

2020-09-22

# Independent finite approximations for Bayesian nonparametric inference: construction, error bounds, and practical implications

---

Tin Nguyen, Jonathan Huggins, Lorenzo Masoero, Lester Mackey, Tamara Broderick. 2020.

"Independent finite approximations for Bayesian nonparametric inference: construction, error bounds, and practical implications." arXiv.org, Volume arXiv:2009.10780 [stat.ME],

<https://hdl.handle.net/2144/42761>

*"Downloaded from OpenBU. Boston University's institutional repository."*

# Independent finite approximations for Bayesian nonparametric inference: construction, error bounds, and practical implications

Tin D. Nguyen<sup>1</sup>, Jonathan Huggins<sup>2</sup>, Lorenzo Masoero<sup>1</sup>, Lester Mackey<sup>3</sup>,  
Tamara Broderick<sup>1</sup>

<sup>1</sup>*CSAIL, MIT, e-mail: tdn@mit.edu; lom@mit.edu; tbroderick@csail.mit.edu*

<sup>2</sup>*Department of Statistics & Mathematics, Boston University, e-mail: huggins@bu.edu*

<sup>3</sup>*Microsoft Research, e-mail: lmackey@microsoft.com*

**Abstract:** Bayesian nonparametrics based on completely random measures (CRMs) offers a flexible modeling approach when the number of clusters or latent components in a dataset is unknown. However, managing the infinite dimensionality of CRMs often leads to slow computation. Practical inference typically relies on either integrating out the infinite-dimensional parameter or using a *finite approximation*: a truncated finite approximation (TFA) or an independent finite approximation (IFA). The atom weights of TFAs are constructed sequentially, while the atoms of IFAs are independent, which (1) make them well-suited for parallel and distributed computation and (2) facilitates more convenient inference schemes. While IFAs have been developed in certain special cases in the past, there has not yet been a general template for construction or a systematic comparison to TFAs. We show how to construct IFAs for approximating distributions in a large family of CRMs, encompassing all those typically used in practice. We quantify the approximation error between IFAs and the target nonparametric prior, and prove that, in the worst-case, TFAs provide more component-efficient approximations than IFAs. However, in experiments on image denoising and topic modeling tasks with real data, we find that the error of Bayesian approximation methods overwhelms any finite approximation error, and IFAs perform very similarly to TFAs.

## 1. Introduction

Many data analysis problems can be seen as discovering a latent set of traits in a population. For instance, we might recover topics or themes from scientific papers, ancestral populations from genetic data, interest groups from social network data, or unique speakers across audio recordings of many meetings (Palla, Knowles and Ghahramani, 2012; Blei, Griffiths and Jordan, 2010; Fox et al., 2010). In all of these cases, we might reasonably expect the number of latent traits present in a data set to grow with the size of the data. One modeling option is to choose a different prior for different data set sizes, but is unwieldy and inconvenient. A simpler option is to choose a single prior that naturally yields different expected numbers of traits for different numbers of data points. In theory, *Bayesian nonparametrics* provides a rich set of priors with exactly this desirable property thanks to a countable infinity of traits, so that there are always more traits to reveal through the accumulation of more data. This latent, infinite-dimensional parameter presents a major practical challenge, though. In what follows, we propose a simple approximation across a wide range of BNP models, which can be seen as a generalization of certain existing special cases. Furthermore, it

is amenable to modern, efficient inference schemes and black-box code; fits easily within complex, potentially deep generative models; and admits straightforward parallelization.

**Background** A particular challenge of the infinite-dimensional parameter is that it is impossible to store an infinity of random variables in memory or learn the distribution over an infinite number of variables in finite time. Some authors have developed conjugate priors and likelihoods (Orbanz, 2010) to circumvent the infinite representation via marginalization and thereby perform exact Bayesian posterior inference (Broderick, Wilson and Jordan, 2018; James, 2017). However, these priors and likelihoods are often just a single piece within a more complex generative model, which is no longer fully conjugate and therefore requires an approximate posterior inference scheme such as Markov Chain Monte Carlo (MCMC) or variational Bayes (VB). Some local steps in, e.g., an MCMC sampler can still take advantage of conditional conjugacy via special marginal forms such the Chinese restaurant process (Teh et al., 2006) or the Indian buffet process (Griffiths and Ghahramani, 2005); see Broderick, Wilson and Jordan (2018) and James (2017) for general treatments. But using these marginal distributions rather than a full and explicit representation of the latent variables typically necessitates a Gibbs sampler, which can be slow to mix and may require special-purpose, model-specific sampling moves. To take advantage of black-box variational inference methods (Ranganath, Gerrish and Blei, 2014; Kucukelbir et al., 2015), modern MCMC methods such as Metropolis-adjusted Langevin algorithm (Roberts and Tweedie, 1996) or Hamiltonian Monte Carlo (HMC) (Neal, 2011; Betancourt, 2017), or modern probabilistic programming systems such as Stan (Carpenter et al., 2017), a full trait representation is generally required.

An alternative approach that still allows use of these convenient inference methods is to approximate the infinite-dimensional prior with a finite-dimensional prior that essentially replaces the infinite collection of random traits by a finite subset of “likely” traits. Unlike a fixed finite-dimensional prior across all data set sizes, this finite dimensional prior is seen as an approximation to the BNP prior and thereby its cardinality is informed directly by the BNP prior. Note that since any moderately complex model will necessitate approximate inference, so long as the approximation error from using the finite-dimensional prior approximation is on the order of the approximation error from MCMC or VB, no inferential quality has been lost.

Much of the previous work on finite approximations developed and analyzed truncations of the random measures underlying the nonparametric prior (Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012; Roychowdhury and Kulis, 2015; Campbell et al., 2019); we call these *truncated finite approximations* (TFAs) and refer to Campbell et al. (2019) for a thorough study of constructions for TFAs. In the present work, we instead consider a finite approximation consisting of independent and identical (i.i.d.) representations of the traits together with their rates within the population; we call these *independent finite approximations* (IFAs). The IFA approach has the potential to be simpler to incorporate in a complex hierarchical model, to exhibit improved mixing, and to be amenable to parallelizing computation during inference. There are not many known finite approximations using i.i.d. random variables and we are unaware of any general-purpose results on constructing them.

**Our Contributions** We propose a construction for IFAs that subsumes a number of special cases which have already been successfully used in applications, with practitioners reporting similar performance to the truncation approach but with faster mixing (Kurihara, Welling and Teh, 2007; Saria, Koller and Penn, 2010; Fox et al., 2010; Johnson and Willsky, 2013). On the other hand, our construction is distinct from that presented in Lee, James

and Choi (2016), which has an arguably smaller scope of application. We propose a broad mechanism for our i.i.d. finite approximation and relate these to existing work. We then quantify the effect of replacing the infinite-dimensional priors with an IFA in probabilistic models, providing interpretable error bounds with explicit dependence on the size of the approximation and the data cardinality. The error bounds reveal that in the worst case, to approximate the target to an accuracy, it is necessary to use a large IFA model while a small TFA model would suffice. However, differences have not been observed in practice, and we confirm through experiments with image denoising and topic modeling that IFAs and TFAs perform similarly on applied problems – IFAs benefit from conceptual ease-of-use.

## 2. Background

We start by summarizing relevant background on nonparametric priors constructed from completely random measures, and how truncated and independent finite approximations for these priors are constructed. Let  $\psi_i$  represent the  $i$ th trait of interest and Let  $\theta_i$  represent the rate, or frequency, of this trait in the population. We can collect the pairs of traits with their frequencies  $(\psi_i, \theta_i)$  in a measure that places non-negative mass  $\theta_i$  at location  $\psi_i$ :  $\Theta := \sum_{i=1}^I \theta_i \delta_{\psi_i}$ .  $I$ , the total number of traits, may be finite or, as in the nonparametric setting, countably infinite. To perform Bayesian inference, we need to choose a prior distribution on  $\Theta$  and a likelihood for the observed data  $Y_{1:N} := \{Y_n\}_{n=1}^N$  given  $\Theta$ , and then we must apply Bayes theorem to obtain the posterior on  $\Theta$  given the observed data.

**Completely random measures** Most common BNP priors can be conveniently formulated as (normalizations of) *completely random measures* (CRMs). CRMs are constructed from Poisson point processes, which are straightforward to manipulate both analytically and algorithmically. Consider a Poisson point process on  $\mathbb{R}_+ := [0, \infty)$  with rate measure  $\nu(d\theta)$  such that  $\nu(\mathbb{R}_+) = \infty$  and  $\int \min(1, \theta) \nu(d\theta) < \infty$ . Such a process generates an infinite number of rates  $(\theta_i)_{i=1}^\infty$ ,  $\theta_i \in \mathbb{R}_+$ , having an almost surely finite sum  $\sum_{i=1}^\infty \theta_i < \infty$ . We assume throughout that  $\psi_i \in \Psi$  for some space  $\Psi$  and  $\psi_i \stackrel{\text{i.i.d.}}{\sim} H$  for some diffuse distribution  $H$ .  $H$  serves as a prior on the trait values: in topic modeling, each topic is a probability vector in the simplex of vocabulary words, and it is typical to use  $H = \text{Dir}$ . The resulting measure  $\Theta$  in this case is a *completely random measure* (CRM) (Kingman, 1967). As shorthand, we will write  $\text{CRM}(H, \nu)$  for the completely random measure generated as just described:  $\Theta := \sum_i \theta_i \delta_{\psi_i} \sim \text{CRM}(H, \nu)$ . The corresponding *normalized CRM* (NCRM) is  $\Xi := \Theta / \Theta(\Psi)$ , which is a discrete probability measure. The set of atom locations of  $\Xi$  is the same as that of  $\Theta$ , while the atom sizes are normalized  $\Xi = \sum_i \xi_i \delta_{\psi_i}$  where  $\xi_i = \theta_i / (\sum_j \theta_j)$ .<sup>1</sup>

**Finite approximations** Since the sequence  $(\theta_i)_{i=1}^\infty$  is countably infinite, it may be difficult to simulate or perform posterior inference in the full model. One approximation scheme is to define the *finite approximation*  $\Theta_K := \sum_{i=1}^K \theta_i \delta_{\psi_i}$ . Since it involves a finite number of parameters,  $\Theta_K$  can be used for efficient posterior inference, including with black-box MCMC and VB algorithms—but some approximation error is introduced by not using the full CRM  $\Theta$ .

A *truncated finite approximation* (TFA) (Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012; Roychowdhury and Kulis, 2015) requires constructing an ordering on the sequence

<sup>1</sup>The possible fixed-location and deterministic components of an (N)CRM (Kingman, 1967) are not considered here for brevity; these components can be added (assuming they are purely atomic) and our analysis modified without undue effort.

$(\theta_i)_{i=1}^\infty$  such that  $\theta_i$  is a function of some auxiliary random variables  $\xi_1, \dots, \xi_i$ ; hence,  $\theta_{i+1}$  reuses the same auxiliary randomness as  $\theta_i$ , plus uses an additional random variable  $\xi_{i+1}$ . Thus, the value of  $\theta_{i+1}$  implicitly depends on the values of  $\theta_1, \dots, \theta_i$ . Truncated finite approximations are attractive because of the nestedness of the approximations  $K$ : in general, the approximation quality increases with  $K$ , and to refine existing truncations, it suffices to generate the next terms in the sequence. On the other hand the complex dependences between the atoms  $\theta_1, \theta_2, \dots$  potentially make inference more challenging.

We here instead pursue what we call an *independent finite approximation* (IFA), which involves choosing a sequence of probability measures  $\nu_1, \nu_2, \dots$  such that for any approximation level  $K$ , we choose  $\theta_1, \dots, \theta_K \stackrel{\text{i.i.d.}}{\sim} \nu_K$ . The  $\nu_K$  are chosen in such a way that  $\Theta_K \xrightarrow{\mathcal{D}} \Theta$  as  $K \rightarrow \infty$  — that is, the IFAs converge in distribution to the CRM. The pros and cons of the IFA invert those of the TFA: the atoms are now i.i.d., potentially making inference easier, but a completely new approximation must be constructed if  $K$  changes. Existing work (Paisley and Carin, 2009; Broderick et al., 2015; Acharya, Ghosh and Zhou, 2015; Lee, James and Choi, 2016; Lee, Miscouridou and Caron, 2019) has only developed i.i.d. finite approximations on a case-by-case basis, where as our focus is a general-purpose mechanism.

For the normalized atom sizes  $\xi_i = \theta_i / \sum_j \theta_j$ , finite approximations also involve random measures with finite support  $\Xi_K = \sum_{i=1}^K \xi_i \delta_{\psi_i}$ . TFAs can be defined in one of two ways. In the first approach, the TFA corresponding to the CRM can be normalized to form the approximation of the NCRM (Campbell et al., 2019). The second approach instead directly constructs an ordering over the sequence  $(\xi_i)_{i=1}^\infty$  and truncate this representation (Ishwaran and James, 2001; Blei and Jordan, 2006). Regarding the independent approach, we will only normalize the IFAs that target a given CRM to form the approximation of the corresponding NCRM.

**The beta process.** For concreteness, we consider the *beta process* (Teh and Görür, 2009; Broderick, Jordan and Pitman, 2012) as a running example of a CRM. We denote its distribution as  $\text{BP}(\gamma, \alpha, d)$ , with discount parameter  $d \in [0, 1)$ , scale parameter  $\alpha > -d$ , mass parameter  $\gamma > 0$ , and rate measure  $\nu(d\theta) = \gamma \frac{\Gamma(\alpha+1)}{\Gamma(1-d)\Gamma(\alpha+1)} \mathbb{1}[\theta \leq 1] \theta^{-d-1} (1-\theta)^{\alpha+d-1} d\theta$ . The case in which  $d = 0$  is the standard beta process (Hjort, 1990; Thibaux and Jordan, 2007). The beta process is typically paired with the Bernoulli likelihood process  $l(x|\theta) = \theta^x (1-\theta)^{1-x}$ ; the combination has been used for factor analysis (Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012) or dictionary learning (Zhou et al., 2009).

### 3. Constructing independent finite approximations

We first show how to easily construct independent finite approximations to a completely random measure. Specifically, our first main result shows how to construct IFAs that converge in distribution to CRMs with rate measures of a particular form. As an important special case, if the CRM is an exponential family CRM (Broderick, Wilson and Jordan, 2018) and the “discount” parameter  $d = 0$ , then the IFA is constructed from random variables in the same exponential family, a connection which is not only useful for approximate inference algorithms, but also for the theoretical analysis of the approximation itself. Finally, we show how normalized IFAs converge to the corresponding NCRM, in the sense that the partition induced by IFA converges to that induced by NCRM.

Formally, IFAs take the following form. For probability measures  $H$  and  $\nu_K$ , write  $\Theta_K \sim$

IFA $_K(H, \nu_K)$  if

$$\Theta_K = \sum_{i=1}^K \theta_{K,i} \delta_{\psi_{K,i}} \quad \theta_{K,i} \stackrel{\text{indep}}{\sim} \nu_K \quad \psi_{K,i} \stackrel{\text{i.i.d.}}{\sim} H.$$

We consider CRMs with rate measures  $\nu$  with densities that, near zero, are (essentially) proportional to  $\theta^{-1-d}$ , where  $d \in [0, 1)$  is the “discount” parameter. The explicit assumptions on  $\nu$  are given in Assumption 1.

**Assumption 1.** For  $d \in [0, 1)$  and  $\eta \in E \subseteq \mathbb{R}^d$ , let  $\Theta \sim \text{CRM}(H, \nu(\cdot; d, \eta))$ , where

$$\nu(d\theta; d, \eta) := \gamma \theta^{-1-d} g(\theta)^{-d} \frac{h(\theta; \eta)}{Z(1-d, \eta)} d\theta.$$

Assume that:

1. for  $\xi > 0$  and  $\eta \in E$ ,  $Z(\xi, \eta) = \int \theta^{\xi-1} g(\theta)^\xi h(\theta; \eta) d\theta < \infty$ ;
2.  $g$  is continuous,  $g(0) = 1$ , and  $\exists 0 < c_* \leq c^* < \infty$  such that  $c_* \leq g(\theta)^{-1} \leq c^*(1+\theta)$ ; and
3. there exists  $\epsilon > 0$  such that for all  $\eta \in E$ ,  $\theta \mapsto h(\theta; \eta)$  is continuous and bounded on  $[0, \epsilon]$ .

Other than the discount  $d$  and mass  $\gamma$ , the rate measure  $\nu$  potentially has additional hyperparameters, which are encapsulated by  $\eta$ . The finiteness of the normalizer  $Z$  is necessary in defining finite-dimensional distributions whose densities are very similar in form to  $\nu$ . The conditions on the behaviors of  $g(\theta)$  and  $h(\theta; \eta)$  imply that the overall rate measure’s behavior near  $\theta = 0$  is dominated by the  $\theta^{-1-d}$  term. These are mild regularity conditions: most popular BNP priors can be cast in such form, and the functions  $g(\theta)$  and  $h(\theta; \eta)$  are such that all three assumptions can be easily verified. Appendix A shows how common process such as beta, gamma (Ferguson and Klass, 1972; Kingman, 1975; Brx, 1999; Titsias, 2008; James, 2013), beta prime (Broderick, Wilson and Jordan, 2018) and generalized gamma process satisfy Assumption 1.

We will now define a sequence of IFAs that converge in distribution to such a CRM. Our IFA construction requires the following definition.

**Definition 3.1.** The parameterized function family  $\{S_b\}_{b \in \mathbb{R}_+}$  are *approximate indicators* if, for any  $b \in \mathbb{R}_+$ ,  $S_b(\theta)$  is a real increasing function such that  $S_b(\theta) = 0$  for  $\theta \leq 0$  and  $S_b(\theta) = 1$  for  $\theta \geq b$ .

Valid examples of approximate indicators are the indicator function  $S_b(\theta) = \mathbf{1}[\theta > 0]$  and the smoothed indicator function

$$S_b(\theta) = \begin{cases} \exp\left(\frac{-1}{1-(\theta-b)^2/b} + 1\right) & \text{if } \theta \in (0, b) \\ \mathbf{1}[\theta > 0] & \text{otherwise.} \end{cases}$$

Our first result now shows how to construct IFAs that provably converge to our family of CRMs.

**Theorem 3.2.** Suppose Assumption 1 hold. Let  $\{S_b\}_{b \in \mathbb{R}_+}$  be a family of approximate indicators. Fix  $a > 0$ , and  $(b_K)_{K \in \mathbb{N}}$ , a decreasing sequence such that  $b_K \rightarrow 0$ . For  $c := \gamma \frac{h(0; \eta)}{Z(1-d, \eta)}$  and  $\kappa = \min(1, \epsilon)$ , let

$$\nu_K(d\theta) := \theta^{-1+cK^{-1}-dS_{b_K}(\theta-aK^{-1})} g(\theta)^{cK^{-1}-d} h(\theta; \eta) Z_K^{-1} d\theta,$$

be a family of probability densities, where  $Z_K$  is chosen such that  $\int \nu_K(d\theta) = 1$ . If  $\Theta_K \sim \text{IFA}_K(H, \nu_K)$ , then  $\Theta_K \xrightarrow{\mathcal{D}} \Theta$  as  $K \rightarrow \infty$ .

The proof can be found in Appendix B.1. The scope of Theorem 3.2 is broader than known i.i.d. finite approximations. Namely, Lee, James and Choi (2016, Theorem 2) designs i.i.d. finite approximations that converge in distribution to either a beta process with  $d > 0$  or a gamma process with  $d > 0$  – as both processes satisfy Assumption 1, our construction can also be applied. The approximation in Lee, James and Choi (2016, Theorem 2) is graceful, as the i.i.d. densities are differentiable, whereas the densities of Theorem 3.2 are only continuous because of the approximate indicators. However, the approximation of Lee, James and Choi (2016, Theorem 2) lacks a well-defined limit in the case of  $d = 0$ , whereas our construction naturally incorporates this situation.

An important corollary of Theorem 3.2 applies to exponential family CRM with  $d = 0$ . In common BNP models, the relationship between the likelihood  $l(\cdot | \theta)$  and the CRM prior is closely related to the well-known conjugacy in exponential families (Broderick, Wilson and Jordan, 2018, Section 4). In particular, the likelihood has an exponential family form

$$l(x|\theta) := \kappa(x)\theta^{\phi(x)} \exp(\langle \mu(\theta), t(x) \rangle - A(\theta)). \quad (1)$$

Here  $x \in \mathbb{N} \cup \{0\}$ ,  $\kappa(x)$  is the base density,  $[t(x), \phi(x)]^T$  is the vector of sufficient statistics,  $A(\theta)$  is the log partition function,  $[\mu(\theta), \log \theta]^T$  is the vector of natural parameters, and  $\langle \mu(\theta), t(x) \rangle$  is an inner product. As for the rate measure, we will analyze those that behave like  $\theta^{-1}$  near 0

$$\nu(\theta) := \gamma' \theta^{-1} \exp \left\{ \left\langle \begin{pmatrix} \psi \\ \lambda \end{pmatrix}, \begin{pmatrix} \mu(\theta) \\ -A(\theta) \end{pmatrix} \right\rangle \right\} \mathbf{1}\{\theta \in U\}, \quad (2)$$

where  $\gamma' > 0$ ,  $\lambda > 0$ ,  $U \subset \mathbb{R}_+$  is the support of  $\nu$ . Eq. (2) leads to the suggestive terminology of *exponential CRMs*. The  $\theta^{-1}$  dependence near 0 means that these models lack power-law behavior e.g., in beta process, see Teh and Görür (2009). Models that can be cast in this form include beta process with Bernoulli likelihood, beta process with negative binomial likelihood (Broderick et al., 2015; Zhou et al., 2012) and gamma process with Poisson likelihood (Acharya, Ghosh and Zhou, 2015; Roychowdhury and Kulis, 2015). For short-hand, we refer to these models as beta–Bernoulli, beta–negative binomial and gamma–Poisson, respectively. The *normalizer*

$$S(\xi, \eta) := \int_U \theta^\xi \exp \left\{ \left\langle \eta, \begin{pmatrix} \mu(\theta) \\ -A(\theta) \end{pmatrix} \right\rangle \right\} d\theta. \quad (3)$$

of the exponential family distribution plays an important role in the sequel. Note that  $S$  is equal to the normalization quantity  $Z$  appearing in Assumption 1, but specialized for the exponential family rate measure.

We now state the simple form taken by  $\text{IFA}_K$  for exponential family CRMs. The assumptions are the natural analogues of Assumption 1, specialized for exponential family rate measures.

**Corollary 3.3.** *Let  $\nu$  be of the form Eq. (2), and assume that:*

1.  $S(\xi, \eta) < \infty$  for  $\xi > -1$ ;
2. There exists  $\epsilon > 0$  such that for any  $\psi, \lambda$ ,  $\theta \mapsto \exp \left\{ \left\langle \eta, \begin{pmatrix} \mu(\theta) \\ -A(\theta) \end{pmatrix} \right\rangle \right\} \mathbf{1}\{\theta \in U\}$  is a continuous and bounded function of  $\theta$  on  $[0, \epsilon]$ .

For  $c := \gamma' \exp \left\{ \langle \eta, \begin{pmatrix} \mu(0) \\ -A(0) \end{pmatrix} \rangle \right\}$ , let

$$\nu_K(\theta) := \frac{\mathbf{1}\{\theta \in U\}}{S(c/K - 1, \eta)} \theta^{c/K-1} \exp \left\{ \langle \eta, \begin{pmatrix} \mu(\theta) \\ -A(\theta) \end{pmatrix} \rangle \right\}. \quad (4)$$

If  $\Theta_K \sim \text{IFA}_K(H, \nu_K)$ , then  $\Theta_K \xrightarrow{\mathcal{D}} \Theta$ .

Corollary 3.3 is sufficient to recover known IFA results for  $\text{BP}(\gamma, \alpha, 0)$  (Doshi-Velez et al., 2009; Paisley and Carin, 2009; Griffiths and Ghahramani, 2011). Appendix A uses Corollary 3.3 to construct IFAs for more example CRMs.

**Example 3.1** (Beta process). When  $d = 0$ , the rate measure of the beta process is  $\nu(\theta) = \gamma\alpha\theta^{-1} \exp((\alpha - 1) \log(1 - \theta)) \mathbf{1}\{0 \leq \theta \leq 1\}$ . The normalizer depends only on  $\xi$  and  $\alpha - 1$ :  $S_{\text{BP}} = \int_0^1 \theta^\xi (1 - \theta)^{\alpha-1} \exp(0) d\theta = B(\xi + 1, \alpha)$ . The assumptions in Corollary 3.3 can be quickly verified.  $S_{\text{BP}} < \infty$  for  $\xi > -1$  is evident as  $B(\xi + 1, \alpha) < \infty$  for  $\xi + 1 > 0, \alpha > 0$ . The function  $\theta \mapsto (1 - \theta)^{\alpha-1}$  is clearly bounded and continuous on the interval  $[0, 0.5]$  for any  $\alpha > 0$ . Therefore  $\nu_K = \text{Beta}(\gamma\alpha/K, \alpha)$ .

