Imperial College London

Department of Mathematics

Bayesian Repulsive Mixtures for Probabilistic Topic Modelling

Jake Hobson

CID: 01716344

Supervised by Riccardo Passeggeri

September 7, 2023

Submitted in partial fulfilment of the requirements for the MSc in Statistics of Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: Jake Hobson Date: September 7, 2023

Abstract

Bayesian mixture models are ubiquitous in modern statistics and are appropriate for a wide range of modelling tasks. When the underlying mixing measure is almost surely discrete, we obtain a clustering behaviour, with each observation being associated with one atom of the mixing measure. Recently, much work has been focused on the case when the atoms of an almost surely discrete mixing measure are allowed to interact. In particular, if the atoms exhibit a repulsive behaviour, then we can expect more well-separated and pronounced clusters. When constructing mixture models with these repulsive properties, the typical approach is to appeal to the vast theory of random measures, which provides many powerful analytical tools for constructing appropriate mixing measures.

In this work, we consider an application of Bayesian repulsive mixture models to the field of probabilistic topic modelling. Probabalistic topic models are designed to model large text-based datasets by uncovering latent themes in the texts. To be specific, this work will introduce a novel topic model, which utilises a Bayesian mixture component with a repulsive mixing measure. We perform Bayesian analysis on the model to investigate its theoretical properties, leveraging several results from the theory of random measures and point processes. These properties show that the novel model is much more expressive than many existing topic models. The drawbacks of the new model, namely the difficulties when performing posterior inference, are also discussed.

Acknowledgements

I would first and foremost like to thank my supervisor, Dr Riccardo Passeggeri, for his invaluable help in navigating what proved to be a challenging but very enjoyable project.

I would also like to thank my parents for their continuous support, both during the completion of this work and at all other times in my life.

I would finally like to thank Mengqi Chen for her unwavering love and encouragement.

Contents

1. Introduction

We live in a world in which large, text-based datasets are ever-more common. As a result, the demand for probabilistic models able to uncover latent structure in such datasets has never been greater. Imagine if, instead of searching through a dataset by searching for individual words or phrases, we could instead infer latent topics that pervade the text and use these to gain an understanding of its structure. We would then be able to search the dataset to find passages on a certain topic.

Probabilistic topic models are a class of generative models used to model a text corpus, specifically designed to uncover any themes that are shared by documents it contains. By a corpus, we mean an unordered collection of documents, and by a document, we mean an unordered collection of words. A word is the most fundamental object in a topic model and the full collection of possible words is assumed to be known a priori. The task of a topic model is to decide by what probabilistic procedure words will be combined to make documents.

One of the most prevalent probabilistic topic models is Latent Dirichlet Allocation (LDA), first introduced in Blei et al. (2003). Under this regime, the model first randomly draws topics, which are distributions over the collection of words - a higher weighting for a given word means that the word is more strongly associated with the topic. Each document is then independently assigned a topic mixture, which is a distribution over the set of all topics - a higher weighting for a given topic means that the topic is more strongly associated with the document. Documents are then generated by a process depending on both their topic mixture and the set of topics.

The use of topics and topic mixtures as latent variables is at the core of most modern probabilistic topic models. However, there is much scope to vary how these latent variables are generated. Many models assume that the topic mixtures for each document are generated independently, which may not be appropriate in some settings. This motivates us to investigate a different approach in this work. We assume that the latent topic mixtures arise from a Bayesian mixture model with an almost surely discrete mixing measure. Moreover, we assume that the atoms of the mixing measure are repulsive, which we model using a construction originally described in Beraha et al. (2023). The hope is that this will lead to the documents being divided into well-defined clusters based on the topics that they are written about. Since the mixing measure is itself random in this Bayesian setting, it belongs to the class of random measures. The theoretical tools derived from the theory of random measures will prove integral to understanding the properties of our new model.

A random measure can be thought of as a random element in the space of all measures on a given measurable space. Since a measurable space in general admits a vast variety of measures on it, the class of random measures is very rich. A typical example that is often used as a mixing measure is the Dirichlet process, which has i.i.d. atom locations. In order to perform inference under our Bayesian mixture model, we will need some notion of conditioning the distribution of a random measure on the observed data. This notion is provided by Palm theory, which is given a thorough treatment in this work.

This work has three main objectives and each of the next three chapters are assigned to one of these objectives:

In Chapter 2, we aim to provide a more approachable introduction to the theory of random measures. The contents of many of the standard references on the subject, in particular Kallenberg (2017), are quite inaccessible. Palm theory in particular, on which later parts of this work will depend heavily, demands a careful exposition. The objective of this chapter is therefore to present the most important results from the field in a more gentle and concise way.

In Chapter 3, we give a survey of some probabilistic topic models and their properties. This includes their specification, hyperparemeter choices, and posterior inference procedures. We delve in particular into LDA, a few of its more inventive generalisations, and an alternative nonparametric approach based on hierarchies of Dirichlet processes. Our objective here is to highlight the existing topic models that are most comparable with our novel topic model.

In Chapter 4, we introduce a completely novel topic model, incorporating a Bayesian repulsive mixture component. This topic model is far removed from those discussed in the previous chapter, since it chooses to leverage the more contemporary framework of Beraha et al. (2023) to enhance the generative procedure for the latent topic mixtures. It is far more flexible than many of its predecessors, and as a result its theoretical behaviours are drastically different. Our objective in this chapter is to present the new model and investigate some of these theoretical behaviours.

2. Theory of Random Measures

Kallenberg (2017) provides a complete but difficult reference on random measures, the first six sections of which form the basis of this chapter. We will also include some results from other textbooks, namely Daley & Vere-Jones (2008) and Baccelli et al. (2020). This more approachable introduction to the theory should leave the reader with a more than sufficient understanding for this work and the references herein.

2.1. Introduction to Random Measures

2.1.1. Definitions

A *Borel space* is a measurable space $(S, \mathcal{S})$ which bijects bi-measurably with some Borel set $B \subset \mathbb{R}$. A *localising ring* on such a space S is a system $\hat{\mathcal{S}} \subset \mathcal{S}$ such that:

1. $\hat{\mathcal{S}}$ is a sub-ring of $\mathcal{S}$,
2. If $B \in \hat{\mathcal{S}}$ and $C \in \mathcal{S}$, then $B \cap C \in \hat{\mathcal{S}}$,
3. There exists a sequence (S_n) in $\hat{\mathcal{S}}$ such that $S_n \uparrow S$ and, for all $B \in \mathcal{S}$, we have $B \in \hat{\mathcal{S}}$ if and only if $B \subset S_n$ for some n.

We call the triple $(S, \mathcal{S}, \hat{\mathcal{S}})$ a *localised Borel space.* Let S be such a space unless otherwise stated. The localising ring allows us to introduce a notion of boundedness on S: a set $B \in \mathcal{S}$ is (i) *bounded* if $B \in \hat{\mathcal{S}}$ and (ii) *locally finite* if $|B \cap M| < \infty$ for all $M \in \hat{\mathcal{S}}$.

Example 2.1.1. When S is a separable and complete metric space, we can take $\mathcal{S}$ to its Borel sets and $\hat{\mathcal{S}}$ to be its metrically bounded Borel sets.

A measure μ on S is *locally finite* if $\mu(B) < \infty$ for all $B \in \hat{\mathcal{S}}$. We write $\mathbb{M}_S$ for the set of all locally finite measures on S.

A *random measure* on S is a map $\xi : \Omega \times \mathcal{S} \to [0, \infty]$ for some arbitrary probability space $(\Omega, \mathcal{A}, \mathbb{P})$ such that:

1. $\xi(\,\cdot\,, B)$ is $\mathcal{A}$-measurable for all $B \in \mathcal{S}$,
2. $\xi(\omega, \,\cdot\,)$ is a locally finite measure on S for all $w \in \Omega$.

We can view a random measure ξ on S as a random element $\xi : \Omega \to \mathbb{M}_S$. This motivates us to introduce some additional structure on $\mathbb{M}_S$. For each $B \in \mathcal{S}$, define the projection map $\pi_B : \mathbb{M}_S \to [0, \infty]$ via $\pi_B(\mu) = \mu(B)$ for $\mu \in \mathbb{M}_S$. Now let

$$\mathcal{M}_S = \sigma\left(\{\pi_B : B \in \mathcal{S}\}\right) .$$

This is a σ-algebra on $\mathbb{M}_S$. Equipping $\mathbb{M}_S$ with such a σ-algebra allows it to enjoy two important properties, which are highlighted by the following results.

Theorem 2.1.2. The measurable space $(\mathbb{M}_S, \mathcal{M}_S)$ is a Borel space.

Proof. See Kallenberg (2017), Theorem 1.5. □

Theorem 2.1.3. Consider a map $\xi : \Omega \times \mathcal{S} \to [0, \infty]$. Then the following are equivalent:

1. ξ is a random measure on S,
2. ξ is a random variable on $(\mathbb{M}_S, \mathcal{M}_S)$.

Proof. See Kallenberg (2017), Lemma 1.14. □

2.1.2. Classes of random measure

A measure μ on S is *integer-valued* if $\mu(B) \in \mathbb{Z}_{\geqslant 0} \cup \{+\infty\}$ for all $B \in \mathcal{S}$ (by $\mathbb{Z}_{\geqslant 0}$ we mean the set of non-negative integers). We write $\mathbb{N}_S$ for the set of integer-valued locally finite measures on S. A *point process* on S is a random measure ξ on S such that $\xi \in \mathbb{N}_S$ a.s..

Remark 2.1.4. Given that S is Borel, one shows that $\{s\} \in \mathcal{S}$ for all $s \in S$[1]. This fact will be integral when making the following definitions.

A measure μ on S is *simple* if it is integer-valued and has $\mu(\{s\}) \leqslant 1$ for all $s \in S$. We write $\mathbb{N}^*_S$ for the set of simple locally finite measures on S. A point process ξ on S is *simple* if $\xi \in \mathbb{N}^*_S$ a.s..

A measure μ on S is *diffuse* if it is non-atomic, that is $\mu\left(\{s\}\right) = 0$ for all $s \in S$. We write $\mathbb{M}^*_S$ for the set of all diffuse locally finite measures on S. A random measure ξ on S is *diffuse* if $\xi \in \mathbb{M}^*_S$ a.s..

To interpret these classes better, we can appeal to a result from measure theory. There exists a decomposition

$$\mu = \alpha + \sum_{k=1}^{\kappa} \beta_k \delta_{\sigma_k} \, , \qquad \mu \in \mathbb{M}_S \, , \tag{2.1}$$

where $\alpha \in \mathbb{M}^*_S$, $\kappa \in \mathbb{Z}_{\geqslant 0} \cup \{+\infty\}$, $\sigma_1, \ldots, \sigma_\kappa \in S$ distinct, and $\beta_1, \ldots, \beta_\kappa \in (0, \infty)$ are all measurable functions of μ. See, for example, Kallenberg (2017), Lemma 1.6. We can easily extend this decomposition to random measures.

Theorem 2.1.5. Let ξ be a random measure on S. Then ξ admits a decomposition of the form

$$\xi = \alpha + \sum_{k=1}^{\kappa} \beta_k \delta_{\sigma_k} \quad \text{a.s.} \, , \tag{2.2}$$

where α is a diffuse random measure on S and:

1. κ is a random variable in $\mathbb{Z}_{\geqslant 0} \cup \{+\infty\}$,
2. $\sigma_1, \ldots, \sigma_\kappa$ are (conditional on κ) a.s. distinct random variable in S,
3. $\beta_1, \ldots, \beta_\kappa$ are (conditional on κ) random variables in $(0, \infty)$.

Proof. See Daley & Vere-Jones (2008), Proposition 9.3.IV.. □

A diffuse random measure on S has $\kappa = 0$ a.s. in any decomposition of the form in (2.2). A point process on S has $\alpha \equiv 0$ and $\beta_1, \ldots, \beta_\kappa \in \mathbb{Z}_{>0}$ a.s. (by $\mathbb{Z}_{>0}$ we mean the set of positive integers). Moreover, if such a point process is simple, then we further have $\beta_1, \ldots, \beta_\kappa = 1$ a.s..

[1] There exists a bi-measurable bijection $f : B \to S$ for some Borel set $B \subset \mathbb{R}$. Thus, given $s \in S$, we have $\{s\} = f\left(f^{-1}(\{s\})\right) = f(\{x\})$, where $x = f^{-1}(s) \in B$. The $\mathcal{S}$-measurability of $\{s\}$ follows from the Borel measurability of $\{x\}$.

2.1.3. Moment measures

The *intensity measure* of a random measure ξ on S is the measure $\mathbb{E}\xi$ on S defined by $\mathbb{E}\xi(B) = \mathbb{E}\left(\xi\left(B\right)\right)$ for $B \in \mathcal{S}$. This definition is justified by the following result.

Lemma 2.1.6. Let ξ be a random measure on S. Then there exists an a.s. unique measure $\mathbb{E}\xi$ on S such that $\mathbb{E}\xi(B) = \mathbb{E}\left(\xi\left(B\right)\right)$ for all $B \in \mathcal{S}$. Moreover, $\mathbb{E}\xi$ is s-finite.

Proof. See Kallenberg (2017), Lemma 2.4. □

For localised Borel spaces $(S, \mathcal{S}, \hat{\mathcal{S}})$ and $(T, \mathcal{T}, \hat{\mathcal{T}})$, the product space $(S \times T, \mathcal{S} \otimes \mathcal{T})$ is also Borel. Moreover, $\hat{\mathcal{S}}$ and $\hat{\mathcal{T}}$ induce a localising ring in $S \times T$: a set $B \in \mathcal{S} \otimes \mathcal{T}$ is bounded if and only if it has bounded projections on S and T. Equipping $S \times T$ with this localising ring makes it a localised Borel space on which we can define random measures. Repeating this process iteratively, given $n \in \mathbb{N}$ and localised Borel spaces $S_1, \ldots, S_n$, we can construct a localised Borel space $\bigtimes_{k=1}^{n} S_k$.

Let ξ be a point process on S and fix $n \in \mathbb{N}$. Then its (random) product measure ξ^n is itself a random measure on S^n. Hence we can define the *n'th order moment measure* of ξ as the intensity $\mathbb{E}\xi^n$ of ξ^n.

Let ξ be a point process on S and fix $n \in \mathbb{N}$. Then $\xi = \sum_{i \in I} \delta_{\sigma_i}$ for some set $I = \{1, 2, \ldots, \kappa\}$ with κ a random variable in $\mathbb{Z}_{\geqslant 0} \cup \{+\infty\}$ and $(\sigma_i)_{i \in I}$ a collection of random variables in S, as per Theorem 2.1.5. The associated *factorial process* on S^n is

$$\xi^{(n)} = \sum_{i \in I^{(n)}} \delta_{\sigma_{i_1}, \ldots, \sigma_{i_n}} ,$$

where $I^{(n)}$ is the non-diagonal part of I^n. One shows that this is well defined with respect to the choice of decomposition of ξ[2]. Moreover, $\xi^{(n)}$ is a random measure on S^n. The *n'th order factorial moment measure* of ξ is the intensity $\mathbb{E}\xi^{(n)}$ of $\xi^{(n)}$.

We can already learn a lot about the distribution of a random measure from only its second order (factorial) moment measure.

Lemma 2.1.7. Let D denote the diagonal part of S^2. Then:

1. A random measure ξ on S is diffuse if and only if $\mathbb{E}\xi^2(D) = 0$,
2. A point process ξ on S is simple if and only if $\mathbb{E}\xi^{(2)}(D) = 0$.

Proof. See Kallenberg (2017), Lemma 2.7. □

Let ξ be a random measure on S and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$. The *conditional intensity* of ξ given $\mathcal{F}$ is the $\mathcal{F}$-measurable process $\mathbb{E}(\xi|\mathcal{F})$ on $\mathcal{S}$ such that $\mathbb{E}(\xi|\mathcal{F})(B) = \mathbb{E}\left(\xi(B)|\mathcal{F}\right)$ a.s. for all $B \in \hat{\mathcal{S}}$. This definition is justified by the following result.

Lemma 2.1.8. Let ξ be a random measure on S and let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{A}$. Then there exists an a.s. unique $\mathcal{F}$-measurable process $\mathbb{E}(\xi|\mathcal{F})$ on $\mathcal{S}$ such that

$$\mathbb{E}(\xi|\mathcal{F})(B) = \mathbb{E}\left(\xi(B)|\mathcal{F}\right) \text{ a.s. for all } B \in \hat{\mathcal{S}} .$$

If $\mathbb{E}(\xi|\mathcal{F})$ is locally finite, then it is an $\mathcal{F}$-measurable random measure on S.

Proof. See Kallenberg (2017), Lemma 2.10. □

[2] For each $\omega \in \Omega$, $\kappa(\omega) = \xi(\omega, S)$ is unique and the $(\sigma_i(\omega))_{i \in I(\omega)}$ are unique up to permutations. Such permutations have no effect on the form in (2.2) due to symmetry.

2.1.4. Uniqueness and extension results

Fix a collection $\mathcal{I} \subset \mathcal{S}$. Then $\mathcal{I}$ is a *generating ring* (respectively *generating semiring*) if it is a sub-ring (respectively sub-semiring) of $\mathcal{S}$ and generates $\mathcal{S}$. Also, we write $\hat{\mathcal{I}}_+$ for the set of simple, $\mathcal{I}$-measurable functions $f \geqslant 0$ on S. These definitions allow us to establish a basic uniqueness criteria on the distribution of random measures on S.

Theorem 2.1.9. Let ξ and η be random measures on S. Fix a generating semiring $\mathcal{I} \subset \hat{\mathcal{S}}$. Then $\xi \overset{\mathrm{d}}{=} \eta$ if and only if

$$(\xi(I_1), \ldots, \xi(I_n)) \overset{\mathrm{d}}{=} (\eta(I_1), \ldots, \eta(I_n)) \text{ for all } n \in \mathbb{N},\ I_1, \ldots, I_n \in \mathcal{I}.$$

Proof. See Daley & Vere-Jones (2008), Proposition 9.2.III. □

Let ξ be a random measure on S. Write $\mathcal{F}_+(S)$ for the space of measurable functions $f : S \to [0, \infty]$. Then, given $f \in \mathcal{F}_+(S)$, the integral $\xi(f) = \int_S f(s)\ \xi(\mathrm{d}s)$ is a random variable in $\mathbb{R}$. We can hence define the *Laplace functional* of ξ be the functional $\mathcal{L}_\xi : \mathcal{F}_+(S) \to \mathbb{R}$ defined by

$$\mathcal{L}_\xi(f) = \mathbb{E}\left(e^{-\xi(f)}\right).$$

This is analogous to the moment generating function of a random variable.

Theorem 2.1.10. Let ξ and η be random measures on S. Fix a generating semiring $\mathcal{I} \subset \hat{\mathcal{S}}$. Then $\xi \overset{\mathrm{d}}{=} \eta$ if and only if

$$\mathcal{L}_\xi(f) = \mathcal{L}_\eta(f) \text{ for all } f \in \hat{\mathcal{I}}_+ .$$

Proof. See Kallenberg (2017), Corollary 2.3. □

The *avoidance probability function* of a random measure ξ on S is the set function $\nu_\xi : \mathcal{S} \to [0, 1]$ defined by

$$\nu_\xi(B) = \mathbb{P}(\xi(B) = 0), \quad B \in \mathcal{S}.$$

For simple point processes, we can recast these uniqueness results in terms of vectors of avoidance probabilities on generating sets.

