+2\alpha\mathbb{E}[\mathbf{z}_i \mid \mathbf{v}_i]+\mathbb{E}[\mathbf{z}_i^2 \mid \mathbf{v}_i]}{(\alpha+\theta+\kappa)^2}\right] \\
&= \frac{\alpha^2+2\alpha\kappa\mathbb{E}[\mathbf{v}_i]+\kappa\mathbb{E}[\mathbf{v}_i]+\kappa(\kappa-1)\mathbb{E}[\mathbf{v}_i^2]}{(\alpha+\theta+\kappa)^2} \\
&= \left[\alpha^2+\frac{\kappa(2\alpha^2+\alpha)}{\alpha+\theta}+\frac{\kappa(\kappa-1)\alpha(\alpha+1)}{(\alpha+\theta)(\alpha+\theta+1)}\right](\alpha+\theta+\kappa)^{-2},
\end{aligned}$$

hence

$$\begin{aligned}
\mathsf{Cov}(\mathbf{v}_i,\mathbf{v}_{i+1}) &= \mathbb{E}[\mathbf{v}_i\mathbf{v}_{i+1}]-\mathbb{E}[\mathbf{v}_i]\mathbb{E}[\mathbf{v}_{i+1}] \\
&= (\alpha+\theta+\kappa)^{-2}\left[\alpha^2+\frac{\kappa(2\alpha^2+\alpha)}{\alpha+\theta}+\frac{\kappa(\kappa-1)\alpha(\alpha+1)}{(\alpha+\theta)(\alpha+\theta+1)}\right]-\frac{\alpha^2}{(\alpha+\theta)^2} \\
&= \frac{\kappa\alpha\theta}{(\alpha+\theta)^2(\alpha+\theta+1)(\alpha+\theta+\kappa)}.
\end{aligned}$$

d) The correlation simplifies as follows

$$\mathsf{Corr}(\mathbf{v}_i,\mathbf{v}_{i+1}) = \frac{\mathsf{Cov}(\mathbf{v}_i,\mathbf{v}_{i+1})}{\sqrt{\mathsf{Var}(\mathbf{v}_i)}\sqrt{\mathsf{Var}(\mathbf{v}_{i+1})}} = \frac{\kappa\alpha\theta(\alpha+\theta)^2(\alpha+\theta+1)}{\alpha\theta(\alpha+\theta)^2(\alpha+\theta+1)(\alpha+\theta+\kappa)} = \frac{\kappa}{\alpha+\theta+\kappa}.$$

□

D.9 Proof of Proposition 4.20

After possibly expanding the original probability space we might construct a chain $(\mathbf{z}_i)_{i\geq 1}$ such that for every $i \geq 1$, $\{\mathbf{z}_i \mid \mathbf{v}_i\} \sim \mathsf{Bin}(\kappa,\mathbf{v}_i)$, where $\mathbf{z}_i$ is conditionally independent of $(\mathbf{v}_1,\mathbf{z}_1,\ldots,\mathbf{v}_{i-1},\mathbf{z}_{i-1})$, given $\mathbf{v}_i$, and $\{\mathbf{v}_{i+1} \mid \mathbf{z}_i\} \sim \mathsf{Be}(\alpha+\mathbf{z}_i,\theta+\kappa-\mathbf{z}_i)$, where $\mathbf{v}_{i+1}$ is conditionally independent of $(\mathbf{v}_1,\mathbf{z}_1,\ldots,\mathbf{v}_{i-1},\mathbf{z}_{i-1},\mathbf{v}_i)$ given $\mathbf{z}_i$. Note that from the Beta-Binomial conjugate model and since $\mathbf{v}_i \sim \mathsf{Be}(\alpha,\theta)$, we also have $\{\mathbf{v}_i \mid \mathbf{z}_i\} \sim \mathsf{Be}(\alpha+\mathbf{z}_i,\theta+\kappa-\mathbf{z}_i)$. Moreover, for every $n \geq 1$, $(\mathbf{v}_1,\ldots,\mathbf{v}_n)$ are conditionally independent given $(\mathbf{z}_1,\ldots,\mathbf{z}_n)$ and we have $\{\mathbf{v}_1 \mid \mathbf{z}_1,\ldots,\mathbf{z}_n\} \sim \mathsf{Be}(\alpha+\mathbf{z}_1,\theta+\kappa-\mathbf{z}_1)$, and for $2 \leq i \leq n$,

$$\{\mathbf{v}_i \mid \mathbf{z}_1,\ldots,\mathbf{z}_n\} \sim \mathsf{Be}(\alpha+\mathbf{z}_{i-1}+\mathbf{z}_i,\theta+2\kappa-\mathbf{z}_{i-1}-\mathbf{z}_i).$$

It is also easy to see that $(\mathbf{z}_i)_{i\geq 1}$ is an stationary Markov chain itself with initial distribution

$$\mathbb{P}[\mathbf{z}_1 = z] = \binom{\kappa}{z}\frac{(\alpha)_z(\theta)_{\kappa-z}}{(\alpha+\theta)_\kappa},$$

and one-step transition

$$\mathbb{P}[\mathbf{z}_{i+1} = z \mid \mathbf{z}_i] = \binom{\kappa}{z}\frac{(\alpha+\mathbf{z}_i)_z(\theta+\kappa-\mathbf{z}_i)_{\kappa-z}}{(\alpha+\theta+\kappa)_\kappa},$$

for every $z \in \{0, \ldots, \kappa\}$. We this considerations in mind, we can easily compute

$$\begin{aligned}
\mathbb{E}\left[\prod_{j=1}^{n} \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{j=1}^{n} \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j} \middle| \mathbf{z}_1, \ldots, \mathbf{z}_n\right]\right] \\
&= \mathbb{E}\left[\frac{(\alpha+\mathbf{z}_1)_{a_1}(\theta+\kappa-\mathbf{z}_1)_{b_1}}{(\alpha+\theta+\kappa)_{a_1+b_1}} \prod_{i=2}^{n} \frac{(\alpha+\mathbf{z}_{i-1}+\mathbf{z}_i)_{a_i}(\theta+2\kappa-\mathbf{z}_{i-1}-\mathbf{z}_i)_{b_i}}{(\alpha+\theta+2\kappa)_{a_i+b_i}}\right] \\
&= \sum_{z_1=0}^{\kappa} \cdots \sum_{z_n=0}^{\kappa} \left\{\left[\prod_{i=1}^{n} \frac{(\alpha_i+z_i)_{a_i}(\theta_i+\kappa-z_i)_{b_i}}{(\alpha_i+\theta_i+\kappa)_{a_i+b_i}}\right]\left[\prod_{i=1}^{n} \binom{\kappa}{z_i} \frac{(\alpha_i)_{z_i}(\theta_i)_{\kappa-z_i}}{(\alpha_i+\theta_i)_\kappa}\right]\right\} \\
&= \sum_{z_1=0}^{\kappa} \cdots \sum_{z_n=0}^{\kappa} \left\{\prod_{i=1}^{n} \binom{\kappa}{z_i} \frac{(\alpha_i)_{a_i+z_i}(\theta_i)_{b_i+\kappa-z_i}}{(\alpha_i+\theta_i)_{a_i+b_i+\kappa}}\right\}
\end{aligned}$$