In comparison, Doshi-Velez et al. (2009) approximates  $\text{BP}(\gamma, 1, 0)$  with each  $\nu_K$  is a  $\text{Beta}(\gamma/K, 1)$  distribution. Griffiths and Ghahramani (2011) also approximates  $\text{BP}(\gamma, \alpha, 0)$  with  $\nu_K$  being  $\text{Beta}(\gamma\alpha/K, \alpha)$ . Lastly, Paisley and Carin (2009) approximates  $\text{BP}(\gamma, \alpha, 0)$  with  $\nu_K$  being  $\text{Beta}(\gamma\alpha/K, \alpha(1 - 1/K))$  distribution, with the difference between  $\text{Beta}(\gamma\alpha/K, \alpha)$  and  $\text{Beta}(\gamma\alpha/K, \alpha(1 - 1/K))$  being not substantive.

Given that  $\text{IFA}_K$  is a converging approximation to the corresponding target CRM, it is natural to ask if the normalization of  $\text{IFA}_K$  converges to the corresponding normalization of CRM i.e., NCRM. Our next result shows that normalized IFA indeed converges, in the sense of *exchangeable partition probability functions*, or EPPF (Pitman, 1995). The EPPF of a NCRM  $\Xi$  gives the probability of partitions of  $\{1, 2, \dots, N\}$  induced by sampling from  $\Xi$ . In particular, under the model  $\Xi \sim \text{NCRM}$ ,  $V_n | \Xi \stackrel{\text{i.i.d.}}{\sim} \Xi$  for  $1 \leq n \leq N$  with the effect of  $\Xi$  marginalized out, the ties among the  $V_n$ 's induce a partition over the set  $\{1, 2, \dots, N\}$ . Let there be  $t \leq N$  distinct values among the  $V_n$ 's, and let  $n_i$  be the number of elements in the  $i$ -th block of the partition induced by sampling from  $\Xi$ , so that  $n_i \geq 1, \sum_{i=1}^t n_i = N$ . The probability of the induced partition is a symmetric function  $p(n_1, n_2, \dots, n_t)$  that depends only on the frequencies  $n_i$  of each block. The EPPF of  $\text{IFA}_K$  is defined analogously.

**Theorem 3.4.** *Suppose Assumption 1 holds, and let  $\Theta_K$  be as in Theorem 3.2. Let  $p(n_1, n_2, \dots, n_t)$  be the EPPF of a NCRM  $\Xi$  where  $\Xi := \Theta/\Theta(\Psi)$  and let  $p_K(n_1, n_2, \dots, n_t)$  be the EPPF of normalized IFA  $\Xi_K$  where  $\Xi_K := \Theta_K/\Theta_K(\Psi)$ . Then, for any  $N$ , for any  $n_i \geq 1, \sum_{i=1}^t n_i = N$ ,*

$$\lim_{K \rightarrow \infty} p_K(n_1, n_2, \dots, n_t) = p(n_1, n_2, \dots, n_t).$$

The proof can be found in Appendix B.2. Since the EPPF gives the probability of each partition, the point-wise convergence in Theorem 3.4 certifies that the distribution over partitions induced by the normalized  $\text{IFA}_K$  converges to that induced by the target NCRM, for any finite data cardinality  $N$ .

#### 4. Non-asymptotic error bounds for CRM-based models

Theorem 3.2 justifies the use of  $\text{IFA}_K$  in the asymptotic limit  $K \rightarrow \infty$  but does not provide guidance on choosing an appropriate approximation level for modeling a data process with a

given cardinality  $N$ . In this section, we quantify the effect of replacing CRM with IFA $_K$  (for finite  $K$ ) in probabilistic models using error bounds that are simple to manipulate, easily yielding recommendation of the appropriate  $K$  for a given  $N$  and accuracy level.

The CRM prior on  $\Theta$  is typically combined with a likelihood that generates trait counts for each data point. Let  $l(\cdot|\theta)$  be a proper probability mass function on  $\mathbb{N}\cup\{0\}$  for all  $\theta$  in the support of  $\nu$ . Then a collection of conditionally independent observations  $X_{1:N}$  given  $\Theta$  are distributed according to the *likelihood process*  $\text{LP}(l, \Theta)$  – i.e.,  $X_n := \sum_i x_{ni} \delta_{\psi_i} \stackrel{\text{i.i.d.}}{\sim} \text{LP}(l, \Theta)$  – if  $x_{ni} \sim l(\cdot|\theta_i)$  independently across  $i$  and i.i.d. across  $n$ . Since the trait counts are typically latent in a full generative model specification, define the observed data  $Y_n | X_n \stackrel{\text{indep}}{\sim} f(\cdot | X_n)$  for a probability kernel  $f$ . For instance, if the sequence  $(\theta_i)_{i=1}^\infty$  represents the topic rates in a document corpus,  $X_n$  might capture how many words in document  $n$  are generated from each topic and  $Y_n$  might be the observed collection of words for that document. The target nonparametric model can thus be summarized as

$$\Theta \sim \text{CRM}(H, \nu), \quad X_n | \Theta \stackrel{\text{i.i.d.}}{\sim} \text{LP}(l; \Theta), \quad Y_n | X_n \stackrel{\text{indep}}{\sim} f(\cdot | X_n) \quad n = 1, 2, \dots, N. \quad (5)$$

The approximating finite-dimensional model, with  $\nu_K$  being given in Theorem 3.2 (or Corollary 3.3), is

$$\Theta_K \sim \text{IFA}_K(H, \nu_K), \quad Z_n | \Theta_K \stackrel{\text{i.i.d.}}{\sim} \text{LP}(l; \Theta_K), \quad W_n | Z_n \stackrel{\text{indep}}{\sim} f(\cdot | Z_n) \quad n = 1, 2, \dots, N. \quad (6)$$

Let  $P_{N,\infty}$  be the distribution of the observations  $Y_{1:N}$ , and  $P_{N,K}$  be the distribution of the observations  $W_{1:N}$ . We define *approximation error* to be the total variation distance  $d_{TV}(P_{N,K}, P_{N,\infty})$  between two observational processes, one using the CRM and the other one using the approximate IFA $_K$  as the prior (Ishwaran and Zarepour, 2002; Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012; Campbell et al., 2019). Recall that total variation distance is the supremum difference in probability mass over measurable sets  $d_{TV}(P_{N,K}, P_{N,\infty}) := \sup_A |P_{N,K}(A) - P_{N,\infty}(A)|$ .

#### 4.1. Assumptions

We restrict attention to exponential family CRM-likelihood pairs. We require Definition 4.1 to express our the assumptions on the target model.

**Definition 4.1.** Suppose  $l(\cdot|\theta)$  has the form Eq. (1) and  $\nu(\theta)$  has the form Eq. (2). For  $n \in \mathbb{N}$ ,  $x_{1:(n-1)} \in (\mathbb{N}\cup\{0\})^{n-1}$ , define shorthands  $T_{1:n} := \sum_{m=1}^{n-1} t(x_m)$  and  $\Phi_{1:n} := \sum_{m=1}^{n-1} \phi(x_m)$ . For  $x \in \mathbb{N}\cup\{0\}$ , let

$$h_c(x|x_{1:(n-1)}) := \kappa(x) \frac{S\left(-1 + \Phi_{1:n} + \phi(x), \eta + \binom{T_{1:n} + t(x)}{n}\right)}{S\left(-1 + \Phi_{1:n}, \eta + \binom{T_{1:n}}{n-1}\right)}$$

and

$$\tilde{h}_c(x|x_{1:(n-1)}) := \kappa(x) \frac{S\left(c/K - 1 + \Phi_{1:n} + \phi(x), \eta + \binom{T_{1:n} + t(x)}{n}\right)}{S\left(c/K - 1 + \Phi_{1:n}, \eta + \binom{T_{1:n}}{n-1}\right)}$$

and

$$M_{n,x} := \gamma' \kappa(0)^{n-1} \kappa(x) S \left( c/K - 1 + (n-1)\phi(0) + \phi(x), \eta + \binom{(n-1)t(0) + t(x)}{n} \right).$$

We show in Appendix C that the functions  $h_c, \tilde{h}_c, M_{n,x}$  govern the *marginal process* representation of the probabilistic models (Broderick, Wilson and Jordan, 2018, Section 6). Namely, the joint distribution of  $X_{1:N}$  can be expressed in terms of the conditionals  $X_n | X_{1:(n-1)}$ , with  $M_{n,x}$  and  $h_c$  governing this process. Similarly, the joint distribution  $Z_{1:N}$  can be expressed in terms of the conditionals  $Z_n | Z_{1:(n-1)}$ , with  $\tilde{h}_c$  governing this process. For the beta-Bernoulli process with  $d = 0$ , the functions have particularly simple forms.

**Example 4.1** (Beta-Bernoulli with  $d = 0$ ). For the beta-Bernoulli model with  $d = 0$ , we have

$$\begin{aligned} h_c(x|x_{1:(n-1)}) &= \frac{\sum_{i=1}^{n-1} x_i}{\alpha - 1 + n} \mathbf{1}\{x = 1\} + \frac{\alpha + \sum_{i=1}^{n-1} (1 - x_i)}{\alpha - 1 + n} \mathbf{1}\{x = 0\}. \\ \tilde{h}_c(x|x_{1:(n-1)}) &= \frac{\sum_{i=1}^{n-1} x_i + \gamma\alpha/K}{\alpha - 1 + n + \gamma\alpha/K} \mathbf{1}\{x = 1\} + \frac{\alpha + \sum_{i=1}^{n-1} (1 - x_i)}{\alpha - 1 + n + \gamma\alpha/K} \mathbf{1}\{x = 0\}, \\ M_{n,1} &= \frac{\gamma\alpha}{\alpha - 1 + n}, \quad M_{n,x} = 0 \text{ for } x > 1. \end{aligned}$$

We now formulate the conditions which can be used to show that  $d_{TV}(P_{N,K}, P_{N,\infty})$  is small.

**Assumption 2.** There exist constants  $\{C_i\}_{i=1}^5$  such that the following hold.

1. For all  $n \in \mathbb{N}$ ,

$$\sum_{x=1}^{\infty} M_{n,x} \leq \frac{C_1}{n - 1 + C_1}. \quad (7)$$

2. For all  $n \in \mathbb{N}$ ,

$$\sum_{x=1}^{\infty} h(x|x_{1:(n-1)} = 0) \leq \frac{1}{K} \frac{C_1}{n - 1 + C_1}. \quad (8)$$

3. For any  $n \in \mathbb{N}$ , for any  $\{x_i\}_{i=1}^{n-1}$ ,

$$\sum_{x=0}^{\infty} \left| h_c(x|x_{1:(n-1)}) - \tilde{h}_c(x|x_{1:(n-1)}) \right| \leq \frac{1}{K} \frac{C_1}{n - 1 + C_1}. \quad (9)$$

4. For all  $n \in \mathbb{N}$ , for any  $K \geq C_2(\ln n + C_3)$ ,

$$\sum_{x=1}^{\infty} \left| M_{n,x} - K \tilde{h}_c(x|x_{1:(n-1)} = 0) \right| \leq \frac{1}{K} \frac{C_4 \ln n + C_5}{n - 1 + C_1}. \quad (10)$$

Note that the conditions depend only on the functions in Definition 4.1 and not on the observational likelihood  $f(\cdot)$  which maps the latent states to the observations. The first condition constrains the growth rate of the target model.  $\sum_{n=1}^N \sum_{x=1}^{\infty} M_{n,x}$  is the expected number of components for data cardinality  $N$  – since each  $\sum_{x=1}^{\infty} M_{n,x}$  is at most  $O(1/n)$ , the total number of components is  $O(\ln N)$ . The second condition means that  $\tilde{h}_c$  is a very good

approximation of  $h_c$  in total variation distance; furthermore, the longer the vector  $\{x_i\}_{i=1}^{n-1}$ , the smaller the error. Similarly, the third condition means that  $K\tilde{h}_c(\cdot|0)$  is a very accurate approximation of  $M_{n,\cdot}$ , and there is also a reduction in the error as  $n$  increases. The set of constants  $C_i$  which satisfy Assumption 2 is not unique: we are in general not interested in the best constants  $C_i$ , rather that they exist. We speculate that such assumptions can be made more explicit in the normalizer  $S$ . For instance, the  $1/K$  dependence is due to smoothness of  $S$  in its first argument, while the dependence on  $n$  is due to some inherent notion of scale dictated by the second and third arguments.

Assumption 2 can be verified for the most important CRM models. In Example 4.2 we verify it for the beta-Bernoulli model, and in Appendix E, we verify it for beta-negative binomial and gamma-Poisson models.

**Example 4.2** (Beta-Bernoulli with  $d = 0$ , continued). The growth rate of the target model is

$$\sum_{x=1}^{\infty} M_{n,x} = M_{n,1} = \frac{\gamma\alpha}{n-1+\alpha}.$$

Since  $\tilde{h}_c$  is supported on  $\{0, 1\}$ , the growth rate of the approximate model is

$$\tilde{h}_c(1|x_{1:(n-1)} = 0) = \frac{\gamma\alpha/K}{\alpha-1+n+\gamma\alpha/K} \leq \frac{1}{K} \frac{\gamma\alpha}{n-1+\alpha}.$$

Since both  $h_c$  and  $\tilde{h}_c$  are supported on  $\{0, 1\}$ , Eq. (9) becomes

$$\left| h_c(1|x_{1:(n-1)}) - \tilde{h}_c(1|x_{1:(n-1)}) \right| = \left| \frac{\sum_{i=1}^{n-1} x_i + \gamma\alpha/K}{\alpha-1+n+\gamma\alpha/K} - \frac{\sum_{i=1}^{n-1} x_i}{\alpha-1+n} \right| \leq \frac{\gamma\alpha}{K} \frac{1}{n-1+\alpha}.$$

Again, because  $M_{n,x} = \tilde{h}_c(x|\cdot) = 0$  for  $x > 1$ , Eq. (10) becomes

$$\left| M_{n,1} - K\tilde{h}_c(1|x_{1:(n-1)} = 0) \right| = \left| \frac{\gamma\alpha}{\alpha-1+n} - \frac{\gamma\alpha}{\alpha-1+n+\frac{\gamma\alpha}{K}} \right| \leq \frac{\gamma^2\alpha}{K} \frac{1}{n-1+\alpha}.$$

Calibrating  $\{C_i\}$  based on these inequalities is straightforward.

## 4.2. Upper bound

Under the aforementioned assumptions, Theorem 4.2 upper bounds the approximation error.

**Theorem 4.2** (Upper bound for exponential family CRMs). *If Assumption 2 holds, then there exist positive constants  $C', C'', C'''$  depending only on  $\{C_i\}_{i=1}^5$  such that*

$$d_{TV}(P_{N,\infty}, P_{N,K}) \leq \frac{C' + C'' \ln^2 N + C''' \ln N \ln K}{K}.$$

The proof can be found in Appendix F.1. Theorem 4.2 states that the IFA approximation error grows as  $O(\ln^2 N)$  with fixed  $K$ , and as decreases as  $O(\frac{\ln K}{K})$  for fixed  $N$ . On the one hand, for fixed  $K$ , it is expected that the error increases as  $N$  increases: with more data, the number of latent components in the data increases, demanding finite approximations of increasingly larger sizes. In particular,  $O(\ln N)$  is the standard Bayesian nonparametric growth rate for non-power law models (Griffiths and Ghahramani, 2011). It is likely that

the  $O(\ln^2 N)$  factor can be improved to  $O(\ln N)$  – more generally, we *conjecture* that the error directly depends on the expected number of latent components in a model for  $N$  observations. On the other hand, for fixed  $N$ , the error goes to zero at least as fast as  $O(\frac{\ln K}{K})$ . We also suspect the  $\ln K$  factor in the numerator can be removed.

### 4.3. Lower bounds

As Theorem 4.2 is only an upper bound, a natural question to investigate is the tightness of the bound in terms of  $N, K$ . In this section, we focus on the beta-Bernoulli process with  $d = 0$ , i.e.,  $P_{N,\infty}$  refers to the observational process coming from  $\text{BP}(\gamma, \alpha, 0)$  and  $P_{N,K}$  refers to the observational process  $\text{IFA}_K$  with  $\nu_K$  as in Example 3.1.

We first look at the dependence of the error bound in terms of  $\ln N$ . For any  $N \in \mathbb{N}$ ,  $\alpha > 0$ , we define the *growth function*

$$C(N, \alpha) := \sum_{n=1}^N \frac{\alpha}{n-1+\alpha}. \quad (11)$$

It is known that  $C(N, \alpha) = \Omega(\ln N)$  (see Lemma D.9). Theorem 4.3 shows that finite approximations cannot be accurate if the approximation level is too small compared to the growth function  $C(N, \alpha)$ .

**Theorem 4.3** ( $\ln N$  is necessary). *For the beta-Bernoulli model with  $d = 0$ , there exists an observation likelihood  $f$ , independent of  $K$  and  $N$ , such that for any  $N$ , if  $K \leq \frac{1}{2}\gamma C(N, \alpha)$ , then*

$$d_{TV}(P_{N,\infty}, P_{N,K}) \geq 1 - \frac{C}{N^{\gamma\alpha/8}},$$

where  $C$  only depends on hyper-parameters of the beta process i.e.,  $\gamma, \alpha$ .

The proof is given in Appendix F.2. Theorem 4.3 implies that as  $N$  grows, if the approximation level  $K$  fails to surpass the  $\frac{1}{2}\gamma C(N, \alpha) = \Omega(\ln N)$  threshold, then the total variation between the approximate and the target model remains bounded from zero – in fact, the error tends to one.

Now turning to the dependence on  $K$  of the upper bound Theorem 4.2, we discuss a lower bound on the approximation error, which reveals that the  $\frac{1}{K}$  factor in the upper bound is *tight* (modulo logarithmic factors).

**Theorem 4.4** (Lower bound of  $1/K$ ). *For the beta-Bernoulli model with  $d = 0$ , there exists an observation likelihood  $f$ , independent of  $K$  and  $N$ , such that for any  $N$ ,*

$$d_{TV}(P_{N,\infty}, P_{N,K}) \geq C(\gamma) \frac{\gamma^2}{K} \frac{1}{(1 + \gamma/K)^2},$$

where  $C(\gamma) := \frac{1}{8} \frac{1}{\gamma + \exp(-1)(\gamma+1) \max(12\gamma^2, 48\gamma, 28)}$ .

The proof can be found in Appendix F.2. While Theorem 4.2 implies that an IFA with  $K = O(\text{poly}(\ln N)/\epsilon)$  atoms suffices in approximating the target model to less than  $\epsilon$  error, Theorem 4.4 implies that an IFA with  $K = \Omega(1/\epsilon)$  atoms is *necessary* in the worst case. This dependence on the accuracy level means that IFAs are worse than TFAs in theory. For example, consider Bondesson approximations (Bondesson, 1982) of  $\text{BP}(\gamma, \alpha, 0)$ .

**Example 4.3** (Bondesson approximation (Bondesson, 1982)). Let  $\alpha \geq 1$ . Let  $E_l \stackrel{iid}{\sim} \text{Exp}(1)$  and  $\Gamma_k = \sum_{l=1}^k E_l$ . The level  $K$  Bondesson approximation of  $\text{BP}(\gamma, \alpha, 0)$  is a TFA  $\sum_{k=1}^K \theta_k \delta_{\psi_k}$  where  $\theta_k = V_k \exp(-\Gamma_k/\gamma\alpha)$ ,  $V_k \stackrel{iid}{\sim} \text{Beta}(1, \alpha - 1)$  and  $\psi_k \stackrel{iid}{\sim} H$ .

The following result gives a bound on the error of the Bondesson approximation:

**Proposition 4.5.** (Campbell et al., 2019) For  $\gamma > 0, \alpha \geq 1$ , let  $\Theta_K$  be distributed according to a level  $K$  Bondesson approximation of  $\text{BP}(\gamma, \alpha, 0)$ ,  $R_n | \Theta_K \stackrel{iid}{\sim} \text{LP}(l; \Theta_K)$ ,  $T_n | R_n \stackrel{indep}{\sim} f(\cdot | R_n)$  with  $N$  observations. Let  $Q_{N,K}$  be the distribution of the observations  $T_{1:N}$ . Then:

$$d_{TV}(P_{N,\infty}, P_{Q_{N,K}}) \leq N\gamma \left( \frac{\gamma\alpha}{1 + \gamma\alpha} \right)^K.$$

Proposition 4.5 implies that a TFA with  $K = O(\ln(N/\epsilon))$  atoms suffices in approximating the target model to less than  $\epsilon$  error. Modulo log factors, comparing the necessary  $\frac{1}{\epsilon}$  level for IFA and the sufficient  $\ln(\frac{1}{\epsilon})$  level for TFA, we conclude that the necessary size for IFA is exponentially larger than the sufficient size for TFA, in the worst case.

## 5. Non-asymptotic error bounds for Dirichlet process-based models

Having analyzed the error incurred by  $\text{IFA}_K$  in CRM-based models like beta-Bernoulli, gamma-Poisson and beta-negative binomial, we now turn the approximation error in NRCM-based models. Our notion of approximation error remains the total variation distance between the target and the approximate observational processes. The forms of the upper and lower bounds are very similar to Theorems 4.2, 4.3 and 4.4. We leave to future work to derive bounds for more general NCRMs.

We focus on the Dirichlet process [DP] (Ferguson, 1973; Sethuraman, 1994) – which is the normalization of a non-power law gamma process – and the finite symmetric Dirichlet [FSD] distribution – which is the normalization of the IFA for gamma process. The Dirichlet process is one of the most widely used nonparametric priors. The gamma process CRM has rate measure  $\nu(d\theta) = \gamma \frac{\lambda^{1-d}}{\Gamma(1-d)} \theta^{-d-1} e^{-\lambda\theta} d\theta$ . We denote its distribution as  $\text{GP}(\gamma, \lambda, d)$ . The normalization of  $\text{GP}(\gamma, 1, 0)$  is a Dirichlet process with mass parameter  $\gamma$  (Kingman, 1975; Ferguson, 1973). By Corollary 3.3,  $\text{IFA}_K(H, \nu_K)$  (where  $\nu_K(\theta) = \text{Gam}(\theta; \gamma/K, 1)$ ) converges to  $\text{GP}(\gamma, 1, 0)$ . Because the normalization of independent gamma random variables is a Dirichlet random variable, the normalization of  $\text{IFA}_K(H, \nu_K)$  is equal in distribution to  $\sum_{i=1}^K p_i \delta_{\psi_i}$  where  $\psi_i \stackrel{i.i.d.}{\sim} H$  and  $\{p_i\}_{i=1}^K \sim \text{Dir}(\frac{\gamma}{K} \mathbf{1}_K)$ . We denote this as  $\text{FSD}_K(\gamma, H)$ .

We consider Dirichlet process mixture models (Antoniak, 1974)

$$\Theta \sim \text{DP}(\alpha, H), \quad X_n | \Theta \stackrel{i.i.d.}{\sim} \Theta, \quad Y_n | X_n \stackrel{i.i.d.}{\sim} f(\cdot | X_n) \quad (12)$$

with corresponding approximation

$$\Theta_K \sim \text{FSD}_K(\alpha, H), \quad Z_n | \Theta_K \stackrel{i.i.d.}{\sim} \Theta_K, \quad W_n | Z_n \stackrel{i.i.d.}{\sim} f(\cdot | Z_n). \quad (13)$$

Let  $P_{N,\infty}$  be the distribution of the observations  $Y_{1:N}$ . Let  $P_{N,K}$  be the distribution of the observations  $W_{1:N}$ .

### 5.1. Upper bound

Upper bounds on the error made by  $\text{FSD}_K$  can be used to determine the sufficient  $K$  to approximate the target process for a given  $N$  and accuracy level. We upper bound  $d_{TV}(P_{N,\infty}, P_{N,K})$  in Theorem 5.1.