Theorem 2.1.11. Let ξ and η be simple point processes on S. Fix a generating ring $\mathcal{U} \subset \hat{\mathcal{S}}$. Then $\xi \overset{\mathrm{d}}{=} \eta$ if and only if

$$\nu_\xi(U) = \nu_\eta(U) \text{ for all } U \in \mathcal{U}.$$

Proof. See Kallenberg (2017), Theorem 2.2. □

Random measures are often constructed by extension of a non-negative set function defined on an underlying generating ring, just as with Carathédory's theorem.

Theorem 2.1.12. Fix a generating ring $\mathcal{U} \subset \hat{\mathcal{S}}$. Let $\eta \geqslant 0$ be a process on $\mathcal{U}$. Then there exists a random measure ξ on S with $\xi(U) = \eta(U)$ a.s. for all $U \in \mathcal{U}$ if and only if

1. For all $A, B \in \mathcal{U}$ disjoint, we have $\eta(A \cup B) = \eta(A) + \eta(B)$,
2. For all sequences (A_n) in $\mathcal{U}$ with $A_n \downarrow \varnothing$, we have $\eta(A_n) \xrightarrow{\mathbb{P}} 0$.

In this case, ξ is a.s. unique.

Proof. See Kallenberg (2017), Theorem 2.15. □

Under an additional tightness condition, we can also construct a simple point process by extension of a vector of avoidance probabilities on an underlying generating ring. This extension is only unique up to distribution, however. For a collection $\mathcal{I} \subset \mathcal{S}$ and given $A \in \mathcal{I}$, we write $\mathcal{P}_A^{(\mathcal{I})}$ for the set of finite partitions π of A into sets $B \in \mathcal{I}$.

Theorem 2.1.13. Fix a generating ring $\mathcal{U} \subset \hat{\mathcal{S}}$. Let η be a $\{0,1\}$-valued process on $\mathcal{U}$. Then there exists a random measure ξ on S with $\nu_\xi(U) = \nu_\eta(U)$ for all $U \in \mathcal{U}$ if and only if

1. For all $A, B \in \mathcal{U}$ disjoint, we have $\eta(A \cup B) = \eta(A) \vee \eta(B)$,
2. For all sequences (A_n) in $\mathcal{U}$ with $A_n \downarrow \varnothing$, we have $\eta(A_n) \xrightarrow{\mathbb{P}} 0$,
3. For all $A \in \mathcal{U}$, $\left\{\sum_{B\in\pi} \eta(B) : \pi \in \mathcal{P}_A^{(\mathcal{U})}\right\}$ is tight.

In this case, the distribution of ξ is unique.

Proof. See Kallenberg (2017), Theorem 2.18 and Corollary 2.21. □

2.1.5. Absolute continuity and differentiation

Recall that two measures μ and ν on S are equivalent if $\mu(A) = 0$ if and only if $\nu(A) = 0$ for $A \in \mathcal{S}$. A *supporting measure* of a random measure ξ on S is a (deterministic) measure ν on S which is equivalent to ξ a.s..

Lemma 2.1.14. Let ξ be a random measure on S. Then a supporting measure of ξ exists and can be chosen to be bounded.

Proof. See Kallenberg (2017), Corollary 2.5. □

A *fixed atom* of a random measure ξ on S is a point $s \in S$ with $\mathbb{P}(\xi(\{s\}) > 0) > 0$. A random measure ξ on S must have countably many fixed atoms. Indeed, let ν be a supporting measure of ξ. Then, for $s \in S$,

$$\mathbb{P}(\xi(\{s\}) > 0) > 0 \iff \mathbb{P}(\xi(\{s\}) = 0) < 1 \iff \nu(\{s\}) > 0\,,$$

since ν is equivalent to ξ a.s.. Hence any fixed atom of ξ is an atom of ν, and the claim follows since ν has countably many atoms (see the decomposition in (2.1)).

Let η be a random measure on S. Let $X \geqslant 0$ be a measurable process on S. Then the integral $\xi(X) = \int_S X(s)\, \eta(\mathrm{d}s)$ is a random variable on $\mathbb{R}$. Additionally, by $X \cdot \eta$ we mean the process on $\mathcal{S}$ given by

$$(X \cdot \eta)(B) = \eta(X \mathbb{1}_B)\,, \quad B \in \mathcal{S}\,. \tag{2.3}$$

There is more we can say about such a process.

Lemma 2.1.15. Let η be a random measure on S and $X \geqslant 0$ be a measurable process on S. Then the process $\xi = X \cdot \eta$ exists and is measurable on $\hat{\mathcal{S}}$. Moreover, when ξ is locally finite, it is also a random measure on S.

Proof. See Kallenberg (2017), Lemma 2.10. □

Surprisingly, knowledge of the joint distribution of (X, η) in (2.3) does not uniquely determine the distribution of $X \cdot \eta$ in general. We can guarantee this uniqueness under an additional condition, however.

Theorem 2.1.16. Let ξ and η be random measures on S and let $X, Y \geqslant 0$ be measurable processes on S. Suppose that $(\xi, X) \overset{\mathrm{d}}{=} (\eta, Y)$. If $\xi \ll \mathbb{E}\xi$ a.s., then

$$(\xi, X, X \cdot \xi) \overset{\mathrm{d}}{=} (\eta, X, X \cdot \eta) .$$

Proof. See Kallenberg (2017), Theorem 2.11. □

We return briefly to a result involving intensity measures, which can be thought of as a generalisation of the celebrated Campbell's theorem.

Theorem 2.1.17. Let η be a random measure on S and $f \geqslant 0$ be a measurable function on S. Suppose that $f \cdot \xi$ is locally finite, such that it defines a random measure. Then

$$\mathbb{E}(f \cdot \xi) = f \cdot \mathbb{E}\xi .$$

Proof. See Kallenberg (2017), Lemma 2.4. □

The *density* of a random measure ξ on S with respect to another random measure η on S with $\xi \ll \eta$ a.s. is the measurable process $X \geqslant 0$ on S such that $\xi = X \cdot \eta$ a.s.. This definition is justified by the following result.

Lemma 2.1.18. Let ξ and η be random measures on S with $\xi \ll \eta$ a.s.. Then there exists an a.s. unique measurable process $X \geqslant 0$ on S such that $\xi = X \cdot \eta$ a.s..

Proof. See Kallenberg (2017), Lemma 2.10. □

We show finally that the relation $\xi \ll \eta$ a.s. for random measures ξ and η on S can be extended to include certain conditional intensities.

Theorem 2.1.19. Let ξ and η be random measures on S with $\xi \ll \eta$ a.s.. Then, for any sub-σ-algebra $\mathcal{F}$ of $\mathcal{A}$ with $\mathcal{F} \supset \sigma(\eta)$,

$$\xi \ll \mathbb{E}(\xi|\mathcal{F}) \ll \eta \text{ a.s.} .$$

Proof. See Kallenberg (2017), Theorem 2.12. □

2.2. Independence of Increments and Infinite Divisibility

2.2.1. Poisson and related processes

Let ξ be a point process on S. Then ξ is *Poisson* with intensity $\mu \in \mathbb{M}_S$ if, for all $n \in \mathbb{N}$ and $B_1, \ldots, B_n \in \hat{\mathcal{S}}$ disjoint, we have $\xi(B_i) \sim \text{Poisson}(\mu(B_i))$ independently for each $i = 1, \ldots, n$. Generalising this, we say that a ξ is a *Cox process* based on a random measure η on $\mathcal{S}$ if $\xi|\eta$ is conditionally Poisson with intensity η.

Theorem 2.2.1. For any random measure η on S, there exists a Cox process ξ directed by η. In particular, for all $\mu \in \mathbb{M}_S$, there exists a Poisson process on S with intensity μ.

Proof. See Kallenberg (2017), Theorem 3.5. □

Via the Cox transformation, we can establish a one-to-one correspondence between simple point processes on S and diffuse random measures on S.

Theorem 2.2.2. Let ξ be a Cox process directed by a random measure η on S. Then:

1. ξ is simple if and only if η is diffuse,
2. The distributions of ξ and η determine each other uniquely.

Proof. See Kallenberg (2017), Theorem 3.3; Lemma 3.6. □

We can use this new-found correspondence to extend the uniqueness results of the previous chapter in certain special cases of interest.

Theorem 2.2.3. Let ξ and η be either (i) simple point processes on S or (ii) diffuse random measures on S. Fix a generating ring $\mathcal{U} \subset \hat{\mathcal{S}}$ and a constant $c > 0$. Then $\xi \overset{\mathrm{d}}{=} \eta$ if and only if

$$\mathbb{E}\left(e^{-c\xi(U)}\right) = \mathbb{E}\left(e^{-c\eta(U)}\right) \text{ for all } U \in \mathcal{U}\,.$$

Proof. See Kallenberg (2017), Theorem 3.8. □

Theorem 2.2.4. Let η be a random measure on S. Let ξ be either (i) a simple point process on S or (ii) a diffuse random measure on S. Fix a generating ring $\mathcal{U} \subset \hat{\mathcal{S}}$. Then $\xi \overset{\mathrm{d}}{=} \eta$ if and only if

$$\xi(U) \overset{\mathrm{d}}{=} \eta(U) \text{ for all } U \in \mathcal{U}\,.$$

Proof. See Kallenberg (2017), Theorem 3.8. □

The second result permits us to require equality in distribution of only one-dimensional distributions. By contrast, Theorem 2.1.9 required equality in distribution of all finite-dimensional distributions.

2.2.2. Independence of increments

A random measure ξ on S has *pairwise independent increments* if, for all $A, B \in \mathcal{S}$ disjoint, we have $\xi(A)$ and $\xi(B)$ independent. It is easy to characterise simple point processes and diffuse random measures with pairwise independent increments.

Theorem 2.2.5.

1. Let ξ be a simple point process on S with no fixed atoms. Then ξ has pairwise independent increments if and only if it is Poisson.
2. Let ξ be a diffuse random measure on S. Then ξ has pairwise independent increments if and only if it is a.s. non-random.

Proof. See Kallenberg (2017), Theorem 3.17. □

A *completely random measure* on S is a random measure ξ on S such that for all $n \in \mathbb{N}$ and all $B_1, \ldots, B_n \in \mathcal{S}$ disjoint, we have $\{\xi(B_1), \ldots, \xi(B_n)\}$ mutually independent. Henceforth, we will use the terms "completely random measure on S" and "random measure on S with independent increments" interchangeably.

Example 2.2.6. Let ξ be a completely random measure on S. Then ξ has pairwise independent increments, so:

1. If ξ is a simple point process with no fixed atoms, then it is necessarily Poisson,
2. If ξ is diffuse, then it is necessarily a.s. non-random.

Through the following representation theorem, we obtain a decomposition of any completely random measure in terms of a Poisson integral.

Theorem 2.2.7. Let ξ be a completely random measure on S. Then

$$\xi = \nu + \sum_{k=1}^{\kappa} \beta_k \delta_{s_k} + \int_0^\infty y \, N(\,\cdot \times \mathrm{d}y) \text{ a.s.}\,, \tag{2.4}$$

where $\nu \in \mathbb{M}_S^*$, N is a Poisson process on $S \times (0, \infty)$ (not necessarily equipped with the standard product localising ring), and:

1. κ is a random variable in $\mathbb{Z}_{\geqslant 0} \cup \{+\infty\}$,
2. $s_1, \ldots, s_\kappa$ are (conditional on κ) fixed points in S,
3. $\beta_1, \ldots, \beta_\kappa$ are (conditional on κ) independent random variables in $(0, \infty)$, independent of N.

Moreover, the three random measures constituting each term in (2.4) are a.s. unique.

Proof. See Daley & Vere-Jones (2008), Theorem 10.1.III. □

2.2.3. Infinite divisibility

A random measure ξ on S is *infinitely divisible* if for all $n \in \mathbb{N}$ we can write

$$\xi \overset{\mathrm{d}}{=} \sum_{j=1}^{n} \xi_{nj}\,, \tag{2.5}$$

where $\xi_{n1}, \ldots, \xi_{nn}$ are i.i.d. random measures on S. A point process ξ on S is infinitely divisible as a point process if for all $n \in \mathbb{N}$ we can write (2.5) with $\xi_{n1}, \ldots, \xi_{nn}$ are i.i.d. point processes on S. When we say that a point process is infinitely divisible, we take this to mean that it is infinitely divisible as a point process.

Theorem 2.2.8. Let ξ be a random measure on S. Fix a generating semiring $\mathcal{I} \subset \hat{\mathcal{S}}$. Then:

1. ξ is infinitely divisible if and only if $(\xi(I_1), \ldots, \xi(I_n))$ is infinitely divisible (as a random variable) for all $n \in \mathbb{N}$ and all $I_1, \ldots, I_n \in \mathcal{I}$,
2. If ξ is a point process, then it is infinitely divisible if and only if $\xi(f)$ is infinitely divisible for all $f \in \hat{\mathcal{I}}_+$.

Proof. See Kallenberg (2017), Theorem 3.24. □

Let $\mathbb{M}_S'$ denote $\mathbb{M}_S$ without the zero measure and let $\mathcal{M}_S'$ be the natural σ-algebra on $\mathbb{M}_S'$ as a subset of $\mathbb{M}_S$. Define analogously $\mathbb{N}_S'$ for $\mathbb{N}_S$. We can characterise the distribution of any infinitely divisible random measure ξ on S uniquely via a measure $\alpha \in \mathbb{M}_S$ and its *Lévy measure* λ on $\mathbb{M}_S'$. We say that ξ is the infinitely divisible process *directed by* (α, λ). The following result makes this more concrete.

Theorem 2.2.9. Let ξ be an infinitely divisible random measure on S. Then

$$-\log \mathcal{L}_\xi(f) = \alpha(f) + \int_{\mathbb{M}'_S} \left[1 - e^{-\mu(f)}\right] \lambda(\mathrm{d}\mu), \quad f \in \mathcal{F}_+(S),$$

where $\alpha \in \mathbb{M}_S$ and λ is a measure on $\mathbb{M}'_S$ satisfying

$$\int_{\mathbb{M}'_S} (\mu(B) \wedge 1)\, \lambda(\mathrm{d}\mu) < \infty \text{ for all } B \in \hat{\mathcal{S}}.$$

Moreover, the pair (α, λ) is unique and any α and λ with the stated properties may occur. For point processes, the result remains valid with $\alpha \equiv 0$ and λ restricted to $\mathbb{N}'_S$.

Proof. See Kallenberg (2017), Theorem 3.20[3]. □

This result has some immediate consequences. We call a measure $\mu \in \mathbb{M}'_S$ *degenerate* if $\mu = r\delta_s$ for some $r > 0$ and $s \in S$.

Corollary 2.2.10. Let η be an infinitely divisible random measure on S. Then any Cox process directed by η is also infinitely divisible.

Proof. See Kallenberg (2017), Corollary 3.22. □

Corollary 2.2.11. Let ξ be an infinitely divisible random measure on S with Lévy measure λ. Then ξ has independent increments if and only if λ is supported by the set of all degenerate measures on S.

Proof. See Kallenberg (2017), Corollary 3.21. □

2.3. Modes of Convergence

Assume for this section that S is a complete, separable metric space and let $\mathcal{S}$ be the corresponding Borel sets of S. Let $\hat{\mathcal{S}}$ be the set of bounded Borel sets in S and $\mathcal{K}$ be the set of compact sets in S.

2.3.1. Weak and vague topologies

We first define two important classes of functions. Let:

1. C_S be the set of bounded, continuous functions $f \geqslant 0$ on S,
2. $\hat{C}_S$ be the set of bounded, continuous functions $f \geqslant 0$ on S with bounded support.

The *vague topology* on $\mathbb{M}_S$ is that generated by the collection of functions on $\mathbb{M}_S$

$$\pi_f : \mu \mapsto \mu(f), \qquad f \in \hat{C}_S.$$

In other words, it is the smallest topology on $\mathbb{M}_S$ such that all members of this collection are continuous. A sequence (μ_n) in $\mathbb{M}_S$ *converges vaguely* to a fixed $\mu \in \mathbb{M}_S$ if

$$\mu_n(f) \to \mu(f) \text{ for all } f \in \hat{C}_S.$$

We write this as $\mu_n \xrightarrow{\mathrm{v}} \mu$.

[3] Kallenberg proves an a.s. decomposition of ξ as a sum of α the expectation Poisson process on $\mathbb{M}_S$ with intensity λ. Since we have not introduced any notion of boundedness on $\mathbb{M}_S$, defining a Poisson process on it would be unintuitive. We instead favour this Laplacian functional representation.

A *Polish space* is a topological space which is separable and completely metrisable. Clearly S is a Polish space and we should like that $\mathbb{M}_S$ is also.

Theorem 2.3.1. The space $\mathbb{M}_S$ endowed with the vague topology is a Polish space. Moreover, the σ-algebra induced by the vague topology coincides with $\mathcal{M}_S$.

Proof. See Kallenberg (2017), Theorem 4.2. □

Write $\hat{\mathbb{M}}_S$ for the set of finite measures on S. The *weak topology* on $\hat{\mathbb{M}}_S$ is that generated by the collection of functions on $\hat{\mathbb{M}}_S$

$$\pi_f : \mu \mapsto \mu(f), \qquad f \in C_S\,.$$

A sequence (μ_n) in $\hat{\mathbb{M}}_S$ *converges weakly* to a fixed $\mu \in \hat{\mathbb{M}}_S$ if

$$\mu_n(f) \to \mu(f) \text{ for all } f \in C_S\,.$$

We write this as $\mu_n \xrightarrow{\text{w}} \mu$.

Theorem 2.3.2. The space $\hat{\mathbb{M}}_S$ endowed with the weak topology is a Polish space.

Proof. See Kallenberg (2017), Lemma 4.5. □

It is easy to characterise vaguely and weakly relatively compact sets of measures. Recall that a set in a topological space is relatively compact if its closure is compact.

Theorem 2.3.3. A set $E \subset \mathbb{M}_S$ is vaguely relatively compact if and only if

1. $\sup_{\mu \in E} \mu(B) < \infty$ for all $B \in \hat{\mathcal{S}}$,
2. $\inf_{K \in \mathcal{K}} \sup_{\mu \in E} \mu(B \backslash K) = 0$ for all $B \in \hat{\mathcal{S}}$.

Moreover, a set $E \subset \hat{\mathbb{M}}_S$ is weakly relatively compact if and only if the above conditions hold with $B = S$.