where $\alpha_1 = \alpha$, $\theta_1 = \theta$ and for $2 \le i \le k$, $\alpha_i = \alpha + z_{i-1}$ and $\theta_i = \theta + \kappa - z_{i-1}$. □

D.10 Proof of Corollary 4.22

Define the probability kernel $\boldsymbol{\nu}^{(n)} : [0,1] \to [0,1]$ by

$$\boldsymbol{\nu}^{(n)}(v; \cdot) = \mathrm{p}_n \delta_v + (1-\mathrm{p}_n)\nu_0^{(n)}$$

for every $n \ge 1$. From Theorem 4.14, it suffices to show that for every $v_n \to v$ in $[0,1]$ we get $\boldsymbol{\nu}^{(n)}(v_n; \cdot) \xrightarrow{w} \nu_0$, whenever $\mathrm{p}_n \to 0$ and $\boldsymbol{\nu}^{(n)}(v_n; \cdot) \xrightarrow{w} \delta_v$, whenever $\mathrm{p}_n \to 1$, as $n \to \infty$. So fix $v_n \to v$ in $[0,1]$ and a continuous an bounded function $f : [0,1] \to \mathbb{R}$ and note that

$$\boldsymbol{\nu}^{(n)}(v_n; f) = \int f(x) \boldsymbol{\nu}^{(n)}(v_n; dx) = \mathrm{p}_n f(v_n) + (1-\mathrm{p}_n)\nu_0^{(n)}(f).$$

By hypothesis we know $\nu_0^{(n)}(f) \to \nu_0(f)$, hence if $\mathrm{p}_n \to 0$ we clearly get $\boldsymbol{\nu}^{(n)}(v_n; f) \to \nu_0(f)$ which yields $\boldsymbol{\nu}^{(n)}(v_n; \cdot) \xrightarrow{w} \nu_0$. Alternatively if $\mathrm{p}_n \to 1$, since f is continuous, we obtain $\boldsymbol{\nu}^{(n)}(v_n; f) \to f(v) = \delta_v(f)$, which implies $\boldsymbol{\nu}^{(n)}(v_n; \cdot) \xrightarrow{w} \delta_v$. □

D.11 Proof of Proposition 4.24

Define $\boldsymbol{\tau}_0^* = 0$ and for every $j \ge 0$ set $\boldsymbol{\tau}_{j+1}^* = \min\{i > \boldsymbol{\tau}_j^* : \mathbf{v}_i \ne \mathbf{v}_{\boldsymbol{\tau}_j^*}\}$. Also define

$$\boldsymbol{\tau}_j = \begin{cases} \boldsymbol{\tau}_j^* & \text{if } \boldsymbol{\tau}_j^* < k+1 \\ k+1 & \text{if } \boldsymbol{\tau}_j^* \ge k+1 \end{cases}$$

and $m = \min\{j \ge 0 : \boldsymbol{\tau}_j = k+1\}$. Note that $\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{m-1}$ indicate the indexes at which the chain $(\mathbf{v}_i)_{i\ge1}$ changes up to index $i = k$. Then we can easily compute

$$\mathbb{E}\left[\prod_{j=1}^{k} \mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j} \middle| (\boldsymbol{\tau}_0, \ldots, \boldsymbol{\tau}_m)\right] = \prod_{j=0}^{m-1} \left\{\int_{[0,1]} (v)^{\sum_{i\in\mathbf{A}_j} a_i}(1-v)^{\sum_{i\in\mathbf{A}_j} b_i} \nu_0(dv)\right\}$$

where $\mathbf{A}_j = \{\boldsymbol{\tau}_j, \ldots, \boldsymbol{\tau}_{j+1} - 1\}$. Finally note that for any sequence $(\tau_0, \ldots, \tau_m)$ with $\tau_0 = 1 < \tau_1 < \cdots < \tau_m = k+1$,

$$\mathbb{P}[(\boldsymbol{\tau}_0, \ldots, \boldsymbol{\tau}_m) = (\tau_0, \ldots, \tau_m)] = \mathrm{p}^{m-1}(1-\mathrm{p})^{k-m},$$

which yields

$$\mathbb{E}\left[\prod_{j=1}^{k}\mathbf{v}_j^{a_j}(1-\mathbf{v}_j)^{b_j}\right] = \sum_{(\tau_0,\dots,\tau_m)} \mathrm{p}^{m-1}(1-\mathrm{p})^{k-m}\times$$
$$\times\prod_{j=0}^{m-1}\left\{\int_{[0,1]}(v)^{\sum_{i\in A_j}a_i}(1-v)^{\sum_{i\in A_j}b_i}\,\nu_0(dv)\right\},$$

where $A_j=\{\tau_j,\dots,\tau_{j+1}-1\}$, and the sum ranges over all sequences $(\tau_0,\dots,\tau_m)$ with $\tau_0=1<\tau_1<\dots<\tau_m=k+1$. The rest of the proof follows by computing the integrals

$$\int_{[0,1]}(v)^{\sum_{i\in\mathbf{A}_j}a_i}(1-v)^{\sum_{i\in\mathbf{A}_j}b_i}\,\nu_0(dv) = \frac{\Gamma(\alpha+\theta)\Gamma(\alpha+\sum_{i\in A_j}a_i)\Gamma(\theta+\sum_{i\in A_j}b_i)}{\Gamma(\alpha)\Gamma(\theta)\Gamma(\alpha+\theta+\sum_{i\in A_j}(a_i+b_i))},$$