**Theorem 5.1** (Upper bound for DP mixture model). *For some constants  $C_1, C_2, C_3$  that only depend on  $\alpha$ ,*

$$d_{TV}(P_{N,\infty}, P_{N,K}) \leq \frac{C_1 + C_2 \ln^2 N + C_3 \ln N \ln K}{K}.$$

The proof is given in Appendix G.1. Theorem 5.1 is similar to Theorem 4.2. The  $O(\ln^2 N)$  growth of the bound for fixed  $N$  can likely be reduced to  $O(\ln N)$ , the inherent growth rate of DP mixture models (Miller and Harrison, 2013). The  $O(\frac{\ln K}{K})$  rate of decrease to zero is tight because of a  $\frac{1}{K}$  lower bound on the approximation error. Theorem 5.1 is an improvement over the existing theory for  $\text{FSD}_K$ , in the sense that Ishwaran and Zarepour (2002, Theorem 4) provides an upper bound on  $d_{TV}(P_{N,\infty}, P_{N,K})$  that lacks an explicit dependence on  $K$  or  $N$  – that bound cannot be inverted to determine the sufficient  $K$  to approximate the target to a given accuracy, while it is simple to determine using Theorem 5.1.

Theorem 5.1 can also be used to analyze models with additional hierarchical structure. For instance, the hierarchical Dirichlet process [HDP] and variants are important use cases of DP and have demonstrated great practical use. We will analyze the error made by  $\text{FSD}_K$  for a variant of HDP we call modified HDP. In HDP, there is a population measure generated by DP,  $G_0 \sim \text{DP}(\omega, H)$ , and for each sub-population indexed by  $d$ , the sub-population measure is generated as  $G_d | G_0 \sim \text{DP}(\alpha, G_0)$ . In modified HDP, the sub-population measure is instead distributed as  $G_d | G_0 \sim \text{TSB}_T(\alpha, G_0)$  where the TSB distribution is explained in Example 5.1.

**Example 5.1** (Stick-breaking approximation (Sethuraman, 1994)). For  $i = 1, 2, \dots, K-1$ , let  $v_i \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha)$ . Set  $v_K = 1$ . Let  $\xi_i = v_i \prod_{j=1}^{i-1} (1 - v_j)$ . Let  $\psi_k \stackrel{\text{i.i.d.}}{\sim} H$ , and  $\Xi_K = \sum_{k=1}^K \xi_k \delta_{\psi_k}$ . We denote the distribution of  $\Xi_K$  as  $\text{TSB}_K(\alpha, H)$ .

In all, the generative process of modified HDP is

$$\begin{aligned} G &\sim \text{DP}(\omega, H) \\ H_d | G &\stackrel{\text{indep}}{\sim} \text{TSB}_T(\alpha, G_0) && \text{across } d \\ \beta_{dn} | H_d &\stackrel{\text{indep}}{\sim} H_d(\cdot), W_{dn} | \beta_{dn} &\stackrel{\text{indep}}{\sim} f(\cdot | \beta_{dn}) && \text{across } d, n \end{aligned} \quad (14)$$

Observation groups are indexed by  $d$  and individual observations are indexed by  $n, d$ . Each group manifests at most  $T$  distinct atoms of the population-level measure in the style of Example 5.1. The number of groups is  $D$ , and the number of observations in each group is  $N$ .

The finite approximation we consider replaces the population level DP with  $\text{FSD}_K$ , keeping the other conditionals intact

$$\begin{aligned} G_K &\sim \text{FSD}_K(\omega, H) \\ F_d | G_K &\stackrel{\text{indep}}{\sim} \text{TSB}_T(\alpha, G_K) && \text{across } d \\ \psi_{dn} | F_d &\stackrel{\text{indep}}{\sim} F_d(\cdot), Z_{dn} | \psi_{dn} &\stackrel{\text{indep}}{\sim} f(\cdot | \psi_{dn}) && \text{across } d, n \end{aligned} \quad (15)$$

Let  $P_{(N,D),\infty}$  be the distribution of the observations  $\{W_{dn}\}$ . Let  $P_{(N,D),K}$  be the distribution of the observations  $\{Z_{dn}\}$ . We have the following corollary to Theorem 5.1.

**Corollary 5.2** (Upper bound for modified HDP). *For some constants  $C_1, C_2, C_3$  which depend only on  $\omega$ ,*

$$d_{TV}(P_{(N,D),\infty}, P_{(N,D),K}) \leq \frac{C_1 + C_2 \ln^2(DT) + C_3 \ln(DT) \ln K}{K}.$$

The proof can be found in Appendix G.1. For fixed  $K$ , Corollary 5.2 is independent of  $N$ , the number of observations in each group, but grows like  $O(\text{poly}(\ln D))$  with the number of groups  $D$ . For fixed  $D$ , the approximation error decrease to zero at rate no slower than  $O\left(\frac{\ln K}{K}\right)$ .

## 5.2. Lower bounds

As Theorem 5.1 is only an upper bound, we now investigate the tightness of the inequality in terms of  $N$  and  $K$ . We return to DP mixture models. We first look at the dependence of the error bound in terms of  $\ln N$ . Theorem 5.3 shows that finite approximations cannot be accurate if the approximation level is too small compared to the growth rate  $\ln N$ .

**Theorem 5.3** ( $\ln N$  is necessary). *There exists a probability kernel  $f(\cdot)$ , independent of  $K, N$ , such that for any  $N \geq 2$ , if  $K \leq \frac{1}{2}C(N, \alpha)$ , then*

$$d_{TV}(P_{N,\infty}, P_{N,K}) \geq 1 - \frac{C'}{N^{\alpha/8}}$$

where  $C'$  is a constant only dependent on  $\alpha$ .

The proof is given in Appendix G.2. Theorem 5.3 implies that as  $N$  grows, if the approximation level  $K$  fails to surpass the  $\frac{1}{2}C(N, \alpha)$  threshold, then the total variation between the approximate and the target model remains bounded from zero – in fact, the error tends to one. Recall that  $C(N, \alpha) = \Omega(\ln N)$ , so the necessary approximation level is  $\Omega(\ln N)$ . Theorem 5.3 is the analog of Theorem 4.3.

We also investigate the tightness of Theorem 5.1 in terms of  $K$ . In Theorem 5.4, our lower bound indicates that the  $\frac{1}{K}$  factor in Theorem 5.1 is tight (up to log factors).

**Theorem 5.4** ( $1/K$  lower bound). *There exists a probability kernel  $f(\cdot)$ , independent of  $K, N$ , such that for any  $N \geq 2$ ,*

$$d_{TV}(P_{N,\infty}, P_{N,K}) \geq \frac{\alpha}{1 + \alpha} \frac{1}{K}.$$

The proof is given in Appendix G.2. While Theorem 5.1 implies that the normalized IFA $_K$  with  $K = O(\text{poly}(\ln N)/\epsilon)$  atoms suffices in approximating the DP mixture model to less than  $\epsilon$  error, Theorem 5.4 implies that a normalized IFA with  $K = \Omega(1/\epsilon)$  atoms is *necessary* in the worst case. This worst-case behavior is analogous to Theorem 4.4 for DP-based models.

The  $\frac{1}{\epsilon}$  dependence means that IFAs are worse than TFAs in theory. It is known that small TFA models are already excellent approximations of the DP. Example 5.1 is a very well-known finite approximation whose error is upper bounded in Proposition 5.5.

**Proposition 5.5.** (*Ishwaran and James, 2001, Theorem 2*) Let  $\Xi_K \sim \text{TSB}_K(\alpha, H)$ ,  $R_n | \Xi_K \stackrel{i.i.d.}{\sim} \Xi_K, T_n | R_n \stackrel{indep}{\sim} f(\cdot | R_n)$  with  $N$  observations. Let  $Q_{N,K}$  be the distribution of the observations  $T_{1:N}$ . Then

$$d_{TV}(P_{N,\infty}, Q_{N,K}) \leq 2N \exp\left(-\frac{K-1}{\alpha}\right).$$

Proposition 5.5 implies that a TFA with  $K = O(\ln(N/\epsilon))$  atoms suffices in approximating the DP mixture model to less than  $\epsilon$  error. Modulo log factors, comparing the necessary  $\frac{1}{\epsilon}$  level for IFA and the sufficient  $\ln(\frac{1}{\epsilon})$  level for TFA, we conclude that the necessary size for normalized IFA is exponentially larger than the sufficient size for TFA, in the worst case.

## 6. Conceptual benefits of finite approximations

As part of Bayesian inference, we need to compute the posterior over the latent variables in our finite-dimensional probabilistic models (Eq. (6)). To set up notation, we denote by  $\theta = (\theta_i)_{i=1}^K$  the collection atom sizes,  $\psi = (\psi_i)_{i=1}^K$  the collection of atom locations and  $x = (x_{n,i})$  the trait count of each observation.

Standard tools to explore or approximate the posterior distribution  $\mathbb{P}(\theta, \psi, x | \text{data})$  require easy-to-simulate Gibbs conditional distributions or tractable expectations. On the one hand, because of the discreteness of the trait counts  $x$ , even with the recent advances in Hamiltonian Monte Carlo ([Hoffman and Gelman, 2014](#)), successful Markov chain Monte Carlo (MCMC) algorithms have been based largely on Gibbs sampling ([Geman and Geman, 1984](#)). In particular, blocked Gibbs sampling utilizing the natural Markov blanket structure is straightforward to implement when the complete conditionals  $\mathbb{P}(\theta | x, \psi, \text{data})$ ,  $\mathbb{P}(x | \psi, \theta, \text{data})$  or  $\mathbb{P}(\psi | x, \theta, \text{data})$  are easy to simulate from. On the other hand, variational inference using mean-field approximation and KL divergence ([Wainwright and Jordan, 2008](#)) requires analytical expectations. The variational distributions are typically chosen to match the parametric form of the complete conditionals – information about the latent variables is easily summarized, and the divergence between approximation and target is (locally) optimized using coordinate ascent updates. Such updates require expectations of the form  $\mathbb{E}_{\theta \sim q}[\ln l(x | \theta)]$  where  $q(\theta)$  is the variational distribution over atom sizes.

Since finite approximations (IFAs/TFAs) with the same number of atoms  $K$  only differ in the prior  $\mathbb{P}(\theta)$ , to compare the ease-of-use between IFAs and TFAs, it suffices to compare the tractability of  $\mathbb{P}(\theta | x, \psi, \text{data})$  under different approximations. For exponential family CRMs with  $d = 0$ , IFAs are highly compatible with standard inference schemes, because the Gibbs conditional  $\mathbb{P}(\theta | x, \psi, \text{data})$  comes from the same exponential family as the prior  $\nu_K$ .

**Lemma 6.1** (Conditional conjugacy of IFA). *Suppose the likelihood is Eq. (1) and the IFA prior  $\nu_K$  is as in Corollary 3.3. Then the complete conditional of atom sizes factorizes across atoms*

$$\mathbb{P}(\theta | x, \psi, \text{data}) = \prod_{k=1}^K \mathbb{P}(\theta_k | x_{\cdot,k}).$$

Furthermore, each  $\mathbb{P}(\theta_k | x_{\cdot,k})$  is in the same exponential family as the IFA prior, with density proportional to

$$\mathbf{1}\{\theta \in U\} \theta^{c/K + \sum_{n=1}^N \phi(x_{n,k}) - 1} \exp\left(\langle \psi + \sum_{n=1}^N t(x_{n,k}), \mu(\theta) \rangle + (\lambda + N)[-A(\theta)]\right) d\theta. \quad (16)$$

The proof follows from the results in Appendix C. Lemma 6.1 implies that the derivation of simulation steps/expectation equations for IFAs of common models such as beta-Bernoulli, gamma-Poisson and beta-negative binomial is straightforward. The complete conditionals over atom sizes are easy-to-simulate because they are well-known exponential families (beta and gamma). Also, the expectations of  $\ln l(x|\theta)$  when  $\theta$  has the exponential family distribution (Eq. (16)) are tractable because of the exponential family algebra between log-likelihood and prior. Finally, a parallelizing strategy to utilize the factorization structure across atoms can yield user-time speed up, with the gains being greatest when there are many instantiated atoms.

There are many different types of TFAs, but in general the derivation of simulation steps/expectation equations are much more involved than for IFAs. While the prior  $\mathbb{P}(\theta)$  can be reasonably easy to sample from, the incorporation of trait counts leads to intractable conditionals  $\mathbb{P}(\theta|x)$ . We consider two illustrative examples, both for exponential CRMs with  $d = 0$ . In Example 6.1, the complete conditional of atom size is both hard to sample from and leads to analytically intractable expectations. In Example 6.2, the complete conditional of atom sizes can be sampled from without introducing auxiliary variables, but important expectations are not analytically tractable.

**Example 6.1** (Stick-breaking approximation (Broderick, Jordan and Pitman, 2012; Paisley, Carin and Blei, 2011)). The following finite approximation is a TFA for BP( $\gamma, \alpha, 0$ )

$$\Theta_K = \sum_{i=1}^K \sum_{j=1}^{C_i} V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)}) \delta_{\psi_{i,j}}$$

where  $C_i \stackrel{iid}{\sim} \text{Poisson}(\gamma)$ ,  $V_{i,j}^{(l)} \stackrel{iid}{\sim} \text{Beta}(1, \alpha)$  and  $\psi_{i,j} \stackrel{iid}{\sim} H$ . A priori, the atom sizes  $V_{i,j}^{(i)} \prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)})$  can be sampled (using stick-breaking proportions  $V_{i,j}$ ), but there is no tractable way to sample from/compute expectations with respect to the conditional distribution  $\mathbb{P}(\theta|x)$  because of the dependence on  $C_i$  as well as the entangled form of each  $\theta$ . Strategies to make the model more tractable include introducing auxiliary round indicator variables  $r_k$  (Broderick, Jordan and Pitman, 2012; Paisley, Carin and Blei, 2011), marginalizing out the stick-breaking proportions (Broderick, Jordan and Pitman, 2012) or replacing the product  $\prod_{l=1}^{i-1} (1 - V_{i,j}^{(l)})$  with more succinct representation (Paisley, Carin and Blei, 2011). However, the final model from these attempts all contain at least one Gibbs conditional that is either difficult to sample from (Broderick, Jordan and Pitman, 2012, Equation 37) or lacks tractable expectations (Paisley, Carin and Blei, 2011, Section 3.3).

Other superposition-based approximations, like decoupled Bondesson or power-law (Campbell et al., 2019), will similarly struggle with the number of atoms per round variables  $C_i$  and the entanglement among the atom sizes.

**Example 6.2** (Bondesson approximation (Doshi-Velez et al., 2009; Teh, Görür and Ghahramani, 2007)). When  $\alpha = 1$ , the Bondesson approximation in Example 4.3 becomes

$$\Theta_K = \sum_{i=1}^K \left( \prod_{j=1}^i p_j \right) \delta_{\psi_i}$$

where  $p_j \stackrel{i.i.d.}{\sim} \text{Beta}(\gamma, 1)$  and  $\psi_i \stackrel{iid}{\sim} H$ . The atom sizes are tangled by the  $p_j$ 's,  $\theta_i = \prod_{j=1}^i p_j$ , but the complete conditional of atom sizes  $\mathbb{P}(\theta|x)$  admits a density with respect to Lebesgue,

and it is proportional to

$$\mathbf{1}\{0 \leq \theta_K \leq \theta_{K-1} \leq \dots \leq \theta_1 \leq 1\} \prod_{j=1}^K \theta_j^{\gamma \mathbf{1}\{j=K\} + \sum_{n=1}^N x_{n,j} - 1} (1 - \theta_j)^{N - \sum_{n=1}^N x_{n,j}}.$$

The conditional distributions  $\mathbb{P}(\theta_i | \theta_{-i}, x)$  are truncated betas, so adaptive rejection sampling (Gilks and Wild, 1992) can be used as a sub-routine to sample each  $\mathbb{P}(\theta_i | \theta_{-i}, x)$ , and then sweep over all atom sizes. However, for this exponential family, expectations of the sufficient statistics  $\ln \theta_i$  and  $\ln(1 - \theta_i)$  are not tractable: variational inference as conducted in Doshi-Velez et al. (2009) required additional approximations.

Other series-based approximations, like thinning or rejection sampling (Campbell et al., 2019), have more intractable dependencies between atom sizes in both the prior and the conditional  $\mathbb{P}(\theta | x)$ .

## 7. Empirical evaluation

We compare the practical performance of IFAs and TFAs on two real-data examples: an image denoising application using the beta-Bernoulli model and topic modeling using the modified HDP. Existing empirical work (e.g., Doshi-Velez et al. (2009, Table 1,2) and Kurihara, Welling and Teh (2007, Figure 4)) suggests two patterns: that the approximations improve in performance as the number of instantiated atoms  $K$  increase, and for the same  $K$ , normalized IFA and TFA have similar performance. Our experiments confirm and expand upon these previous findings.

### 7.1. Image denoising with beta-Bernoulli

Image denoising through dictionary learning is an application where finite approximations of BNP model - in particular beta-Bernoulli with  $d = 0$  - have proven useful (Zhou et al., 2009). The goal is recovering the original noiseless image (left of Fig. 1) from a corrupted one (right of Fig. 1). To do so, the input image is deconstructed into small contiguous patches and we postulate that each patch is a combination of underlying *basis elements*. By estimating the coefficients expressing the combination, possibly in addition to estimating the basis elements themselves, one can denoise the individual patches and ultimately the overall image. The beta-Bernoulli process allow simultaneous estimation of basis elements and basis assignments. The nonparametric nature sidesteps the cumbersome problem of calibrating the number of basis elements. The number of extracted patches depends on both the patch size and the dimensions of the input image: even on the same input image, the analysis might process a varying number of “observations.” Better denoised images have high peak signal-to-noise-ratio, or PSNR (Hore and Ziou, 2010), with respect to the noiseless image: the PSNR between two identical images is  $\infty$ .

To compare IFA and TFA, we considered beta process  $\text{BP}(\gamma, 1, 0)$  due to past work which suggests that the hyper-parameters  $\gamma, \alpha$  do not play a large role (Zhou et al., 2009). Each configuration of the latent variables  $x, \psi, \theta$  leads to a candidate denoised image. By default, a sequential<sup>2</sup> Gibbs sampler traverses the posterior over latent variables - the final denoised

<sup>2</sup>Patches i.e., observations are gradually introduced in epochs, and the sampler only modifies the latent variables of the current epoch’s observations.

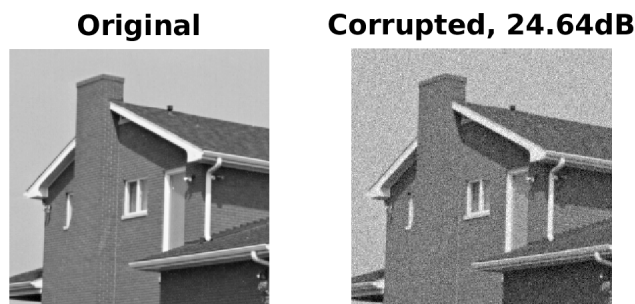


Fig 1: Original versus corrupted images. The number plotted on top of the noisy image is peak signal-to-noise-ratio, or PSNR, with respect to the noiseless image.

image is a weighted average of the candidate images encountered during the sampler run. There is randomness in how the latent variables are initialized, as well as in the simulation of the Gibbs conditionals. The gradual data introduction employed in the Gibbs sampler can be thought of as a way to initialize the latent variables for the entire set of observations. For a  $256 \times 256$  image like the right panel of Fig. 1, the number of extracted patches,  $N$ , is about  $60k$ . More details about the finite approximations, hyper-parameter settings and inference can be found in Appendix H.1.

In Fig. 2, the quality of denoised images improves with increasing  $K$  – furthermore, the quality is very similar across the two types of approximation. Both kinds perform much better than the baseline i.e., noisy input image. The improvement with  $K$  is largest for small  $K$ , and plateaus for larger values of  $K$ . For a given approximation level, the quality of TFA denoising and that of IFA are almost the same. Furthermore, the denoised image from TFA is more similar to the denoised image from IFA than it is similar to the original image, indicated by the large gap in PSNR. The error bars reflect randomness in both initialization and simulation of the conditionals across 5 trials.

Fig. 3 shows that the modes of TFA posterior are centers of regions of attraction in IFA posterior, and vice-versa. For both kinds of approximation,  $K = 60$ . Rather than randomly initializing the latent variables at the beginning of the Gibbs sampler of one model i.e., cold start, we can use the last configuration of latent variables visited in the other model as the initial state of the Gibbs sampler – i.e., warm start. To isolate the effect of the initial conditions, all the patches are available from the start as opposed to being gradually introduced. For both kinds of approximation, the Gibbs sampler initialized at the warm start visits candidate images that basically have the same PSNR as the starting configuration. The early iterates of cold-start Gibbs sampler are noticeably lower in quality compared to the warm-start iterates, and the quality at the plateau is still lower than that of the warm start. Each trace of PSNR of cold-start Gibbs corresponds to a random seed in initialization and simulation of the conditionals, while each trace of warm-start PSNR corresponds to a different final state of the alternative model’s training. The variation across warm starts is tiny – the variation across cold starts is larger but still very small.

Experiments on other noisy images can be found in Appendix I; the trends are the same.

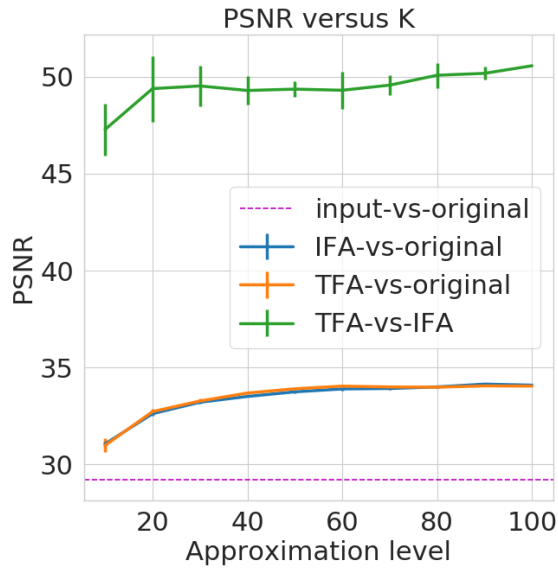


Fig 2: Finite approximations have similar performance across approximation levels. For each  $K$ , the final denoised image is a weighted average of candidate images encountered during Gibbs sampling.

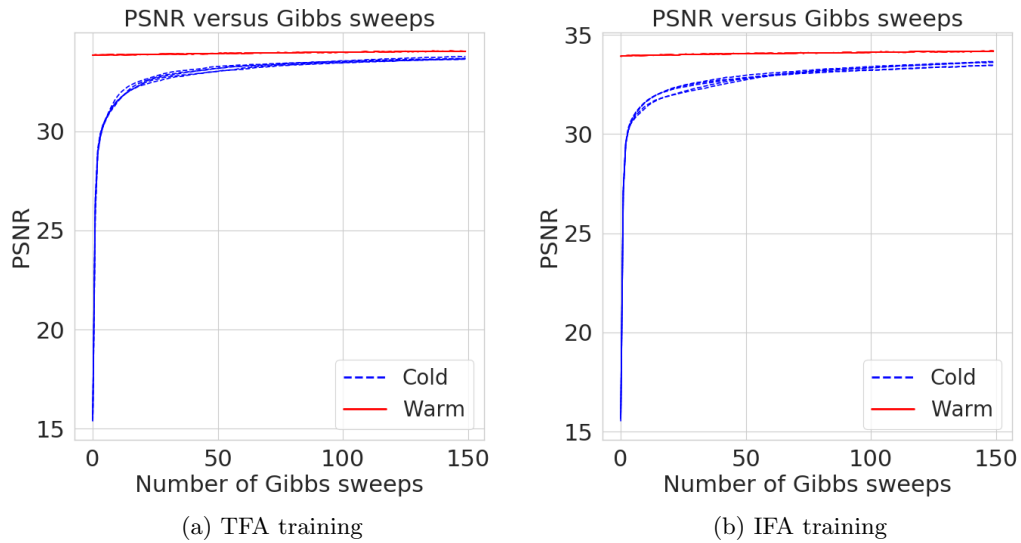


Fig 3: The output of one model is a good initialization for the training of the other one.