Proof. See Kallenberg (2017), Theorem 4.2; Lemma 4.4. □

Let ξ, $(\xi)_n$ be random measures on S. We introduce the following types of convergence of the sequence (ξ_n) towards ξ:

1. $\xi_n \xrightarrow{\text{v}} \xi$ a.s. if $\mathbb{P}(\xi_n \xrightarrow{\text{v}} \xi) = 1$,
2. $\xi_n \xrightarrow{\text{v}\mathbb{P}} \xi$ if $\xi_n(f) \xrightarrow{\mathbb{P}} \xi(f)$ for all $f \in \hat{\mathcal{C}}_S$,
3. $\xi_n \xrightarrow{\text{v}} \xi$ in L^1 if $\xi_n(f) \to \xi(f)$ in L^1 for all $f \in \hat{\mathcal{C}}_S$.

Lemma 2.3.4. Let ξ, (ξ_n) be random measures on S. Then:

1. $\xi_n \xrightarrow{\text{v}} \xi$ a.s. if and only if $\xi_n(f) \to \xi(f)$ a.s. for all $f \in \hat{\mathcal{C}}_S$,
2. $\xi_n \xrightarrow{\text{v}\mathbb{P}} \xi$ if and only if, for any subsequence $N' \subset \mathbb{N}$, there exists a further subsequence $N'' \subset N'$ such that $\xi_n \xrightarrow{\text{v}} \xi$ a.s. along N'',
3. $\xi_n \xrightarrow{\text{v}} \xi$ in L^1 if and only if $\mathbb{E}\xi \in \mathbb{M}_S$ and $\mathbb{E}\xi_n \xrightarrow{\text{v}} \mathbb{E}\xi$ and $\xi_n \xrightarrow{\text{v}\mathbb{P}} \xi$.

Proof. See Kallenberg (2017), Lemma 4.8. □

2.3.2. Convergence in distribution

A sequence (ξ_n) of random measures on S *converges vaguely in distribution* to a random measure ξ on S if

$$\mathbb{E}\left(g(\xi_n)\right) \to \mathbb{E}\left(g(\xi)\right) \text{ for all bounded vaguely continuous } g : \mathbb{M}_S \to \mathbb{R}\,.$$

We write this as $\xi_n \overset{\text{vd}}{\to} \xi$.

Theorem 2.3.5. Let ξ, (ξ_n) be random measures on S. Then the following are equivalent:

1. $\xi_n \overset{\text{vd}}{\to} \xi$,
2. $\xi_n(f) \overset{\text{d}}{\to} \xi(f)$ for all $f \in \hat{\mathcal{C}}_S$,
3. $\mathcal{L}_{\xi_n}(f) \to \mathcal{L}_\xi(f)$ for all $f \in \hat{\mathcal{C}}_S$.

Proof. See Kallenberg (2017), Theorem 4.11. □

Unsurprisingly, similar characterisations exists in terms of avoidance probabilities of point processes on a dissecting ring and in terms of exponential moments. For a random measure ξ on S, let $\hat{\mathcal{S}}_\xi$ be set of $B \in \hat{\mathcal{S}}$ such that $\xi(\partial B) = 0$ a.s..

Theorem 2.3.6. Let ξ, (ξ_n) be point processes on S such that ξ is simple. Fix a dissecting ring $\mathcal{U} \subset \hat{\mathcal{S}}_\xi$ and a dissecting semiring $\mathcal{I} \subset \mathcal{U}$. Then $\xi_n \overset{\text{vd}}{\to} \xi$ if and only if

1. $\nu_{\xi_n}(U) \to \nu_\xi(U)$ for all $U \in \mathcal{U}$,
2. $\limsup_{n\to\infty} \mathbb{P}(\xi_n(I) > 1) \leqslant \mathbb{P}(\xi(I) > 1)$ for all $I \in \mathcal{I}$.

Proof. See Kallenberg (2017), Theorem 4.15. □

Theorem 2.3.7. Let ξ, (ξ_n) be random measures (respectively point processes) on S such that ξ is diffuse (respectively simple). Fix $t > s > 0$ and a dissecting ring $\mathcal{U} \subset \hat{\mathcal{S}}_\xi$ and a dissecting semiring $\mathcal{I} \subset \mathcal{U}$. Then $\xi_n \overset{\text{vd}}{\to} \xi$ if and only if

1. $\mathbb{E}\left(e^{-t\xi_n(U)}\right) \to \mathbb{E}\left(e^{-t\xi(U)}\right)$ for all $U \in \mathcal{U}$,
2. $\liminf_{n\to\infty} \mathbb{E}\left(e^{-s\xi_n(I)}\right) \geqslant \mathbb{E}\left(e^{-s\xi(I)}\right)$ for all $I \in \mathcal{I}$.

Proof. See Kallenberg (2017), Theorem 4.16. □

A sequence (ξ_n) of a.s. finite random measures on S *converges weakly in distribution* to an a.s. finite random measure ξ on S if

$$\mathbb{E}\left(g(\xi_n)\right) \to \mathbb{E}\left(g(\xi)\right) \text{ for all bounded weakly continuous } g : \hat{\mathbb{M}}_S \to \mathbb{R}\,.$$

We write this as $\xi_n \overset{\text{wd}}{\to} \xi$.

Theorem 2.3.8. Let ξ, (ξ_n) be a.s. bounded random measures on S. Then the following are equivalent:

1. $\xi_n \overset{\text{wd}}{\to} \xi$,
2. $\xi_n \overset{\text{vd}}{\to} \xi$ and $\xi_n(S) \to \xi(S)$,
3. $\xi_n \overset{\text{vd}}{\to} \xi$ and $\inf_{B\in\hat{\mathcal{S}}} \limsup_{n\to\infty} \mathbb{E}\left(\xi_n(B^c) \wedge 1\right) = 0$.

Proof. See Kallenberg (2017), Theorem 4.19. □

2.4. Palm theory

As part of this work's effort to provide an introduction to the theory random measures, we must now turn our attention to Palm theory. This section is based primarily on the treatments of the subject in Kallenberg (2017) and Baccelli et al. (2020) - both texts are somewhat difficult, which presents an opportunity to give a more gentle introduction. We return to the general framework introduced in Section 2.1.

2.4.1. Palm distributions

Let ξ be a random measure on S. The *Campbell measure* associated with ξ is the measure C_ξ on $S \times \mathbb{M}_S$ defined by

$$C_\xi(A \times U) = \mathbb{E}\left[\xi(A) 1_U(\xi)\right] = \int_U \int_A \xi(\mathrm{d}s)\ \mathcal{P}_\xi(\mathrm{d}\mu), \qquad A \in \mathcal{S},\ U \in \mathcal{M}_S\ .$$

where $\mathcal{P}_\xi$ is the law of ξ as a random variable on $\mathbb{M}_S$. By standard measure theoretic arguments (i.e. Carathédory's Extension Theorem), this extends uniquely to a (σ-finite) measure on $S \times \mathbb{M}_S$. We note an important special case:

$$C_\xi(A \times \mathbb{M}_S) = \mathbb{E}[\xi(A) 1_{\mathbb{M}_S}(\xi)] = \mathbb{E}[\xi(A)] = \mathbb{E}\xi(A), \qquad A \in \mathcal{S}\ .$$

In this sense, the Campbell measure is a generalisation of the intensity measure.

Let ξ be a random measure on S such that $\mathbb{E}\xi \in \mathbb{M}_S$. Then the *Palm kernel* associated with ξ is the (measurable) function $\mathcal{P}_\xi : S \times \mathcal{M}_S \to [0,1]$ which satisfies:

1. $\mathcal{P}_\xi(\,\cdot\,, U)$ is a measurable function on S and is $\mathbb{E}\xi$-integrable with
$$\int_A \mathcal{P}_\xi(s, U)\ \mathbb{E}\xi(\mathrm{d}s) = C_\xi(A \times U) \text{ for all } A \in \mathcal{S}$$
for each $U \in \mathcal{M}_S$,
2. $\mathcal{P}_\xi(s, \,\cdot\,)$ is a probability measure on $\mathbb{M}_S$ for all $s \in S$.

We will write $\mathcal{P}_\xi^{(s)}(\,\cdot\,)$ for $\mathcal{P}_\xi(s, \,\cdot\,)$. This definition is justified by the following result.

Lemma 2.4.1. Let ξ be a random measure on S with $\mathbb{E}\xi \in \mathbb{M}_S$. Then its Palm kernel as defined above exists and $\{\mathcal{P}_\xi^{(s)}\}_{s \in S}$ is unique up to a $\mathbb{E}\xi$-null set.

Proof. See Daley & Vere-Jones (2008), Proposition 13.1.IV. □

Let ξ be a random measure on S such that $\mathbb{E}\xi \in \mathbb{M}_S$. A family of *Palm versions* of ξ is a collection $\{\xi_s\}_{s \in S}$ of random measure on S such that $\xi_s \sim \mathcal{P}_\xi^{(s)}$ for $\mathbb{E}\xi$-almost all $s \in S$.[4]

Remark 2.4.2. The distribution of a Palm version $\{\xi_s\}_{s \in S}$ as above is unique up to an $\mathbb{E}\xi$-null set. There is no notion of a.s. uniqueness in the Palm versions of ξ, however, so a random measure will in general have uncountably many families of Palm versions.

[4] It may be necessary to extend the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ in the event that it is not coarse enough to accommodate the Palm distributions of ξ. This a technicality that will have no effect on our analysis as it does not change any distributions.

The principle result arising from Palm theory is the *Campbell-Little-Mecke formula.* This allows us in certain situations to exchange expectations with integrals in a similar way as with Fubini's Theorem. The precise statement of the formula is given below.

Theorem 2.4.3. Let ξ be a random measure on S such that $\mathbb{E}\xi \in \mathbb{M}_S$. Then, for all $g \geqslant 0$ measurable on $S \times \mathbb{M}_S$,

$$\mathbb{E}\left[\int_S g(s,\xi)\,\xi(\mathrm{d}s)\right] = \int_S \mathbb{E}[g(s,\xi_s)]\,\mathbb{E}\xi(\mathrm{d}s)\,,$$

where $\{\xi_s\}_{s\in S}$ is a family of Palm versions of ξ.

Proof. See Daley & Vere-Jones (2008), Proposition 13.1.IV. □

As it stands, it is not easy to compute the Palm distributions associated with a general random measure on S. This is usually done instead using the following result. Let $\mathcal{L}_\xi^{(s)} = \mathcal{L}_{\xi_s}$ for each $s \in S$, where $\{\xi_s\}_{s\in S}$ is a family of Palm versions of ξ, noting that $\{\mathcal{L}_\xi^{(s)}\}_{s\in S}$ is unique up to an $\mathbb{E}\xi$-null set.

Theorem 2.4.4. Let ξ be a random measure on S such that $\mathbb{E}\xi$ is finite. Then, for all measurable functions $f, g \geqslant 0$ on S,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_\xi(f+tg)\bigg|_{t=0} = -\int_S g(s)\mathcal{L}_\xi^{(s)}(f)\,\mathbb{E}\xi(\mathrm{d}s)\,.$$

Moreover, if a family $\{\Phi_s\}_{s\in S}$ of functionals on $\mathcal{F}_+(S)$ satisfies this relation in place of $\{\mathcal{L}_\xi^{(s)}\}_{s\in S}$, then $\Phi_s = \mathcal{L}_\xi^{(s)}$ for $\mathbb{E}\xi$-almost all $s \in S$.

Proof. See Daley & Vere-Jones (2008) Proposition 13.1.VI. □

An interesting property of Palm distributions is that they preserve any almost-sure properties of the original distribution. This is made explicit below.

Lemma 2.4.5. Let ξ be a random measure on S such that $\mathbb{E}\xi \in \mathbb{M}_S$. Then for all $T \in \mathcal{M}_S$ with $\mathcal{P}_\xi(T) = 1$, we have $\mathcal{P}_\xi^{(s)}(T) = 1$ for $\mathbb{E}\xi$-almost all $s \in S$.

Proof. A number of proofs are available. We argue the following: for all $A \in \mathcal{S}$,

$$C_\xi(A \times T) = \mathbb{E}[\xi(A)1_T(\xi)] = \mathbb{E}[\xi(A)] = \mathbb{E}\xi(A)\,.$$

whence

$$\mathbb{E}\xi(A) = C_\xi(A \times T) = \int_A \mathcal{P}_\xi^{(s)}(T)\,\mathbb{E}\xi(\mathrm{d}s)\,.$$

Since $A \in \mathcal{S}$ was arbitrary, we must have $\mathcal{P}_\xi^{(s)}(T) = 1$ for $\mathbb{E}\xi$-almost all $s \in S$. □

This result has a particularly important consequence.

Corollary 2.4.6. Let ξ be a random measure on S with $\mathbb{E}\xi \in \mathbb{M}_S$. Let $\{\xi_s\}_{s\in S}$ be a family of Palm versions of ξ. If ξ is a (simple) point process, then ξ_s is also a (simple) point process for $\mathbb{E}\xi$-almost all $s \in S$.

Proof. This follows immediately from Lemma 2.4.5 by taking $T = \mathbb{N}_S$ (or $T = \mathbb{N}_S^*$ for the simple case). □

It turns out that Palm distributions have a very natural interpretation when we restrict our attention to point processes. The following result captures this intuition.

Lemma 2.4.7. Let ξ be a point process on S with $\mathbb{E}\xi \in \mathbb{M}_S$. Let $\{\xi_s\}_{s\in S}$ be a family of Palm distributions of ξ. Then $\xi_s(s) \geqslant 1$ a.s. for $\mathbb{E}\xi$-almost all $s \in S$.

Proof. This follows by setting $n = 1$ in Lemma 2.4.10 below (see alternatively Baccelli et al. (2020), for example). □

Intuitively, the Palm distribution of a point process ξ on S at $s \in S$ (when it exists) is also a point process on S but with a fixed atom at s. This motivates the interpretation of the Palm distribution of ξ at s as the distribution of ξ conditional on the fact that it has an atom at s.

Let ξ be a point process on S such that $\mathbb{E}\xi \in \mathbb{M}_S$. Motivated by this analysis, we define a family of *reduced Palm versions* of ξ as a collection $\{\xi_s^!\}_{s\in S}$ of random measures on S such that

$$\xi_s^! + \delta_s \sim \mathcal{P}_\xi^{(s)} \quad \text{for } \mathbb{E}\xi\text{-almost all } s \in S\,.$$

We can construct a family of reduced Palm versions of ξ from a family $\{\xi_s\}_{s\in S}$ of Palm versions of ξ. Indeed, $\xi_s(\{s\}) \geqslant 1$ a.s. for $\mathbb{E}\xi$-almost all $s \in S$, so for $\mathbb{E}\xi$-almost all $s \in S$ there exists a random measure $\xi_s^!$ on S with

$$\xi_s^! = \xi_s - \delta_s \text{ a.s.}$$

This is sufficient to define a family $\{\xi_s^!\}_{s\in S}$ of reduced Palm versions.

Example 2.4.8. Let ξ be a Poisson process on S with $\mathbb{E}\xi \in \mathbb{M}_S$. Let $\{\xi_s^!\}_{s\in S}$ be a family of reduced Palm versions of ξ. Then one shows (Baccelli et al. 2020) that

$$\xi_s^! \overset{\mathrm{d}}{=} \xi \quad \text{for } \mathbb{E}\xi\text{-almost all } s \in S\,.$$

This is not surprising, since the Poisson process is completely random, so conditioning on the presence of an atom and then removing it should have no effect on the distribution.

2.4.2. Higher order Palm distributions

Let ξ be a random measure on S. For $n \in \mathbb{N}$, the *n-th order Campbell measure* associated with ξ is the measure C_ξ^n on $S^n \times \mathbb{M}_S$ defined by

$$C_\xi^n(A \times U) = \mathbb{E}\left[\xi^n(A)1_U(\xi)\right] = \int_U \int_A \xi^n(\mathrm{d}s)\, \mathcal{P}_\xi(\mathrm{d}\mu), \qquad A \in \mathcal{S}^{\otimes n},\ U \in \mathcal{M}_S\,.$$

Again, the full definition of C_ξ^n is obtained by extension of this formula.

Let ξ be a random measure on S with $\mathbb{E}\xi^n \in \mathbb{M}_{S^n}$ for some $n \in \mathbb{N}$. The *n-th Palm kernel* associated with ξ is the function $\mathcal{P}_\xi^n : S^n \times \mathcal{M}_S \to [0,1]$ such that

1. $\mathcal{P}_\xi^n(\,\cdot\,, U)$ is a measurable function on S^n and is $\mathbb{E}\xi^n$-integrable with

$$\int_A \mathcal{P}_\xi^n(\mathbf{s}, U)\, \mathbb{E}\xi^n(\mathrm{d}\mathbf{s}) = C_\xi^n(A \times U) \text{ for all } A \in \mathcal{S}^{\otimes n}$$

for each $U \in \mathcal{M}_S$,

2. $\mathcal{P}_\xi^n(\mathbf{s}, \cdot\,)$ is a probability measure on $\mathbb{M}_S$ for all $\mathbf{s} \in S^n$.

As before, we write $\mathcal{P}_\xi^{(\mathbf{s})}(\,\cdot\,)$ for $\mathcal{P}_\xi^n(\mathbf{s}, \,\cdot\,)$. One shows that $\{\mathcal{P}_\xi^{(\mathbf{s})}\}_{\mathbf{s}\in S^n}$ exists and is $\mathbb{E}\xi^n$-a.e. unique. See Baccelli et al. (2020) for example.

Let ξ be a random measure on S such that $\mathbb{E}\xi^n \in \mathbb{M}_{S^n}$ for some $n \in \mathbb{N}$. A family of *n-th Palm versions* of ξ is a collection $\{\xi_{\mathbf{s}}\}_{\mathbf{s}\in S^n}$ of random measure on S such that $\xi_{\mathbf{s}} \sim \mathcal{P}_{\xi}^{(\mathbf{s})}$ for $\mathbb{E}\xi^n$-almost all $\mathbf{s} \in S^n$.

We can use the n-th Palm distributions to establish a generalised version of the Campbell-Little-Mecke formula, which will prove invaluable in the sequel.

Theorem 2.4.9. Let ξ be a random measure on S such that $\mathbb{E}\xi^n \in \mathbb{M}_{S^n}$ for some $n \in \mathbb{N}$. Then, for all $g \geqslant 0$ measurable on $S^n \times \mathbb{M}_S$,

$$\mathbb{E}\left[\int_{S^n} g(\mathbf{s},\xi)\,\xi^n(\mathrm{d}\mathbf{s})\right] = \int_{S^n} \mathbb{E}[g(\mathbf{s},\xi_{\mathbf{s}})]\,\mathbb{E}\xi^n(\mathrm{d}\mathbf{s})\ ,$$

where $\{\xi_{\mathbf{s}}\}_{\mathbf{s}\in S^n}$ is a family of n-th Palm versions of ξ.

Proof. See Baccelli et al. (2020), Theorem 3.3.2. □

We now prove the general form of Lemma 2.4.7 for higher order Palm distributions.