when $\nu_0=\mathsf{Be}(\alpha,\theta)$, and then simplifying. □

D.12 Proof of Theorem 4.25

Let $\left(\mathbf{v}_i^{(\mathrm{p})}\right)_{i\ge1}$ be the non-homogeneous Markov length variables of $\boldsymbol{\mu}^{(\mathrm{p})}$. So that $\left(\mathbf{v}_i^{(0)}\right)_{i\ge1}$ is an independent sequence with $\mathbf{v}_i^{(0)}\sim\nu_i$, and $\left(\mathbf{v}_i^{(1)}\right)_{i\ge1}=\left(\Upsilon^{(j)}(\mathbf{v})\right)_{j\ge0}$, where $\mathbf{v}\sim\nu_1$, $\Upsilon^{(0)}$ denotes the identity function, and for every $j\ge1$, $\Upsilon^{(j)}=\Upsilon_j\circ\dots\circ\Upsilon_1$. Now, from Theorem 3.14 we know $\boldsymbol{\mu}^{(\mathrm{p})}$ is proper if and only if

$$\lim_{j\to\infty}\mathbb{E}\left[\prod_{i=1}^{j}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right]=0.$$

For $\mathrm{p}\in(0,1)$ and $j\ge1$ it is easy to compute

$$\begin{aligned}\mathbb{E}\left[f\left(\mathbf{v}_{j+1}^{(\mathrm{p})}\right)\,\middle|\,\mathbf{v}_1^{(\mathrm{p})},\dots,\mathbf{v}_j^{(\mathrm{p})}\right] &= \mathbb{E}\left[f\left(\mathbf{v}_{j+1}^{(\mathrm{p})}\right)\,\middle|\,\mathbf{v}_j^{(\mathrm{p})}\right]\\ &= \mathrm{p}f\left(\Upsilon_j\left(\mathbf{v}_j^{(\mathrm{p})}\right)\right)+(1-\mathrm{p})\int f(x)\nu_{i+1}(dx) \qquad \text{(D.8)}\\ &= \mathrm{p}f\left(\Upsilon_j\left(\mathbf{v}_j^{(\mathrm{p})}\right)\right)+(1-\mathrm{p})\mathbb{E}\left[f\left(\mathbf{v}_{j+1}^{(0)}\right)\right],\end{aligned}$$

for every measurable and integrable function $f:[0,1]\to\mathbb{R}$. This yields

$$\mathbb{E}\left[\prod_{i=1}^{j+1}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right]=\mathbb{E}\left[\left\{1-\mathrm{p}\Upsilon_j\left(\mathbf{v}_j^{(\mathrm{p})}\right)-(1-\mathrm{p})\mathbb{E}\left[\mathbf{v}_{j+1}^{(0)}\right]\right\}\prod_{i=1}^{j}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right], \qquad \text{(D.9)}$$

which in turn implies

$$0\le\mathbb{E}\left[\prod_{i=1}^{j+1}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right]\le\left\{1-(1-\mathrm{p})\mathbb{E}\left[\mathbf{v}_{j+1}^{(0)}\right]\right\}\mathbb{E}\left[\prod_{i=1}^{j}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right].$$

Inductively, we can prove that for every $j\ge1$

$$0\le\mathbb{E}\left[\prod_{i=1}^{j}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right]\le\prod_{i=1}^{j}\left(1-\left\{(1-\mathrm{p})\mathbb{E}\left[\mathbf{v}_i^{(0)}\right]\right\}\right). \qquad \text{(D.10)}$$

Now, if $\boldsymbol{\mu}^{(0)}$ is proper, by Lemma 3.14 we must have

$$0 = \lim_{j\to\infty} \mathbb{E}\left[\prod_{i=1}^{j}\left(1-\mathbf{v}_i^{(0)}\right)\right] = \lim_{j\to\infty} \prod_{i=1}^{j}\left(1-\mathbb{E}\left[\mathbf{v}_i^{(0)}\right]\right).$$

Since $\nu_i(\{1\}) = 0$ we trivially get $0 \leq \mathbb{E}\left[\mathbf{v}_i^{(0)}\right] < 1$ for every $i \geq 1$, hence $\sum_{i\geq 1}\mathbb{E}\left[\mathbf{v}_i^{(0)}\right] = \infty$. This proves $\sum_{i\geq 1}(1-\mathrm{p})\mathbb{E}\left[\mathbf{v}_i^{(0)}\right] = \infty$ for every $\mathrm{p} \in (0,1)$, which is equivalently to $\lim_{j\to\infty}\prod_{i=1}^{j}\left(1-\left\{(1-\mathrm{p})\mathbb{E}\left[\mathbf{v}_i^{(0)}\right]\right\}\right) = 0$ because $0 \leq (1-\mathrm{p})\mathbb{E}\left[\mathbf{v}_i^{(0)}\right] < 1$. Putting this together with equation (D.10) show that $\lim_{j\to\infty}\mathbb{E}\left[\prod_{i=1}^{j}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right] = 0$ and from Lemma 3.14 we obtain $\boldsymbol{\mu}^{(\mathrm{p})}$ is proper.

On other side, we also have that (D.9) implies

$$0 \leq \mathbb{E}\left[\prod_{i=1}^{j+1}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right] \leq \mathbb{E}\left[\left\{1-\mathrm{p}\Upsilon_j\left(\mathbf{v}_j^{(\mathrm{p})}\right)\right\}\left(1-\mathbf{v}_j^{(\mathrm{p})}\right)\prod_{i=1}^{j-1}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right], \tag{D.11}$$

and from equation (D.8), for the choice $f(x) = (1-p\Upsilon_j(x))(1-x)$, we obtain

$$\begin{aligned}\mathbb{E}&\left[\left\{1-\mathrm{p}\Upsilon_j\left(\mathbf{v}_j^{(\mathrm{p})}\right)\right\}\left(1-\mathbf{v}_j^{(\mathrm{p})}\right)\middle|\,\mathbf{v}_{j-1}^{(\mathrm{p})}\right]\\ &\leq \mathrm{p}\left\{1-\mathrm{p}\Upsilon_j\left(\Upsilon_{j-1}\left(\mathbf{v}_{j-1}^{(\mathrm{p})}\right)\right)\right\}\left\{1-\Upsilon_{j-1}\left(\mathbf{v}_j^{(\mathrm{p})}\right)\right\}\\ &\leq \left\{1-\mathrm{p}\Upsilon_j\left(\Upsilon_{j-1}\left(\mathbf{v}_{j-1}^{(\mathrm{p})}\right)\right)\right\}\left\{1-\mathrm{p}\Upsilon_{j-1}\left(\mathbf{v}_j^{(\mathrm{p})}\right)\right\}.\end{aligned}$$