**7.2. Topic modelling with modified hierarchical Dirichlet process**

Finally, we compare the performance of normalized IFA (i.e.,  $FSD_K$ ) and TFA (i.e.,  $TSB_K$ ) when used in DP-based model. In this section, we provide evidence of the same trends in the modified HDP – a more complicated model than a Dirichlet process mixture – when

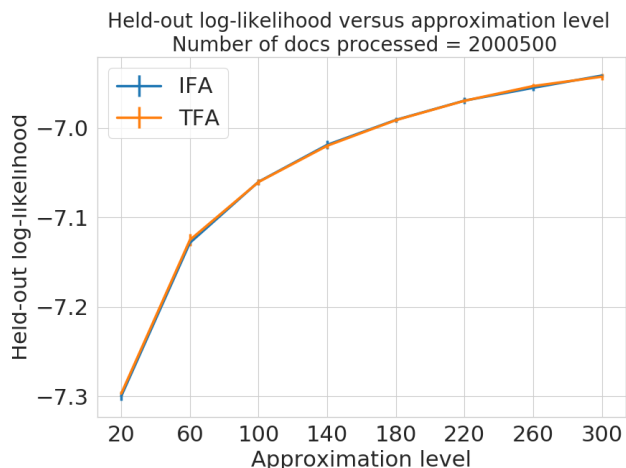


Fig 4: Finite approximations have similar performance across approximation levels.

analyzing Wikipedia documents.

For both IFA and TFA, we use stochastic variational inference with mean-field factorization (Hoffman et al., 2013) to approximate the posterior over the latent topics based on training documents. The training corpus is nearly one million documents from Wikipedia. There is randomness in the initial values of the variational parameters, as well as in the order that data minibatches are processed. The quality of inferred topics is measured by the predictive log-likelihood on a set of  $10k$  held-out documents. More details about the finite approximations, hyper-parameter settings, variational inference and definition of test log-likelihood can be found in Appendix H.2.

In Fig. 4, the quality of the inferred topics improves as the approximation level grows – furthermore, the quality is very similar across the two types of approximation. The improvement with  $K$  is largest for small  $K$ : the slope plateaus for large  $K$ . For a given approximation level, the quality of TFA topics and that of normalized IFA are almost the same. The error bars reflect variation across both the random initialization and the ordering of data minibatches processed by stochastic variational inference.

In Fig. 5, the modes of TFA posterior are centers of regions of attraction in IFA posterior, and vice-versa. The number of topics is fixed to be  $K = 300$ . Rather than randomly initializing the variational parameters at the start of variational inference of one model i.e., cold start, we can use the variational parameters at the end of the other model’s training as the initialization i.e., warm start. The learning rate for warm-start training is slightly different from that for cold start, to reflect the fact that many batches of data had been processed leading up to the warm-start variational parameters. For both kinds of approximation, the test log-likelihood basically stays the same for warm-start training iterates, hinting that such initialization is part of an attractive region. The early iterates of cold start are noticeably lower in quality compared to the warm iterates – however at the end of training, the test log-likelihoods are nearly the same. Each trace of cold start corresponds to a different initialization and ordering of data batches processed. Each trace of warm start corresponds to a different output of the other model’s training and a different ordering of data batches processed. The variation across either cold starts or warm starts is small.

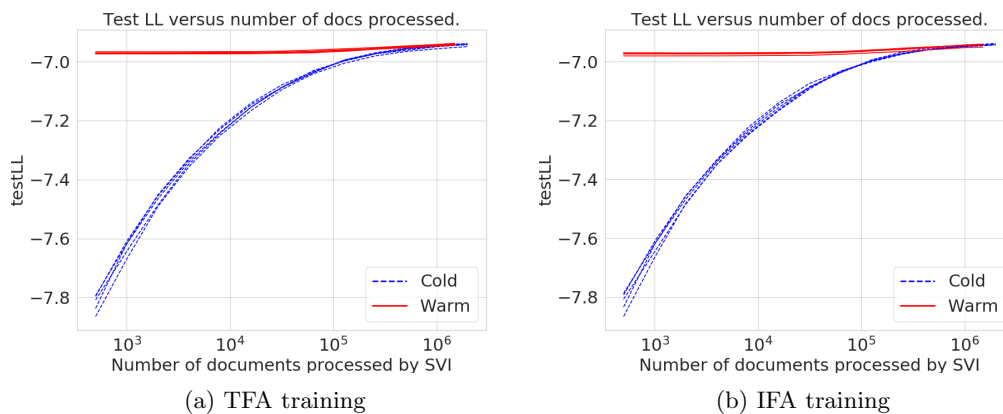


Fig 5: The output of one model is a good initialization for the training of the other one.

## 8. Discussion

We have provided a general construction of independent finite approximations for completely random measures, analyzed error bounds on IFAs for conjugate exponential family CRM with no power law and the Dirichlet process, and investigated how they compare to truncated finite approximations in realistic data applications. Our error bounds reveal that in the worst case, for the same number of atoms instantiated, IFA has larger error than TFA. However, we have not observed the worst case in our experiments, suggesting that either the error bounds can be tightened for relevant conditional densities  $f$  or that additional sources of error, such as those from approximate inference, dominate approximation error made by the finite approximations. From a practical point of view, IFA is easier than TFA to work with.

Our analyses and experiments suggest a number of directions for future work. For example, the error bound analysis could be extended for conjugate family CRM with power-law behavior. We speculate that in such situations, the  $O(\ln N)$  factor appearing in the numerator of the upper bounds will be replaced by  $O(N^a)$  where  $O(N^a)$  is the growth rate of BNP models with power law behavior.

## References

- ACHARYA, A., GHOSH, J. and ZHOU, M. (2015). Nonparametric Bayesian factor analysis for dynamic count matrices. In *AISTATS*.
- ADELL, J. A. and LEKUONA, A. (2005). Sharp estimates in signed Poisson approximation of Poisson mixtures. *Bernoulli* **11** 47–65.
- ALDOUS, D. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983* 1–198.
- ALZER, H. (1997). On some inequalities for the gamma and psi functions. *Mathematics of computation* **66** 373–389.
- ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2** 1152–1174.
- BARBOUR, A. D. and HALL, P. (1984). On the rate of Poisson convergence. In *Mathematical Proceedings of the Cambridge Philosophical Society* **95** 473–480. Cambridge University Press.

- BETANCOURT, M. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv.org*.
- BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *Ann. Statist.* **1** 353–355.
- BLEI, D. M., GRIFFITHS, T. L. and JORDAN, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* **57** 1–30.
- BLEI, D. M. and JORDAN, M. I. (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis* **1** 121–144.
- BONDESSON, L. (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability*.
- BRIX, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* **31** 929–953.
- BRODERICK, T., JORDAN, M. I. and PITMAN, J. (2012). Beta processes, stick-breaking and power laws. *Bayesian analysis* **7** 439–476.
- BRODERICK, T., WILSON, A. C. and JORDAN, M. I. (2018). Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli* **24** 3181–3221.
- BRODERICK, T., MACKEY, L., PAISLEY, J. and JORDAN, M. I. (2015). Combinatorial Clustering and the Beta Negative Binomial Process. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** 290–306.
- CAMPBELL, T., HUGGINS, J. H., HOW, J. P. and BRODERICK, T. (2019). Truncated random measures. *Bernoulli* **25** 1256–1288.
- CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76**.
- DOSHI-VELEZ, F., MILLER, K. T., VAN GAEL, J. and TEH, Y. W. (2009). Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics* 137–144.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 209–230.
- FERGUSON, T. S. and KLASS, M. J. (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics* **43**.
- FOX, E. B., SUDDERTH, E., JORDAN, M. I. and WILLSKY, A. S. (2010). A Sticky HDP-HMM with Application to Speaker Diarization. *The Annals of Applied Statistics* **5** 1020–1056.
- GEMAN, S. and GEMAN, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **6** 721–741.
- GILKS, W. R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **41** 337–348.
- GNEDIN, A. V. (1998). On convergence and extensions of size-biased permutations. *Journal of Applied Probability* **35** 642650.
- GORDON, L. (1994). A Stochastic Approach to the Gamma Function. *The American Mathematical Monthly* **101** 858–865.
- GRIFFITHS, T. L. and GHAHRAMANI, Z. (2005). Infinite Latent Feature models and the Indian Buffet Process. In *Advances in Neural Information Processing Systems*.
- GRIFFITHS, T. L. and GHAHRAMANI, Z. (2011). The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research* **12** 1185–1224.
- HJORT, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models

- for life history data. *the Annals of Statistics* **18** 1259–1294.
- HOFFMAN, M., BACH, F. R. and BLEI, D. M. (2010). Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* 856–864.
- HOFFMAN, M. D. and GELMAN, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593–1623.
- HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research* **14** 1303–1347.
- HORE, A. and ZIOU, D. (2010). Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition* 2366–2369. IEEE.
- ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**.
- ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics* **30** 269–283.
- JAMES, L. F. (2013). Stick-breaking  $PG(\alpha, \zeta)$ -Generalized Gamma Processes. *arXiv.org*.
- JAMES, L. F. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian Buffet Process priors. *The Annals of Statistics* **45** 2016–2045.
- JOHNSON, N. L., KEMP, A. W. and KOTZ, S. (2005). *Univariate Discrete Distributions. Wiley Series in Probability and Statistics*. Wiley.
- JOHNSON, M. J. and WILLSKY, A. S. (2013). Bayesian Nonparametric Hidden Semi-Markov Models. *Journal of Machine Learning Research* **14** 673–701.
- KALLENBERG, O. (2002). *Foundations of modern probability*, 2nd ed. Springer, New York.
- KINGMAN, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21** 59–78.
- KINGMAN, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society B* **37** 1–22.
- KUCUKELBIR, A., RANGANATH, R., GELMAN, A. and BLEI, D. M. (2015). Automatic Variational Inference in Stan. In *Advances in Neural Information Processing Systems*.
- KURIHARA, K., WELLING, M. and TEH, Y. W. (2007). Collapsed Variational Dirichlet Process Mixture Models. In *International Joint Conference on Artificial Intelligence* 2796–2801.
- LAST, G. and PENROSE, M. (2017). *Lectures on the Poisson Process. Institute of Mathematical Statistics Textbooks*.
- LE CAM, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* **10** 1181–1197.
- LEE, J., JAMES, L. F. and CHOI, S. (2016). Finite-dimensional BFRY priors and variational Bayesian inference for power law models. In *Advances in Neural Information Processing Systems* 3162–3170.
- LEE, J., MISCOURIDOU, X. and CARON, F. (2019). A unified construction for series representations and finite approximations of completely random measures. *arXiv preprint arXiv:1905.10733*.
- LOEVE, M. (1956). Ranking Limit Problem. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory* 177–194. University of California Press, Berkeley, Calif.
- MADRAS, N. and SEZER, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli* **16** 882–908.
- MILLER, J. W. and HARRISON, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information*

- Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 199–206. Curran Associates, Inc.
- NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* 113–162. Chapman and Hall/CRC.
- ORBANZ, P. (2010). Conjugate Projective Limits. *arXiv.org*.
- PAISLEY, J., BLEI, D. M. and JORDAN, M. I. (2012). Stick-breaking beta processes and the Poisson process. In *Artificial Intelligence and Statistics* 850–858.
- PAISLEY, J. and CARIN, L. (2009). Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning. ICML '09* 777–784. ACM, New York, NY, USA.
- PAISLEY, J., CARIN, L. and BLEI, D. (2011). Variational inference for stick-breaking beta process priors. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* 889–896.
- PALLA, K., KNOWLES, D. A. and GHAHRAMANI, Z. (2012). An Infinite Latent Attribute Model for Network Data. In *International Conference on Machine Learning*. University of Cambridge.
- PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92** 21–39.
- PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102** 145–158.
- PITMAN, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series* 245–267.
- POLLARD, D. (2001). *A User's Guide to Measure Theoretic Probability* **8**. Cambridge University Press.
- RANGANATH, R., GERRISH, S. and BLEI, D. M. (2014). Black Box Variational Inference. In *International Conference on Artificial Intelligence and Statistics* 814–822.
- ROBERTS, G. O. and TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363.
- ROYCHOWDHURY, A. and KULIS, B. (2015). Gamma processes, stick-breaking, and variational inference. In *Artificial Intelligence and Statistics* 800–808.
- SARIA, S., KOLLER, D. and PENN, A. (2010). Learning individual and population level traits from clinical temporal data Technical Report.
- SETHURAMAN, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica* **4** 639–650.
- TEH, Y. W., GÖRÜR, D. and GHAHRAMANI, Z. (2007). Stick-breaking Construction for the Indian Buffet Process. In *International Conference on Artificial Intelligence and Statistics*.
- TEH, Y. W. and GÖRÜR, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*.
- TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* **101** 1566–1581.
- THIBAU, R. and JORDAN, M. I. (2007). Hierarchical Beta Processes and the Indian Buffet Process. In *International Conference on Artificial Intelligence and Statistics*.
- TITSIAS, M. (2008). The infinite gamma-Poisson feature model. In *Advances in Neural Information Processing Systems*.
- WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning* **1** 1–305.
- WANG, C., PAISLEY, J. and BLEI, D. (2011). Online variational inference for the hier-

archical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 752–760.

- ZHOU, M., CHEN, H., REN, L., SAPIRO, G., CARIN, L. and PAISLEY, J. W. (2009). Non-parametric Bayesian dictionary learning for sparse image representations. In *Advances in Neural Information Processing Systems 22* (Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta, eds.) 2295–2303. Curran Associates, Inc.
- ZHOU, M., HANNAH, L., DUNSON, D. and CARIN, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *Artificial Intelligence and Statistics* 1462–1471.

## Appendix A: Additional examples of IFA construction

Let  $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$  denote the beta function.

**Example A.1** (Beta process). Taking  $E = \mathbb{R}_+$ ,  $g(\theta) = 1$ ,  $h(\theta; \eta) = (1 - \theta)^{\eta-1} \mathbf{1}[\theta \leq 1]$ , and  $Z(\xi, \eta) = B(\xi, \eta)$  in Theorem 3.2 yields the beta process  $\text{BP}(\gamma, \eta - d, d)$ , which has rate measure

$$\nu(d\theta) = \gamma \frac{\mathbf{1}[\theta \leq 1]}{B(\eta, 1-d)} \theta^{-1-d} (1-\theta)^{\eta-1} d\theta.$$

Since  $h$  is continuous and bounded on  $[0, 1/2]$ , Assumption 1 hold.

**Example A.2** (Beta prime process). Taking  $E = \mathbb{R}_+$ ,  $g(\theta) = (1 + \theta)^{-1}$ ,  $h(\theta; \eta) = (1 + \theta)^{-\eta}$ , and  $Z(\xi, \eta) = B(\xi, \eta)$  in Theorem 3.2 yields the beta prime process, which has rate measure

$$\nu(d\theta) = \frac{\gamma}{B(\eta, 1-d)} \theta^{-1-d} (1+\theta)^{-d-\eta} d\theta.$$

Since  $g$  is continuous,  $g(0) = 1$ ,  $1 \leq g(\theta) \leq 1 + \theta$ , and  $h(\theta; \eta)$  is continuous and bounded on  $[0, 1]$ , Assumption 1 hold. In the case of  $d = 0$ ,  $c = \gamma\eta$  and

$$\nu_n(\theta) = \text{Beta}'(\theta; \gamma\eta/n, \eta).$$

**Example A.3** (Gamma process). Taking  $E = \mathbb{R}_+$ ,  $g(\theta) = 1$ ,  $h(\theta; \eta) = e^{-\eta\theta}$ , and  $Z(\xi, \eta) = \Gamma(\xi)\eta^{-\xi}$  in Theorem 3.2 yields the gamma process, with rate measure

$$\nu(d\theta) = \gamma \frac{\lambda^{1-d}}{\Gamma(1-d)} \theta^{-d-1} e^{-\lambda\theta} d\theta.$$

Since  $h(\theta; \eta)$  is continuous and bounded on  $[0, 1]$ , Assumption 1 hold.

**Example A.4** (Generalized gamma process). Taking  $E = \mathbb{R}_+^2$ ,  $g(\theta) = 1$ ,  $h(\theta; \eta) = e^{-(\eta_1\theta)^{\eta_2}}$ , and  $Z(\xi, \eta) = \Gamma(\xi/\eta_2)(\eta_1\eta_2)^{-\xi}$  in Theorem 3.2 yields the generalized gamma distribution  $\text{Gam}(\xi, \eta_1, \eta_2)$ .<sup>3</sup> The corresponding rate measure is

$$\nu(d\theta) = \frac{\gamma(\eta_1\eta_2)^{1-d}}{\Gamma((1-d)/\eta_2)} \theta^{-d-1} e^{-(\eta_1\theta)^{\eta_2}} d\theta,$$

which is the rate measure for the gamma process  $\text{GP}(\gamma, \eta, d)$ . Since  $h(\theta; \eta)$  is continuous and bounded on  $[0, 1]$ , Assumption 1. In the case of  $d = 0$ ,  $c = \frac{\gamma\eta_1\eta_2}{\Gamma(\eta_2^{-1})}$  and

$$\nu_n(\theta) = \text{Gam}\left(\theta; \frac{\gamma\eta_1\eta_2}{n\Gamma(\eta_2^{-1})}, \eta_1, \eta_2\right).$$

<sup>3</sup>[https://en.wikipedia.org/wiki/Generalized\\_gamma\\_distribution](https://en.wikipedia.org/wiki/Generalized_gamma_distribution)

## Appendix B: Proof of IFA convergence

### B.1. IFA converges to CRM in distribution

In order to prove our main result, we require a few auxiliary results.

**Lemma B.1** ((Kallenberg, 2002, Lemmas 12.1 and 12.2)). *Let  $\Theta$  be a random measure and  $\Theta_1, \Theta_2, \dots$  a sequence of random measures. If for all measurable sets  $A$  and  $t > 0$ ,*

$$\lim_{K \rightarrow \infty} \mathbb{E}[e^{-t\Theta_K(A)}] = \mathbb{E}[e^{-t\Theta(A)}],$$

then  $\Theta_K \xrightarrow{\mathcal{D}} \Theta$ .

For a density  $f$ , let  $\mu(t, f) : \theta \mapsto (1 - e^{-t\theta})f(\theta)$ . In results that follow we assume all measures on  $\mathbb{R}_+$  have densities with respect to Lebesgue measure. We abuse notation and use the same symbol to denote the measure and the density.

**Proposition B.2.** *Let  $\Theta \sim \text{CRM}(H, \nu)$  and for  $K = 1, 2, \dots$ , let  $\Theta_K \sim \text{IFA}_K(H, \nu_K)$  where  $\nu$  is a measure and  $\nu_1, \nu_2, \dots$  are probability measures on  $\mathbb{R}_+$ , all absolutely continuous with respect to Lebesgue measure. If  $\|\mu(1, n\nu_K) - \mu(1, \nu)\|_1 \rightarrow 0$ , then  $\Theta_K \xrightarrow{\mathcal{D}} \Theta$ .*

*Proof.* Let  $t > 0$  and  $A$  a measurable set. First, recall that the Laplace functional of the CRM  $\Theta$  is

$$\mathbb{E}[e^{-t\Theta(A)}] = \exp \left\{ -H(A) \int_0^\infty \mu(t, \nu)(\theta) d\theta \right\}.$$

We have

$$\begin{aligned} \mathbb{E}[e^{-t\theta_{K,1} \mathbb{1}(\psi_{K,1} \in A)}] &= \mathbb{P}(\psi_{K,1} \in A) \mathbb{E}[e^{-t\theta_{K,1}}] + \mathbb{P}(\psi_{K,1} \notin A) \\ &= H(A) \mathbb{E}[e^{-t\theta_{K,1}}] + 1 - H(A) \\ &= 1 - H(A)(1 - \mathbb{E}[e^{-t\theta_{K,1}}]) \\ &= 1 - \frac{H(A)}{K} \int_0^\infty \mu(t, K\nu_K)(\theta) d\theta. \end{aligned}$$

Since  $\frac{|1 - e^{-t\theta}|}{|1 - e^{-\theta}|} \leq \max(1, t)$ , it follows by hypothesis that  $\|\mu(t, K\nu_K) - \mu(t, \nu)\|_1 \rightarrow 0$ . Thus, by dominated convergence and the standard exponential limit,

$$\begin{aligned} \lim_{K \rightarrow \infty} \mathbb{E}[e^{-t\theta_{K,1} \mathbb{1}(\psi_{K,1} \in A)}]^K &= \lim_{K \rightarrow \infty} \left( 1 - \frac{H(A)}{K} \int_0^\infty \mu(t, K\nu_K)(\theta) d\theta \right)^K \\ &= \exp \left\{ - \lim_{K \rightarrow \infty} H(A) \int_0^\infty \mu(t, K\nu_K)(\theta) d\theta \right\} \\ &= \exp \left\{ -H(A) \int_0^\infty \mu(t, \nu)(\theta) d\theta \right\}. \end{aligned}$$

Finally, by the independence of the random variables  $\{\theta_{K,i}\}_{i=1}^K$ ,

$$\lim_{K \rightarrow \infty} \mathbb{E}[e^{-t\Theta_K(A)}] = \lim_{K \rightarrow \infty} \mathbb{E}[e^{-t\theta_{K,1} \mathbb{1}(\psi_{K,1} \in A)}]^K,$$

so result follows from Lemma B.1.  $\square$

**Lemma B.3.** *If there exist measures  $\pi(\theta) d\theta$  and  $\pi'(\theta) d\theta$  on  $\mathbb{R}_+$  such that for some  $\kappa > 0$ ,*

1. *the measures  $\mu, \mu_1, \mu_2, \dots$  have densities  $f, f_1, f_2, \dots$  wrt  $\pi$  and densities  $f', f'_1, f'_2, \dots$  wrt  $\pi'$ ,*
2.  $\int_0^\kappa |f'(\theta) - f'_K(\theta)| d\theta \rightarrow 0$ ,
3.  $\sup_{\theta \in [\kappa, \infty)} |f(\theta) - f_K(\theta)| \rightarrow 0$ ,
4.  $\sup_{\theta \in [0, \kappa]} \pi'(\theta) \leq c' < \infty$ , and
5.  $\int_\kappa^\infty \pi(\theta) d\theta \leq c < \infty$ ,

then

$$\|\mu - \mu_K\|_1 \rightarrow 0.$$

*Proof.* We have, using the assumptions and Hölder's inequality,

$$\begin{aligned} \|\mu - \mu_K\|_1 &= \int_0^\kappa |f'(\theta) - f'_K(\theta)| \pi'(d\theta) + \int_\kappa^\infty |f(\theta) - f_K(\theta)| \pi(d\theta) \\ &\leq \left( \sup_{\theta \in [0, \kappa]} \pi'(\theta) \right) \int_0^\kappa |f'(\theta) - f'_K(\theta)| d\theta \\ &\quad + \left( \sup_{\theta \in [\kappa, \infty)} |f(\theta) - f_K(\theta)| \right) \int_\kappa^\infty \pi(d\theta) \\ &\leq c' \int_0^\kappa |f'(\theta) - f'_K(\theta)| d\theta + c \sup_{\theta \in [\kappa, \infty)} |f(\theta) - f_K(\theta)|. \end{aligned}$$

The conclusion follows by dominated convergence.  $\square$

*Proof of Theorem 3.2.* Note that since  $h$  is continuous and bounded on  $[0, \epsilon]$ ,  $c < \infty$ . We will apply Lemma B.3 with  $\kappa$  as given in the theorem statement,  $\mu = \mu(1, \nu)$ ,  $\mu_K = \mu(1, n\nu_K)$ ,

$$\pi(\theta) = p(\theta; 1 - d, \eta) = \frac{\theta^{-d} g(\theta)^{1-d} h(\theta; \eta)}{Z(1 - d, \eta)},$$

and  $\pi'(\theta) := (\theta g(\theta))^d \pi(\theta)$ . Thus,  $f(\theta) = \gamma(1 - e^{-\theta})(\theta g(\theta))^{-1}$ ,

$$f_K(\theta) = n Z_K^{-1} (1 - e^{-\theta}) \theta^{-1+cK^{-1}+d-dS_{b_K}(\theta-aK^{-1})} g(\theta)^{-1+cK^{-1}},$$

and  $f'(\theta) = (\theta g(\theta))^{-d} f(\theta)$ , and  $f'_K(\theta) = (\theta g(\theta))^{-d} f_K(\theta)$ .

We now note a few useful properties that we will use repeatedly in the proof. Observe that  $(a/K)^{cK^{-1}} = 1 + o(1)$ . The assumption that  $h$  is bounded and continuous implies that on  $[0, a/K]$ ,  $h(\theta; \eta) = h(0; \eta) + o(1)$ . Similarly, for any  $\delta > 0$ ,  $g(\theta)$  is bounded and continuous for  $\theta \in [0, \delta]$  and therefore, together with the fact that  $g(0) = 1$ , we can conclude that on  $[0, a/K]$ ,  $g(\theta) = 1 + o(1)$ .