Lemma 2.4.10. Fix $n \in \mathbb{N}$ and let ξ be a point process on S with $\mathbb{E}\xi^n \in \mathbb{M}_{S^n}$. Let $\{\xi_{\mathbf{s}}\}_{\mathbf{s}\in S^n}$ be a family of n-th Palm versions of ξ. Then

$$\xi_{\mathbf{s}}(\{s_i\}) > 0 \text{ for all } i = 1,\dots,n \text{ a.s. for } \mathbb{E}\xi^n\text{-almost all } \mathbf{s} \in S^n\,.^5 \tag{2.6}$$

Proof. We have $\xi = \sum_{i\in I}\delta_{\sigma_i}$ for some set $I = \{1,2,\dots,\kappa\}$ with κ a random variable in $\mathbb{Z}_{\geqslant 0} \cup \{+\infty\}$ and $(\sigma_i)_{i\in I}$ a collection of random variables in S. Let

$$A = \{(\mathbf{s},\mu) \in S^n \times \mathbb{M}_S : \mu(\{s_i\}) = 0 \text{ for some } i = 1,\dots,n\}\,.$$

Then $A \in \mathcal{S}^{\otimes n} \otimes \mathcal{M}_S$, so $(\mathbf{s},\mu) \mapsto 1_A((\mathbf{s},\mu))$ is measurable on $S^n \times M_S$. Hence

$$\begin{aligned}\int_{S^n} \mathbb{P}((\mathbf{s},\xi_{\mathbf{s}}) \in A)\,\mathbb{E}\xi^n(\mathrm{d}\mathbf{s}) &= \int_{S^n} \mathbb{E}[1_A(\mathbf{s},\xi_{\mathbf{s}})]\,\mathbb{E}\xi^n(\mathrm{d}\mathbf{s}) \\ &= \mathbb{E}\left[\int_{S^n} 1_A(\mathbf{s},\xi)\,\xi^n(\mathrm{d}\mathbf{s})\right] \\ &= \mathbb{E}\left[\sum_{i\in I^n} 1_A(\sigma_{i_1},\dots,\sigma_{i_n},\xi)\right] \\ &= \mathbb{E}\left[\sum_{i\in I^n} 1\{\xi(\{\sigma_{i_k}\}) = 0 \text{ for some } k = 1,\dots,n\}\right] \\ &= 0\end{aligned}$$

where the second equality follows from the generalised Campbell-Little-Mecke formula and the last equality follows from the fact that $\{\sigma_i\}_{i\in I}$ is precisely the support of ξ a.s.. It follows that, for $\mathbb{E}\xi^n$-almost all $\mathbf{s} \in S^n$, $(\mathbf{s},\xi_s) \notin A$ a.s., or equivalently $\xi_{\mathbf{s}}(\{s_i\}) > 0$ for all $i = 1,\dots,n$ a.s.. □

Intuitively, (2.6) asserts that the $\xi_{\mathbf{s}}$ has atoms at s_i for each $i = 1,\dots,n$. This motivates a similar interpretation of higher order Palm distributions to that of first order Palm distributions. That is, the distribution of $\xi_{\mathbf{s}}$ is the distribution of the original point process ξ conditional on the fact that it has atoms at s_i for all $i = 1,\dots,n$.

[5] Baccelli et al. (2020) gives a stronger form of this result in Proposition 3.3.4. However, the proof of this result skips a number of computional steps and so is not entirely convincing. The result presented here is more well-established and is proved similarly in Kallenberg (2017).

2.4.3. Higher order reduced Palm distributions

Let ξ be a point process on S. For $n \in \mathbb{N}$, the *n-th order reduced Campbell measure* associated with ξ is the measure $C_\xi^{(n)}$ on $S^n \times \mathbb{M}_S$ defined by

$$C_\xi^{(n)}(A \times U) = \mathbb{E}\left[\int_A 1\left\{\xi - \sum_{i=1}^n \delta_{s_i} \in U\right\} \xi^{(n)}(\mathrm{d}\mathbf{s})\right] \qquad A \in S^{\otimes n},\ U \in \mathcal{M}_S\,.$$

Once more, the full definition of $C_\xi^{(n)}$ is obtained by extension of this formula.

Let ξ be a random measure on S with $\mathbb{E}\xi^{(n)} \in \mathbb{M}_{S^n}$ for some $n \in \mathbb{N}$. The *n-th reduced Palm kernel* associated with ξ is the function $\mathcal{P}_\xi^{!n} : S^n \times \mathcal{M}_S \to [0,1]$ such that

1. $\mathcal{P}_\xi^{!n}(\,\cdot\,, U)$ is a measurable function on S^n and is $\mathbb{E}\xi^{(n)}$-integrable with
$$\int_A \mathcal{P}_\xi^{!n}(\mathbf{s}, U)\ \mathbb{E}\xi^{(n)}(\mathrm{d}\mathbf{s}) = C_\xi^{(n)}(A \times U) \text{ for all } A \in \mathcal{S}^{\otimes n}$$
for each $U \in \mathcal{M}_S$,
2. $\mathcal{P}_\xi^{!n}(\mathbf{s}, \,\cdot\,)$ is a probability measure on $\mathbb{M}_S$ for all $\mathbf{s} \in S^n$.

We write $\mathcal{P}_\xi^{!(\mathbf{s})}(\,\cdot\,)$ for $\mathcal{P}_\xi^{!n}(\mathbf{s}, \,\cdot\,)$. One shows that $\{\mathcal{P}_\xi^{!(\mathbf{s})}\}_{\mathbf{s}\in S^n}$ exists and is $\mathbb{E}\xi^{(n)}$-a.e. unique. See Baccelli et al. (2020) for example.

Remark 2.4.11. When $n = 1$, we have $\xi^{(1)} = \xi$, so $\mathbb{E}\xi^{(1)} = \mathbb{E}\xi$ and $C_\xi^{(1)} = C_\xi$. Hence the 1-st order reduced Palm distributions coincide with the standard Palm distributions.

Let ξ be a random measure on S such that $\mathbb{E}\xi^{(n)} \in \mathbb{M}_{S^n}$ for some $n \in \mathbb{N}$. A family of *n-th reduced Palm versions* of ξ is a collection $\{\xi_\mathbf{s}^!\}_{\mathbf{s}\in S^n}$ of random measure on S such that $\xi_\mathbf{s}^! \sim \mathcal{P}_\xi^{!(\mathbf{s})}$ for $\mathbb{E}\xi^{(n)}$-almost all $\mathbf{s} \in S^n$.

We can restate the Campbell-Little-Mecke formula in terms of the reduced Palm versions of a random measure.

Theorem 2.4.12. Let ξ be a random measure on S such that $\mathbb{E}\xi^{(n)} \in \mathbb{M}_{S^n}$ for some $n \in \mathbb{N}$. Then, for all $g \geqslant 0$ measurable on $S^n \times \mathbb{M}_S$,

$$\mathbb{E}\left[\int_{S^n} g\left(\mathbf{s}, \xi - \sum_{i=1}^n \delta_{s_i}\right)\ \xi^{(n)}(\mathrm{d}\mathbf{s})\right] = \int_{S^n} \mathbb{E}[g(\mathbf{s}, \xi_\mathbf{s}^!)]\ \mathbb{E}\xi^{(n)}(\mathrm{d}\mathbf{s})\ ,$$

where $\{\xi_\mathbf{s}^!\}_{\mathbf{s}\in S^n}$ is a family of n-th reduced Palm versions of ξ.

Proof. See Baccelli et al. (2020), Theorem 3.3.6. □

As in the one dimensional case, there is an intuitive relationship between a point processes' n-th Palm versions and its n-th reduced Palm versions.

Theorem 2.4.13. Let ξ be a point process on S with $\mathbb{E}\xi^{(n)} \in \mathbb{M}_{S^n}$ for some $n \in \mathbb{N}$. Let $\{\xi_\mathbf{s}\}_{\mathbf{s}\in S^n}$ and $\{\xi_\mathbf{s}^!\}_{\mathbf{s}\in S^n}$ be n-th Palm and reduced Palm versions of ξ respectively. Then

$$\xi_\mathbf{s}^! \overset{\mathrm{d}}{=} \xi_\mathbf{s} - \sum_{i=1}^n \delta_{s_i} \text{ for } \mathbb{E}\xi^{(n)}\text{-almost all } \mathbf{s} \in S^{(n)}\,,$$

where $S^{(n)}$ denotes the non-diagonal part of S^n, with $\mathbb{E}\xi^{(n)}$ restricted to it.

Proof. See Baccelli et al. (2020), Corollary 3.3.8. □

3. Probabilistic Topic Models

We now move on to the second objective of this work, which is to provide a survey of the ever-growing suite of probabilistic topic models available to the modern statistician. We first give a more concrete mathematical formulation of the topic modelling problem, before introducing a selection of models aimed at solving it. This will set the stage for the introduction of our novel topic model in the next chapter.

3.1. Problem Outline

When describing the topic modelling problem, this work largely follows the notation and conventions originally established in Blei et al. (2003).

The most fundamental object in a topic model is a *word*. We assume that we know the full collection of words available for the document generation process a priori - this collection is called the *lexicon*. If we let $V \in \mathbb{N}$ be the number of words in the lexicon, then we can assign each word with an index $v = 1, \dots, V$. We represent the v-th word as the v-th standard basis vector of $\mathbb{R}^V$. That is, the v-th word is represented as the vector in $\mathbb{R}^V$ whose components are all zero except for its v-th component, which takes value one.

By a *document*, we mean a collection $\mathbf{w} = (w_1, \dots, w_M)$ of words. The order of words within a document is assumed to be inconsequential, such that the components of $\mathbf{w}$ are exchangeable[1]. A *corpus* is defined as a collection $\mathcal{D} = (\mathbf{w}_1, \dots, \mathbf{w}_N)$ of N documents, where $N \in \mathbb{N}$. The documents also need not contain the same number of words, so we write M_n for the number of words in document n, where $M_n \in \mathbb{N}$ for each $n = 1, \dots, N$.

We are interested in generative models able to model the process by which an N-document corpus is created from only the lexicon and the document lengths $M_1, \dots, M_N$. Probabilistic topic models are a special class of such models that utilise a certain type of latent variables, called *topics*. Mathematically, a topic is a discrete distribution over the lexicon of V words, which we represent as a V-dimensional probability vector ϕ. Although the use of these latent variables in the corpus-generation process varies between models, they typically represent the relevance of each word to a given topic.

3.2. Latent Dirichlet Allocation

3.2.1. Some definitions

First introduced in Blei et al. (2003), Latent Dirichlet Allocation (LDA) is a probabilistic topic model that has garnered widespread attention. We begin this section with some notations and definitions.

[1] This assumption is known in the wider literature as the *bag-of-words assumption*. A reader interested in a full account of the nature of exchangeability is directed to Aldous (1985).

For $d \in \mathbb{N}$, we write Δ^d for the $(d-1)$-dimensional simplex

$$\Delta^d = \left\{ x \in \mathbb{R}^d : x_1, \ldots, x_d \in [0,1],\ \sum_{i=1}^{d} x_i = 1 \right\} .$$

For $d \in \mathbb{N}$ and $a \in (0, \infty)^d$, we write $\text{Dirichlet}_d(\alpha)$ to mean the usual Dirichlet distribution on Δ^d with concentration parameter α. Its density function is

$$f(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{d} x_i^{\alpha_i - 1} , \qquad x \in \Delta^d . \tag{3.1}$$

where B is the multivariate Beta distribution[2].

For $d \in \mathbb{N}$ and $p \in \Delta^d$, we write $\text{Categorical}_d(p)$ to mean the discrete distribution over d outcomes, where the probability of outcome i is p_i for $i = 1, \ldots d$. In what follows, it will be instructive to view this as a multinomial distribution with a single trial and event probabilities given by p. In this way, the support of this distribution is the set of all standard basis vectors of $\mathbb{R}^d$. Note that, when $d = V$, this support is the same as the set of all words under our chosen representation.

3.2.2. Model specification and hyperparameter selection

We are now in a position to specify the LDA model of a corups $\mathcal{D}$. The exact variant we focus on is the so-called "smoothed LDA" (Blei et al. 2003), which is a generative version of LDA. Given a total number of topics T, assume that, independently

$$\begin{aligned}
\phi_t &\sim \text{Dirichlet}_V(\eta) , & t &= 1, \ldots, T \\
\theta_n &\sim \text{Dirichlet}_T(\alpha) , & n &= 1, \ldots, N \\
z_{nm}|\theta_n &\sim \text{Categorical}_T(\theta_n) , & m &= 1, \ldots, M_n, \quad n = 1, \ldots, N \\
w_{nm}|z_{nm}, \boldsymbol{\phi} &\sim \text{Categorical}_V(\phi_{z_{nm}}) , & m &= 1, \ldots, M_n, \quad n = 1, \ldots, N .
\end{aligned}$$

Here, α and η are respectively T- and V-dimensional vectors of non-negative real numbers that must be specified as hyperparameters for the two Dirichlet priors. A graphical representation of the model is shown in Figure 3.1.

In the above, we call $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_T)$ the *topic matrix*. This quantity is shared between documents. We interpret ϕ_{tv} as the probability that a randomly sampled word is the v-th word, given that the word belongs to topic t. This is an important object of inference in the model, as it dictates which words are most strongly associated with which topics.

Given the topic matrix, the generative process of a document $\mathbf{w}_n$ is relatively intuitive. We first sample θ_n, which we call the *topic mixture* associated with the document. We interpret θ_{nt} as the relevance of topic t to the contents of document n. We then sample each of the $m = 1, \ldots, M_n$ words that constitute $\mathbf{w}_n$ independently as follows:

1. Sample the topic z_{nm} associated with w_{nm} according to the probability vector θ_n,
2. Sample a word w_{nm} from the lexicon according to the probability vector $\phi_{z_{nm}}$.

[2] This density function technically only determines $(x_1, \ldots, x_{d-1})$, setting $x_d = 1 - \sum_{i=1}^{d-1} x_i$ after the fact. Indeed, Δ^d is a $(d-1)$-dimensional space, and so any density defined on it should be with respect to the $(d-1)$-dimensional Lebesgue measure, as opposed to the d-dimensional one.

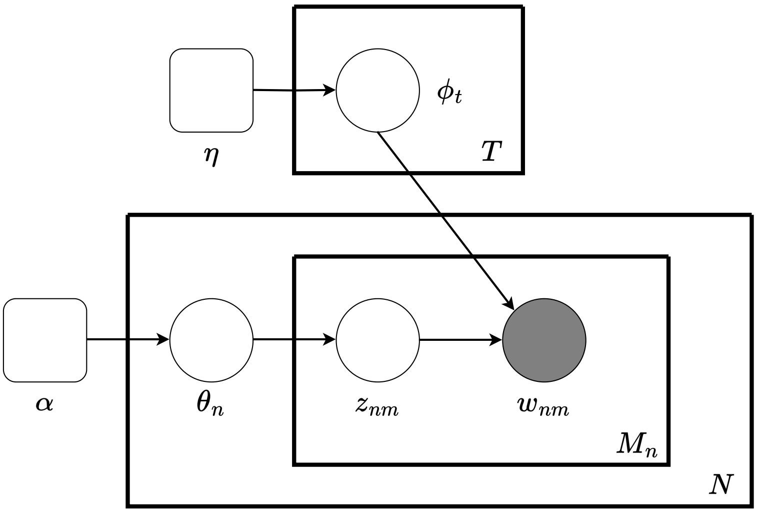

Figure 3.1.: A plate diagram representing LDA. The circles represent random variables and the squares represent hyperparameters. A circle is shaded if that variable is observed. An arrow from one variable to another means that the second's distribution is given conditionally on the first. The plates indicate the number of times that the process inside of them is repeated. Inspired by Figure 7 in Blei et al. (2003).

In terms of selecting hyperparameters for the LDA model, Blei et al. (2003) restricts attention to a symmetric parameter η, where $\eta_t = \eta_0$ for all $t = 1, \ldots, T$ for some $\eta_0 > 0$. This represents a belief that the components of each ϕ_t are exchangeable, with no bias being given towards any given word in the distribution of the topics. The choice of η_0 is more subjective. Choosing $\eta_0 \gg 1$ corresponds to a belief that the topics will be more balanced, whereas a choosing $\eta_0 \ll 1$ corresponds to a belief that the topics will be more polarised. Intuitively, the latter case might be favourable, particularly if we seek topics that are characterised by giving a large weighting to a few words that are most relevant to that topic. No such restriction is made on α.

Unlike many of LDAs predecessors, each word w_{nm} in the corpus is assigned its own topic z_{nm}, allowing for more variety and nuance in document construction. The model is also fully generative: given a new document $\mathbf{w}^*$, its likelihood under the model can be computed (conditional on the number M^* of words it contains). For a more complete survey of the models that influenced LDA, see Blei et al. (2003).

3.2.3. Posterior inference

There are a number of methods available when performing posterior inference under LDA. Since the analytical posterior distribution is not tractable (Blei et al. 2003), only approximate methods are available. We provide a brief outline of two of the more popular approaches here.

The approach proposed originally in Blei et al. (2003) is a variational inference procedure. In this method, we start with a flexible family of parametric distributions. We then optimise its parameters to maximise the similarity between this parametric distribution and the posterior distribution under the LDA model. For particularly large datasets (i.e. corpuses with a large number of documents), we can appeal to the online variational procedure of Hoffman et al. (2010). This avoids parsing every document at every iteration by utilising a stochastic optimisation procedure in place of a deterministic one. They argue this leads to improved performance.

The other approach is a standard MCMC procedure. Griffiths & Steyvers (2004) gives a collapsed Gibbs sampler to provide samples from the posterior distribution of the latent topics. It works by integrating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ out of the posterior joint likelihood. This is possible due to the the conjugacy between the Dirichlet distribution and the multinomial distribution (or categorical distribution, in our case). The algorithm then amounts to repeatedly sampling from the posterior distributions of the z_{nm}, conditional only on $\mathcal{D}$. Based on these posterior samples, estimators for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are constructed.

3.3. Other Topic Models

The novel model introduced in the proceeding chapter shares many features of LDA and can be thought of as an extension of it. It will hence be pertinent to review some existing extensions of LDA to better understand where the new model fits in.

3.3.1. DPP-LDA

Under LDA, the components of the topic matrix $\boldsymbol{\phi}$ are i.i.d.. This can lead to configurations in which two of its components ϕ_t and $\phi_{t'}$ are very similar, i.e. one in which the model selecting two very similar topics. This may not be desirable in many contexts as it can cause high redundancy in the latent variables during the learning process and can hinder their interpretability a posteri (Zou & Adams 2012). A model that generated a more diverse selection of topics, however, would not suffer from these issues.

In Zou & Adams (2012), the assumption that the components of the topic matrix $\boldsymbol{\phi}$ are i.i.d. is relaxed. This permits interactions between the ϕ_t, such that their distributions on Δ^V are correlated in general. Specifically, the set $\{\phi_1, \ldots, \phi_T\}$ is distributed according to a determinantal point process (DPP) - see Section 4.1.4 for an exposition - which has a scalar parameter $\lambda > 0$. The model is called DPP-LDA and is specified as

$$\begin{aligned}
\{\phi_1, \ldots, \phi_T\} &\sim \text{DPP}_V(K_\lambda) \\
\theta_n &\sim \text{Dirichlet}_T(\alpha)\,, && n = 1, \ldots, N \\
z_{nm}|\theta_m &\sim \text{Categorical}_T(\theta_n)\,, && m = 1, \ldots, M_n, \quad n = 1, \ldots, N \\
w_{nm}|z_{nm}, \boldsymbol{\phi} &\sim \text{Categorical}_V(\phi_{z_{nm}})\,, && m = 1, \ldots, M_n, \quad n = 1, \ldots, N\,.
\end{aligned}$$

A graphical representation of DPP-LDA is given in Figure 3.2.