Inserting the last equation into (D.11) we get

$$\begin{aligned}0 &\leq \mathbb{E}\left[\prod_{i=1}^{j+1}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right]\\ &\leq \mathbb{E}\left[\left\{1-\mathrm{p}\Upsilon_j\left(\Upsilon_{j-1}\left(\mathbf{v}_{j-1}^{(\mathrm{p})}\right)\right)\right\}\left\{1-\mathrm{p}\Upsilon_{j-1}\left(\mathbf{v}_j^{(\mathrm{p})}\right)\right\}\prod_{i=1}^{j-1}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right]\end{aligned}$$

Continuing inductively we can prove that for every $j \geq 1$

$$0 \leq \mathbb{E}\left[\prod_{i=1}^{j}\left(1-\mathbf{v}_i^{(\mathrm{p})}\right)\right] \leq \mathbb{E}\left[\prod_{i=0}^{j-1}\left(1-\mathrm{p}\Upsilon^{(i)}(\mathbf{v})\right)\right], \tag{D.12}$$

where $\mathbf{v} \sim \nu_1$. Now, if $\boldsymbol{\mu}^{(1)}$ is proper, then we have that $\lim_{j\to\infty}\mathbb{E}\left[\prod_{i=0}^{j-1}\left(1-\Upsilon^{(i)}(\mathbf{v})\right)\right] = 0$, which is equivalent to $\lim_{j\to\infty}\prod_{i=0}^{j-1}\left(1-\Upsilon^{(i)}(\mathbf{v})\right) = 0$, almost surely, because the random variables $\Upsilon^{(i)}(\mathbf{v})$'s are positive and bounded by 1. Noting that by hypothesis $0 < \Upsilon^{(i)}(\mathbf{v}) < 1$ almost surely, we get $\sum_{i\geq 0}\Upsilon^{(i)}(\mathbf{v}) = \infty$, which yields $\sum_{i\geq 0}\mathrm{p}\Upsilon^{(i)}(\mathbf{v}) = \infty$, for every $\mathrm{p} \in (0,1)$, this in turn is equivalent to $\lim_{j\to\infty}\prod_{i=0}^{j-1}\left(1-\mathrm{p}\Upsilon^{(i)}(\mathbf{v})\right) = 0$ almost surely. This said, by taking limits as $j \to \infty$ in D.12 we finally obtain that $\boldsymbol{\mu}^{(\mathrm{p})}$ is proper.

□

D.13 Proof of Theorem 4.26

For $\mathrm{p}_n \in (0,1)$ and $m \geq 1$, consider the probability kernel $\boldsymbol{\nu}_m^{(\mathrm{p}_n)} : [0,1] \to [0,1]$, given by

$$\boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v;\cdot) = \mathrm{p}_n \delta_{\Upsilon_m(v)} + (1-\mathrm{p}_n)\nu_{m+1}. \tag{D.13}$$

Fix $v_n \to v \in [0,1]$ and a continuous and bounded function $f : [0,1] \to \mathbb{R}$. Note that

$$\boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v_n; f) = \mathrm{p}_n f(\Upsilon_m(v_n)) + (1-\mathrm{p}_n)\nu_{m+1}(f),$$

hence if $\mathrm{p}_n \to 0$ we get $\boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v_n; f) \to \nu_{m+1}(f)$. Alternatively, since Υ_m and f are continuous, we have that $f(\Upsilon_m(v_n)) \to f(\Upsilon_m(v))$, so if $\mathrm{p}_n \to 1$, $\boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v_n; f) \to f(\Upsilon_m(v))$. This is

$$\boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v_n;\cdot) \xrightarrow{w} \nu_{m+1}, \quad \text{and} \quad \boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v_n;\cdot) \xrightarrow{w} \delta_{\Upsilon_m(v)} \tag{D.14}$$

as $\mathrm{p}_n \to 0$ and $\mathrm{p}_n \to 1$, respectively.

(i) For $\mathrm{p} \in (0,1)$ let us $\left(\mathbf{v}_i^{(\mathrm{p})}\right)_{i\geq 1}$ to the length variables of $\mathbf{W}^{(p)}$. Now fix $\mathrm{p}_n \to 0$ in $[0,1]$, and say that for some $m \geq 1$,

$$\left(\mathbf{v}_1^{(\mathrm{p}_n)}, \dots, \mathbf{v}_m^{(\mathrm{p}_n)}\right) \xrightarrow{d} (\mathbf{v}_1, \dots, \mathbf{v}_m) \tag{D.15}$$

as $n \to \infty$. Since $(\mathbf{v}_i)_{i\geq 1}$ and $\left(\mathbf{v}_i^{(\mathrm{p}_n)}\right)_{i\geq 1}$ are Markov chains, we get $\mathbb{P}\left[\mathbf{v}_{m+1} \in \cdot \,\middle|\, \mathbf{v}_1, \dots, \mathbf{v}_m\right] = \nu_{m+1}$ and $\mathbb{P}\left[\mathbf{v}_{m+1}^{(\mathrm{p}_n)} \in \cdot \,\middle|\, \mathbf{v}_1^{(\mathrm{p}_n)}, \dots, \mathbf{v}_m^{(n)}\right] = \boldsymbol{\nu}_m^{(\mathrm{p}_n)}\left(\mathbf{v}_m^{(n)};\cdot\right)$ for $\boldsymbol{\nu}_m^{(\mathrm{p}_n)}$ as in equation (D.13). By equation D.14 we know that that for $v_n \to v \in [0,1]$, $\boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v_n;\cdot) \xrightarrow{w} \nu_{m+1}$, this together with (D.15) and Lemma D.2 yield

$$\left(\mathbf{v}_1^{(\mathrm{p}_n)}, \dots, \mathbf{v}_{m+1}^{(\mathrm{p}_n)}\right) \xrightarrow{d} (\mathbf{v}_1, \dots, \mathbf{v}_{m+1})$$

as $n \to \infty$. This induction argument together with the assumption $\mathbf{v}_1^{(n)} \overset{d}{=} \mathbf{v}_1$, for every $n \geq 1$, show that (D.15) holds for every $m \geq 1$ and we even obtain $\left(\mathbf{v}_i^{(\mathrm{p}_n)}\right)_{i\geq 1} \xrightarrow{d} (\mathbf{v}_i)_{i\geq 1}$. The rest of the proof follows like that of Theorem 4.3, by noting that the mappings $(v_i)_{i\geq 1} \mapsto \mathsf{SB}[(v_i)_{i\geq 1}]$ and

$$[(w_1, w_2, \dots), (s_1, s_2, \dots)] \mapsto \sum_{j\geq 1} w_j \delta_{s_j}$$

are continuous with respect to the product and weak topologies.