For the remainder of the proof we will consider  $K$  large enough that  $aK^{-1} + 2b_K$  and  $cK^{-1}$  are less than  $\kappa$ . The normalizing constant  $Z_K$  can be written as

$$\begin{aligned} Z_K &= \int_0^{a/K} (\theta g(\theta))^{-1+cK^{-1}} \pi'(d\theta) \\ &\quad + \int_{a/K}^\kappa \theta^{-1+cK^{-1}-dS_{b_K}(\theta-aK^{-1})} g(\theta)^{-1+cK^{-1}} \pi'(d\theta) \\ &\quad + \int_\kappa^\infty (\theta g(\theta))^{-1+cK^{-1}-d} \pi'(d\theta). \end{aligned}$$

We rewrite each term in turn. For the first term,

$$\begin{aligned} \int_0^{a/K} \theta^{-1+cK^{-1}} g(\theta)^{-1+cK^{-1}} \pi'(\mathrm{d}\theta) &= (c/\gamma + o(1)) \int_0^{a/K} \theta^{-1+cK^{-1}} \mathrm{d}\theta \\ &= (c/\gamma + o(1)) \frac{K}{c} \left(\frac{a}{K}\right)^{cK^{-1}} \\ &= \frac{K}{\gamma} + o(K). \end{aligned}$$

Since  $\kappa \leq 1$  and  $S_{b_K} \in [0, 1]$ , for  $\theta \in [a/K, \kappa]$ ,  $\theta^{-dS_{b_K}(\theta-aK^{-1})} \leq \theta^{-d}$ . Since  $g(0) = 1$ ,  $c_* \leq 1$  and therefore  $g(\theta)^{-1+cK^{-1}} \leq c_*^{-1+c}$ . Hence the second term is upper bounded by

$$\begin{aligned} c_*^{-1+c} \int_{a/K}^{\kappa} \theta^{-1+cK^{-1}-d} \pi'(\mathrm{d}\theta) &\leq c_*^{-1}(c/\gamma + O(1)) \frac{K^d}{a^d} \frac{K}{c} (\kappa^{cK^{-1}} - (a/K)^{cK^{-1}}) \\ &= O(K^d) \times O(\log K) \\ &= o(K). \end{aligned}$$

For the third term,

$$\begin{aligned} \int_{\kappa}^{\infty} (\theta g(\theta))^{-1+cK^{-1}-d} \pi'(\mathrm{d}\theta) &= \int_{\kappa}^{\infty} (\theta g(\theta))^{-1+cK^{-1}} \pi(\mathrm{d}\theta) \\ &\leq (\kappa c_*)^{-1+cK^{-1}} \int_{\kappa}^{\infty} \pi(\mathrm{d}\theta) \\ &\leq (\kappa c_*)^{-1}. \end{aligned}$$

Hence,  $Z_K = \frac{K}{\gamma} + o(K)$  and  $KZ_K^{-1} = \gamma(1 + e_K)$ , where  $e_K = o(1)$ .

Next, we have

$$\begin{aligned} &\sup_{\theta \in [\kappa, \infty)} |f(\theta) - f_K(\theta)| \\ &= \sup_{\theta \in [\kappa, \infty)} (1 - e^{-\theta})(\theta g(\theta))^{-1} |\gamma - KZ_K^{-1}(\theta g(\theta))^{cK^{-1}}| \\ &\leq \sup_{\theta \in [\kappa, \infty)} \gamma(\theta g(\theta))^{-1} |1 - (1 + e_K)(\theta g(\theta))^{cK^{-1}}| \\ &\leq \gamma \sup_{\theta \in [\kappa, \infty)} (\theta g(\theta))^{-1} |1 - (\theta g(\theta))^{cK^{-1}}| \\ &\quad + \gamma e_K \sup_{\theta \in [\kappa, \infty)} (\theta g(\theta))^{-1+cK^{-1}}. \end{aligned} \tag{B.1}$$

To bound the two terms we will use the fact that if  $\theta \geq \kappa$ , then

$$\theta g(\theta) \geq \frac{\theta}{c^*(1+\theta)} \geq \frac{\kappa}{c^*(1+\kappa)} =: \tilde{\kappa}$$

and if  $\theta \leq 1$  then  $\theta g(\theta) \leq c_* \leq 1$ . Hence, letting  $\psi := \theta g(\theta)$ , for the first term in Eq. (B.1)

we have

$$\begin{aligned}
& \gamma \sup_{\theta \in [\kappa, \infty)} (\theta g(\theta))^{-1} |1 - (\theta g(\theta))^{cK^{-1}}| \\
& \leq \gamma \sup_{\psi \in [\tilde{\kappa}, \infty)} \psi^{-1} |1 - \psi^{cK^{-1}}| \\
& \leq \gamma \sup_{\psi \in [\tilde{\kappa}, 1]} \psi^{-1} |1 - \psi^{cK^{-1}}| + \gamma \sup_{\psi \in [1, \infty)} \psi^{-1} |1 - \psi^{cK^{-1}}| \\
& \leq \gamma \tilde{\kappa}^{-1} \sup_{\psi \in [\tilde{\kappa}, 1]} |1 - \psi^{cK^{-1}}| + \gamma \left( \frac{K-c}{K} \right)^{Kc^{-1}} \left| 1 - \frac{K}{K-c} \right| \\
& \leq \gamma \tilde{\kappa}^{-1} (1 - \tilde{\kappa}^{cK^{-1}}) + O(1) \times \frac{c}{K-c} \\
& = \gamma \tilde{\kappa}^{-1} \times o(1) + O(K^{-1}) \\
& \rightarrow 0.
\end{aligned}$$

Similarly, for the second term in Eq. (B.1) we have

$$\begin{aligned}
\gamma e_K \sup_{\theta \in [\kappa, \infty)} (\theta g(\theta))^{-1+cK^{-1}} & \leq \gamma e_K \sup_{\psi \in [\tilde{\kappa}, \infty)} \psi^{-1+cK^{-1}} \\
& \leq \gamma \tilde{\kappa}^{-1} e_K \\
& \rightarrow 0.
\end{aligned}$$

Since  $g(\theta)$  is bounded on  $[0, \kappa]$ ,  $g(\theta)^{cK^{-1}} = 1 + o(1)$  and therefore  $(1 + e_K)g(\theta)^{cK^{-1}} = 1 + e'_K$ , where  $e'_K = o(1)$ . Using this observation together with the bound  $(1 - e^{-\theta})\theta^{-1} \leq 1$ , we have

$$\begin{aligned}
& \int_0^\kappa |f'(\theta) - f'_K(\theta)| d\theta = \int_0^\kappa (\theta g(\theta))^{-d} |f(\theta) - f_K(\theta)| d\theta \\
& = \int_0^\kappa (1 - e^{-\theta})(\theta g(\theta))^{-1-d} |\gamma - K Z_K^{-1} \theta^{cK^{-1}+d-dS_{b_K}(\theta-aK^{-1})} g(\theta)^{cK^{-1}}| d\theta \\
& \leq \gamma [c^*(1 + \kappa)]^{1+d} \int_0^\kappa \theta^{-d} |1 - (1 + e'_K) \theta^{cK^{-1}+d-dS_{b_K}(\theta-aK^{-1})}| d\theta \\
& \leq \gamma \int_0^\kappa \theta^{-d} |1 - \theta^{cK^{-1}+d-dS_{b_K}(\theta-aK^{-1})}| d\theta + \gamma e'_K \int_0^\kappa \theta^{cK^{-1}+d-dS_{b_K}(\theta-aK^{-1})} d\theta. \quad (\text{B.2})
\end{aligned}$$

We bound the first integral in Eq. (B.2) in four parts: from 0 to  $aK^{-1}$ , from  $aK^{-1}$  to  $aK^{-1} + b_K$ , from  $aK^{-1} + b_K$  to  $\kappa - b_K$ , and from  $\kappa - b_K$  to  $\kappa$ . The first part is equal to

$$\begin{aligned}
& \int_0^{aK^{-1}} \theta^{-d} |1 - \theta^{d+cK^{-1}}| d\theta \leq \int_0^{aK^{-1}} \theta^{-d} + \theta^{cK^{-1}} d\theta \\
& = \frac{\theta^{1-d}}{1-d} + \frac{K}{c+K} \theta^{1+cK^{-1}} \Big|_0^{aK^{-1}} \\
& = \frac{1}{1-d} (aK^{-1})^{1-d} + \frac{K}{c+K} (aK^{-1})^{1+cK^{-1}} \\
& \rightarrow 0.
\end{aligned}$$

The second part is equal to

$$\begin{aligned}
\int_{aK^{-1}}^{aK^{-1}+b_K} \theta^{-d} |1 - \theta^{cK^{-1}+d-dS_{b_K}(\theta-aK^{-1})}| d\theta &\leq \int_{aK^{-1}}^{aK^{-1}+b_K} \theta^{-d} + \theta^{cK^{-1}-d} d\theta \\
&\leq 2 \int_{aK^{-1}}^{aK^{-1}+b_K} \theta^{-d} d\theta \\
&= \frac{2}{1-d} \theta^{1-d} \Big|_{aK^{-1}}^{aK^{-1}+b_K} \\
&= \frac{2}{1-d} ((aK^{-1} + b_K)^{1-d} - (aK^{-1})^{1-d}) \\
&\rightarrow 0.
\end{aligned}$$

The third part is equal to

$$\begin{aligned}
\int_{aK^{-1}+b_K}^{\kappa-b_K} \theta^{-d} |1 - \theta^{cK^{-1}}| d\theta &= \int_{aK^{-1}+b_K}^{\kappa-b_K} \theta^{-d} - \theta^{cK^{-1}-d} d\theta \\
&= \frac{1}{1-d} \theta^{1-d} - \frac{K}{c+K(1-d)} \theta^{1-d+cK^{-1}} \Big|_{aK^{-1}+b_K}^{\kappa-b_K} \\
&= \frac{(\kappa-b_K)^{1-d}}{1-d} - \frac{K}{c+K(1-d)} (\kappa-b_K)^{1-d+cK^{-1}} \\
&\quad - \frac{(aK^{-1}+b_K)^{1-d}}{1-d} + \frac{K}{c+K} (aK^{-1}+b_K)^{1-d+cK^{-1}} \\
&\rightarrow 0.
\end{aligned}$$

The fourth part is equal to

$$\begin{aligned}
\int_{\kappa-b_K}^{\kappa} \theta^{-d} |1 - \theta^{cK^{-1}}| d\theta &\leq \int_{\kappa-b_K}^{\kappa} \theta^{-d} + \theta^{cK^{-1}-d} d\theta \\
&\rightarrow 0
\end{aligned}$$

using the same argument as the second part. The second integral in Eq. (B.2) is upper bounded by

$$\gamma e'_K \int_0^{\kappa} \theta^{cK^{-1}-dS_{b_K}(\theta-aK^{-1})} d\theta \leq \gamma e'_K \int_0^{\kappa} \theta^{-d} d\theta = \gamma e'_K \frac{\kappa^{1-d}}{1-d} = o(K).$$

Since  $\sup_{\theta \in [0, \kappa]} \pi'(\theta) < \infty$  by the boundedness of  $g$  and  $h$  and  $\pi$  is a probability density by construction, conclude using Lemma B.3 that  $\|\mu - \mu_K\|_1 \rightarrow 0$ . It then follows from Lemma B.1 that  $\Theta_K \xrightarrow{\mathcal{D}} \Theta$ .  $\square$

## B.2. Normalized IFA EPPF converges to NCRM EPPF

*Proof of Theorem 3.4.* First, we show that the total mass of IFA converges in distribution to the total mass of CRM. Through Appendix B.1, we have shown that for all measurable sets  $A$  and  $t > 0$ , the Laplace functionals converge:

$$\lim_{K \rightarrow \infty} \mathbb{E}[e^{-t\Theta_K(A)}] = \mathbb{E}[e^{-t\Theta(A)}],$$

By choosing  $A = \Psi$  i.e. the ground space, we have that  $\Theta_K(\Psi)$  is the total mass of IFA and  $\Theta(\Psi)$  is the total mass of CRM

$$\Theta_K(\Psi) = \sum_{i=1}^K \theta_{K,i}, \quad \Theta(\Psi) = \sum_{i=1}^{\infty} \theta_i.$$

Since for any  $t > 0$ , the Laplace transform of  $\Theta_K(\Psi)$  converges to that of  $\Theta(\Psi)$ , we conclude that  $\Theta_K(\Psi)$  converges to  $\Theta(\Psi)$  in distribution (Kallenberg, 2002, Theorem 5.3):

$$\sum_{i=1}^K \theta_{K,i} \xrightarrow{\mathcal{D}} \Theta(\Psi). \quad (\text{B.3})$$

Second, we show that the decreasing order statistics of IFA atom sizes converges (in finite-dimensional distributions i.e., in f.d.d) to the decreasing order statistics of CRM atom sizes. For each  $K$ , the decreasing order statistics of IFA atoms is denoted by  $\{\theta_{K,(i)}\}_{i=1}^K$ :

$$\theta_{K,(1)} \geq \theta_{K,(2)} \geq \cdots \geq \theta_{K,(K)}.$$

We will leverage (Loeve, 1956, Theorem 4 and page 191) to find the limiting distribution  $\{\theta_{K,(i)}\}_{i=1}^K$  as  $K \rightarrow \infty$ . It is easy to verify the conditions to use the theorem: because the sums  $\sum_{i=1}^K \theta_{K,i}$  converge in distribution to a limit, we know that all the  $\theta_{K,i}$ 's are uniformly asymptotically negligible (Kallenberg, 2002, Lemma 15.13). Now, we discuss what the limits are. It is well-known that  $\Theta(\Psi)$  is an infinitely divisible positive random variable with no drift component and Levy measure exactly  $\nu(d\theta)$  Perman, Pitman and Yor (1992). In the terminology of (Loeve, 1956, Equation 2), the characteristics of  $\Theta(\Psi)$  are  $a = b = 0$  (no drift or Gaussian parts),  $L(x) := 0$  (because nonnegative random variable, see for instance Proposition 2 from <http://www.math.utah.edu/~davar/ps-pdf-files/Levy.pdf>) and:

$$M(x) := -\nu([x, \infty)).$$

Let  $I$  be a counting process *in reverse* over  $(0, \infty)$  defined based on the Poisson point process  $\{\theta_i\}_{i=1}^{\infty}$  in the following way. For any  $x$ ,  $I(x)$  is the number of points  $\theta_i$  exceeding the threshold  $x$ :

$$I(x) := |\{i : \theta_i \geq x\}|.$$

We augment  $I(0) = \infty$  and  $I(\infty) = 0$ . As a stochastic process,  $I$  has independent increments, in that for all  $0 = t_0 < t_1 < \cdots < t_k$ , the increments  $I(t_i) - I(t_{i-1})$  are independent, furthermore the law of the increments is  $I(t_{i-1}) - I(t_i) \sim \text{Poisson}(M(t_i) - M(t_{i-1}))$ . These properties are simple consequences of the counting measure induced by the Poisson point process. According to (Loeve, 1956, Page 191), the limiting distribution of  $\{\theta_{K,(i)}\}_{i=1}^K$  is governed by  $I$ , in the sense that for any fixed  $t \in \mathbb{N}$ , for any  $x_1, x_2, \dots, x_t \in [0, \infty)$ :

$$\begin{aligned} \lim_{K \rightarrow \infty} \mathbb{P}(\theta_{K,(1)} < x_1, \theta_{K,(2)} < x_2, \dots, \theta_{K,(t)} < x_t) \\ = \mathbb{P}(I(x_1) < 1, I(x_2) < 2, \dots, I(x_t) < t). \end{aligned} \quad (\text{B.4})$$

Because the  $\theta_i$ 's induce  $I$ , we can relate the left hand side to the order statistics of the Poisson point process. We denote the decreasing order statistic of the  $\{\theta_i\}_{i=1}^{\infty}$  as:

$$\theta_{(1)} \geq \theta_{(2)} \geq \cdots \geq \theta_{(n)} \geq \cdots$$

Clearly, for any  $t \in \mathbb{N}$ , the event that  $I(x)$  exceeds  $t$  is the same as the top  $t$  jumps among the  $\{\theta_i\}_{i=1}^\infty$  exceed  $x$ :  $I(x) \geq t \iff \theta_{(t)} \geq x$ . Therefore Eq. (B.4) can be rewritten as, for any fixed  $t \in \mathbb{N}$ , for any  $x_1, x_2, \dots, x_t \in [0, \infty)$ :

$$\lim_{K \rightarrow \infty} \mathbb{P}(\theta_{K,(1)} < x_1, \theta_{K,(2)} < x_2, \dots, \theta_{K,(t)} < x_t) = \mathbb{P}(\theta_{(1)} < x_1, \theta_{(2)} < x_2, \dots, \theta_{(t)} < x_t) \tag{B.5}$$

It is well-known that convergence of the distribution function imply weak convergence – for instance, see Problem 1 of [https://link.springer.com/content/pdf/10.1007/978-1-4612-5254-2\\_3.pdf](https://link.springer.com/content/pdf/10.1007/978-1-4612-5254-2_3.pdf). Actually, from (Loeve, 1956, Theorem 5 and page 194), for any fixed  $t \in \mathbb{N}$ , the convergence in distribution of  $\{\theta_{K,(i)}\}_{i=1}^t$  to  $\{\theta_i\}_{i=1}^t$  holds jointly with the convergence of  $\sum_{i=1}^K \theta_{K,(i)}$  to  $\sum_{i=1}^\infty \theta_i$ : the two conditions of the theorem, which are continuity of the distribution function of each  $\theta_{K,i}$  and  $M(0) = -\infty$  (there is a typo in Loeve (1956)), are easily verified. Therefore, by continuous mapping theorem, if we define the normalized atom sizes:

$$p_{K,(s)} := \frac{\theta_{K,(s)}}{\sum_{i=1}^K \theta_{K,i}} \quad p_{(s)} := \frac{\theta_{(s)}}{\sum_{i=1}^\infty \theta_i}$$

we also have that the normalized decreasing order statistics converge:

$$(p_{K,i})_{i=1}^K \xrightarrow{f.d.d.} (p_{(i)})_{i=1}^\infty$$

Finally we show that the EPPFs converge. In addition, if we define the *size-biased permutation* (in the sense of (Gnedin, 1998, Section 2) ) of the normalized atom sizes:

$$\{\tilde{p}_{K,i}\} \sim \text{SBP}(p_{K,(s)}) \quad \{\tilde{p}_i\} \sim \text{SBP}(p_{(s)})$$

then by (Gnedin, 1998, Theorem 1), the finite-dimensional distributions of the size-biased permutation also converges:

$$(\tilde{p}_{K,i})_{i=1}^K \xrightarrow{f.d.d.} (\tilde{p}_i)_{i=1}^\infty \tag{B.6}$$

From here, we fix the number of samples  $N$ , the number of components  $t$  and the size of the clusters  $n_i$ . (Pitman, 1996, Equation 45) gives the EPPF of  $\Xi = \Theta/\Theta(\Psi)$ :

$$p(n_1, n_2, \dots, n_t) = \mathbb{E} \left( \prod_{i=1}^t \tilde{p}_i^{n_i-1} \prod_{i=1}^{t-1} \left( 1 - \sum_{j=1}^i \tilde{p}_j \right) \right),$$

Likewise, the EPPF of  $\Xi_K = \Theta_K/\Theta_K(\Psi)$  is:

$$p_K(n_1, n_2, \dots, n_t) = \mathbb{E} \left( \prod_{i=1}^t \tilde{p}_{K,i}^{n_i-1} \prod_{i=1}^{t-1} \left( 1 - \sum_{j=1}^i \tilde{p}_{K,j} \right) \right)$$

Since  $t$  is fixed, and each  $p_j$  is  $[0, 1]$  valued, the mapping from the  $t$ -dimensional vector  $p$  to the product  $\prod_{i=1}^t p_i^{n_i-1} \prod_{i=1}^{t-1} \left( 1 - \sum_{j=1}^i p_j \right)$  is continuous and bounded. The choice of  $N$ ,  $t$ ,  $n_i$  have been fixed but arbitrary. Hence, the convergence in finite-dimensional distributions of in Eq. (B.6) imply that the EPPFs converge.  $\square$

### Appendix C: Marginal processes of exponential CRMs

The marginal process characterization describes the probabilistic model not through the two-stage sampling  $\Theta \sim \text{CRM}(H, \nu)$  and  $X_n | \Theta \stackrel{iid}{\sim} \text{LP}(l; \Theta)$ , but through the conditional distributions  $X_n | X_{n-1}, X_{n-2}, \dots, X_1$  i.e. the underlying  $\Theta$  has been *marginalized out*. This perspective removes the need to infer a countably infinite set of target variables. In addition, the *exchangeability* between  $X_1, X_2, \dots, X_N$  i.e. the joint distribution's invariance with respect to ordering of observations Aldous (1985), often enables the development of inference algorithms, namely Gibbs samplers.

(Broderick, Wilson and Jordan, 2018, Corollary 6.2) derives the conditional distributions  $X_n | X_{n-1}, X_{n-2}, \dots, X_1$  for general exponential family CRMs Eqs. (1) and (2).

**Proposition C.1** (Target's marginal process (Broderick, Wilson and Jordan, 2018, Corollary 6.2)). *For any  $n$ ,  $X_n | X_{n-1}, \dots, X_1$  is a random measure with finite support.*

1. Let  $\{\zeta_i\}_{i=1}^{K_n-1}$  be the union of atom locations in  $X_1, X_2, \dots, X_{n-1}$ . For  $1 \leq m \leq n-1$ , let  $x_{m,j}$  be the atom size of  $X_m$  at atom location  $\zeta_j$ . Denote  $x_{n,i}$  to be the atom size of  $X_n$  at atom location  $\zeta_i$ . The  $x_{n,i}$ 's are independent across  $i$  and the p.m.f. of  $x_{n,i}$  at  $x$  is:

$$\kappa(x) \frac{S\left(-1 + \sum_{m=1}^{n-1} \phi(x_{m,i}) + \phi(x), \eta + \binom{\sum_{m=1}^{n-1} t(x_{m,i}) + t(x)}{n}\right)}{S\left(-1 + \sum_{m=1}^{n-1} \phi(x_{m,i}), \eta + \binom{\sum_{m=1}^{n-1} t(x_{m,i})}{n-1}\right)}.$$

2. For each  $x \in \mathbb{N}$ ,  $X_n$  has  $p_{n,x}$  atoms whose atom size is exactly  $x$ . The locations of each atom are iid  $H$ : as  $H$  is diffuse, they are disjoint from the existing union of atoms  $\{\zeta_i\}_{i=1}^{K_n-1}$ .  $p_{n,x}$  is Poisson-distributed, independently across  $x$ , with mean:

$$\gamma' \kappa(0)^{n-1} \kappa(x) S\left(c/K - 1 + (n-1)\phi(0) + \phi(x), \eta + \binom{(n-1)t(0) + t(x)}{n}\right).$$

In Proposition C.2, we state a similar characterization of  $Z_n | Z_{n-1}, Z_{n-2}, \dots, Z_1$  for finite-dimensional model Eq. (6) and give the proof.

**Proposition C.2** (Approximation's marginal process). *For any  $n$ ,  $Z_n | Z_{n-1}, \dots, Z_1$  is a random measure with finite support.*

1. Let  $\{\zeta_i\}_{i=1}^{K_n-1}$  be the union of atom locations in  $Z_1, Z_2, \dots, Z_{n-1}$ . For  $1 \leq m \leq n-1$ , let  $z_{m,j}$  be the atom size of  $Z_m$  at atom location  $\zeta_j$ . Denote  $z_{n,i}$  to be the atom size of  $Z_n$  at atom location  $\zeta_i$ .  $z_{n,i}$ 's are independently across  $i$  and the p.m.f. of  $z_{n,i}$  at  $x$  is:

$$\kappa(x) \frac{S\left(c/K - 1 + \sum_{m=1}^{n-1} \phi(z_{m,i}) + \phi(x), \eta + \binom{\sum_{m=1}^{n-1} t(z_{m,i}) + t(x)}{n}\right)}{S\left(c/K - 1 + \sum_{m=1}^{n-1} \phi(z_{m,i}), \eta + \binom{\sum_{m=1}^{n-1} t(z_{m,i})}{n-1}\right)}.$$

2.  $K - K_{n-1}$  atom locations are generated iid from  $H$ .  $Z_n$  has  $p_{n,x}$  atoms whose size is exactly  $x$  (for  $x \in \mathbb{N} \cup \{0\}$ ) over these  $K - K_{n-1}$  atom locations (the  $p_{n,0}$  atoms whose atom size is 0 can be interpreted as not present in  $Z_n$ ). The joint distribution of  $p_{n,x}$  is

a Multinomial with  $K - K_{n-1}$  trials, with success of type  $x$  having probability:

$$\kappa(x) \frac{S\left(c/K - 1 + (n-1)\phi(0) + \phi(x), \eta + \binom{(n-1)t(0) + t(x)}{n}\right)}{S\left(c/K - 1 + (n-1)\phi(0), \eta + \binom{(n-1)t(0)}{n-1}\right)}.$$

*Proof of Proposition C.2.* We only need to prove the conditional distributions for the atom sizes: that the  $K$  distinct atom locations are generated iid from the base measure is clear.