The key property of the DPP is that its point locations have a tendency to repel one another and spread out across the sample space. For DPP-LDA, this encourages the topics to be diverse, attributing a low a priori probability to topic matrices with similar topics. One effect of this is that words that are very common in the corpus but do not have a particular topic associated with them (usually called stop words - examples include “and”, “the” and “to”) are grouped into their own topics. Thus there is no need to remove these words before fitting, unlike when fitting LDA (Blei & Lafferty 2005).

Posterior inference under DPP-LDA is conducted in Blei & Lafferty (2005) via a variational Bayes approach. The procedure is similar to that under LDA. The most important alteration amounts to the inclusion of a penalty term for the similarity of the components of $\boldsymbol{\phi}$ in the optimisation step of the procedure. The scalar parameter λ controls the strength of this penalty.

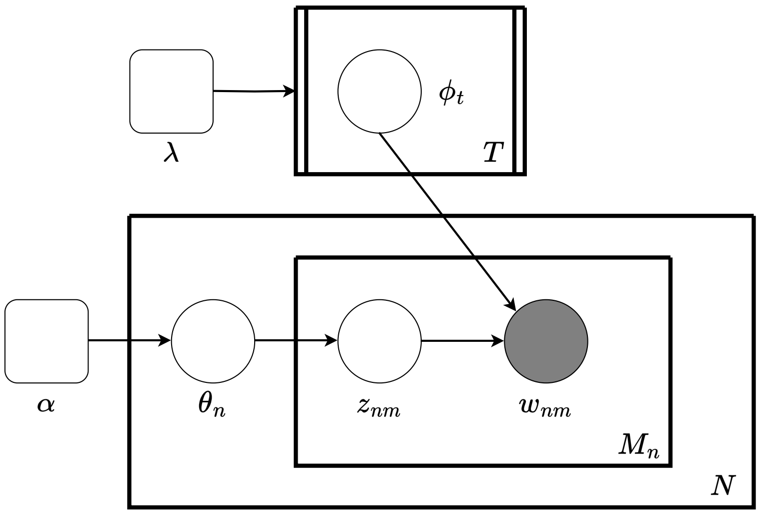

Figure 3.2.: A plate diagram representing DPP-LDA. The double-struck plate indicates that the variables are sampled from a DPP as opposed to via i.i.d. sampling. Inspired by Figure 2 in Zou & Adams (2012).

3.3.2. Correlated topic models

Another target for refinement in the LDA model is the Dirichlet prior on the topic mixtures θ_n. While the simplicity of the Dirichlet distribution is often an advantage, in the topic modelling context we may desire a more expressive prior distribution on the θ_n. For instance, we may seek greater control over the correlations between the components of θ_n, since such correlations may be very nuanced in some corpuses.

In Blei & Lafferty (2005), the Dirichlet prior on the θ_n is replaced by the logistic normal distribution, defined as follows. Let $v \sim \text{Normal}_d(\mu, \Sigma)$ for some $\mu \in \mathbb{R}^d$ and some (positive definite) $\Sigma \in \mathbb{R}^{d \times d}$, where $\text{Normal}_d(\mu, \Sigma)$ denotes a d-dimensional multivariate normal distribution with mean μ and covariance matrix Σ. Define u on Δ^d by setting

$$u_i = \frac{\exp v_i}{\sum_{j=1}^{d} \exp v_j} \,,$$

for $i = 1, \ldots, d$. Then u is distributed according to a logistic normal distribution with parameters (μ, Σ). We write this as $u \sim \text{Logistic-Normal}_d(\mu, \Sigma)$.

The correlated topic model (CTM) is specified as

$$\begin{aligned} \theta_n &\sim \text{Logistic-Normal}_T(\mu, \Sigma)\,, && n = 1, \ldots, N \\ z_{nm}|\theta_n &\sim \text{Categorical}_T(\theta_n)\,, && m = 1, \ldots, M_n, \quad n = 1, \ldots, N \\ w_{nm}|z_{nm}, \boldsymbol{\phi} &\sim \text{Categorical}_V(\phi_{z_{nm}})\,, && m = 1, \ldots, M_n, \quad n = 1, \ldots, N\,. \end{aligned}$$

A graphical representation of the model is given in Figure 3.3. It is important to note that the CTM is not fully generative. Indeed, it does not specify a prior on the topic matrix ϕ, nor on the new logistic normal distribution parameters μ and Σ. In this sense, it does not belong to the category of probabilistic topic models discussed thus far, but it still makes for an interesting comparison with our novel topic model in the sequel.

In Blei & Lafferty (2005), a variational Bayes method is used to optimise $\boldsymbol{\phi}$, μ and Σ. This is more complex than that used for LDA, since the distribution of θ_n is no longer conjugate to that of z_{mn}.

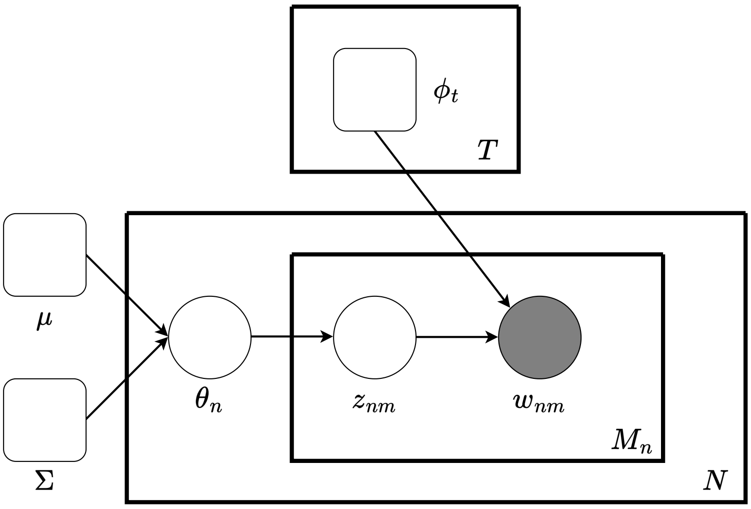

Figure 3.3.: A plate diagram representing the CTM. Inspired by Figure 1 in Blei & Lafferty (2005).

3.3.3. Hierarchical Dirichlet Processes

A final drawback of LDA that we look to address is that we must fix the number of topics beforehand, as opposed to inferring it from the corpus. In order to avoid this, we can consider probabilistic topic models with a nonparametric component, which can relieve us of our finite bound on the number of topics. Here, we explore a nonparametric topic model based on the Hierarchical Dirichlet Process (HDP) of Teh et al. (2006).

Let S be a complete, separable metric space as in Section 2.3. Fix $\alpha > 0$ and $H \in \mathbb{P}_S$. We say that a random probability distribution G on S is distributed according to a *Dirichlet process* (DP) with scale $\alpha > 0$ and base distribution H if, for any partition $(A_i)_{i=1}^m$ of S such that $A_i \in \mathcal{S}$ for all $i = 1, \dots, m$, we have

$$(G(A_1), \dots, G(A_m)) \sim \text{Dirichlet}_m\left(\alpha H(A_1), \dots, \alpha H(A_m)\right) .$$

We write this as $G \sim \text{DP}(\alpha, H)$. It is well known that, if $G \sim \text{DP}(\alpha, H)$, then G is purely atomic with infinitely many atoms a.s.. It follows by Theorem 2.1.5 that G admits a decomposition of the form

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\sigma_k} ,$$

for $\beta_1, \beta_2, \dots$ and $\sigma_1, \sigma_2, \dots$ random variables on $(0, \infty)$ and S respectively. The σ_k are in fact i.i.d. samples from H, whereas $(\beta_1, \beta_2, \dots)$ is generated via a stick-breaking construction (for more details, see Teh et al. (2006) and the references therein).

The DP is applied to topic modelling via the following hierarchical model. Fix $\eta > 0$, $\alpha > 0$ and a distribution H on Δ^V. We assume that

$$\begin{aligned}
G_0 &\sim \text{DP}(\eta, H) \\
G_n | G_0 &\sim \text{DP}(\alpha, G_0) , && n = 1, \dots, N \\
z_{nm} | G_n &\sim G_n , && m = 1, \dots, N_m, \quad n = 1, \dots, N \\
w_{nm} | z_{nm} &\sim \text{Categorical}_V(z_{nm}) , && m = 1, \dots, N_m, \quad n = 1, \dots, N .
\end{aligned}$$

See Figure 3.4 for a graphical representation of this model

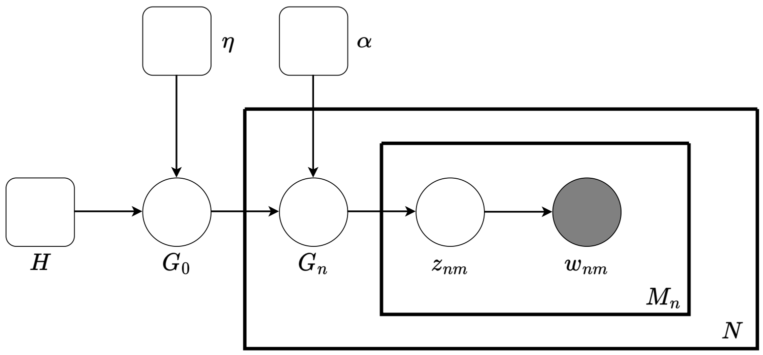

Figure 3.4.: A plate diagram representing the HDP topic model. Inspired by Figure 1 in Teh et al. (2006).

At first, this model may seem very far removed from LDA. Indeed, the latent topic mixture θ_n for the n-th document is replaced with a Dirichlet distribution, from which the z_{nm} are sampled directly. However, let us view the atom locations of G_0 as an infinite collection of latent topics $\phi_1, \phi_2, \ldots$. Since G_0 is a.s. discrete, the points of G_n must be the same as the points of G_0 a.s., so the atom locations of G_n are also these latent topics. Sampling G_n is thus equivalent to sampling a new weight for each topic, i.e. sampling an infinite-dimensional probability vector, with each component corresponding to one of the ϕ_t. Note that this is exactly what θ_n was under LDA, so z_{nm} has exactly the same interpretation under this model as it did under LDA. Thus we can view this model as an infinite-dimensional analogue of LDA.

We must appeal to MCMC methods to sample from the posterior of this model - its nonparametric nature makes variational inference difficult. Teh et al. (2006) gives a number of algorithms to sample from the joint posterior distributions of the latent variables in the model. Its performance is also compared to that of LDA.

4. A Novel Topic Model

In this chapter, we present our novel probabilistic topic model. We first introduce the Bayesian modelling framework of Beraha et al. (2023), which we will use as a base for the model. We then discuss how this framework can be applied in topic modelling, leading into the full specification of our novel model. At the end of this section, we recount back to the existing topic models discussed in the previous chapter.

4.1. Bayesian Mixtures with Interacting Atoms

4.1.1. Introduction

The novel topic model we investigate is an application of the framework of Beraha et al. (2023). This is a type of Bayesian mixture model, where the mixing probability measure is obtained by normalising an almost surely discrete random measure. We discuss this framework in its full generality in the first part of this section.

Beraha et al. (2023) provides some remarkably general theoretical properties of models that use their framework. We will need these results later in the chapter, and so this section shall also provide a treatment of them in its second part.

The advantage of the framework of Beraha et al. (2023) is its flexibility in how the atoms of the mixing measure can be distributed. In our topic model, the atoms will be distributed according to a determinantal point process, a class of repulsive point processes. We discuss the relevant properties of such processes at the end of the section.

4.1.2. Normalising random measures

Let S be a complete, separable metric space - we carry over the nomenclature introduced in Section 2.3 to the remainder of this chapter. A measure μ on S is a *probability measure* if $\mu(S) = 1$. We write $\mathbb{P}_S$ for the space of all locally finite probability measures on S. A *random probability measure* on S is a random measure ξ on S such that $\xi \in \mathbb{P}_S$ a.s.. We also endow $\mathbb{P}_S$ with its σ-algebra as a subset of $\mathbb{M}_S$.

Let $\mathcal{Q}$ be a distribution on $\mathbb{P}_S$. For some fixed $N \in \mathbb{N}$, let also $Y_1, \ldots, Y_N$ be random variables on S and $X_1, \ldots, X_N$ be random variables on $\mathbb{R}^d$. This section is concerned with models of the form

$$\begin{aligned} \rho &\sim \mathcal{Q} \\ Y_n|\rho &\sim \rho, && n = 1, \ldots, N \\ X_n|Y_n &\sim f(\,\cdot\,|Y_n), && n = 1, \ldots, N\,. \end{aligned} \tag{4.1}$$

where $f(\,\cdot\,|y)$ is a probability density function on $\mathbb{R}^d$ (with respect to the Lebesgue measure on $\mathbb{R}^d$) for each $y \in S$. Here, the X_n are observed, while the corresponding Y_n are hidden latent variables.

We now concern ourselves with the selection of the distribution $\mathcal{Q}$. The approach we focus on is to take ρ to be a normalisation of an a.s. discrete random measure φ on S. To this end, we define an operator t, acting on the subset of $\mathbb{M}_S$ containing all μ with $0 < \mu(S) < \infty$, which is given by

$$t(\mu)(B) = \frac{\mu(B)}{\mu(S)}, \qquad \mu \in \mathbb{M}_S,\ 0 < \mu(S) < \infty, \quad B \in \mathcal{S}\,.$$

Now, provided that $0 < \varphi(S) < \infty$ a.s., we can derive a random probability measure $t(\varphi)$ and set $\rho \overset{\mathrm{d}}{=} t(\varphi)$ to define $\mathcal{Q}$. The problem remains to choose such a φ.

From an analytical standpoint, an attractive choice for φ is to let it be a completely random measure that is a.s. discrete and has no fixed atoms. This allows us to exploit Theorem 2.2.7 and write

$$\varphi = \int_0^\infty r\, N(\,\cdot \times \mathrm{d}r) \text{ a.s.}\,,$$

for some Poisson process N on $S \times (0, \infty)$. This is a choice that has been widely explored in the literature (e.g. James et al. (2009)) and posterior inference is tractable.

The framework set out in Beraha et al. (2023) is more general. First, let $\Phi = \sum_{k=1}^{\kappa} \delta_{\sigma_k}$ be any simple point process on S (see the representation in Theorem 2.1.5) with $\kappa < \infty$ a.s. and let $\mathcal{P}_\Phi$ denote its law on $\mathbb{M}_S$. From this, we construct a new simple point process Ψ on $S \times (0, \infty)$ by sampling $\beta_1, \ldots, \beta_\kappa$ from some fixed distribution H on $(0, \infty)$ and setting $\Psi = \sum_{k=1}^{\kappa} \delta_{(\sigma_k, \beta_k)}$. We finally define φ from Ψ according to

$$\varphi = \int_0^\infty r\, \Psi(\,\cdot \times \mathrm{d}r) = \sum_{k=1}^{\kappa} \beta_k \delta_{\sigma_k}\,.$$

Here, φ has inherited the atom locations from a general simple point process Φ on S, with each atom being endowed with a weight sampled independently from H. As in Beraha et al. (2023), we write $\varphi \sim \mathrm{RM}(\mathcal{P}_\Phi, H)$ when φ is distributed as above.

Remark 4.1.1. If Φ is a Poisson process, then so is Ψ, in which case we recover the completely random measure framework of James et al. (2009).

To summarise, the modelling framework set out in Beraha et al. (2023) is specified, given distributions $\mathcal{P}_\Phi$ and H on $\mathbb{M}_S$ and $(0, \infty)$ respectively, by

$$\begin{aligned} \varphi &\sim \mathrm{RM}(\mathcal{P}_\Phi, H) \\ Y_n \mid \varphi &\sim t(\varphi), && n = 1, \ldots, N \\ X_n \mid Y_n &\sim f(\,\cdot\,|Y_n), && n = 1, \ldots, N\,. \end{aligned} \tag{4.2}$$

This model is remarkably general, with the theoretical results of the next section holding for any choice of simple point process Φ and distribution H. We conclude by discussing an interesting choice of H.

Example 4.1.2. Fix $\gamma > 0$ and take $H \sim \mathrm{Gamma}(\gamma, 1)$. Here, $\mathrm{Gamma}(\alpha, \beta)$ for $\alpha > 0$ and $\beta > 0$ means the Gamma distribution with shape-rate parametrisation (see Section 4.1.3). We know that $t(\varphi) = \sum_{k=1}^{\kappa} \tilde{\beta}_k \delta_{\sigma_k}$, where $\tilde{\beta}_k = \frac{\beta_k}{\sum_{c=1}^{\kappa} \beta_c}$ for $k = 1, \ldots, \kappa$. It follows from standard properties of the Gamma distribution that, conditional on κ,

$$(\tilde{\beta}_1, \ldots, \tilde{\beta}_\kappa) \sim \mathrm{Dirichlet}_\kappa(\gamma, \ldots, \gamma)\,.$$

Hence the weights of the mixing distribution are distributed according to a Dirichlet distribution with symmetric parameter. This will be highly relevant in the sequel.

4.1.3. Theoretical results

In this section, we discuss two of the theoretical results of Beraha et al. (2023). To do so, we must first introduce an auxiliary variable U_N. Given $\alpha > 0$ and $\beta > 0$, we write Gamma(α, β) to mean the Gamma distribution with the shape-rate parametrisation, that is the distribution with density

$$f(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0 .$$

where $\Gamma(\,\cdot\,)$ is the univariate Gamma function. Our auxiliary variable then satisfies $U_N|\xi \sim \text{Gamma}\,(N, \xi(S))$. Following the exposition of Beraha et al. (2023), write $\mathbf{Y} = (Y_1, \dots, Y_N)$, and write $\mathcal{P}_\varphi$ for the distribution of φ. Then we have

$$\mathbb{P}(\mathbf{Y} \in \mathrm{d}\mathbf{y}, U_N \in \mathrm{d}u, \varphi \in \mathrm{d}\mu) = \frac{u^{N-1}}{\Gamma(N)} e^{-\mu(S)u} \mathrm{d}u \prod_{n=1}^{N} \mu(\mathrm{d}y_n)\, \mathcal{P}_\varphi(\mathrm{d}\mu) . \tag{4.1}$$

As in Beraha et al. (2023), we characterise a vector $\mathbf{y} \in S^N$ by its L unique values $\mathbf{y}^* = (y_1^*, \dots, y_L^*)$ and by a partition $\pi = \pi(\mathbf{y})$ of $\{1, \dots, N\}$, such that i and j belong to the same partition if and only if $y_i = y_j$. Let also $n_l = \#\{n : y_n = y_l^*\}$ for $l = 1, \dots, L$. Since $(\mathbf{Y}^*, \pi(\mathbf{Y}))$ characterises $\mathbf{Y}$, we can instead write (4.1) as

$$\begin{aligned}\mathbb{P}(\mathbf{Y} \in \mathrm{d}\mathbf{y}, U_N \in \mathrm{d}u, \varphi \in \mathrm{d}\mu) &= \mathbb{P}(\mathbf{Y}^* \in \mathrm{d}\mathbf{y}^*, \pi(\mathbf{Y}) = \pi(\mathbf{y}), U_N \in \mathrm{d}u, \varphi \in \mathrm{d}\mu) \\ &= \frac{u^{N-1}}{\Gamma(N)} e^{-\mu(S)u} \mathrm{d}u \prod_{l=1}^{L} \mu(\mathrm{d}y_l^*)^{n_l}\, \mathcal{P}_\varphi(\mathrm{d}\mu) .\end{aligned}$$

Next, for each $n \in \mathbb{N}$, let $\{\Psi^!_{(\mathbf{y},\mathbf{r})}\}_{(\mathbf{y},\mathbf{r}) \in S^n \times (0,\infty)^n}$ be a family of reduced Palm versions of Ψ, where Ψ is the random measure on $S \times (0, \infty)$ derived from Φ in the previous section. For an arbitrarily fixed $\mathbf{r}_0 \in (0,\infty)^n$, we introduce a collection $\{\tilde{\varphi}_\mathbf{y}\}_{\mathbf{y} \in S^n}$ of random measures on S via

$$\tilde{\varphi}_\mathbf{y} = \int_0^\infty t\ \Psi^!_{(\mathbf{y},\mathbf{r}_0)}(\,\cdot \times \mathrm{d}t) . \tag{4.2}$$

The collection $\{\tilde{\varphi}_\mathbf{y}\}_{\mathbf{y} \in S^n}$ can be thought of as a family of reduced Palm versions of φ (Beraha et al. 2023), although no such notion exists in the theory (φ is not a point process in general and need not be integer-valued).