(ii) Fix $\mathrm{p}_n \to 1$, and assume that for some $m \geq 1$,

$$\left(\mathbf{v}_1^{(\mathrm{p}_n)}, \dots, \mathbf{v}_m^{(\mathrm{p}_n)}\right) \xrightarrow{d} \left(\Upsilon^{(0)}(\mathbf{v}), \dots, \Upsilon^{(m-1)}(\mathbf{v})\right) \tag{D.16}$$

as $n \to \infty$. Realize that $\left(\Upsilon^{(i)}\right)_{i\geq 1}$ and $\left(\mathbf{v}_i^{(\mathrm{p}_n)}\right)_{i\geq 1}$ are Markov chains, with $\mathbb{P}\left[\Upsilon^{(m)}(\mathbf{v}) \in \cdot \,\middle|\, \Upsilon^{(0)}(\mathbf{v}), \dots, \Upsilon^{(m-1)}(\mathbf{v})\right] = \delta_{\Upsilon_m(\Upsilon^{(m-1)}(\mathbf{v}))}$ and $\mathbb{P}\left[\mathbf{v}_{m+1}^{(\mathrm{p}_n)} \in \cdot \,\middle|\, \mathbf{v}_1^{(\mathrm{p}_n)}, \dots, \mathbf{v}_m^{(n)}\right] = \boldsymbol{\nu}_m^{(\mathrm{p}_n)}\left(\mathbf{v}_m^{(n)};\cdot\right)$ for $\boldsymbol{\nu}_m^{(\mathrm{p}_n)}$ as in equation (D.13). By equation D.14 we know that that for $v_n \to v \in [0,1]$, $\boldsymbol{\nu}_m^{(\mathrm{p}_n)}(v_n;\cdot) \xrightarrow{w} \delta_{\Upsilon_m(v)}$, this together with (D.15) and Lemma D.2 imply

$$\left(\mathbf{v}_1^{(\mathrm{p}_n)},\dots,\mathbf{v}_{m+1}^{(\mathrm{p}_n)}\right) \xrightarrow{d} \left(\Upsilon^{(0)}(\mathbf{v}),\dots,\Upsilon^{(m)}(\mathbf{v})\right)$$

as $n \to \infty$. The rest of the proof of (ii), with the exception of the ordering of the limiting weights follows identically as that of (i). To check that the limiting weights are decreasingly ordered under the stated conditions, if $\Upsilon_i(v) < v$, for every $v \in [0,1]$ and $i \geq 1$, then for every random variable $\mathbf{v} \sim \nu_1$ we have that $\mathbf{v} \geq \Upsilon^{(1)}(\mathbf{v}) \geq \Upsilon^{(2)}(\mathbf{v}) \geq \cdots$. Further, as $\nu_1(\{0\}) = 0$ and Υ_i maps $(0,1)$ into $(0,1)$ we obtain $0 < \Upsilon^{(i)}(\mathbf{v})$ almost surely. This implies that for $(\mathbf{w}_j)_{j\geq1} = \mathsf{SB}\left[\left(\Upsilon^{(i)}(\mathbf{v})\right)_{i\geq1}\right]$,

$$\mathbf{w}_{j+1} = \frac{\Upsilon^{(j)}(\mathbf{v})\left[1-\Upsilon^{(j-1)}(\mathbf{v})\right]}{\Upsilon^{(j-1)}(\mathbf{v})}\mathbf{w}_j \leq \mathbf{w}_j$$

which shows $(\mathbf{w}_j)_{j\geq1}$ is decreasingly ordered □

D.14 Proof of Theorem 4.29

For every $\alpha,\beta > 0$ the mapping $x \mapsto \mathcal{I}_x(\alpha,\beta)$ is increasing, continuous, and maps $[0,1]$ into $[0,1]$, hence it has an inverse function $\mathcal{I}^{-1}_{(\cdot)}(\alpha,\beta)$ which is also increasing and continuous. Since the composition of increasing and continuous functions is another function of this king we get that

$$x \mapsto \Upsilon_i(x) = \mathcal{I}^{-1}_{\mathcal{I}_x(1-\sigma,\theta+i\sigma)}(1-\sigma,\theta+(i+1)\sigma),$$

is increasing and continuous for every $0 \leq \sigma < 1$, $\theta > -\sigma$, and $i \geq 1$. Trivially $\Upsilon_i(x) = 1$ if and only if $x = 1$ and $\Upsilon_i(x) = 0$ if and only if $x = 0$, so it is clear that Υ_i maps $(0,1)$ into $(0,1)$. This proves (a). (b) and (c) follow immediately by simple composition of the corresponding functions. To prove (d) fix $i \geq 1$ and $\mathbf{v}_i \sim \mathsf{Be}(1-\sigma,\theta+i\sigma)$. Then by (a) and (b) we get

$$\mathbb{P}[\Upsilon_i(\mathbf{v}_i) \leq x] = \mathbb{P}[\mathbf{v}_i \leq \Upsilon_i^{-1}(x)] = \mathcal{I}_{\Upsilon_i^{-1}(x)}(1-\sigma,\theta+i\sigma) = \mathcal{I}_x(1-\sigma,\theta+(i+1)\sigma),$$

which is the distribution function of a $\mathsf{Be}(1-\sigma,\theta+(i+1)\sigma)$ distribution. That is $\Upsilon_i(\mathbf{v}_i) \sim \mathsf{Be}(1-\sigma,\theta+(i+1)\sigma)$. The second statement of (d) follows by a simple induction argument. It remains to prove (e) and (f).