First we consider  $n = 1$ . By construction Corollary 3.3, a priori, the trait frequencies  $\{\theta_i\}_{i=1}^K$  are independent, each following the distribution:

$$\mathbb{P}(\theta_i \in d\theta) = \frac{\mathbf{1}\{\theta \in U\}}{S(c/K - 1, \eta)} \theta^{c/K-1} \exp\left(\langle \eta, \binom{\mu(\theta)}{-A(\theta)} \rangle\right).$$

Conditioned on  $\{\theta_i\}_{i=1}^K$ , the atom sizes  $z_{1,i}$  that  $Z_1$  puts on the  $i$ th atom location are independent across  $i$  and each is distributed as:

$$\mathbb{P}(z_{1,i} = x | \theta_i) = \kappa(x) \theta^{\phi(x)} \exp(\langle \mu(\theta_i), t(x) \rangle - A(\theta_i)).$$

Integrating out  $\theta_i$ , the marginal distribution for  $z_{1,i}$  is:

$$\begin{aligned} \mathbb{P}(z_{1,i} = x) &= \int \mathbb{P}(z_{1,i} = x | \theta_i = \theta) \mathbb{P}(\theta_i \in d\theta) \\ &= \frac{\kappa(x)}{S(c/K - 1, \eta)} \int_U \theta^{c/K-1+\phi(x)} \exp\left(\langle \eta + \binom{t(x)}{1}, \binom{\mu(\theta)}{-A(\theta)} \rangle\right) d\theta \\ &= \kappa(x) \frac{S\left(c/K - 1 + \phi(x), \eta + \binom{t(x)}{1}\right)}{S(c/K - 1, \eta)}, \end{aligned}$$

by definition of  $S$  as the normalizer Eq. (3).

Now we consider  $n \geq 2$ . The distribution of  $z_{n,i}$  only depends on the distribution of  $z_{n-1,i}, z_{n-2,i}, \dots, z_{1,i}$  since the atom sizes across different atoms are independent of each other both a priori and a posteriori. The predictive distribution is an integral:

$$\mathbb{P}(z_{n,i} = x | z_{1:(n-1),i}) = \int \mathbb{P}(z_{n,i} = x | \theta_i) \mathbb{P}(\theta_i \in d\theta | z_{1:(n-1),i}).$$

Because the prior over  $\theta_i$  is conjugate for the likelihood  $z_{i,j} | \theta_i$ , and the observations  $z_{i,j}$  are conditionally independent given  $\theta_i$ , the posterior  $\mathbb{P}(\theta_i \in d\theta | z_{1:(n-1),i})$  is in the same exponential family but with different natural parameters:

$$\mathbf{1}\{\theta \in U\} \frac{\theta^{c/K-1+\sum_{m=1}^{n-1} \phi(z_{m,i})} \exp\left(\langle \eta + \binom{\sum_{m=1}^{n-1} t(z_{m,i})}{n-1}, \binom{\mu(\theta)}{-A(\theta)} \rangle\right) d\theta}{S\left(c/K - 1 + \sum_{m=1}^{n-1} \phi(z_{m,i}), \eta + \binom{\sum_{m=1}^{n-1} t(z_{m,i})}{n-1}\right)}.$$

This means that the predictive distribution  $\mathbb{P}(z_{n,i} = x | z_{1:(n-1),i})$  equals:

$$\begin{aligned} & \frac{\int_U \theta^{c/K-1+\sum_{m=1}^{n-1} \phi(z_{m,i})+\phi(x)} \exp\left(\left\langle \eta + \left(\frac{\sum_{m=1}^{n-1} t(z_{m,i}) + t(x)}{n}\right), \begin{pmatrix} \mu(\theta) \\ -A(\theta) \end{pmatrix} \right\rangle\right) d\theta}{S\left(c/K-1+\sum_{m=1}^{n-1} \phi(z_{m,i}), \eta + \left(\frac{\sum_{m=1}^{n-1} t(z_{m,i})}{n-1}\right)\right)} \\ &= \kappa(x) \frac{S\left(c/K-1+\sum_{m=1}^{n-1} \phi(z_{m,i}) + \phi(x), \eta + \left(\frac{\sum_{m=1}^{n-1} t(z_{m,i}) + t(x)}{n}\right)\right)}{S\left(c/K-1+\sum_{m=1}^{n-1} \phi(z_{m,i}), \eta + \left(\frac{\sum_{m=1}^{n-1} t(z_{m,i})}{n-1}\right)\right)}. \end{aligned}$$

The predictive distribution  $\mathbb{P}(z_{n,i} = x | z_{1:(n-1),i})$  govern both the distribution of atom sizes for known atom locations and new atom locations.  $\square$

## Appendix D: Technical lemmas

### D.1. Concentration

**Lemma D.1** (Modified upper tail Chernoff bound). *Let  $X = \sum_{i=1}^n X_i$ , where  $X_i = 1$  with probability  $p_i$  and  $X_i = 0$  with probability  $1 - p_i$ , and all  $X_i$  are independent. Let  $\mu$  be an upper bound on  $E(X) = \sum_{i=1}^n p_i$ . Then for all  $\delta > 0$ :*

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\delta^2}{2 + \delta}\mu\right).$$

*Proof of Lemma D.1.* The proof relies on the regular upper tail Chernoff bound <http://math.mit.edu/~goemans/18310S15/chernoff-notes.pdf> and an argument using stochastic domination. Truly, we pad the first  $n$  Poisson trials that define  $X$  with additional trials  $X_{n+1}, X_{n+2}, \dots, X_{n+m}$  where  $m$  is the smallest natural number such that  $\frac{\mu - E[X]}{m} \leq 1$ , each  $X_{n+i}$  is a Bernoulli with probability  $\frac{\mu - E[X]}{m}$ , and the trials are independent. Then  $Y = X + \sum_{j=1}^m X_{n+j}$  is itself the sum of Poisson trials with mean exactly  $\mu$ , so the regular Chernoff bound applies:

$$\mathbb{P}(Y \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\delta^2}{2 + \delta}\mu\right).$$

However by construction,  $X$  is stochastically dominated by  $Y$ , so the tail probabilities of  $X$  is bounded by the tail probabilities of  $Y$ .  $\square$

**Lemma D.2** (Lower tail Chernoff bound). *Let  $X = \sum_{i=1}^n X_i$ , where  $X_i = 1$  with probability  $p_i$  and  $X_i = 0$  with probability  $1 - p_i$ , and all  $X_i$  are independent. Let  $\mu := E(X) = \sum_{i=1}^n p_i$ . Then for all  $\delta \in (0, 1)$ :*

$$\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp(-\mu\delta^2/2).$$

**Lemma D.3** (Tail bounds for Poisson distribution). *If  $X \sim \text{Poisson}(\lambda)$  then for any  $x > 0$ :*

$$\mathbb{P}(X \geq \lambda + x) \leq \exp\left(-\frac{x^2}{2(\lambda + x)}\right),$$

and for any  $0 < x < \lambda$ :

$$\mathbb{P}(X \leq \lambda - x) \leq \exp\left(-\frac{x^2}{2\lambda}\right).$$

*Proof of Lemma D.3.* For  $x \geq -1$ , let  $\psi(x) := 2((1+x)\ln(1+x) - x)/x^2$ .

We first inspect the upper tail bound. If  $X \sim \text{Poisson}(\lambda)$ , for any  $x > 0$ , (Pollard, 2001, Exercise 3 p.272) implies that:

$$\mathbb{P}(Z \geq \lambda + x) \leq \exp\left(-\frac{x^2}{2\lambda}\psi\left(\frac{x}{\lambda}\right)\right).$$

To show the upper tail bound, it suffices to prove that  $\frac{x^2}{2\lambda}\psi\left(\frac{x}{\lambda}\right)$  is greater than  $\frac{x^2}{2(\lambda+x)}$ . In general, we show that for  $u \geq 0$ :

$$(u+1)\psi(u) - 1 \geq 0. \quad (\text{D.1})$$

The denominator of  $(u+1)\psi(u) - 1$  is clearly positive. Consider the numerator of  $(u+1)\psi(u) - 1$ , which is  $g(u) := 2((u+1)^2 \ln(u+1) - u(u+1) - u^2)$ . Its 1st and 2nd derivatives are:

$$\begin{aligned} g'(u) &= 4(u+1)\ln(u+1) - 2u + 1 \\ g''(u) &= 4\ln(u+1) + 2. \end{aligned}$$

Since  $g''(u) \geq 0$ ,  $g'(u)$  is monotone increasing. Since  $g'(0) = 1$ ,  $g'(u) > 0$  for  $u \geq 0$ , hence  $g(u)$  is monotone increasing. Because  $g(0) = 0$ , we conclude that  $g(u) \geq 0$  for  $u > 0$  and Eq. (D.1) holds. Plugging in  $u = x/\lambda$ :

$$\psi\left(\frac{x}{\lambda}\right) \geq \frac{1}{1 + \frac{x}{\lambda}} = \frac{\lambda}{x + \lambda},$$

which shows  $\frac{x^2}{2\lambda}\psi\left(\frac{x}{\lambda}\right) \geq \frac{x^2}{2(\lambda+x)}$ .

Now we inspect the lower tail bound. We follow the proof of <http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf>. We first argue that:

$$\mathbb{P}(X \leq \lambda - x) \leq \exp\left(-\frac{x^2}{2\lambda}\psi\left(-\frac{x}{\lambda}\right)\right). \quad (\text{D.2})$$

For any  $\theta$ , the moment generating function  $\mathbb{E}[\exp(\theta X)]$  is well-defined and well-known:

$$\mathbb{E}[\exp(\theta X)] := \exp(\lambda(\exp(\theta) - 1)).$$

Therefore:

$$\begin{aligned} \mathbb{P}(X \leq \lambda - x) &\leq \mathbb{P}(\exp(\theta X) \leq \exp(\theta(\lambda - x))) \leq \mathbb{P}(\exp(\theta(\lambda - x - X)) \geq 1) \\ &\leq \exp(\theta(\lambda - x))\mathbb{E}[\exp(-\theta X)], \end{aligned}$$

where we have used Markov's inequality. We now aim to minimize  $\exp(\theta(\lambda - x))\mathbb{E}[\exp(-\theta X)]$  as a function of  $\theta$ . Its logarithm is:

$$\lambda(\exp(-\theta) - 1) + \theta(\lambda - x).$$

This is a convex function, whose derivative vanishes at  $\theta = -\ln\left(1 - \frac{x}{\lambda}\right)$ . Overall this means the best upper bound on  $\mathbb{P}(X \leq \lambda - x)$  is:

$$\exp\left(-\lambda\left(\frac{x}{\lambda} + \left(1 - \frac{x}{\lambda}\right)\ln\left(1 - \frac{x}{\lambda}\right)\right)\right),$$

which is exactly the right hand side of Eq. (D.2). Hence to demonstrate the lower tail bound, it suffices to show that:

$$\psi\left(-\frac{x}{\lambda}\right) \geq 1.$$

More generally, we show that for  $-1 \leq u \leq 0$ ,  $\psi(u) - 1 \geq 0$ . Consider the numerator of  $\psi(u) - 1$ , which is  $h(u) := 2((1+u)\ln(1+u) - u) - u^2$ . The first two derivatives are:

$$\begin{aligned} h'(u) &= 2(1 + \ln(1+u)) - 2u \\ h''(u) &= \frac{2}{1+u} - 2 \end{aligned}$$

Since  $h''(u) \geq 0$ ,  $h(u)$  is convex on  $[-1, 0]$ . Note that  $h(0) = 0$ . Also, by simple continuity argument,  $h(-1) = 2$ . Therefore,  $h$  is non-negative on  $[0, 1]$ , meaning that  $\psi(u) \geq 1$ .  $\square$

**Lemma D.4** (Multinomial-Poisson approximation). *Let  $\{p_i\}_{i=1}^{\infty}$ ,  $p_i \geq 0$ ,  $\sum_{i=1}^{\infty} p_i < 1$ . Suppose there are  $n$  independent trials: in each trial, success of type  $i$  has probability  $p_i$ . Let  $X = \{X_i\}_{i=1}^{\infty}$  be the number of type  $i$  successes after  $n$  trial. Let  $Y = \{Y_i\}_{i=1}^{\infty}$  be independent Poisson random variables, where  $Y_i$  has mean  $np_i$ . Then:*

$$d_{TV}(X, Y) \leq n \left( \sum_{i=1}^{\infty} p_i \right)^2.$$

*Proof of Lemma D.4.* First we remark that both  $X$  and  $Y$  can be sampled in two-steps.

- Regarding  $X$ , first sample  $N_1 \sim \text{Binom}(n, \sum_{i=1}^{\infty} p_i)$ . Then, for each  $1 \leq k \leq N_1$ , sample  $Z_k$  where  $\mathbb{P}(Z_k = i) = \frac{p_i}{\sum_{j=1}^{\infty} p_j}$ . Then,  $X_i = \sum_{k=1}^{N_1} \mathbf{1}\{Z_k = i\}$  for each  $i$ .
- Regarding  $Y$ , first sample  $N_2 \sim \text{Poisson}(n \sum_{i=1}^{\infty} p_i)$ . Then, for each  $1 \leq k \leq N_2$ , sample  $T_k$  where  $\mathbb{P}(T_k = i) = \frac{p_i}{\sum_{j=1}^{\infty} p_j}$ . Then,  $Y_i = \sum_{k=1}^{N_2} \mathbf{1}\{T_k = i\}$  for each  $i$ .

The two-step sampling perspective for  $X$  comes from rejection sampling: to generate a success of type  $k$ , we first generate some type of success, and then re-calibrate to get the right proportion for type  $k$ . The two-step perspective for  $Y$  comes from the thinning property of Poisson distribution (Last and Penrose, 2017, Exercise 1.5). The thinning property implies that for any finite index set  $\mathcal{K}$ , all  $\{Y_i\}$  for  $i \in \mathcal{K}$  are mutually independent and marginally,  $Y_i \sim \text{Poisson}(np_i)$ . Hence the whole collection  $\{Y_i\}_{i=1}^{\infty}$  are independent Poissons and the mean of  $Y_i$  is  $np_i$ .

Observing that the conditional  $X|N_1 = n$  is the same as  $Y|N_2 = n$ , we use propagation rule Lemma D.7:

$$d_{TV}(X, Y) \leq d_{TV}(N_1, N_2).$$

Total variation between  $N_1$  and  $N_2$  is just the classic Binomial-Poisson approximation Le Cam (1960).

$$d_{TV}(N_1, N_2) \leq n \left( \sum_{i=1}^{\infty} p_i \right)^2.$$

$\square$

**Lemma D.5** (Total variation between Poissons (Adell and Lekuona, 2005, Corollary 3.1)). *Let  $P_1$  be the Poisson distribution with mean  $s$ ,  $P_2$  the Poisson distribution with mean  $t$ . Then:*

$$d_{TV}(P_1, P_2) \leq 1 - \exp(-|s - t|) \leq |s - t|.$$

## D.2. Total variation

First is the chain rule, which will be applied to compare joint distributions that admit densities.

**Lemma D.6** (Chain rule). *Suppose  $(X_1, Y_1)$  and  $(X_2, Y_2)$  are two distributions, over  $\mathcal{A} \times \mathcal{B}$ , that have densities w.r.t a common measure. Then:*

$$d_{TV}(P_{X_1, Y_1}, P_{X_2, Y_2}) \leq d_{TV}(P_{X_1}, P_{X_2}) + \sup_{a \in \mathcal{A}} d_{TV}(P_{Y_1|X_1=a}, P_{Y_2|X_2=a}).$$

*Proof of Lemma D.6.* Because both  $P_{X_1, Y_1}$  and  $P_{X_2, Y_2}$  have densities, total variation distance is half of  $L_1$  distance between the densities:

$$\begin{aligned} d_{TV}(P_{X_1, Y_1}, P_{X_2, Y_2}) &= \frac{1}{2} \int_{(a,b) \in \mathcal{A} \times \mathcal{B}} |P_{X_1, Y_1}(a, b) - P_{X_2, Y_2}(a, b)| da db \\ &= \frac{1}{2} \int_{(a,b) \in \mathcal{A} \times \mathcal{B}} |P_{X_1, Y_1}(a, b) - P_{X_2}(a)P_{Y_1|X_1}(b|a) + P_{X_2}(a)P_{Y_1|X_1}(b|a) - P_{X_2, Y_2}(a, b)| da db \\ &\leq \frac{1}{2} \int_{(a,b) \in \mathcal{A} \times \mathcal{B}} (P_{Y_1|X_1}(b|a)|P_{X_1}(a) - P_{X_2}(a)| + P_{X_2}(a)|P_{Y_1|X_1}(b|a) - P_{Y_2|X_2}(b|a)|) da db \\ &= \frac{1}{2} \int_{(a,b) \in \mathcal{A} \times \mathcal{B}} P_{Y_1|X_1}(b|a)|P_{X_1}(a) - P_{X_2}(a)| da db + \frac{1}{2} \int_{(a,b) \in \mathcal{A} \times \mathcal{B}} P_{X_2}(a)|P_{Y_1|X_1}(b|a) - P_{Y_2|X_2}(b|a)| da db. \end{aligned}$$

where we have used triangle inequality. Regarding the first term, using Fubini:

$$\begin{aligned} &\frac{1}{2} \int_{(a,b) \in \mathcal{A} \times \mathcal{B}} P_{Y_1|X_1}(b|a)|P_{X_1}(a) - P_{X_2}(a)| da db \\ &= \frac{1}{2} \int_{a \in \mathcal{A}} \left( \int_{b \in \mathcal{B}} P_{Y_1|X_1}(b|a) db \right) |P_{X_1}(a) - P_{X_2}(a)| da \\ &= \frac{1}{2} \int_{a \in \mathcal{A}} |P_{X_1}(a) - P_{X_2}(a)| da \\ &= d_{TV}(P_{X_1}, P_{X_2}). \end{aligned}$$

Regarding the second term:

$$\begin{aligned} &\frac{1}{2} \int_{(a,b) \in \mathcal{A} \times \mathcal{B}} P_{X_2}(a)|P_{Y_1|X_1}(b|a) - P_{Y_2|X_2}(b|a)| da db \\ &= \int_{a \in \mathcal{A}} \left( \frac{1}{2} \int_{b \in \mathcal{B}} |P_{Y_1|X_1}(b|a) - P_{Y_2|X_2}(b|a)| db \right) P_{X_2}(a) da \\ &\leq \left( \sup_{a \in \mathcal{A}} d_{TV}(P_{Y_1|X_1=a}, P_{Y_2|X_2=a}) \right) \int_{a \in \mathcal{A}} P_{X_2}(a) da \\ &= \sup_{a \in \mathcal{A}} d_{TV}(P_{Y_1|X_1=a}, P_{Y_2|X_2=a}) \end{aligned}$$

Sum of the first and second upper bound give the total variation chain rule.  $\square$

Second is the propagation rule, which applies even if distributions don't have densities.

**Lemma D.7** (Propagation rule). *Suppose  $(X_1, Y_1)$  and  $(X_2, Y_2)$  are two distributions over  $\mathcal{A} \times \mathcal{B}$ . Suppose the conditional  $Y_2|X_2 = a$  is the same as the conditional  $Y_1|X_1 = a$ , which we just denote as  $Y|X = a$ . Then:*

$$d_{TV}(P_{Y_1}, P_{Y_2}) \leq d_{TV}(P_{X_1}, P_{X_2}).$$

*Proof of Lemma D.7.* It is well-known that total variation between  $P_U$  and  $P_V$  is the infimum of  $\mathbb{P}(U \neq V)$  over all couplings  $(U, V)$  where  $U \sim P_U$  and  $V \sim P_V$  ((Madras and Sezer, 2010, Equation 13)). For any joint distribution of  $(X_1, Y_1, X_2, Y_2)$  where marginally  $(X_1, Y_1) \sim P_{X_1, Y_1}$  and  $(X_2, Y_2) \sim P_{X_2, Y_2}$ ,  $(Y_1, Y_2)$  is a coupling where  $Y_1 \sim P_{Y_1}$  and  $Y_2 \sim P_{Y_2}$ . Therefore:

$$d_{TV}(P_{Y_1}, P_{Y_2}) \leq \mathbb{P}(Y_1 \neq Y_2) = \mathbb{P}(Y_1 \neq Y_2, X_1 \neq X_2) + \mathbb{P}(Y_1 \neq Y_2, X_1 = X_2).$$

Now suppose the joint distribution over  $(X_1, Y_1, X_2, Y_2)$  is such that, conditioned on  $X_1 = X_2 = a$  for any  $a$ ,  $\mathbb{P}(Y_1 = Y_2 | X_1 = X_2 = a) = 1$  (when  $X_1 \neq X_2$ , it doesn't matter the relationship between  $Y_1 | X_1 = a$  and  $Y_2 | X_2 = b$ ). This is possible since the conditional  $Y_2 | X_2 = a$  is the same as the conditional  $Y_1 | X_1 = a$ . For such a distribution,  $\mathbb{P}(Y_1 \neq Y_2, X_1 = X_2) = 0$ . Hence:

$$d_{TV}(P_{Y_1}, P_{Y_2}) \leq \mathbb{P}(Y_1 \neq Y_2, X_1 \neq X_2) \leq \mathbb{P}(X_1 \neq X_2).$$

Now, we recognize that  $(X_1, X_2)$  is an arbitrary coupling between  $P_{X_1}$  and  $P_{X_2}$ . Taking infimum over all couplings, we arrive at the propagation rule.  $\square$

Third is the product rule.

**Lemma D.8** (Product rule).  $Z_1 = (X_1, Y_1)$  and  $Z_2 = (X_2, Y_2)$  are two distributions over  $\mathcal{A} \times \mathcal{B}$ . Suppose  $P_{X_1, Y_1}$  factorizes into  $P_{X_1} P_{Y_1}$  and similarly  $P_{X_2, Y_2} = P_{X_2} P_{Y_2}$ . Then:

$$\inf_{\text{coupling } P_{Z_1}, P_{Z_2}} \mathbb{P}(Z_1 \neq Z_2) \leq \inf_{\text{coupling } P_{X_1}, P_{X_2}} \mathbb{P}(X_1 \neq X_2) + \inf_{\text{coupling } P_{Y_1}, P_{Y_2}} \mathbb{P}(Y_1 \neq Y_2)$$

*Proof of Lemma D.8.* Consider any  $(X_1, X_2)$  that is a coupling of  $P_{X_1}$  and  $P_{X_2}$ , and any  $(Y_1, Y_2)$  that is a coupling of  $P_{Y_1}$  and  $P_{Y_2}$ . Because of the factorization structure between the  $X$ 's and the  $Y$ 's, we can construct  $(X'_1, X'_2, Y'_1, Y'_2)$  such that  $(X'_1, X'_2) \stackrel{D}{=} (X_1, X_2)$ ,  $(Y'_1, Y'_2) \stackrel{D}{=} (Y_1, Y_2)$ ,  $(X'_1, Y'_1) \sim P_{X_1, Y_1}$ ,  $(X'_2, Y'_2) \sim P_{X_2, Y_2}$ . By union bound:

$$\mathbb{P}((X'_1, Y'_1) \neq (X'_2, Y'_2)) \leq \mathbb{P}(X'_1 \neq X'_2) + \mathbb{P}(Y'_1 \neq Y'_2)$$

Because  $\inf_{\text{coupling } P_{Z_1}, P_{Z_2}} \mathbb{P}(Z_1 \neq Z_2) \leq \mathbb{P}((X'_1, Y'_1) \neq (X'_2, Y'_2))$ , we have:

$$\inf_{\text{coupling } P_{Z_1}, P_{Z_2}} \mathbb{P}(Z_1 \neq Z_2) \leq \mathbb{P}(X'_1 \neq X'_2) + \mathbb{P}(Y'_1 \neq Y'_2).$$

We finish the proof by taking the infimum over couplings  $(X_1, X_2)$  and  $(Y_1, Y_2)$  of the RHS.  $\square$

### D.3. Miscellaneous

**Lemma D.9** (Order of growth of harmonic-like sums).

$$\sum_{n=1}^N \frac{\alpha}{n-1+\alpha} \geq \alpha(\ln N - \psi(\alpha) - 1).$$

where  $\psi$  is the digamma function.