We require one final definition, which is the family of Laplace transforms of the distribution H given by

$$\kappa(u, n) = \int_0^\infty e^{-ur} r^n H(\mathrm{d}r), \qquad u > 0, \quad n \in \mathbb{N} .$$

These functions will appear in many results for the remainder of this work.

The first result from Beraha et al. (2023) that we choose to highlight is a characterisation of the marginal distribution of $\mathbf{Y}$ under the model.

Theorem 4.1.3. The marginal distribution of $\mathbf{Y}$ is given for $\mathbf{y} \in S^n$ by

$$\begin{aligned}\mathbb{P}(\mathbf{Y} \in \mathrm{d}\mathbf{y}) &= \mathbb{P}(\mathbf{Y}^* \in \mathrm{d}\mathbf{y}^*, \pi(\mathbf{Y}) = \pi(\mathbf{y})) \\ &= \int_0^\infty \frac{u^{N-1}}{\Gamma(N)} \mathcal{L}_{\tilde{\varphi}_{\mathbf{y}^*}}(u) \prod_{l=1}^{L} \kappa(u, n_l)\, \mathrm{d}u\ \mathbb{E}\Phi^L(\mathrm{d}\mathbf{y}^*) .\end{aligned}$$

Proof. See Beraha et al. (2023), Theorem 4.2. □

The second result we choose to highlight is a characterisation of the predictive distribution of a new observation Y_{N+1} under the model, given we already know $\mathbf{Y}$.

Theorem 4.1.4. Fix a non-atomic reference measure $P_0 \in \mathbb{P}_S$. Assume that $\mathbb{E}\Phi^l \ll P_0^l$, such that $\mathbb{E}\Phi^l$ admits a density m_{Φ^l} with respect to P_0^l, for each $l = 1, \ldots, L+1$. Then, for $\mathbf{y} \in S^N$ and $u > 0$, the predictive distribution of Y_{N+1} given $\mathbf{Y} = \mathbf{y}$ and $U_N = u$ is given for $y \in S$ by

$$\begin{aligned}&\mathbb{P}(Y_{N+1} \in \mathrm{d}y | \mathbf{Y} = \mathbf{y}, U_N = u) \\ &\quad \propto \sum_{l=1}^{L} \frac{\kappa(u, n_l + 1)}{\kappa(u, n_l)} \delta_{y_l^*}(\mathrm{d}y) + \kappa(u, 1) \frac{\mathcal{L}_{\tilde{\varphi}_{(\mathbf{y}^*, y)}}(u)}{\mathcal{L}_{\tilde{\varphi}_{\mathbf{y}^*}}(u)} \frac{m_{\Phi^{L+1}}(\mathbf{y}^*, y)}{m_{\Phi^L}(\mathbf{y}^*)} P_0(\mathrm{d}y) \,. \qquad (4.3)\end{aligned}$$

Proof. See Beraha et al. (2023), Theorem 4.3. □

4.1.4. Determinantal point processes

It remains to consider candidates for the simple point process Φ which will share the distribution of its atom locations with our mixture measure. We will focus on the case when S is a compact subset of $\mathbb{R}^p$ for some $p \in \mathbb{N}$ and fix a reference measure χ on S. We consider determinantal point processes, first introduced in Macchi (1975).

Let ξ be a point process on S. Suppose that $\mathbb{E}\xi^{(n)} \ll \chi^n$ for each $n \in \mathbb{N}$. Then, for each $n \in \mathbb{N}$, there exists a χ^n-a.e. unique function $\rho^{(n)} : S^n \to \mathbb{R}$ with

$$\mathbb{E}\xi^{(n)}(\mathrm{d}\mathbf{s}) = \rho^{(n)}(s_1, \ldots, s_n)\, \chi^n(\mathrm{d}\mathbf{s}) \,, \qquad \mathbf{s} = (s_1, \ldots, s_n) \in S^n \,.$$

We call $\rho^{(n)}$ the *n-th order joint intensity* of ξ with respect to χ. The 1st order joint intensity $\rho^{(1)}$ is referred to as the *intensity function* of ξ with respect to χ.

By a *kernel* on S we mean a symmetric, non-negative definite, measurable function $K : S \times S \to \mathbb{R}$. Fixing $n, m \in \mathbb{N}$, we introduce the following notation: for $\mathbf{s} \in S^n$ and $\mathbf{t} \in S^m$, write $K(\mathbf{s}, \mathbf{t})$ for the $n \times m$ real matrix with entries

$$[K(\mathbf{s}, \mathbf{t})]_{ij} = K(s_i, t_j) \,, \qquad i = 1, \ldots, n \,, \quad j = 1, \ldots, m \,.$$

Let ξ be a point process on S. We say that ξ is a *determinantal point process* (DPP) with reference measure χ and kernel K if, for all $n \in \mathbb{N}$, its n-th order joint intensity with respect to χ exists and has

$$\rho^{(n)}(s_1, \ldots, s_n) = \det\left(K(\mathbf{s}, \mathbf{s})\right) \text{ for } \chi^n\text{-almost all } \mathbf{s} = (s_1, \ldots, s_n) \in S^n \,.$$

We write $\xi \sim \mathrm{DPP}(K)$. Henceforth, the dependency on the measure χ will be implicit.

Remark 4.1.5. Fix a kernel K on S. Then a DPP with kernel K is also a DPP with kernel K', where K' is any kernel on S such that

$$K'(s, t) = K(s, t) \text{ for all } s, t \in S_0$$

where $S_0 \in \mathcal{S}$ has $\chi(S \backslash S_0) = 0$. See Baccelli et al. (2020) Section 5.1.2.

We first have the following result, which is essential to our objective of constructing a mixing measure from a DPP.

Theorem 4.1.6. Let ξ be a DPP on S with kernel K. Then ξ is a simple point process.

Proof. See Baccelli et al. (2020), Lemma 5.1.4. □

Example 4.1.7. Let K be a kernel on S such that $K(s,t) = 0$ whenever $s \neq t$. Set $\lambda(s) = K(s,s)$ for $s \in S$. Assume that there exists a DPP ξ with kernel K. Then, for χ^n-almost all $\mathbf{s} = (s_1, \ldots, s_n) \in S^n$,

$$\rho^{(n)}(s_1, \ldots, s_n) = \det\left(K(\mathbf{s},\mathbf{s})\right) = \prod_{i=1}^{n} \lambda(s_i)\,,$$

whence

$$\mathbb{E}\xi^{(n)}(\mathrm{d}s_1 \times \cdots \times \mathrm{d}s_n) = \prod_{i=1}^{n} \lambda(s_i)\,\chi(\mathrm{d}s_1)\cdots\chi(\mathrm{d}s_n) = \prod_{i=1}^{n} \Lambda(\mathrm{d}s_i)\,,$$

where $\Lambda(\mathrm{d}s) = \lambda(s)\,\chi(\mathrm{d}s)$. This characterises the Poisson process on S (see, for example, Daley & Vere-Jones (2008), Example 9.5(d)), hence ξ is a Poisson process on S with intensity measure Λ. Thus the Poisson process is a special case of the DPP.

The most important property of the DPP is its repulsiveness. If ξ is a DPP on S with kernel K, then the atoms of ξ will in general repel one another and spread out to fill S. The exact strength of the repulsion between two locations $s, t \in S$ is characterised via the kernel as $K(s,t)$. To be precise, we compute

$$\rho^{(2)}(s,t) = \det\left(\begin{bmatrix} K(s,s) & K(s,t) \\ K(t,s) & K(t,t) \end{bmatrix}\right) = K(s,s)K(t,t) - K(s,t)^2 \leqslant \rho^{(1)}(s)\rho^{(2)}(t)\,,$$

where the inequality can be strict in the non-Poisson case. Intuitively, the likelihood of two atoms occurring in the same infinitesimal area is smaller than the likelihood if the points were i.i.d. distributed.

Another advantage of using a DPP is that its theoretical properties are well-understood. In particular, a reduced Palm distribution of a DPP is itself a DPP.

Theorem 4.1.8. Let ξ be a DPP on S with kernel K and fix $n \in \mathbb{N}$. Let $\{\xi^!_{\mathbf{s}}\}_{\mathbf{s} \in S^n}$ be a family of n-th reduced Palm versions of ξ. Then, for $\mathbb{E}\xi^{(n)}$-almost all $\mathbf{s} \in S^n$, we have $\xi^!_{\mathbf{s}} \sim \mathrm{DPP}(K_{\mathbf{s}})$, where

$$K_{\mathbf{s}}(u,v) = K(u,v) - K(u,\mathbf{s})K(\mathbf{s},\mathbf{s})^{-1}K(\mathbf{s},v), \qquad u, v \in S\,. \tag{4.4}$$

Proof. See Baccelli et al. (2020), Theorem 5.5.2. □

Remark 4.1.9. In the context of the above theorem, note that $K_{\mathbf{s}}$ exists if and only if $K(\mathbf{s},\mathbf{s})$ is non-singular. However,

$$\mathbb{E}\xi^{(n)}(\{\mathbf{s} \in S^n : \det(K(\mathbf{s},\mathbf{s})) = 0\}) = \int_{\{\mathbf{s} \in S^n : \det(K(\mathbf{s},\mathbf{s}))=0\}} \det(K(\mathbf{s},\mathbf{s}))\,\chi^n(\mathrm{d}\mathbf{s}) = 0\,,$$

so $K(\mathbf{s},\mathbf{s})$ is non-singular for $\mathbb{E}\xi^{(n)}$-almost all $\mathbf{s} \in S^n$. This is sufficient to define the required family of reduced Palm versions.

As it turns out, the class of DPPs is on its own too general to be useful for our novel topic model, so we elect to focus on a subclass of the class of all DPPs. For $J \in \mathbb{N}$, we define a *J-kernel* on S to be a kernel K on S of the form

$$K(s,t) = \sum_{j=1}^{J} \omega_j(s)\omega_j(t)\,, \qquad s, t \in S\,, \tag{4.5}$$

where $\omega_1, \ldots, \omega_J \in L^2(\chi)$ are orthonormal. A DPP ξ on S with kernel K is a *J-DPP* if K is a J-kernel. We write this as $\xi \sim \mathrm{DPP}_J(K)$ to emphasise the dependence on J.

We finally give two important results about J-DPPs. The first establishes their existence and the second that their number of atoms is a.s. non-random. Some additional results, which are not as instructional, are also given in Appendix A.

Theorem 4.1.10. Let K be a J-kernel on S for some $J \in \mathbb{N}$. Then there exists a J-DPP on S with kernel K.

Proof. See Baccelli et al. (2020), Lemma 5.2.3, which gives a construction. □

Theorem 4.1.11. Let ξ be a J-DPP on S for some $J \in \mathbb{N}$. Then $\xi(S) = J$ a.s.. That is, the number of atoms of ξ is J a.s..

Proof. See Hough et al. (2009), Lemma 4.4.1. □

4.2. Formulating a New Model

4.2.1. Motivation

We now return to the topic modelling problem. Recall that we are interested in probabilistic topic models capable of modelling collections of documents and uncovering latent themes in the texts. We saw in Section 3.3 how relaxing some of the assumptions of LDA can pave the way for more nuanced and expressive topic models.

Here, we take an a priori belief distinct from those of the models in Section 3.3. We believe that there are a finite number of topic mixtures associated with the corpus and that the content of each document is generated based on one of these topic mixtures. In this way, the documents all share the same set of possible topic mixtures, with some topic mixtures perhaps being used in the generation of multiple documents and some perhaps not being used in the generation of any documents.

As for the distribution of the finitely many possible topic mixtures, we assume further that they have a tendency to be diverse. That is, we believe a priori that it is much more likely for two given topic mixtures to be different than it is for them to be similar.

The effect of this structure is that the documents in the corpus will be clustered. Documents associated with the same topic mixture will be similar, since their individual words will all have the same marginal distribution. By contrast, documents associated with different topic mixtures will likely have vastly different word compositions.

An example of a corpus which may be well-suited to modelling in this way is a collection of news headlines from a news station. The news station will release headlines on a number of stories, with each story being on a certain subject. The station may choose to release more than one headline based on the same story, in which case we can expect each headline to be on similar subjects. In this way, the headlines will be clustered according to what story they are covering. Moreover, the stories that a news station follows are likely to be very diverse and we might expect a low probability that two stories are on very similar subjects.

In the remainder of this chapter, we introduce a novel topic model capable of modelling such structure. We also discuss some of its theoretical properties, so that they can be compared to those of existing topic models. Methods for posterior inference under this model are also given a treatment at the end of this section.

4.2.2. Model specification and hyperparameter selection

We now move to our novel probabilistic topic model of a corpus $\mathcal{D}$. The model specification draws from all of the theory that has been discussed so far. It inherits the general structure of LDA - in particular it inherits how documents are generated from a topic mixture and a topic matrix. To generate the topic mixtures, the new model leverages the Bayesian mixture framework of Beraha et al. (2023), which has the effect of inducing the clustering behaviour discussed in the motivation. The DPP also makes an appearance in the construction of the mixing measure.

Before stating the full model, we must make some preparations. To specify the model, we will need to define a DPP on $S = \Delta^T$, which we endow with its metric as a subset of $\mathbb{R}^T$. For this, we need a reference measure χ on S, which we take to be the law of the $\text{Dirichlet}_T(\beta)$ distribution with $\beta_t = 1$ for $t = 1, \ldots, T$. Examining (3.1), we see that this is the law of the uniform distribution on Δ^T.

Fix $J \in \mathbb{N}$, which represents the total number of topic mixtures that documents in the corpus can choose from. Let φ be a J-DPP on Δ^T with reference measure χ and J-kernel K. Let finally H_γ be the $\text{Gamma}(\gamma, 1)$ distribution, where $\gamma > 0$ is a hyperparameter.

We are now in a position to state our novel topic model fully. We assume that

$$\begin{aligned}
\phi_t &\sim \text{Dirichlet}_V(\eta)\,, && t = 1, \ldots, T \\
\xi &\sim \text{RM}(\mathcal{P}_\varphi, H_\gamma)\,, && \\
\theta_n | \xi &\sim t(\xi)\,, && n = 1, \ldots, N \\
z_{nm} | \theta_n &\sim \text{Categorical}_T(\theta_n)\,, && m = 1, \ldots, M_n, \quad n = 1, \ldots, N \\
w_{nm} | z_{nm}, \boldsymbol{\phi} &\sim \text{Categorical}_V(\phi_{z_{nm}})\,, && m = 1, \ldots, M_n, \quad n = 1, \ldots, N\,.
\end{aligned}$$

Here, η is a V-dimensional non-negative vector. The full collection of hyperparameters for this model is Dirichlet parameter η, the Gamma distribution parameter γ, and the J-kernel K. A graphical representation of the model is given in Figure 4.1.

The latent variables in this model are readily interpretable. $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_T)$ is the topic matrix, exactly the same as under LDA. Moreover, θ_n is the latent topic mixture associated with document $\mathbf{w}_n$, where $\mathbf{w}_n$ is, conditional on θ_n and $\boldsymbol{\phi}$, generated in the same way as under LDA. The new variable ξ is an a.s. discrete random probability measure. Its atom locations represent the possible topic mixtures which documents in the corpus can inherit and their weights represent the probabilities that a document inherits a given topic mixture.

For the Dirichlet parameter η, we follow Blei et al. (2003) and use a symmetric parameter, with all components equal to some $\eta_0 > 0$. This is to achieve the same exchangeability in the components of the ϕ_t discussed in Section 3.2.2. The choice of η_0 is once again completely subjective, with the effect on the ϕ_t of tuning η_0 being the same as that discussed in Section 3.2.2.

To justify our choice of H, we consider Example 4.1.2, in which we argued that taking $H = H_\gamma$ for some $\gamma > 0$ leads to the weights of the normalised mixing measure having a Dirichlet distribution with symmetric parameter. This is a favourable choice as the properties of the Dirichlet distribution are well-understood. In particular, we know that taking $\gamma \gg 1$ will lead to each topic mixture having similar prior probabilities, whereas taking $\gamma \ll 1$ will lead to some topic mixtures having very high prior probabilities.

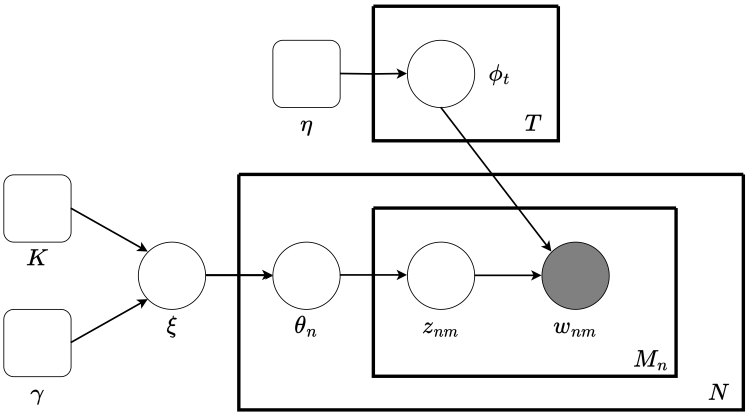

Figure 4.1.: A plate diagram representing the novel topic model.

Since K is a J-kernel, specifying it is a matter of specifying the orthonormal functions $\omega_1, \ldots, \omega_J$ in $L^2(\chi)$ from (4.5). Doing so is somewhat beyond the scope of this work and we discuss it in Chapter 5.