(e) As shown by Karp (2016), for $v \in [0,1]$ and $\alpha > 0$ fixed, the mapping $\beta \mapsto \mathcal{I}_v(\alpha,\beta)$ is log-concave, which implies it is quasi-concave, that is for every $\beta_1,\beta_2 > 0$ and $\lambda \in [0,1]$

$$\mathcal{I}_v(\alpha,\lambda\beta_1+(1-\lambda)\beta_2) \geq \min\{\mathcal{I}_v(\alpha,\beta_1),\mathcal{I}_v(\alpha,\beta_2)\}. \tag{D.17}$$

Further, it is a well-known property of the regularized Beta function that

$$\mathcal{I}_v(\alpha,\beta+1) = \mathcal{I}_v(\alpha,\beta) + \frac{v^\alpha(1-v)^\beta}{\beta\mathcal{B}(\alpha,\beta)} > \mathcal{I}_v(\alpha,\beta), \tag{D.18}$$

where $\mathcal{B}(\alpha,\beta) = \Gamma(\alpha+\beta)/[\Gamma(\alpha)\Gamma(\beta)]$ denotes the Beta function. Hence by (D.17) and (D.18) we obtain that for every $\beta > 0$ and $\varepsilon \in [0,1]$,

$$\mathcal{I}_v(\alpha,\beta+\varepsilon) \geq \min\{\mathcal{I}_v(\alpha,\beta),\mathcal{I}_v(\alpha,\beta+1)\} = \mathcal{I}_v(\alpha,\beta).$$

That is $\beta \mapsto \mathcal{I}_v(\alpha,\beta)$ is monotonically increasing, particularly

$$\mathcal{I}_v(1-\sigma,\theta+i\sigma) \leq \mathcal{I}_v(1-\sigma,\theta+(i+1)\sigma).$$

Finally, since $v \mapsto \mathcal{I}_v^{-1}(\alpha,\beta)$ is increasing, we conclude

$$\Upsilon_i(v) = \mathcal{I}^{-1}_{\mathcal{I}_v(1-\sigma,\theta+i\sigma)}(1-\sigma,\theta+(i+1)\sigma) \leq \mathcal{I}^{-1}_{\mathcal{I}_v(1-\sigma,\theta+(i+1)\sigma)}(1-\sigma,\theta+(i+1)\sigma) = v$$

for every $v \in [0,1]$.

To prove (f) we will require preliminary Lemmas.

Lemma D.4. *Let $0 < \alpha \leq 1$ and $\alpha+\beta > 2$, then $v \mapsto \mathcal{I}_v(\alpha,\beta)$ is concave.*

Proof: It can be easily seen that

$$\frac{\partial^2 \mathcal{I}_v(\alpha,\beta)}{\partial v^2} = \frac{v^{\alpha-2}(1-v)^{\beta-2}}{\mathcal{B}(\alpha,\beta)}\{(\alpha-1)(1-v)-(\beta-1)v\},$$

which is non-positive if and only if $(\alpha-1)/(\beta+\alpha-2) \leq v$. This holds for every $v \in (0,1)$, as $(\alpha-1)/(\beta+\alpha-2) \leq 0$. Thus $v \mapsto \mathcal{I}_v(\alpha,\beta)$ is concave. □

Lemma D.5. *Let $0 < \alpha \leq 1$ and $\beta \geq 1$. Then for every $n \in \{1,2,...\}$ and $v \in (0,1)$*

$$\mathcal{I}^{-1}_{\mathcal{I}_v(\alpha,\beta+n)}(\alpha,\beta+n+1) \geq \frac{n}{n+1}v.$$

Proof: Fix $v \in (0,1)$ and $n \in \{1,2,...\}$. By the mean value theorem we know that there exist u satisfying $nv/(n+1) < u < v$, such that

$$\left.\frac{\partial \mathcal{I}_x(\alpha,\beta+n+1)}{\partial x}\right|_u = \mathcal{I}'_u(\alpha,\beta+n+1) = \frac{\mathcal{I}_v(\alpha,\beta+n+1)-\mathcal{I}_{nv/(n+1)}(\alpha,\beta+n+1)}{v-nv/(n+1)}.$$

By Lemma D.4 we have that $x \mapsto \mathcal{I}_x(\alpha,\beta+n+1)$ is concave, which implies $\mathcal{I}'_u(\alpha,\beta+n+1) \geq \mathcal{I}'_v(\alpha,\beta+n+1)$. That is

$$\frac{\mathcal{I}_v(\alpha,\beta+n+1)-\mathcal{I}_{nv/(n+1)}(\alpha,\beta+n+1)}{v(n+1)^{-1}} \geq \mathcal{I}'_v(\alpha,\beta+n+1) = \frac{v^{\alpha-1}(1-v)^{\beta+n}}{\mathcal{B}(\alpha,\beta+n+1)},$$

where $\mathcal{B}$ denotes the beta function. Evidently, $(\alpha+\beta+n) \geq (n+1)$, thus

$$(\alpha+\beta+n)\left\{\frac{\mathcal{I}_v(\alpha,\beta+n+1)-\mathcal{I}_{nv/(n+1)}(\alpha,\beta+n+1)}{v}\right\} \geq \frac{v^{\alpha-1}(1-v)^{\beta+n}}{\mathcal{B}(\alpha,\beta+n+1)}.$$

Recalling that $\mathcal{B}(a,b+1) = b\mathcal{B}(a,b)/(a+b)$ for $a,b > 0$, the above equation can be written as

$$\mathcal{I}_v(\alpha,\beta+n+1)-\mathcal{I}_{nv/(n+1)}(\alpha,\beta+n+1) \geq \frac{v^{\alpha}(1-v)^{\beta+n}}{(\beta+n)\mathcal{B}(\alpha,\beta+n)}.$$

Further, recalling that $\mathcal{I}_v(a,b+1) = \mathcal{I}_v(a,b) + \{v^a(1-v)^b\}/\{b\mathcal{B}(a,b)\}$, for $a,b > 0$, we obtain

$$\mathcal{I}_v(\alpha,\beta+n) \geq \mathcal{I}_{nv/(n+1)}(\alpha,\beta+n+1).$$

Finally, since the mapping $x \mapsto \mathcal{I}_x^{-1}(\alpha,\beta+n+1)$ is increasing we get

$$\mathcal{I}^{-1}_{\mathcal{I}_v(\alpha,\beta+n)}(\alpha,\beta+n+1) \geq \frac{n}{n+1}v.$$

□

Lemma D.6. *Let* $0 < \alpha \leq 1$ *and* $\beta \geq 1$*. For every* $n \in \{1, 2, ...\}$ *and* $v \in [0, 1]$ *define*

$$\Psi^{(n)}(v) = \mathcal{I}^{-1}_{\mathcal{I}_v(\alpha,\beta)}(\alpha, \beta + n).$$