*Proof of Lemma D.9.* It is well-known (for instance [https://en.wikipedia.org/wiki/Chinese\\_restaurant\\_process](https://en.wikipedia.org/wiki/Chinese_restaurant_process)) that:

$$\sum_{n=1}^N \frac{\alpha}{n-1+\alpha} = \alpha[\psi(\alpha+N) - \psi(\alpha)]$$

(Gordon, 1994, Theorem 5) says that

$$\psi(\alpha+N) \geq \ln(\alpha+N) - \frac{1}{2(\alpha+N)} - \frac{1}{12(\alpha+N)^2} \geq \ln N - 1.$$

□

We list a collection of technical lemmas that are used when verifying Assumption 2 for the recurring examples.

The first set assists in the beta-Bernoulli model.

- For  $\alpha > 0$  and  $i = 1, 2, 3, \dots$ :

$$\frac{1}{i+\alpha-1} \leq 2 \left( \frac{1}{2\alpha} \mathbf{1}\{i=1\} + \frac{1}{i} \mathbf{1}\{i>1\} \right). \quad (\text{D.3})$$

- For  $m, x, y > 0$ ,  $m \leq y$ :

$$\left| \frac{m+x}{y+x} - \frac{m}{y} \right| \leq \frac{x}{y}. \quad (\text{D.4})$$

*Proof of Eq. (D.3).* If  $i = 1$ ,  $\frac{1}{i+\alpha-1} = \frac{1}{\alpha}$ . If  $i \geq 1$ ,  $\frac{1}{i+\alpha-1} \leq \frac{1}{i-1} \leq \frac{2}{i}$ . □

*Proof of Eq. (D.4).*

$$\left| \frac{m+x}{y+x} - \frac{m}{y} \right| = \left| \frac{(m+x)y - m(y+x)}{y(y+x)} \right| = \left| \frac{x(y-m)}{y(y+x)} \right| \leq \frac{x}{y}.$$

□

The second set aid in the gamma-Poisson model.

- For  $x \in [0, 1)$ ;

$$(1-x) \ln(1-x) + x \geq 0. \quad (\text{D.5})$$

- For  $x \in (0, 1)$ , for  $p \geq 0$ :

$$(1-x)^p + p \frac{x}{1-x} \geq 1. \quad (\text{D.6})$$

- For  $\lambda > 0$ , for  $m > 0, t > 1, x > 0$ :

$$d_{TV}(\text{NB}(m, t^{-1}), \text{NB}(m+x, t^{-1})) \leq x \frac{1/t}{1-1/t}. \quad (\text{D.7})$$

- For  $y \geq 1, m > 0, K > 0$ :

$$\left| \frac{m}{y} - K \frac{\Gamma(m/K + y)}{\Gamma(m/K)y!} \right| \leq e \frac{m^2}{K}. \quad (\text{D.8})$$

where  $e$  is the Euler constant.

*Proof of Eq. (D.5).* Set  $g(x)$  to be the function on the right hand side. Then its derivative is  $g'(x) = -\ln(1-x) \geq 0$ , meaning the function is monotone increasing. Since  $g(0) = 0$ , it's true that  $g(x) \geq 0$  over  $[0, 1)$ .  $\square$

*Proof of Eq. (D.6).* Let  $f(p) = (1-x)^p + p\frac{x}{1-x} - 1$ . Then  $f'(p) = \ln(1-x)(1-x)^p + \frac{x}{1-x}$ . Also  $f''(p) = (\ln(1-x))^2(1-x)^p > 0$ . So  $f'(p)$  is monotone increasing. At  $p = 0$ ,  $f'(0) = \ln(1-x) + \frac{x}{1-x} > 0$ . Therefore  $f'(p) \geq 0$  for all  $p$ . So  $f(p)$  is increasing. Since  $f(0) = 0$ , it's true that  $f(p) \geq 0$  for all  $p$ .  $\square$

*Proof of Eq. (D.7).* It is known that  $\text{NB}(r, \theta)$  is a Poisson stopped sum distribution ([Johnson, Kemp and Kotz, 2005](#), Equation 5.15):

- $N \sim \text{Poisson}(-r \ln(1-\theta))$ .
- $Y_i \stackrel{iid}{\sim} \text{Log}(\theta)$  where the  $\text{Log}(\theta)$  distribution's pmf at  $k$  equals  $\frac{-\theta^k}{k \ln(1-\theta)}$ .
- $\sum_{i=1}^N Y_i \sim \text{NB}(r, \theta)$ .

Therefore, by total variation's chain rule Lemma D.6, to compare  $\text{NB}(m, t^{-1})$  with  $\text{NB}(m + \gamma/K, t^{-1})$  it suffices to compare the two generating Poissons.

$$\begin{aligned} d_{TV}(\text{NB}(m, t^{-1}), \text{NB}(m + \gamma/K, t^{-1})) \\ \leq d_{TV}(\text{Poisson}(-m \ln(1-t^{-1})), \text{Poisson}(-(m + \gamma\lambda/K) \ln(1-t^{-1}))) \\ \leq -\ln(1-t^{-1}) \frac{\gamma\lambda}{K} \leq \frac{t^{-1}}{1-t^{-1}} \frac{\gamma\lambda}{K}. \end{aligned}$$

We have used the fact that total variation distance between Poissons is dominated by their different in means Lemma D.5 and Eq. (D.5) where  $x = (\lambda + i)^{-1}$ .  $\square$

*Proof of Eq. (D.8).* Since  $\Gamma\left(\frac{m}{K} + y\right) = \left(\prod_{j=0}^{y-1} \left(\frac{m}{K} + j\right)\right) \Gamma\left(\frac{m}{K}\right) = \Gamma\left(\frac{m}{K}\right) \frac{m}{K} \prod_{j=1}^{y-1} \left(\frac{m}{K} + j\right)$ , we have:

$$\left| \frac{m}{y} - K \frac{\Gamma(m/K + y)}{\Gamma(m/K)y!} \right| = \frac{m}{y} \left( \prod_{j=1}^{y-1} \frac{m/K + j}{j} - 1 \right).$$

We inspect the product in more detail.

$$\begin{aligned} \prod_{j=1}^{y-1} \frac{m/K + j}{j} &= \prod_{j=1}^{y-1} \left( 1 + \frac{m/K}{j} \right) \leq \prod_{j=1}^{y-1} \exp\left(\frac{m/K}{j}\right) \\ &= \exp\left(\frac{m}{K} \sum_{j=1}^{y-1} \frac{1}{j}\right) \leq \exp\left(\frac{m}{K} (\ln y + 1)\right) = (ey)^{m/K}. \end{aligned}$$

where the  $(y-1)$ th Harmonic sum is bounded by  $\ln y + 1$ . In all:

$$\left| \frac{m}{y} - K \frac{\Gamma(m/K + y)}{\Gamma(m/K)y!} \right| \leq \frac{m}{y} \left( (ey)^{m/K} - 1 \right) \leq \frac{m}{y} \frac{m}{K} (ey - 1) \leq e \frac{m^2}{K}.$$

$\square$

The third set aid in the beta-negative binomial model.

- For  $x > 0, z \geq y \geq 1$ :

$$B(x, y) - B(x, z) \leq (z - y)B(x + 1, y - 0.5) \leq (z - y)B(x + 1, y - 1). \quad (\text{D.9})$$

- For any  $\theta \in [0, 1], r > 0, b \geq 1$ :

$$\sum_{y=1}^{\infty} \frac{\Gamma(y+r)}{y!\Gamma(r)} B(y, b+r) \leq \frac{r}{b-0.5}. \quad (\text{D.10})$$

- For  $b \geq 1$ , for any  $c > 0$ , for any  $K \geq c$ :

$$\left| 1 - \frac{\Gamma(b)}{\Gamma(b+c/K)} \right| \leq \frac{c}{K} (2 + \ln b). \quad (\text{D.11})$$

- There exists a constant  $D''$  such that for all  $b > 1, c > 0, K \geq 2c(\ln b + 2)$ :

$$\left| c - \frac{K}{B(c/K, b)} \right| \leq \frac{c}{K} (3 \ln b + 8). \quad (\text{D.12})$$

*Proof of Eq. (D.9).* First we prove that for any  $x \in [0, 1)$ :

$$\sqrt{1-x} \ln(1-x) + x \geq 0.$$

Truly, let  $g(x)$  to be the function on the right hand side. Then its derivative is:

$$g'(x) = \frac{2\sqrt{1-x} - \ln(1-x) - 2}{2\sqrt{1-x}}.$$

Denote the numerator function by  $h(x)$ . Its derivative is:

$$h'(x) = \frac{1}{1-x} - \frac{1}{\sqrt{1-x}} \geq 0,$$

since  $x \in [0, 1]$  meaning  $h$  is monotone increasing. Since  $h(0) = 0$ , it means  $h(x) \geq 0$ . This means  $g'(x) \geq 0$  i.e.  $g$  itself is monotone increasing. Since  $g(0) = 0$  it's true that  $g(x) \geq 0$  for all  $x \in [0, 1)$ .

Second we prove that for all  $x \in [0, 1]$ , for all  $p \geq 0$ :

$$(1-x)^p + p \frac{x}{\sqrt{1-x}} - 1 \geq 0. \quad (\text{D.13})$$

Truly, let  $f(p) = (1-x)^p + p \frac{x}{\sqrt{1-x}} - 1$ . Then  $f'(p) = \ln(1-x)(1-x)^p + \frac{x}{\sqrt{1-x}}$ . Also  $f''(p) = (\ln(1-x))^2(1-x)^p > 0$ . So  $f'(p)$  is monotone increasing. At  $p = 0$ ,  $f'(0) = \ln(1-x) + \frac{x}{\sqrt{1-x}} > 0$ . Therefore  $f'(p) \geq 0$  for all  $p$ . So  $f(p)$  is increasing. Since  $f(0) = 0$ , it's true that  $f(p) \geq 0$  for all  $p$ .

We finally prove the inequality about beta functions.

$$\begin{aligned} B(x, y) - B(x, z) &= \int_0^1 \theta^{x-1} (1-\theta)^{y-1} (1 - (1-\theta)^{z-y}) d\theta \\ &\leq \int_0^1 \theta^{x-1} (1-\theta)^{y-1} (z-y)\theta(1-\theta)^{-0.5} d\theta \\ &= (z-y) \int_0^1 \theta^x (1-\theta)^{y-1.5} d\theta = (z-y)B(x+1, y-0.5). \end{aligned}$$

where we have used  $1 - (1-\theta)^{z-y} \leq (z-y)\theta(1-\theta)^{-1/2}$ . As for  $B(x+1, y-0.5) \leq B(x+1, y-1)$ , it is because of the monotonicity of the beta function.  $\square$

*Proof of Eq. (D.10).*

$$\begin{aligned}
\sum_{y=1}^{\infty} \frac{\Gamma(y+r)}{y! \Gamma(r)} B(y, b+r) &= \int_0^1 \sum_{y=1}^{\infty} \frac{\Gamma(y+r)}{y! \Gamma(r)} \theta^{y-1} (1-\theta)^{b+r-1} d\theta \\
&= \int_0^1 \theta^{-1} \left( \sum_{y=1}^{\infty} \frac{\Gamma(y+r)}{y! \Gamma(r)} \theta^y \right) (1-\theta)^{b+r-1} d\theta \\
&= \int_0^1 \left( \theta^{-1} \left( \frac{1}{(1-\theta)^r} - 1 \right) \right) (1-\theta)^{b+r-1} d\theta \\
&= \int_0^1 (\theta^{-1} (1 - (1-\theta)^r)) (1-\theta)^{b-1} d\theta \\
&\leq \int_0^1 \theta^{-1} r \frac{\theta}{\sqrt{1-\theta}} (1-\theta)^{b-1} d\theta \\
&= r \int_0^1 (1-\theta)^{b-1.5} d\theta = \frac{r}{b-0.5},
\end{aligned}$$

where the identity  $\sum_{y=1}^{\infty} \frac{\Gamma(y+r)}{y! \Gamma(r)} \theta^y = \frac{1}{(1-\theta)^r} - 1$  is due to the normalization constant for negative binomial distributions, and we also used Eq. (D.13) on  $1 - (1-\theta)^r$ .  $\square$

*Proof of Eq. (D.11).* First we prove that:

$$1 - \frac{\Gamma(b)}{\Gamma(b+c/K)} \leq \frac{c}{K} (2 + \ln b).$$

The recursion defining  $\Gamma(b)$  allows us to write:

$$1 - \frac{\Gamma(b)}{\Gamma(b+c/K)} = 1 - \left( \prod_{i=1}^{\lfloor b \rfloor - 1} \frac{b-i}{b+c/K-i} \right) \frac{\Gamma(b - \lfloor b \rfloor + 1)}{\Gamma(b+c/K - \lfloor b \rfloor + 1)}.$$

The argument proceeds in one of two ways. If  $\frac{\Gamma(b - \lfloor b \rfloor + 1)}{\Gamma(b+c/K - \lfloor b \rfloor + 1)} \geq 1$ , then we have:

$$\begin{aligned}
1 - \frac{\Gamma(b)}{\Gamma(b+c/K)} &\leq 1 - \prod_{i=1}^{\lfloor b \rfloor - 1} \frac{b-i}{b+c/K-i} \\
&= \left( 1 - \frac{b-1}{b+c/K-1} \right) + \frac{b-1}{b+c/K-1} - \left( \prod_{i=1}^{\lfloor b \rfloor - 1} \frac{b-i}{b+c/K-i} \right) \\
&= \frac{c}{K} \frac{1}{b+c/K-1} + \frac{b-1}{b+c/K-1} \left( 1 - \prod_{i=2}^{\lfloor b \rfloor - 1} \frac{b-i}{b+c/K-i} \right) \\
&\leq \frac{c}{K} \frac{1}{b-1} + \left( 1 - \prod_{i=2}^{\lfloor b \rfloor - 1} \frac{b-i}{b+c/K-i} \right) \\
&\leq \dots \leq \frac{c}{K} \sum_{i=1}^{\lfloor b \rfloor - 1} \frac{1}{b-i} \leq \frac{c}{K} (\ln b + 1).
\end{aligned}$$

Else,  $\frac{\Gamma(b-\lfloor b \rfloor+1)}{\Gamma(b+c/K-\lfloor b \rfloor+1)} < 1$  and we write:

$$\begin{aligned} & 1 - \frac{\Gamma(b)}{\Gamma(b+c/K)} \\ &= 1 - \frac{\Gamma(b-\lfloor b \rfloor+1)}{\Gamma(b+c/K-\lfloor b \rfloor+1)} + \frac{\Gamma(b-\lfloor b \rfloor+1)}{\Gamma(b+c/K-\lfloor b \rfloor+1)} \left( 1 - \prod_{i=1}^{\lfloor b \rfloor-1} \frac{b-i}{b+c/K-i} \right) \\ &\leq \left( 1 - \frac{\Gamma(b-\lfloor b \rfloor+1)}{\Gamma(b+c/K-\lfloor b \rfloor+1)} \right) + \frac{c}{K}(\ln b + 1). \end{aligned}$$

We now argue that for all  $x \in [1, 2)$ , for all  $K \geq c$ ,  $1 - \frac{\Gamma(x)}{\Gamma(x+c/K)} \leq \frac{c}{K}$ . By convexity of  $\Gamma(x)$ , we know that  $\Gamma(x) \geq \Gamma(x+c/K) - \frac{c}{K}\Gamma'(x+c/K)$ . Hence  $\frac{\Gamma(x)}{\Gamma(x+c/K)} \geq 1 - \frac{c}{K} \frac{\Gamma'(x+c/K)}{\Gamma(x+c/K)}$ . Since  $x+c/K \in [1, 3)$  and  $\psi(y) = \frac{\Gamma'(y)}{\Gamma(y)}$ , the digamma function, is a monotone increasing function (it is the derivative of a  $\ln \Gamma(x)$ , which is also convex),  $\left| \frac{\Gamma'(x+c/K)}{\Gamma(x+c/K)} \right| \leq \left| \frac{\Gamma'(3)}{\Gamma(3)} \right| \leq 1$ . Applying this to  $x = b - \lfloor b \rfloor + 1$ , we conclude that:

$$1 - \frac{\Gamma(b)}{\Gamma(b+c/K)} \leq \frac{c}{K}(2 + \ln b).$$

We now show that:

$$\frac{\Gamma(b)}{\Gamma(b+c/K)} - 1 \geq -\frac{c}{K}(\ln b + \ln 2).$$

Convexity of  $\Gamma(y)$  means that:

$$\Gamma(b) \geq \Gamma(b+c/K) - \frac{c}{K}\Gamma'(b+c/K) \rightarrow \frac{\Gamma(b)}{\Gamma(b+c/K)} - 1 \geq -\frac{c}{K} \frac{\Gamma'(b+c/K)}{\Gamma(b+c/K)}.$$

From (Alzer, 1997, Equation 2.2), we know that  $\psi(x) \leq \ln(x)$  for positive  $x$ . Therefore:

$$-\frac{c}{K} \frac{\Gamma'(b+c/K)}{\Gamma(b+c/K)} \geq -\frac{c}{K} \ln(b+c/K) \geq -\frac{c}{K}(\ln b + \ln 2)$$

since  $b + \frac{c}{K} \leq 2b$ .

We combine two sides of the inequality to conclude that the absolute value is at most  $\frac{c}{K}(2 + \ln b)$ .  $\square$

*Proof of Eq. (D.12).*

$$\begin{aligned} \left| c - \frac{K}{B(c/K, b)} \right| &= c \left| \frac{K/c}{\Gamma(c/K)} \frac{\Gamma(c/K+b)}{\Gamma(b)} - 1 \right| \\ &= c \left| \frac{K/c}{\Gamma(c/K)} \left( \frac{\Gamma(c/K+b)}{\Gamma(b)} - 1 \right) + \left( \frac{K/c}{\Gamma(c/K)} - 1 \right) \right| \\ &\leq c \left( \frac{K/c}{\Gamma(c/K)} \left| \frac{\Gamma(c/K+b)}{\Gamma(b)} - 1 \right| + \left| \frac{K/c}{\Gamma(c/K)} - 1 \right| \right). \end{aligned}$$

On the one hand:

$$\frac{K/c}{\Gamma(c/K)} = \frac{\Gamma(1)}{\Gamma(1+c/K)}.$$

From Eq. (D.11), we know:

$$\left| \frac{\Gamma(1)}{\Gamma(1 + c/K)} - 1 \right| \leq \frac{2c}{K}.$$

On the other hand, let  $y = \frac{\Gamma(c/K+b)}{\Gamma(b)}$ . Then:

$$\left| \frac{\Gamma(c/K+b)}{\Gamma(b)} - 1 \right| = \left| \frac{1}{y} - 1 \right| = \frac{|1-y|}{y} \leq \frac{2c}{K}(2 + \ln b).$$

Again using Eq. (D.11),  $|1-y| \leq \frac{c}{K}(2 + \ln b)$ . Since  $K \geq 2c(\ln b + 2)$ , this is at most 0.5, meaning  $y \geq 0.5$ . In all:

$$\begin{aligned} \left| c - \frac{K}{B(c/K, b)} \right| &\leq c \left( \left( 1 + \frac{2c}{K} \right) 2 \frac{c}{K} (2 + \ln b) + \frac{2c}{K} \right) \\ &\leq \frac{c}{K} (3 \ln b + 8). \end{aligned}$$

□

## Appendix E: Verification of upper bound's assumptions for additional examples

### E.1. Gamma-Poisson

First we write down the functions in Definition 4.1 for gamma-Poisson. This requires expressing the rate measure and likelihood in exponential-family form:

$$h(x|\theta) = \frac{1}{x!} \theta^x \exp(-\theta), \quad \nu(d\theta) = \gamma \lambda \theta^{-1} \exp(-\lambda \theta),$$

which means that  $\kappa(x) = 1/x!$ ,  $\phi(x) = x$ ,  $\mu(\theta) = 0$ ,  $A(\theta) = \theta$ . This leads to the normalizer:

$$S = \int_0^\infty \theta^\xi \exp(-\lambda \theta) d\theta = \Gamma(\xi + 1) \lambda^{-(\xi+1)}.$$

Therefore,  $h_c$  is:

$$\begin{aligned} h_c(x_n = x | x_{1:(n-1)}) &= \frac{1}{x!} \frac{\Gamma(-1 + \sum_{i=1}^{n-1} x_i + x + 1) (\lambda + n)^{-1 + \sum_{i=1}^{n-1} x_i + x + 1}}{\Gamma(-1 + \sum_{i=1}^{n-1} x_i + 1) (\lambda + n - 1)^{-1 + \sum_{i=1}^{n-1} x_i + 1}} \\ &= \frac{1}{x!} \frac{\Gamma(\sum_{i=1}^{n-1} x_i + x)}{\Gamma(\sum_{i=1}^{n-1} x_i)} \left( \frac{1}{\lambda + n} \right)^x \left( 1 - \frac{1}{\lambda + n} \right)^{\sum_{i=1}^{n-1} x_i}, \end{aligned}$$

and similarly  $\tilde{h}_c$  is:

$$\begin{aligned} \tilde{h}_c(x_n = x | x_{1:(n-1)}) &= \frac{1}{x!} \frac{\Gamma(-1 + \sum_{i=1}^{n-1} x_i + x + 1 + \gamma \lambda / K) (\lambda + n)^{-1 + \sum_{i=1}^{n-1} x_i + x + 1 + \gamma \lambda / K}}{\Gamma(-1 + \sum_{i=1}^{n-1} x_i + 1 + \gamma \lambda / K) (\lambda + n - 1)^{-1 + \sum_{i=1}^{n-1} x_i + 1 + \gamma \lambda / K}} \\ &= \frac{1}{x!} \frac{\Gamma(\sum_{i=1}^{n-1} x_i + x + \gamma \lambda / K)}{\Gamma(\sum_{i=1}^{n-1} x_i + \gamma \lambda / K)} \left( \frac{1}{\lambda + n} \right)^x \left( 1 - \frac{1}{\lambda + n} \right)^{\sum_{i=1}^{n-1} x_i + \gamma \lambda / K}, \end{aligned}$$

and  $M_{n,x}$  is:

$$M_{n,x} = \gamma\lambda \frac{1}{x!} \Gamma(x)(\lambda+n)^{-x} = \frac{\gamma\lambda}{x(\lambda+n)^x}.$$

Now, we state the constants so that gamma-Poisson satisfies Assumption 2, and give the proof.

**Proposition E.1** (Gamma-Poisson satisfies Assumption 2). *The following hold for arbitrary  $\gamma, \lambda > 0$ . For any  $n$ :*

$$\sum_{x=1}^{\infty} M_{n,x} \leq \frac{\gamma\lambda}{n-1+\lambda}.$$

$$\sum_{x=1}^{\infty} \tilde{h}_c(x|x_{1:(n-1)}=0) \leq \frac{\gamma\lambda}{n-1+\lambda}.$$

For any  $K$ :

$$\sum_{x=0}^{\infty} \left| h_c(x|x_{1:(n-1)}) - \tilde{h}_c(x|x_{1:(n-1)}) \right| \leq \frac{2\gamma\lambda}{K} \frac{1}{n-1+\lambda}.$$

For any  $K$ :

$$\sum_{x=1}^{\infty} \left| M_{n,x} - K\tilde{h}_c(x|x_{1:(n-1)}=0) \right| \leq \frac{\gamma^2\lambda + e\gamma^2\lambda^2}{K} \frac{1}{n-1+\lambda}.$$

*Proof of Proposition E.1.* The growth rate condition of target model is simple:

$$\sum_{x=1}^{\infty} M_{n,x} = \gamma\lambda \sum_{x=1}^{\infty} \frac{1}{x(\lambda+n)^x} \leq \gamma\lambda \sum_{x=1}^{\infty} \frac{1}{(\lambda+n)^x} = \frac{\gamma\lambda}{n-1+\lambda}.$$

The growth rate condition of approximate model is also simple:

$$\begin{aligned} \sum_{x=1}^{\infty} \tilde{h}_c(x|x_{1:(n-1)}=0) &= 1 - \tilde{h}_c(0|x_{1:(n-1)}=0) = \left(1 - \frac{1}{\lambda+n}\right)^{\gamma\lambda/K} \\ &\leq \frac{\gamma\lambda}{K} \frac{(\lambda+n)^{-1}}{1 - (\lambda+n)^{-1}} = \frac{1}{K} \frac{\gamma\lambda}{n-1+\lambda}, \end{aligned}$$

where we have used Eq. (D.6) with  $p = \frac{\gamma\lambda}{K}$ ,  $x = (\lambda+n)^{-1}$ .