With reference to the motivation underpinning the model, we see that the locations of the J atoms of ξ are members of Δ^T and they represent our full collection of possible topic mixtures. When we sample θ_n from ξ, we are selecting one of these possible topic mixtures to generate our document $\mathbf{w}_n$ from, which coincides with our motivating assumption on how the topic mixtures arise. Moreover, the atom locations of ξ are inherited from a DPP, so they are repulsive and have a tendency to spread out. This corresponds to our a priori belief that the topic mixtures have a tendency to be diverse. Our hope is that these two effects will combine to induce the well-separated clustering behaviour that we targeted in the motivation.

A crucial observation is that, if we marginalise over the z_{nm} (for fixed n), the likelihood of the document $\mathbf{w}_n$ can be expressed purely in terms of θ_n and $\boldsymbol{\phi}$. If we denote this density by $f(\,\cdot\,|\theta_n, \boldsymbol{\phi})$, then we can write this topic model as

$$\begin{aligned} \phi_t &\sim \text{Dirichlet}_V(\eta)\,, && t = 1, \ldots, T \\ \xi &\sim \text{RM}(\mathcal{P}_\varphi, H_\gamma)\,, && \\ \theta_n|\xi &\sim t(\xi)\,, && n = 1, \ldots, N \\ \mathbf{w}_n &\sim f(\,\cdot\,|\theta_n, \boldsymbol{\phi})\,, && n = 1, \ldots, N\,. \end{aligned}$$

Hence, conditional on the additional parameter $\boldsymbol{\phi}$, this model takes the form of that discussed in Section 4.1. This will provide the main avenue through which we can investigate the model's theoretical properties in the proceeding section.

4.2.3. Theoretical properties

In this section, we state the two main results of this work, giving analytical forms of the marginal and predictive distributions of the latent topics $\boldsymbol{\theta}$ under our novel model. We also discuss some important special cases of these results.

We first seek the form of the marginal distribution of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. This will be our main point of comparison with the existing topic models, so it is important that we understand it. We write $\boldsymbol{\theta}^*$ for the L unique components of $\boldsymbol{\theta}$.

Theorem 4.2.1. Assume that $J \geqslant L$. Then, for $\mathbf{y} \in (\Delta^T)^{\otimes N}$, the marginal distribution of the vector $\boldsymbol{\theta}$ of topic mixtures is given by

$$\begin{aligned}\mathbb{P}(\boldsymbol{\theta} \in \mathrm{d}\mathbf{y}) &= \mathbb{P}(\boldsymbol{\theta}^* \in \mathrm{d}\mathbf{y}^*, \pi(\boldsymbol{\theta}) = \pi(\mathbf{y})) \\ &= \frac{1}{\Gamma(N)} \prod_{l=1}^{L} \frac{\Gamma(n_l+\gamma)}{\Gamma(\gamma)} \int_0^\infty \frac{u^{N-1}}{(u+1)^{\gamma J+N}} \, \mathrm{d}u \, \det\left(K(\mathbf{y}^*, \mathbf{y}^*)\right) \chi^L(\mathrm{d}\mathbf{y}^*) \,. \end{aligned} \tag{4.6}$$

Proof. See Appendix C. □

This rather complicated formula demands a few remarks, as there are two special cases that are of independent interest.

Remark 4.2.2. The first terms of (4.6) depend on $\mathbf{y}$ only through $\pi(\mathbf{y})$. Thus

$$\mathbb{P}(\boldsymbol{\theta}^* \in \mathrm{d}\mathbf{y}^*) \propto \det\left(K(\mathbf{y}^*, \mathbf{y}^*)\right) \chi^L(\mathrm{d}\mathbf{y}^*) \,.$$

Hence the kernel K is the only model parameter that controls the marginal distribution of the locations of the topic mixtures.

Remark 4.2.3. To find the marginal distribution of one topic mixture under the model, we can set $N = 1$. This implies that $L = 1$ and $n_1 = 1$. Now

$$\begin{aligned}\mathbb{P}(\theta_1 \in \mathrm{d}y) &= \frac{\Gamma(\gamma+1)}{\Gamma(\gamma)} \int_0^\infty \frac{1}{(u+1)^{\gamma J+1}} \, \mathrm{d}u \, \det\left(K(y,y)\right) \chi(\mathrm{d}y) \\ &= \gamma \int_0^\infty (u+1)^{-(\gamma J+1)} \, \mathrm{d}u \, K(y,y) \chi(\mathrm{d}y) \\ &= \frac{1}{J} K(y,y) \chi(\mathrm{d}y) \,. \end{aligned}$$

Hence the marginal distribution of a single topic θ_1 has density with respect to χ, and moreover this density is proportional to $y \mapsto K(y,y)$.

We now move on to the predictive distribution of θ_{N+1} given $\theta_1, \ldots, \theta_N$. This will offer us another point of comparison with the existing topic models.

Theorem 4.2.4. Assume that $J > L$. Then the predictive distribution of a new topic mixture θ_{N+1} given the vector $\boldsymbol{\theta}$ of observed topic mixtures is specified for $y \in \Delta^T$ as:

1. If $y = y_l^*$ for some $l = 1, \ldots, L$, then
$$\mathbb{P}(\theta_{N+1} = y | \boldsymbol{\theta} = \mathbf{y}, U_N = u) \propto \frac{n_l + \gamma}{u+1} \,.$$
2. If $y \neq y_l^*$ for all $l = 1, \ldots, L$, then
$$\mathbb{P}(\theta_{N+1} \in \mathrm{d}y | \boldsymbol{\theta} = \mathbf{y}, U_N = u) \propto \frac{\gamma}{(u+1)^\gamma} K_{\mathbf{y}^*}(y,y) \, \chi(\mathrm{d}y) \,.$$
where $K_{\mathbf{y}^*}$ is the kernel defined in (4.4).

Moreover, all of the constants of proportionality above are equal.

Proof. See Appendix D. □

Remark 4.2.5. Consider the case $y \neq y_l^*$ for all $l = 1, \ldots, L$. After integrating out u, the distribution of θ_{N+1} is the same (up to a constant) as the marginal distribution found in Remark 4.2.3 but with $K_{\mathbf{y}}$ in place of K. Since $K_{\mathbf{y}}$ appears in the specification of the reduced Palm distributions of Φ at $\mathbf{y}$, we can interpret this difference as θ_{N+1} having a tendency to be spread apart from the y_n.

4.3. Inference via Gibbs Sampling

The cost of the sophistication of this model is that posterior inference is difficult. While implementation of posterior inference procedures is beyond the scope of this work, we will discuss in this section some potential approaches.

It is tempting to perform variational inference on the novel model, as this is the method of choice for LDA and many of its extensions. However, in replacing the Dirichlet prior on $\boldsymbol{\theta}$ with our repulsive mixture component, we have lost the Dirichlet-multinomial conjugacy that made this approach so tractable. In this way, we surrender the ability to use the variational Bayes algorithms developed for the models in Chapter 3.

A more promising approach is MCMC. As remarked earlier in the chapter, if we condition on the topic matrix $\boldsymbol{\phi}$, then our model follows the framework of Beraha et al. (2023) exactly. Hence any Gibbs sampling procedure that can be used for this framework is a candidate to be applied to our novel model, provided an additional step is added to sample from the conditional distribution of $\boldsymbol{\phi}$. It is this observation that we will leverage to propose posterior inference schemes under our model.

In the next two subsections, we discuss two broad classes of Gibbs sampler that we could employ for inference under our model. Following Beraha et al. (2023), we term these "conditional samplers" and "marginal samplers". A conditional sampler is designed to sample from the full joint posterior distribution of the model, i.e.

$$\mathbb{P}(U_N \in \mathrm{d}u, \xi \in \mathrm{d}\mu, \boldsymbol{\theta} \in \mathrm{d}\mathbf{y}, \boldsymbol{\phi} \in \mathrm{d}\boldsymbol{\nu} | \mathcal{D}) \,, \quad u > 0\,,\ \mu \in \mathbb{M}_S\,,\ \mathbf{y} \in (\Delta^T)^N\,,\ \boldsymbol{\nu} \in (\Delta^V)^T\,. \tag{4.7}$$

A marginal sampler instead works by integrating out the latent mixing measure ξ before the posterior sampling distributions are computed. That is, it samples from the joint posterior distribution of the form

$$\mathbb{P}(U_N \in \mathrm{d}u, \boldsymbol{\theta} \in \mathrm{d}\mathbf{y}, \boldsymbol{\phi} \in \mathrm{d}\boldsymbol{\nu} | \mathcal{D}) \,, \quad u > 0\,,\ \mathbf{y} \in (\Delta^T)^N\,,\ \boldsymbol{\nu} \in (\Delta^V)^T\,. \tag{4.8}$$

Note that, by using a marginal algorithm over a conditional one, we forfeit the ability to sample from the posterior of the mixing measure ξ.

4.3.1. Conditional algorithms

Section 5.1 of Beraha et al. (2023) provides a sampling algorithm able to sample from the full joint posterior of their Bayesian mixture model. If we were to apply it to our topic model, we could then obtain samples from the joint distribution

$$\mathbb{P}(U_n \in \mathrm{d}u, \xi \in \mathrm{d}\mu, \boldsymbol{\theta} \in \mathrm{d}\mathbf{y} | \boldsymbol{\phi}, \mathcal{D}) \,, \qquad u > 0\,, \quad \mu \in \mathbb{M}_S\,, \quad \mathbf{y} \in (\Delta^T)^N\,. \tag{4.9}$$

Combining this with a simple Metropolis-Hastings MCMC algorithm to sample from

$$\mathbb{P}(\boldsymbol{\phi} \in \mathrm{d}\boldsymbol{\nu} | U_N, \xi, \boldsymbol{\theta}, \mathcal{D}) \,, \qquad \boldsymbol{\nu} \in (\Delta^V)^T\,,$$

we would have a full Gibbs algorithm able to sample from (4.7).

The issue arises in sampling from the distribution (4.9). The step which updates the current value of ξ by sampling from the conditional distribution

$$\mathbb{P}(\xi \in \mathrm{d}\mu | U_N, \boldsymbol{\theta}, \boldsymbol{\phi}, \mathcal{D}) \,, \qquad \mu \in \mathbb{M}_S\,,$$

is very difficult to implement. In short, it involves evaluating a Laplace functional (see Beraha et al. (2023), equation (13)) which has no easy closed form. Whether it is possible to overcome this and implement the algorithm remains an unanswered question.

There are other candidate algorithms that could be adapted to sample from (4.9). For example, Xie & Xu (2020) give a reversible-jump MCMC algorithm for posterior sampling under a similar DPP mixture model. Although this model assumes Gaussian observations (with the latent variables being their mean parameter), it may be possible to adapt their approach to a topic modelling setting.

Many conditional MCMC methods will require sampling from a DPP. A suite of both exact and approximate algorithms exist for this purpose - see Lavancier & Rubak (2023) for a very recent survey. Interestingly (and rather fortunately), a particularly efficient algorithm exists for sampling from the reduced Palm distributions of a J-DPP.

4.3.2. Marginal algorithms

The marginal algorithm provided in Section 5.2 Beraha et al. (2023) trades the ability to sample from the posterior distribution of the mixing measure in order to obtain more tractable update steps. If we are able to apply it to our topic model, then we would be able to sample from the distribution

$$\mathbb{P}(U_N \in \mathrm{d}u, \boldsymbol{\theta} \in \mathrm{d}\mathbf{y} | \boldsymbol{\phi}, \mathcal{D}) \,, \qquad u > 0 \,, \quad \mathbf{y} \in (\Delta^T)^N \,. \tag{4.10}$$

As in the conditional case, if we were to then combine this with an algorithm to sample from the conditional distribution

$$\mathbb{P}(\boldsymbol{\phi} \in \mathrm{d}\boldsymbol{\nu} | U_N, \boldsymbol{\theta}, \mathcal{D}) \,, \qquad \boldsymbol{\nu} \in (\Delta^V)^T \,.$$

then this will leave us with a complete Gibbs sampler, able to give samples from (4.8).

Unlike in the conditional case, sampling from the distribution (4.10) is much more feasible. Indeed, as proposed in Beraha et al. (2023), we can utilise Theorem 4.1.4 (which specialises to Theorem 4.2.4 under our model) to sample for each $n = 1, \ldots, N$ from the conditional distribution

$$\mathbb{P}(\theta_n \in \mathrm{d}y | U_N, \theta_1, \ldots, \theta_{n-1}, \theta_{n+1}, \ldots, \theta_N, \boldsymbol{\phi}) \,, \qquad y \in \Delta^T \,.$$

Sampling from the distribution of U_N conditional on the other variables is also straightforward. Putting all of these pieces together would give us our marginal sampler.

The disadvantage of this approach is that we learn nothing about the posterior distribution of the mixing measure ξ. This is somewhat disappointing, as we cannot learn the locations of all the atoms of ξ. In particular, we can only the composition of the topic mixtures that are associated with at least one document in the corpus. However, we are still able to obtain posterior samples for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, which are the more important targets for inference under this model.

4.4. Comparison with Existing Models

4.4.1. With LDA

The novel topic model borrows many aspects from LDA. Most importantly, both models use the same method for generating documents by combining a topic matrix and a topic mixture. This allows documents to exhibit multiple topics in different proportions, which was the main innovation of LDA when it was introduced. The key innovation of our model is the inclusion of a new layer in the hierarchy, allowing topics to be shared between documents in a way that they cannot be under LDA.

Some effects of this innovation on the theoretical behaviour of the model is clear upon analysis of the marginal distributions of the topic mixtures. Under LDA, the components of $\boldsymbol{\theta}^*$ are i.i.d., which is not the case for our novel model. Indeed, we see from Remark 4.2.2 that the marginal density of $\boldsymbol{\theta}^*$ is proportional to $\mathbf{y}^* \mapsto \det(K(\mathbf{y}^*, \mathbf{y}^*))$, which does not factorise. In fact, the components $\boldsymbol{\theta}^*$ tend to repel one another, which is captured most intuitively in the following. Take $L = 2$ and assume that K is continuous (with respect to the inner product on Δ^T). Then $\det\left(K\left((y, y), (y, y)\right)\right) = 0$ for any $y \in \Delta^T$, so $\lim_{y_1^* \to y_2^*} \mathbb{P}(\theta_1^* \in \mathrm{d}y_1^*, \theta_2^* \in \mathrm{d}y_2^*) = 0$. It follows that

$$\mathbb{P}(\theta_1^* \in \mathrm{d}y_1^*, \theta_2^* \in \mathrm{d}y_2^*) < \mathbb{P}(\theta_1^* \in \mathrm{d}y_1^*)\mathbb{P}(\theta_2^* \in \mathrm{d}y_2^*)$$

for all $y_1^*, y_2^* \in S$ sufficiently close. Intuitively, distinct topic mixtures under our novel topic are less likely to be located near to each other than under LDA.

We also cannot ignore the fact that the novel model assigns positive probability to documents sharing topic mixtures, whereas LDA does not. As has been discussed already, this leads to a clustering effect, with documents in the same cluster sharing the same topic mixture. Under LDA, many of the latent topic mixtures may be very similar, leading to a high number of redundant latent variables. Our model penalises this undesirable behaviour by encouraging documents to share topic mixtures instead. The result is a model which is more interpretable, as similar documents are grouped together in a clear way, and the topic mixtures are well-spaced.

The novel model also offers unprecedented control over the marginal distribution of a single topic mixture. Under LDA the topic mixtures are all distributed according to a Dirichlet distribution, whose density is parametrised by the T-dimensional parameter α. By contrast, in 4.2.3 we saw that the marginal density of a single topic mixture under the novel model has distribution with density $y \mapsto \frac{1}{J}K(y, y)$. This makes our novel model much more expressive, as this marginal distribution can vary dramatically with the nonparametric model parameter K. We can in theory tune the kernel K so that the marginal density of a single topic mixture has a desirable density, rather than being restricted to the finite-dimensional class of Dirichlet distributions, as under LDA.

This innovation also has a crucial effect on the predictive abilities of the model. Under LDA, the predictive posterior distribution of θ_{N+1} given $\boldsymbol{\theta}$ is the same as the marginal distribution of a single topic mixture, that is a Dirichlet distribution. This is independent of $\boldsymbol{\theta}$, rendering LDA completely unable to predict new topic mixtures given the set of topic mixtures associated with documents in the corpus. Under the novel topic model, however, we have a non-trivial predictive posterior distribution. This distribution assigns positive probability to each of these two cases:

1. θ_{N+1} is equal exactly to one of the existing topic mixtures $\theta_1^*, \ldots, \theta_L^*$. Intuitively, the model is respecting the possibility that a new document will be very similar to an existing document in the corpus, so much so that it should be placed in an existing cluster. In this way, the observed topic mixtures $\boldsymbol{\theta}$ are incorporated heavily into the predictive posterior.

2. θ_{N+1} is equal to a new topic mixture. We argued in Remark 4.2.5 that this topic mixture is encouraged to be spaced apart from the existing topic mixtures $\theta_1^*, \ldots, \theta_L^*$. When the model makes a prediction about a topic mixture not observed, it tends to choose one that is very dissimilar to those already observed.

However, what the novel topic model gains in flexibility, it loses in tractability of the posterior. While approximate posterior inference under LDA is easy via a variational Bayes approach, posterior inference under the novel topic model is difficult.

4.4.2. With other topic models

The DPP-LDA model discussed in Section 3.3.1 also utilises a DPP prior to enhance LDA. This model, however, assumes instead that the topics themselves have a repulsive property, as opposed to the topic mixtures. Moreover, DPP-LDA models the topics as the atoms of the DPP itself, rather than using the DPP to construct a mixture measure and then sampling the topics from this. Although this model is very different from our novel topic model, it is interesting to see how the repulsiveness of the DPP can be introduced in various ways and at various points in the Bayesian hierarchy of LDA to achieve different effects. It is also worth noting that introducing a DPP-based prior on the topic matrix does not alter posterior inference in any major way, whereas under our model the introduction of the DPP forces us to resort to MCMC procedures.

We saw in Section 3.3.2 how some of the shortcomings of LDA can be addressed by introducing a more flexible prior on the topic distributions. In particular, by using a logistic-normal distribution in place of the Dirichlet distribution as the prior on $\boldsymbol{\theta}$, the model is able to specify a much more flexible prior distribution. In terms of flexibility in the marginal distribution of the topic mixtures, however, our model is superior. The logistic-normal distribution that governs these marginal distributions is still specified by finite-dimensional parameters (namely a mean vector and covariance matrix). By contrast, our model allows near-complete flexibility, since we can specify the marginal density function exactly. In terms of posterior tractability, the situation is similar to that with LDA. While approximate posterior inference under the CTM is feasible via a variational Bayes approach (Blei & Lafferty 2005), posterior inference under the novel topic model is highly impractical.