Then $\sum_{n=1}^{\infty} \Psi^{(n)}(v) = \infty$*, for every* $v \in (0, 1)$*.*

Proof: Let $0 < \alpha \leq 1$ and $\beta \geq 1$. For $j \geq 1$, set

$$\Psi_j(v) = \mathcal{I}^{-1}_{\mathcal{I}_v(\alpha,\beta+j-1)}(\alpha, \beta + j),$$

so that for every $v \in [0, 1]$, $\Psi^{(1)}(v) = \Psi_1(v)$ and $\Psi^{(n)}(v) = (\Psi_n \circ \cdots \circ \Psi_1)(v)$ for $n \geq 2$. Fix $v \in (0, 1)$, we first prove by induction that

$$\Psi^{(n)}(v) \geq \frac{\Psi_1(v)}{n}, \tag{D.19}$$

for every $n \geq 1$. For the inductive base we trivially have that $\Psi^{(1)}(v) = (1/1)\Psi_1(v)$. Now, assume that $\Psi^{(n-1)}(v) \geq \Psi_1(v)/(n-1)$ for some $n \geq 2$. Then as $\Psi^{(n)}$ is increasing, and by Lemma D.5 we obtain

$$\Psi^{(n)}(v) = \Psi_n\{\Psi^{(n-1)}(v)\} \geq \Psi_n\{\Psi_1(v)/(n-1)\} \geq \left(\frac{n-1}{n}\right)\left(\frac{\Psi_1(v)}{n-1}\right) = \frac{\Psi_1(v)}{n}.$$

This proves (D.19) for every $n \geq 1$, hence $\sum_{n=1}^{\infty} \Psi^{(n)}(v) \geq \Psi_1(v) \sum_{n=1}^{\infty} 1/n = \infty$. □

Proof of Theorem 4.29 (f): First note that if $\sigma = 0$, $\Upsilon^{(n)}$ is the identity function and the result is trivial. Otherwise, there exists $m \in \{1, 2, ...\}$ such that $\theta + m\sigma \geq 1$. Set $\alpha = 1 - \sigma$, $\beta = \theta + m\sigma$ and $\hat{\Upsilon}^{(i)}(v) = \mathcal{I}^{-1}_{\mathcal{I}_v(\alpha,\beta)}(\alpha, \beta + i\sigma)$ for every $i \geq 1$ and $v \in [0, 1]$. Fix $v \in (0, 1)$ and define $\hat{v} = \Upsilon^{(m-1)}(v) = \mathcal{I}^{-1}_{\mathcal{I}_v(\alpha,\theta+\sigma)}(\alpha, \theta + m\sigma)$. This way, for every $n > m$, $\Upsilon^{(n)}(v) = \hat{\Upsilon}^{(n-m)}(\hat{v})$.

As the mapping $b \mapsto \mathcal{I}_{\hat{v}}(\alpha, b)$ is monotonically increasing, we have that $b \mapsto \mathcal{I}^{-1}_{\hat{v}}(\alpha, b)$ is monotonically decreasing. Hence, since $\sigma < 1$, we get $\hat{\Upsilon}^{(i)}(\hat{v}) = \mathcal{I}^{-1}_{\mathcal{I}_{\hat{v}}(\alpha,\beta)}(\alpha, \beta + i\sigma) \geq \mathcal{I}^{-1}_{\mathcal{I}_{\hat{v}}(\alpha,\beta)}(\alpha, \beta + i)$ and by Lemma D.6 we obtain $\sum_{i=1}^{\infty} \hat{\Upsilon}^{(i)}(\hat{v}) = \infty$. Finally we note that $\sum_{n=0}^{\infty} \Upsilon^{(n)}(v) \geq \sum_{n=m+1}^{\infty} \Upsilon^{(n)}(v) = \sum_{n=m+1}^{\infty} \hat{\Upsilon}^{(n-m)}(\hat{v})$, from which the result follows. □

References

Aldous, D. J. (1981). Representations of partially exchangeable arrays of random variables, *Journal of Multivariate Analysis* **11**: 581–598.

Aldous, D. J. (1985). Exchangeability and related topics, *École d'été de probabilités de Saint-Flour, XIII—1983*, Vol. 1117 of *Lecture Notes in Math.*, Springer, Berlin, pp. 1–198.

Anderson, P. W. (1972). More is different, *Science* **177**: 393–396.

Barry, D. and Hartigan, J. (1993). A Bayesian analysis for change point problems, *Journal of the American Statistical Association* **88**: 309–319.

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley series in probability and statistics, John Wiley and Sons Inc.

Bissiri, P. and Ongaro, A. (2014). On the topological support of species sampling priors, *Electronic Journal of Statistics* **8**(1): 861–882.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes, *The Annals of Statistics* **1**: 353–355.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I. and Rodríguez, A. (2019). Latent nested nonparametric priors (with discussion), *Bayesian Analysis* **14**(4): 1303–1356.

Castillo, I. (2017). Pólya tree posterior distributions on densities, *Annales de l'Institut Herni Poincaré - Probabilités et Statistiques* **53**(4): 2074–2102.

Daley, D. J. and Vere-Jones, D. (2008). *Introduction to the Theory of Point Processes. Volume II: General Theory and Structure*, Probability and Its Applications, Springer.

Datta, S. (1991). On the consistency of posterior mixtures and its applications, *The Annals of Statistics* **19**: 338–353.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalizations of Dirichlet processes?, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**: 823–858.

De Blasi, P., Martínez, A. F., , Mena, R. H. and Prünster, I. (2020). On inferential implications of decreasing weights structures in mixture models., *Computational Statistics and Data Analysis* **147**(106940).

de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio., *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale* **4**: 251–299.

Diaconis, P. and Freedman, D. (1980). De finetti's theorem for Markov chain, *Annals of Probability* **8**: 115–130.

Escobar, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means, *unpublished Ph.D. thesis, Yale University, Dept. of Statistics* .

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior, *Journal of the American Statistical Association* **89**: 268–277.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**(430): 577–588.

Ewens, W. (1972). The sampling theory of selectively neutral alleles, *Theor. Popul. Biol.* **3**: 87–112.

Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Prünster, I. and Teh, Y. (2016). On the stick-breaking representation for homogeneous NRMIs, *Bayesian Analysis* **11**: 697–724.

Favaro, S., Lijoi, A. and Prünster (2012). On the stick-breaking representation of normalized inverse Gaussian priors, *Biometrika* **99**: 663–674.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, third edn, John Wiley.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems, **1**(2): 209–230.

Fuentes-García, R., Mena, R. H. and Walker, S. G. (2010). A new Bayesian nonparametric mixture model, *Communications in Statistics - Simulation and Computation* **39**(4): 669–682.

Fuentes-García, R., Mena, R. H. and Walker, S. G. (2019). Modal posterior clustering motivated by hopfield's networkl, *CSDA* **137**: 92–100.

Ghosal, S., Ghosh, J. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation, *The Annals of Statistics* **27**: 143–158.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Gil-Leyva, M. F. (2021). Bayesian non-parametric priors based on random sets, *in* D. Hernández, F. Leonardi, R. H. Mena and J. C. a. Pardo (eds), *Special Volume of the XV CLAPEM, Advances in Probability and Mathematical Statistics Series (In press)*, Birkhäuser.

Gil-Leyva, M. F. and Mena, R. H. (2021). Stick-breaking processes with exchangeable length variables, *Submitted manuscript* .

Gil-Leyva, M. F., Mena, R. H. and Nicoleris, T. (2020). Beta-Binomial stick-breaking non-parametric prior, *Electronic Journal of Statistics* **14**: 1479 – 1507.

Gilks, W., Best, N. and Tan, K. (1995). Adaptive Rejection Metropolis Sampling, *Applied Statistics* **44**: 455–472.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles, *Journal of Mathematical Sciences* **138**: 5674–5685.

Hansen, B. and Pitman, J. (2000). Prediction rule for exchangeable sequences related to species sampling, *Statistics & Probability Letters* **46**: 251–256.

Hartigan, J. A. (1990). Partition models, *Communications in Statistics - Theory and Methods* **19**(8): 2745–2756.

Hewitt, E. and Savage, L. (1955). Symmetric measures on Cartesian products, *Trans. Am. Math. Soc.* **80**: 470–501.

Hjort, N., Holmes, C., Müller, P. and Walker, S. G. (2010). *Bayesian Nonparametrics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association* **96**(453): 161–173.

James, L. F., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **36**(1): 76–97.

Kallenberg, O. (1989). On the representation theorem for exchangeable arrays, *Z. Wahrscheinlichkeitstheorie verw.* **30**: 137–154.

Kallenberg, O. (2002). *Foundations of Modern Probability*, second edn, Springer, New York.

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*, first edn, Springer.

Kallenberg, O. (2017). *Random Measures, Theory and Applications*, Vol. 77, first edn, Springer.

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

Karlin, S. and Taylor, H. (1975). *A first couse in Stochastic processes*, second edn, Elsevier.

Karp, D. B. (2016). Normalized incomplete beta function: Log-concavity in parameters and other properties, *Journal of Mathematical Sciences* **217**: 91–107.

Kingman, J. F. (1975). Random discrete distributions, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* pp. 1–22.

Kingman, J. F. (1978a). Random partitions in population genetics, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **361**: 1–20.

Kingman, J. F. (1978b). The representation of partition structures, *Journal of the London Methematical Society* **18**: 374–380.

Kingman, J. F. (1982). The coalescent, *Stochastic Processes and their Applications* **13**: 235–248.

Kingman, J. F. (1993). *Poisson Processes*, Oxford University Press, New York, NY.

Lijoi, A., Mena, R. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse gaussian priors, *Journal of the American Statistical Association* **100**(472): 1278–1291.

MacEachern, S. N. (1994). Estimating Normal means with a conjugate style dirichlet process priors, *Communications in Statistics–Simulations* **23**: 727–741.

MacEachern, S. N. and Müller, P. (1998). Estimating mixtures of Dirichlet process model, *Journal of Computational and Graphical Statistics* **7**: 223–238.

Mena, R. and Walker, S. G. (2012). An eppf from independent sequences of geometric random variables, *Statistics and Probability Letters* **82**: 1059–1066.

Mena, R. and Walker, S. G. (2015). On the Bayesian mixture model and identifiability, *Journal of Computational and Graphical Statistics* **24**: 1155–1169.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics* **9**: 349–265.

Parthasarathy, K. R. (1967). *Probability measures on metric spaces.*, Academic press, New York.

Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions, *Probability Theory and Related Fields* **92**(1): 21–39.

Phadia, E. G. (2016). *Prior processes and their applications (Nonparametric Bayesian estimation)*, Springer Series in Statistics, second edn, Springer International Publishing, Switzerland.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions, *Probability Theory and Related Fields* **102**: 145–158.

Pitman, J. (1996a). Random discrete distributions invariant under size-biased permutation, *Advances in Applied Probability* **28**(2): 525–539.

Pitman, J. (1996b). Some developments of the Blackwell-MacQueen urn scheme, *in* T. F. et al. (ed.), *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, Vol. 30 of *Lecture Notes-Monograph Series*, Institute of Mathematical Statistics, Hayward, California, pp. 245–267.

Pitman, J. (2006). *Combinatorial stochastic processes*, Vol. 1875 of *École d'été de probabilités de Saint-Flour*, first edn, Springer-Verlag Berlin Heidelberg, New York.

Pitman, J. and Yakubovich, Y. (2017). An ergodic theorem for partially exchangeable random partitions, *Electronic Communications in Probability* **22**: 1–10.

Pitman, J. and Yor, M. (1992). Arcsine laws and interval partitions derived from a stable subordinator, *Proceedings of the London Mathematical Society* **s3-65**(2): 326–356.

Prünster, I. (2003). *Random probability measures derived from increasing additive processes and their application to Bayesian statistics*, Ph.D. thesis.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**: 557–574.

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of random measures with independent increments, *The Annals of Statistics* **31**: 560–585.

Ryll-Nardzewski, C. (1957). On stationary sequences of random variables and the de Finetti's equivalence, *Colloquium Mathematicae* **4**: 149–156.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica* **4**: 639–650.

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

Walker, S. G. and Damien, P. (1998). Sampling methods for Bayesian nonparametric inference involving stochastic processes, *in* D. Dey, P. Müller and D. Sinha (eds), *Practical Nonparametric and Semiparametric Bayesian Statistics*, Springer–Verlag, pp. 243–254.

West, M., Müller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with applcations in regression and density estimation, *in* A. F. M. Smith and P. R. Freeman (eds), *A tribute to D. V. Lindley*, Wiley.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in bayesian density estimation, *Electronic Journal of Statistics* **2**: 298–331.

Yamato, H. (1984). Expectations of functions of samples from distributions chosen from dirichlet processes., *Fac. Sci. Kagoshima Univ. Math. Phys. Chem.* **17**: 1–8.