The hierarchical Dirichlet process approach discussed in Section 3.3.3 strives to improve upon LDA by introducing a nonparameteric mixture component. This is also what our novel model is aiming for, but the two resulting models are in fact very different. Most importantly, under the HDP model it is not possible for two different documents to share the same latent topic distribution G_n, whereas this is an integral feature of our model. Moreover, the equivalent of the latent topic mixtures under the HDP model (i.e. the weights for the Dirichlet distributions G_n) do not share the same repulsive property as under the novel topic model and are sampled independently. An important similarity between the HDP approach and ours, however, is the need to use MCMC procedures for posterior inference.

5. Conclusion

In this work, we have proposed a novel topic model which utilises a Bayesian repulsive mixture component. We have given an exposition of all of the required theory to make full use of the framework of Beraha et al. (2023) and showed how it is used to construct our model. We have also seen how our novel topic model compares to other similar probabilistic topic models.

The first contribution of this work is the textbook-style first chapter, which gives a gentle introduction to the theory of random measures. This chapter is heavily inspired by popular references on the subject, but with the results carefully curated and only the most important points included. Given an individual with sufficient knowledge of measure-theoretic probability theory, following this chapter should give them a working knowledge of the theory of random measures. In particular, it will be more than sufficient to appreciate the following chapters of this work and the references therein.

The second contribution of this work is a survey of some modern probabilistic topic models. We first discuss LDA, the most canonical probabilistic topic model, before moving on to some of its extensions. These extensions each relax the assumptions of LDA in different and inventive ways, leading to a variety of behaviours.

The third and most important contribution of this work is the specification and Bayesian analysis of our novel topic model. The repulsive Bayesian mixture framework that our novel model utilises has never before been applied in a topic modelling setting to the best of the writer's knowledge, so the proposed model is a genuine innovation. This becomes even more clear when it is compared to a suite of existing topic models - we argue that our novel model is very distinct from all of these.

There is still significant scope for future work in this area. The most notable obstacle that this work has not addressed is the selection of a kernel for the DPP prior in the novel topic model. Given that this kernel must be a J-kernel and must be defined on the simplex (as opposed to $\mathbb{R}^d$), constructing concrete examples is highly non-trivial. Moreover, the efficacy of the model when applied to real-world datasets, in particular compared to existing probabalistic topic models, has not been confirmed empirically. Time constraints for this research staved off any opportunity to refine and implement the MCMC algorithms discussed - this certainly warrants further investigation.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII-1983*, pp. 1–198.

Baccelli, F., Blaszczyszyn, B. & Karray, M. (2020). Random Measures, Point Processes, and Stochastic Geometry. Inria, 2020. `https://inria.hal.science/hal-02460214`.

Beraha, M., Argiento, R., Camerlenghi, F. & Guglielmi, A. (2023). Normalized random meaures with interacting atoms for bayesian nonparametric mixtures. *arXiv.* [Preprint] `https://arxiv.org/abs/2302.09034` [Accessed 06/09/2023].

Blei, D. M. & Lafferty, J. D. (2005). Correlated topic models. In *Advances in Neural Information Processing Systems*, Vol. 18, pp. 147–154.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(4-5), 993–1022.

Daley, D. J. & Vere-Jones, D. (2008). *An Introduction to the Theory of Point Processes Volume II: General Theory and Structure.* 2nd ed. New York, NY, Springer New York.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, pp. 5228–5235.

Hoffman, M. D., Blei, D. M. & Bach, F. (2010). Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, Vol. 23, pp. 856–864.

Hough, J., Krishnapur, M., Peres, Y. & Virág, B. (2009). Zeros of gaussian analytic functions and determinantal point processes. Provided by American Mathematical Society.

James, L. F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1), 76–97.

Kallenberg, O. (2017). *Random Measures, Theory and Applications.* 1st ed. Cham, Springer International Publishing.

Lavancier, F. & Rubak, E. (2023). On simulation of continuous determinantal point processes. *Statistics and Computing*, 33(5).

Macchi, O. (1975). The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7, 83–122.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566–1581.

Xie, F. & Xu, Y. (2020). Bayesian repulsive gaussian mixture model'. *Journal of the American Statistical Association*, 115(529), 187–203.

Zou, J. Y. & Adams, R. P. (2012). Priors for diversity in generative latent variable models. In *Advances in Neural Information Processing Systems*, Vol. 25, pp. 2996–3004.

6. Appendices

A. Additional properties of DPPs

There are a few more properties of the DPP that were omitted for sake of brevity, which we instead discuss here. We inherit the context and notation from Section 4.1.3.

Lemma A.1. Let ξ be a DPP on S with reference measure χ and kernel K. Fix $n \in N$ and let $\mathbf{s} = (s_1, \ldots, s_n) \in S^{(n)}$, where $S^{(n)}$ is the non-diagonal part of S^n. Then the n-th moment measure of ξ is evaluated at $\mathbf{s}$ is

$$\mathbb{E}\xi^n(\mathrm{d}\mathbf{s}) = \det\left(K(\mathbf{s}, \mathbf{s})\right)\chi(\mathrm{d}\mathbf{s})\,.$$

Proof. Studying the definitions of ξ^n and $\xi^{(n)}$, we see that they coincide on $S^{(n)}$ a.s.. Hence we have

$$\mathbb{E}\xi^n(\mathrm{d}\mathbf{s}) = \mathbb{E}\xi^{(n)}(\mathrm{d}\mathbf{s}) = \det\left(K(\mathbf{s}, \mathbf{s})\right)\chi(\mathrm{d}\mathbf{s})$$

as required. □

Lemma A.2. For $J \in \mathbb{N}$, let ξ be a J-DPP on S with reference measure χ and J-kernel K. Fix $n \in \mathbb{N}$ and let $\{\xi^!_{\mathbf{s}}\}_{\mathbf{s} \in S^n}$ be a family of n-th reduced Palm versions of ξ. Then

$$\xi^!_{\mathbf{s}}(S) = J - n \text{ a.s. for } \mathbb{E}\xi^{(n)}\text{-almost all } \mathbf{s} \in S^n\,.$$

Proof. Let $A = \{(\mathbf{s}, \mu) \in S^n \times \mathbb{M}_S : \mu(S) \neq J - n\}$. Then $A \in \mathcal{S}^{\otimes n} \otimes \mathcal{M}_S$, so $(\mathbf{s}, \mu) \mapsto 1_A(\mathbf{s}, \mu)$ is measurable on $S^n \times \mathbb{M}_S$. Hence, by the Campbell-Little-Mecke formula for reduced Palm measures (Theorem 2.4.12), we have

$$\begin{aligned}
\int_{S^n} \mathbb{P}((\mathbf{s}, \xi^!_{\mathbf{s}}) \in A)\,\mathbb{E}\xi^{(n)}(\mathrm{d}\mathbf{s}) &= \int_{S^n} \mathbb{E}[1_A(\mathbf{s}, \xi^!_{\mathbf{s}})]\,\mathbb{E}\xi^{(n)}(\mathrm{d}\mathbf{s}) \\
&= \mathbb{E}\left[\int_{S^n} 1_A\left(\mathbf{s}, \xi - \sum_{i=1}^n \delta_{s_i}\right)\xi^{(n)}(\mathrm{d}\mathbf{s})\right] \\
&= \mathbb{E}\left[\int_{S^n} 1\left\{\left[\xi - \sum_{i=1}^n \delta_{s_i}\right](S) \neq J - n\right\}\xi^{(n)}(\mathrm{d}\mathbf{s})\right] \\
&= \mathbb{E}\left[\int_{S^n} 1\left\{\xi(S) - n \neq J - n\right\}\xi^{(n)}(\mathrm{d}\mathbf{s})\right] \\
&= \mathbb{E}\left[1\left\{\xi(S) \neq J\right\}\xi^{(n)}(S^n)\right] \\
&= 0
\end{aligned}$$

since $\xi(S) = J$ a.s. by Theorem 4.1.11. It follows that, for $\mathbb{E}\xi^{(n)}$-almost all $\mathbf{s} \in S^n$, $(\mathbf{s}, \xi^!_s) \notin A$ a.s., or equivalently $\xi_{\mathbf{s}}(J) = J - n$ a.s.. □

B. Some lemmas

Here we establish some necessary results to prove the main theorems. For this section, we inherit context and notation from Section 4.1.

Lemma B.1. Let Φ be a simple point process on S and let H be a distribution on $(0, \infty)$. Suppose that a random measure φ on S has $\varphi \sim \mathrm{RM}(\mathcal{P}_\Phi, H)$. For $n \in \mathbb{N}$, let $\{\tilde{\varphi}_{\mathbf{y}}\}_{\mathbf{y} \in S^n}$ be the collection of random measures defined in (4.2). Then

$$\tilde{\varphi}_{\mathbf{y}} \sim \mathrm{RM}(\mathcal{P}_{\Phi^!_{\mathbf{y}}}, H) \text{ for } \mathbb{E}\Phi^{(n)}\text{-almost all } \mathbf{y} \in S^n ,$$

where $\{\Phi^!_{\mathbf{y}}\}_{\mathbf{y} \in S^n}$ is a family of n-th reduced Palm versions of Φ.

Proof. Consider the family $\{\Psi^!_{(\mathbf{y}, \mathbf{r}_0)}\}_{\mathbf{y} \in S^n}$ appearing in (4.2). For $\mathbb{E}\Phi^{(n)}$-almost all $\mathbf{y} \in S^n$, Beraha et al. (2023) Lemma B.2 asserts that the distribution of $\Psi^!_{(\mathbf{y}, \mathbf{r}_0)}$ can be constructed by equipping each atom of $\Phi^!_{\mathbf{y}}$ with an independent draw from H. But now, considering (4.2), we see that $\tilde{\varphi}_{\mathbf{y}}$ is constructed from $\mathcal{P}_{\Phi^!_{\mathbf{y}}}$ and H according to the regime set out in Section 4.1.2. □

Lemma B.2. Let Φ be a simple point process on S and let H be a distribution on $(0, \infty)$. Suppose that a random measure φ on S has $\varphi \sim \mathrm{RM}(\mathcal{P}_\Phi, H)$. Then

$$\mathcal{L}_\varphi(f) = \mathbb{E}\left[\exp\left\{\int_S \log \psi(f(s))\, \Phi(\mathrm{d}s)\right\}\right], \qquad f \in \mathcal{F}_+(S),$$

where $\psi(x) = \int_0^\infty e^{-rx}\, H(\mathrm{d}x)$ for $x \in \mathbb{R}$ is the Laplace transform of H.

Proof. See Beraha et al. (2023), Lemma B.3. □

Lemma B.3. Let H be the Gamma$(\gamma, 1)$ distribution for some $\gamma > 0$. Then

$$\psi(u) = \frac{1}{(u+1)^\gamma}, \quad \kappa(u, n) = \frac{\Gamma(n+\gamma)}{\Gamma(\gamma)(u+1)^{n+\gamma}}, \qquad u > 0,\ n \in \mathbb{N}.$$

Proof. The form of ψ follows immediately from the Laplace transform of the Gamma distribution. For the form of κ, we have

$$\begin{aligned}
\kappa(u, n) &= \int_0^\infty e^{-ur} r^n \cdot \frac{1}{\Gamma(\gamma)} r^{\gamma-1} e^{-r}\, \mathrm{d}r \\
&= \frac{1}{\Gamma(\gamma)} \int_0^\infty r^{n+\gamma-1} e^{-(u+1)r}\, \mathrm{d}r \\
&= \frac{1}{\Gamma(\gamma)} \frac{\Gamma(n+\gamma)}{(u+1)^{n+\gamma}},
\end{aligned}$$

as required. □

C. Proof of Theorem 4.2.1

Proof of Theorem 4.2.1. Proving this theorem is simply a matter of applying Theorem 4.1.3 in our specific setting. First, we note that the forms of ψ and κ are as given in Lemma B.3. Next, we note that $\tilde{\varphi}_{\mathbf{y}*} \sim \mathrm{RM}(\mathcal{P}_{\Phi^!_{\mathbf{y}*}}, H_\gamma)$ by Lemma B.1. Thus, applying Lemma B.2 and then Lemma A.2, we obtain

$$\begin{aligned}
\mathcal{L}_{\tilde{\varphi}_{\mathbf{y}*}}(u) &= \mathbb{E}\left[\exp\left\{\int_{\Delta^T} \log\psi(u)\,\Phi^!_{\mathbf{y}*}(\mathrm{d}s)\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{\Phi^!_{\mathbf{y}*}(\Delta^T)\log\psi(u)\right\}\right] \\
&= \mathbb{E}\left[\exp\left\{(J-L)\log\psi(u)\right\}\right] \\
&= \psi(u)^{J-L} \\
&= \frac{1}{(u+1)^{\gamma(J-L)}}\,.
\end{aligned}$$

Now it follows that

$$\begin{aligned}
&\int_0^\infty \frac{u^{N-1}}{\Gamma(N)}\mathcal{L}_{\tilde{\varphi}_{\mathbf{y}*}}(u)\prod_{l=1}^{L}\kappa(u,n_l)\,\mathrm{d}u \\
&\quad= \int_0^\infty \frac{u^{N-1}}{\Gamma(N)}\cdot\frac{1}{(u+1)^{\gamma(J-L)}}\cdot\prod_{l=1}^{L}\frac{\Gamma(n_l+\gamma)}{\Gamma(\gamma)(u+1)^{n_l+\gamma}}\,\mathrm{d}u \\
&\quad= \frac{1}{\Gamma(N)}\prod_{l=1}^{L}\frac{\Gamma(n_l+\gamma)}{\Gamma(\gamma)}\int_0^\infty u^{N-1}\cdot\frac{1}{(u+1)^{\gamma(J-L)}}\cdot\frac{1}{(u+1)^{\sum_{l=1}^{L}(n_l+\gamma)}}\,\mathrm{d}u \\
&\quad= \frac{1}{\Gamma(N)}\prod_{l=1}^{L}\frac{\Gamma(n_l+\gamma)}{\Gamma(\gamma)}\int_0^\infty \frac{u^{N-1}}{(u+1)^{\gamma(J-L)+N+\gamma L}}\,\mathrm{d}u \\
&\quad= \frac{1}{\Gamma(N)}\prod_{l=1}^{L}\frac{\Gamma(n_l+\gamma)}{\Gamma(\gamma)}\int_0^\infty \frac{u^{N-1}}{(u+1)^{\gamma J+N}}\,\mathrm{d}u\,.
\end{aligned}$$

Appealing to Lemma A.1 gives the form for $\mathbb{E}\Phi^L(\mathrm{d}\mathbf{y}^*)$, since, by construction, the components of $\mathbf{y}^*$ are distinct. We can now apply Theorem 4.1.3, combining these two formula to obtain the result. □

D. Proof of Theorem 4.2.4

Proof of Theorem 4.2.4. As above, we evaluate each term in (4.3) in our specific setting. In particular, we can take P_0 to be χ, which is indeed a non-atomic probability measure on Δ^T. Note that the functions ψ and κ are as given in Lemma B.3. Let finally α be the constant of proportionality in (4.3).

Now suppose that $y = y_l^*$ for some $l = 1, \ldots, L$. Then

$$\begin{aligned}\mathbb{P}(\theta_{N+1} = y | \boldsymbol{\theta} = \mathbf{y}, U_N = u) &= \int_{\{y\}} \mathbb{P}(\theta_{N+1} \in \mathrm{d}z | \boldsymbol{\theta} = \mathbf{y}, U_N = u)\, \mathrm{d}z \\ &= \alpha \cdot \frac{\kappa(u, n_l + 1)}{\kappa(u, n_l)} \\ &= \alpha \cdot \frac{\Gamma(\gamma)}{\Gamma(\gamma)} \cdot \frac{\Gamma(n_l + \gamma + 1)}{\Gamma(n_l + \gamma)} \cdot \frac{(u+1)^{n_l + \gamma}}{(u+1)^{n_l + \gamma + 1}} \\ &= \alpha \cdot \frac{n_l + \gamma}{u + 1}\,.\end{aligned}$$

Suppose alternatively that $y \neq y_l^*$ for all $l = 1, ..., L$, which we note is possible since $J > L$. Then the first term in (4.3) vanishes, leaving

$$\mathbb{P}(\theta_{N+1} \in \mathrm{d}y | \boldsymbol{\theta} = \mathbf{y}, U_N = u) = \alpha \cdot \kappa(u, 1) \frac{\mathcal{L}_{\tilde{\varphi}_{(\mathbf{y}^*, y)}}(u)}{\mathcal{L}_{\tilde{\varphi}_{\mathbf{y}^*}}(u)} \frac{m_{\Phi^{L+1}}(\mathbf{y}^*, y)}{m_{\Phi^L}(\mathbf{y}^*)} \chi(\mathrm{d}y)\,. \tag{6.1}$$

To evaluate this, we first note that, for $u > 0$,

$$\mathcal{L}_{\tilde{\varphi}_{\mathbf{y}^*}}(u) = \psi(u)^{J-L}, \qquad \mathcal{L}_{\tilde{\varphi}_{(\mathbf{y}^*, y)}}(u) = \psi(u)^{J-L-1}$$

by the same argument as in the preceding proof. Also, the definition of the DPP yields $m_{\Phi^L}(\mathbf{y}^*) = \det(K(\mathbf{y}^*, \mathbf{y}^*))$ and

$$\begin{aligned}m_{\Phi^{L+1}}(\mathbf{y}^*, y) &= \det\left(K((\mathbf{y}^*, y), (\mathbf{y}^*, y))\right) \\ &= \det(K(\mathbf{y}^*, \mathbf{y}^*)) \det(K(y, y) - K(y, \mathbf{y}^*) K(\mathbf{y}^*, \mathbf{y}^*)^{-1} K(\mathbf{y}^*, y)) \\ &= m_{\Phi^L}(\mathbf{y}^*)[K(y, y) - K(y, \mathbf{y}^*) K(\mathbf{y}^*, \mathbf{y}^*)^{-1} K(\mathbf{y}^*, y)] \\ &= m_{\Phi^L}(\mathbf{y}^*)\, K_{\mathbf{y}^*}(y, y)\,.\end{aligned}$$

where the second equality follows from Schur's determinant identity. Substituting these into (6.1), we have

$$\begin{aligned}&\mathbb{P}(\theta_{N+1} \in \mathrm{d}y | \boldsymbol{\theta} = \mathbf{y}, U_N = u) \\ &\qquad = \alpha \cdot \frac{\Gamma(1 + \gamma)}{\Gamma(\gamma)(u+1)^{\gamma+1}} \cdot \frac{\psi(u)^{J+L-1}}{\psi(u)^{J+L}} \cdot \frac{m_{\Phi^L}(\mathbf{y}^*)\, K_{\mathbf{y}^*}(y, y)}{m_{\Phi^L}(\mathbf{y}^*)} \chi(\mathrm{d}y) \\ &\qquad = \alpha \cdot \frac{\gamma}{(u+1)^{\gamma+1}} \cdot \frac{1}{\psi(u)} \cdot K_{\mathbf{y}^*}(y, y) \chi(\mathrm{d}y) \\ &\qquad = \alpha \cdot \frac{\gamma}{(u+1)^{\gamma}} K_{\mathbf{y}^*}(y, y) \chi(\mathrm{d}y)\,.\end{aligned}$$

We can now apply Theorem 4.1.4, combining these two formulae to obtain the result. We note finally that the constant of proportionality is the same in each case. □