For the total variation between  $h_c$  and  $\tilde{h}_c$  condition, observe that  $h_c$  and  $\tilde{h}_c$  are probability mass functions of negative binomial distributions, namely:

$$h_c(x|x_{1:(n-1)}) = \text{NB} \left( x \mid \sum_{i=1}^{n-1} x_i, (\lambda+n)^{-1} \right)$$

$$\tilde{h}_c(x|x_{1:(n-1)}) = \text{NB} \left( x \mid \sum_{i=1}^{n-1} x_i + \gamma\lambda/K, (\lambda+n)^{-1} \right).$$

The two negative binomial distributions have the same success probability and only differ in the number of trials. Hence using Eq. (D.7), we have:

$$\sum_{x=0}^{\infty} \left| h_c(x|x_{1:(n-1)}) - \tilde{h}_c(x|x_{1:(n-1)}) \right| \leq 2 \frac{\gamma\lambda}{K} \frac{(\lambda+n)^{-1}}{1 - (\lambda+n)^{-1}} = \frac{2\gamma\lambda}{K} \frac{1}{n-1+\lambda}.$$

For the total variation between  $M_{n,\cdot}$  and  $K\tilde{h}_c(\cdot | 0)$  condition:

$$\begin{aligned} & \sum_{x=1}^{\infty} \left| M_{n,x} - K\tilde{h}_c(x|x_{1:(n-1)} = 0) \right| \\ &= \sum_{x=1}^{\infty} \frac{1}{(\lambda+n)^x} \left| \frac{\gamma\lambda}{x} - K \frac{\Gamma(\gamma\lambda/K+x)}{\Gamma(\gamma\lambda/K)x!} \left(1 - \frac{1}{\lambda+n}\right)^{\gamma\lambda/K} \right| \\ &\leq \sum_{x=1}^{\infty} \frac{1}{(\lambda+n)^x} \left( \left| \frac{\gamma\lambda}{x} \left(1 - \left(1 - \frac{1}{\lambda+n}\right)^{\gamma\lambda/K}\right) \right| + \left| \frac{\gamma\lambda}{x} - K \frac{\Gamma(\gamma\lambda/K+x)}{\Gamma(\gamma\lambda/K)x!} \right| \right). \end{aligned}$$

Using Eqs. (D.7) and (D.8) we can upper bound:

$$\begin{aligned} 1 - \left(1 - \frac{1}{\lambda+n}\right)^{\gamma\lambda/K} &\leq \frac{\gamma\lambda}{K} \frac{1}{\lambda+n-1} \\ \left| \frac{\gamma\lambda}{x} - K \frac{\Gamma(\gamma\lambda/K+x)}{\Gamma(\gamma\lambda/K)x!} \right| &\leq \frac{e\gamma^2\lambda^2}{K}. \end{aligned}$$

This means:

$$\begin{aligned} & \sum_{x=1}^{\infty} \left| M_{n,x} - K\tilde{h}_c(x|x_{1:(n-1)} = 0) \right| \\ &\leq \sum_{x=1}^{\infty} \frac{1}{(\lambda+n)^x} \frac{\gamma\lambda}{x} \frac{\gamma\lambda}{K} \frac{1}{\lambda+n-1} + \sum_{x=1}^{\infty} \frac{1}{(\lambda+n)^x} \frac{e\gamma^2\lambda^2}{K} \\ &\leq \frac{\gamma^2\lambda^2}{K} \frac{1}{(\lambda+n-1)^2} + \frac{e\gamma^2\lambda^2}{K} \frac{1}{\lambda+n-1} \\ &\leq \frac{\gamma^2\lambda + e\gamma^2\lambda^2}{K} \frac{1}{n-1+\lambda}. \end{aligned}$$

□

## E.2. Beta-negative binomial

First we write down the functions in Definition 4.1 for beta-negative binomial. This requires expressing the rate measure and likelihood in exponential-family form:

$$\begin{aligned} h(x|\theta) &= \frac{\Gamma(x+r)}{x!\Gamma(r)} \theta^x \exp(r \log(1-\theta)), \\ \nu(d\theta) &= \gamma\alpha\theta^{-1} \exp(\log(1-\theta)(\alpha-1)) \mathbf{1}\{\theta \leq 1\}, \end{aligned}$$

which means that  $\kappa(x) = \Gamma(x+r)/\Gamma(r)x!$ ,  $\phi(x) = x$ ,  $\mu(\theta) = 0$ ,  $A(\theta) = -r \log(1-\theta)$ . This leads to the normalizer:

$$S = \int_0^1 \theta^\xi (1-\theta)^{r\lambda} d\theta = B(\xi+1, r\lambda+1).$$

To match the parametrizations, we need to set  $\lambda = \frac{\alpha-1}{r}$  i.e.  $r\lambda = \alpha - 1$ . Therefore,  $h_c$  is:

$$h_c(x_n = x|x_{1:(n-1)}) = \frac{\Gamma(x+r)}{x!\Gamma(r)} \frac{B(\sum_{i=1}^{n-1} x_i + x, rn + \alpha)}{B(\sum_{i=1}^{n-1} x_i, r(n-1) + \alpha)},$$

and  $\tilde{h}_c$  is:

$$\tilde{h}_c(x_n = x | x_{1:(n-1)}) = \frac{\Gamma(x+r)}{x!\Gamma(r)} \frac{B(c/K + \sum_{i=1}^{n-1} x_i + x, rn + \alpha)}{B(c/K + \sum_{i=1}^{n-1} x_i, r(n-1) + \alpha)}.$$

and  $M_{n,x}$  is:

$$M_{n,x} = \gamma\alpha \frac{\Gamma(x+r)}{x!\Gamma(r)} B(x, rn + \alpha).$$

Now, we state the constants so that beta-negative binomial satisfies Assumption 2, and give the proof.

**Proposition E.2** (Beta-negative binomial satisfies Assumption 2). *The following hold for any  $\gamma > 0, \alpha \geq 1$ . For any  $n$ :*

$$\sum_{x=1}^{\infty} M_{n,x} \leq \frac{\gamma\alpha}{n-1 + (\alpha - 0.5)/r}.$$

For any  $n$ , any  $K$ :

$$\sum_{x=1}^{\infty} \tilde{h}_c(x | x_{1:(n-1)} = 0) \leq \frac{1}{K} \frac{4\gamma\alpha}{n-1 + (\alpha - 0.5)/r}.$$

For any  $K$ :

$$\sum_{x=0}^{\infty} \left| h_c(x | x_{1:(n-1)}) - \tilde{h}_c(x | x_{1:(n-1)}) \right| \leq \frac{\gamma\alpha}{K} \frac{1}{n-1 + \alpha/r}.$$

For any  $n$ , for  $K \geq \gamma\alpha(3 \ln(r(n-1) + \alpha) + 8)$ :

$$\begin{aligned} & \sum_{x=1}^{\infty} \left| M_{n,x} - K \tilde{h}_c(x | x_{1:(n-1)} = 0) \right| \\ & \leq \frac{\gamma\alpha}{K} \frac{(4\gamma\alpha + 3) \ln(rn + \alpha + 1) + (10 + 2r)\gamma\alpha + 24}{n-1 + (\alpha - 0.5)/r}. \end{aligned}$$

*Proof of Proposition E.2.* The first growth rate condition is easy to verify:

$$\sum_{x=1}^{\infty} M_{n,x} = \gamma\alpha \sum_{x=1}^{\infty} \frac{\Gamma(x+r)}{\Gamma(r)x!} B(x, rn + \alpha) \leq \gamma\alpha \frac{r}{r(n-1) + \alpha - 0.5}.$$

where we have used Eq. (D.10) with  $b = r(n-1) + \alpha$ .

As for the other growth rate condition,

$$\begin{aligned} \sum_{x=1}^{\infty} \tilde{h}_c(x | x_{1:(n-1)} = 0) &= 1 - \tilde{h}_c(0 | x_{1:(n-1)} = 0) = 1 - \frac{B(\gamma\alpha/K, rn + \alpha)}{B(\gamma\alpha/K, r(n-1) + \alpha)} \\ &= \frac{B(\gamma\alpha/K, r(n-1) + \alpha) - B(\gamma\alpha/K, rn + \alpha)}{B(\gamma\alpha/K, r(n-1) + \alpha)}. \end{aligned}$$

The numerator is small because of Eq. (D.9) where  $x = \gamma\alpha/K, y = r(n-1) + \alpha, z = rn + \alpha$ :

$$B(\gamma\alpha/K, r(n-1) + \alpha) - B(\gamma\alpha/K, rn + \alpha) \leq rB(\gamma\alpha/K + 1, r(n-1) + \alpha - 1).$$

The denominator is large because Equation (D.12) with Equation (D.12) with  $c = \gamma\alpha$ ,  $b = r(n-1) + \alpha$ :

$$\frac{1}{B(\gamma\alpha/K, r(n-1) + \alpha)} \leq \frac{4\gamma\alpha}{K}.$$

Combining the two give and using a simple bound on the beta function yields:

$$\sum_{x=1}^{\infty} \tilde{h}_c(x | x_{1:(n-1)} = 0) \leq \frac{1}{K} \frac{4\gamma\alpha}{n-1 + (\alpha - 0.5)/r}.$$

For the total variation between  $h_c$  and  $\tilde{h}_c$  condition, we first discuss how each function can be expressed a probability mass function of so-called beta negative binomial i.e., BNB ((Johnson, Kemp and Kotz, 2005, Section 6.2.3)) distribution. Let  $A = \sum_{i=1}^{n-1} x_i$ . Observe that:

$$\frac{\Gamma(x+r)}{\Gamma(r)x!} \frac{B(A+x, rn+\alpha)}{B(A, r(n-1) + \alpha)} = \frac{\Gamma(A+r)}{\Gamma(A)x!} \frac{B(r+x, A+r(n-1) + \alpha)}{B(r, r(n-1) + \alpha)}. \quad (\text{E.1})$$

The random variable  $V_1$  whose p.m.f at  $x$  appears on the right hand side of Eq. (E.1) is the result of a two-step sampling procedure.

$$P \sim \text{Beta}(r, r(n-1) + \alpha), \quad V_1|P \sim \text{NB}(A; P).$$

We denote such a distribution as  $V_1 \sim \text{BNB}(A; r, r(n-1) + \alpha)$ . An analogous argument applies to  $\tilde{h}_c$ :

$$P \sim \text{Beta}(r, r(n-1) + \alpha), \quad V_2|P \sim \text{NB}\left(A + \frac{\gamma\alpha}{K}; P\right).$$

Therefore:

$$\begin{aligned} h_c(x | x_{1:(n-1)}) &= \text{BNB}(x | A; r, r(n-1) + \alpha) \\ \tilde{h}_c(x | x_{1:(n-1)}) &= \text{BNB}\left(x | A + \frac{\gamma\alpha}{K}; r, r(n-1) + \alpha\right). \end{aligned}$$

We now bound the total variation between the BNB distributions. Because they have a common mixing distribution, we can upper bound the distance with an integral using simple triangle inequalities:

$$\begin{aligned} d_{TV}(h_c, \tilde{h}_c) &= \frac{1}{2} \sum_{x=0}^{\infty} |\mathbb{P}(V_1 = x) - \mathbb{P}(V_2 = x)| \\ &= \frac{1}{2} \sum_{x=0}^{\infty} \left| \int_0^1 (\mathbb{P}(V_1 = x|P=p) - \mathbb{P}(V_2 = x|P=p)) \mathbb{P}(P \in dp) \right| \\ &\leq \int_0^1 \left( \frac{1}{2} \sum_{x=0}^{\infty} |\mathbb{P}(V_1 = x|P=p) - \mathbb{P}(V_2 = x|P=p)| \right) \mathbb{P}(P \in dp) \\ &= \int_0^1 d_{TV}(\text{NB}(A, p), \text{NB}(A + \gamma\alpha/K, p)) \mathbb{P}(P \in dp). \end{aligned}$$

For any  $p$ , Eq. (D.7) is used to upper bound the total variation distance between negative

binomial distributions. Therefore:

$$\begin{aligned} d_{TV} \left( h_c, \tilde{h}_c \right) &\leq \int_0^1 \frac{\gamma\alpha}{K} \frac{p}{1-p} \mathbb{P}(P \in dp) \\ &= \frac{\gamma\alpha}{K} \frac{1}{B(r, r(n-1) + \alpha)} \int_0^1 p^r (1-p)^{r(n-1) + \alpha - 2} dp \\ &= \frac{\gamma\alpha}{K} \frac{B(r+1, r(n-1) + \alpha - 1)}{B(r, r(n-1) + \alpha)} = \frac{\gamma\alpha}{K} \frac{1}{n-1 + \alpha/r}. \end{aligned}$$

Finally, we verify the condition between  $K\tilde{h}_c$  and  $M_{n,\cdot}$ , which is showing that the following sum is small:

$$\sum_{x=1}^{\infty} \frac{\Gamma(x+r)}{x!\Gamma(r)} \left| c\gamma\alpha B(x, rn + \alpha) - K \frac{B(\gamma\alpha/K + x, rn + \alpha)}{B(\gamma\alpha/K, r(n-1) + \alpha)} \right|.$$

We look at the summand for  $x = 1$  and the summation from  $x = 2$  through  $\infty$  separately. For  $x = 1$ , we prove that:

$$\left| \gamma\alpha B(1, rn + \alpha) - K \frac{B(\gamma\alpha/K + 1, rn + \alpha)}{B(\gamma\alpha/K, r(n-1) + \alpha)} \right| \leq \frac{4r\gamma^2\alpha^2}{K} \frac{2 + \ln(rn + \alpha + 1)}{rn + \alpha}. \quad (\text{E.2})$$

Expanding gives:

$$\begin{aligned} &\left| \gamma\alpha B(1, rn + \alpha) - K \frac{B(1 + \gamma\alpha/K, rn + \alpha)}{B(\gamma\alpha/K, r(n-1) + \alpha)} \right| \\ &= \frac{|\gamma\alpha B(1, rn + \alpha) B(\gamma\alpha/K, r(n-1) + \alpha) - KB(1 + \gamma\alpha/K, rn + \alpha)|}{B(\gamma\alpha/K, r(n-1) + \alpha)}. \end{aligned} \quad (\text{E.3})$$

We look at the numerator of the right hand side in Eq. (E.3):

$$\begin{aligned} &\left| \gamma\alpha B(1, rn + \alpha) \frac{\Gamma(\gamma\alpha/K)\Gamma(r(n-1) + \alpha)}{\Gamma(\gamma\alpha/K + r(n-1) + \alpha)} - K \frac{\Gamma(1 + \gamma\alpha/K)\Gamma(rn + \alpha)}{\Gamma(1 + \gamma\alpha/K + rn + \alpha)} \right| \\ &= \gamma\alpha\Gamma(\gamma\alpha/K) \left| \frac{1}{rn + \alpha} \frac{\Gamma(r(n-1) + \alpha)}{\Gamma(\gamma\alpha/K + r(n-1) + \alpha)} - \frac{\Gamma(rn + \alpha)}{\Gamma(\gamma\alpha/K + 1 + rn + \alpha)} \right| \\ &= \frac{\gamma\alpha\Gamma(\gamma\alpha/K)}{rn + \alpha} \left| \frac{\Gamma(r(n-1) + \alpha)}{\Gamma(\gamma\alpha/K + r(n-1) + \alpha)} - \frac{\Gamma(rn + \alpha + 1)}{\Gamma(\gamma\alpha/K + 1 + rn + \alpha)} \right| \\ &\leq \frac{\gamma\alpha\Gamma(\gamma\alpha/K)}{rn + \alpha} \left( \left| \frac{\Gamma(r(n-1) + \alpha)}{\Gamma(\gamma\alpha/K + r(n-1) + \alpha)} - 1 \right| + \left| \frac{\Gamma(rn + \alpha + 1)}{\Gamma(\gamma\alpha/K + 1 + rn + \alpha)} - 1 \right| \right) \\ &\leq \frac{\gamma\alpha\Gamma(\gamma\alpha/K)}{rn + \alpha} \frac{2\gamma\alpha}{K} (2 + \ln(rn + \alpha + 1)). \end{aligned}$$

where we have used Eq. (D.11) with  $c = \gamma\alpha$  and  $b = r(n-1) + \alpha$  or  $b = rn + \alpha + 1$ . In all, Eq. (E.3) is upper bounded by:

$$\begin{aligned} &\frac{2\gamma^2\alpha^2}{rn + \alpha} \frac{2 + \ln(rn + \alpha + 1)}{K} \frac{\Gamma(\gamma\alpha/K)}{B(\gamma\alpha/K, r(n-1) + \alpha)} \\ &= \frac{2\gamma^2\alpha^2}{rn + \alpha} \frac{2 + \ln(rn + \alpha + 1)}{K} \frac{\Gamma(\gamma\alpha/K + r(n-1) + \alpha)}{\Gamma(r(n-1) + \alpha)} \\ &\leq \frac{4\gamma^2\alpha^2}{K} \frac{2 + \ln(rn + \alpha + 1)}{rn + \alpha}, \end{aligned}$$

since  $\frac{\Gamma(r(n-1)+\alpha)}{\Gamma(r(n-1)+\alpha+\gamma\alpha/K)} \geq 1 - \frac{\gamma\alpha}{K}(2+\ln(r(n-1)+\alpha)) \geq 0.5$  with  $K \geq 2\gamma\alpha(2+\ln(r(n-1)+\alpha))$ , and this is the proof of Eq. (E.2).

We now move onto the summands from  $x = 2$  to  $\infty$ . By triangle inequality:

$$\left| \gamma\alpha B(x, rn + \alpha) - K \frac{B(\gamma\alpha/K + x, rn + \alpha)}{B(\gamma\alpha/K, r(n-1) + \alpha)} \right| \leq T_1(x) + T_2(x),$$

where:

$$T_1(x) := B(x, rn + \alpha) \left| \gamma\alpha - \frac{K}{B(\gamma\alpha/K, r(n-1) + \alpha)} \right|,$$

$$T_2(x) := K \frac{|B(x, rn + \alpha) - B(\frac{\gamma\alpha}{K} + x, rn + \alpha)|}{B(\gamma\alpha/K, r(n-1) + \alpha)}.$$

The helper inequalities we have proven once again are useful:

$$\left| \gamma\alpha - \frac{K}{B(\gamma\alpha/K, r(n-1) + \alpha)} \right| \leq \frac{\gamma\alpha}{K}(3\ln(r(n-1) + \alpha) + 8)$$

$$\frac{K}{B(\gamma\alpha/K, r(n-1) + \alpha)} \leq \gamma\alpha + \frac{\gamma\alpha}{K}(3\ln(r(n-1) + \alpha) + 8) \leq 2\gamma\alpha,$$

$$|B(x, rn + \alpha) - B(\gamma\alpha/K + x, rn + \alpha)| \leq \frac{\gamma\alpha}{K} B(x-1, rn + \alpha + 1)$$

since  $K \geq \gamma\alpha(3\ln(r(n-1) + \alpha) + 8)$ , we have applied Eq. (D.12) in the first and second inequality and Eq. (D.9) in the third one. So for each  $x \geq 2$ , each summand is at most:

$$\frac{\Gamma(x+r)}{x!\Gamma(r)} \left| cB(x, rn + \alpha) - K \frac{B(\gamma\alpha/K + x, rn + \alpha)}{B(\gamma\alpha/K, r(n-1) + \alpha)} \right|$$

$$\leq \frac{\gamma\alpha(3\ln(r(n-1) + \alpha) + 8)}{K} \frac{\Gamma(x+r)}{x!\Gamma(r)} B(x, rn + \alpha) + \frac{2\gamma^2\alpha^2}{K} \frac{\Gamma(x+r)}{x!\Gamma(r)} B(x-1, rn + \alpha + 1).$$

To upper bound the summation from  $x = 2$  to  $\infty$ , it suffices to bound:

$$\sum_{x=2}^{\infty} \frac{\Gamma(x+r)}{\Gamma(r)x!} B(x, rn + \alpha) \leq \sum_{x=1}^{\infty} \frac{\Gamma(x+r)}{\Gamma(r)x!} B(x, rn + \alpha) \leq \frac{r}{r(n-1) + \alpha - 0.5},$$

and:

$$\sum_{x=2}^{\infty} \frac{\Gamma(x+r)}{\Gamma(r)x!} B(x-1, rn + \alpha + 1) \leq r \sum_{x=2}^{\infty} \frac{\Gamma(x-1+r+1)}{\Gamma(r+1)(x-1)!} B(x-1, rn + \alpha + 1)$$

$$\leq r \sum_{z=1}^{\infty} \frac{\Gamma(z+r+1)}{\Gamma(r+1)z!} B(z, rn + \alpha + 1)$$

$$\leq \frac{r(r+1)}{r(n-1) + \alpha - 0.5}$$

So the summation from  $x = 2$  to  $\infty$  is upper bounded by:

$$\frac{\gamma\alpha(3\ln(r(n-1) + \alpha) + 8)}{K} \frac{r}{r(n-1) + \alpha - 0.5} + \frac{2\gamma^2\alpha^2}{K} \frac{r(r+1)}{r(n-1) + \alpha - 0.5} \quad (\text{E.4})$$

Eqs. (E.2) and (E.4) combine to give:

$$\begin{aligned} & \sum_{x=1}^{\infty} \left| M_{n,x} - K \tilde{h}_c(x | x_{1:(n-1)} = 0) \right| \\ & \leq \frac{\gamma \alpha (4\gamma \alpha + 3) \ln(rn + \alpha + 1) + (10 + 2r)\gamma \alpha + 24}{K (n - 1 + (\alpha - 0.5)/r)}. \end{aligned}$$

□

## Appendix F: Proofs of CRM bounds

### F.1. Upper bound

*Proof of Theorem 4.2.* Let  $\beta$  be the smallest positive constant where  $\beta^2 C_1 / (1 + \beta) \geq 2$ . We will focus on the case where the approximation level  $K$  is essentially  $\Omega(\ln N)$ :

$$K \geq \max((\beta + 1) \max(C(K, C_1), C(N, C_1)), C_2(\ln N + C_3)). \quad (\text{F.1})$$

To see why it is sufficient, observe that the upper bound in Theorem 4.2 naturally holds for  $K$  smaller than  $\ln N$ . Total variation distance is always upper bounded by 1; if  $K = o(\ln N)$ , then by selecting reasonable constants  $C', C'', C'''$ , we can make the right hand side at least 1, and satisfy the inequality. In the sequel, we will only consider the situation in Eq. (F.1).

First, we argue that it suffices to bound the total variation distance between the *feature-allocation matrices* coming from the target model and the approximate model. Given the latent measures  $X_1, X_2, \dots, X_N$  from the target model, we can read off the feature-allocation matrix  $F$ , which has  $N$  rows and as many columns as there are unique atom locations among the  $X_i$ 's:

1. The  $i$ th row of  $F$  records the atom sizes of  $X_i$ .
2. Each column corresponds to an atom location: the locations are sorted first according to the index of the first measure  $X_i$  to manifest it (counting from 1, 2, ...), and then its atom size in  $X_i$ .

The marginal process that described the atom sizes of  $X_n | X_{n-1}, X_{n-2}, \dots, X_1$  in Proposition C.1 is also the description of how the rows of  $F$  are generated. The joint distribution  $X_1, X_2, \dots, X_n$  can be two-step sampled. First, the feature-allocation matrix  $F$  is sampled. Then, the atom locations are drawn iid from the base measure  $H$ : each column of  $F$  is assigned an atom location, and the latent measure  $X_i$  has atom size  $F_{i,j}$  on the  $j$ th atom location. A similar two-step sampling generates  $Z_1, Z_2, \dots, Z_n$ , the latent measures under the approximate model: the distribution over the feature-allocation matrix  $F'$  follows Proposition C.2 instead of Proposition C.1, but conditioned on the feature-allocation matrix, the process generating atom locations and constructing latent measures is exactly the same. In other words, this implies that the conditional distributions  $Y_{1:N} | F = f$  and  $W_{1:N} | F' = f$  are the same, since both models have the same the observational likelihood  $f$  given the latent measures 1 through  $N$ . Denote  $P_F$  to be the distribution of the feature-allocation matrix under the target model, and  $P_{F'}$  the distribution of the feature-allocation matrix under the approximate model. Lemma D.7 implies that:

$$d_{TV}(P_{W_{1:N}}, P_{Y_{1:N}}) \leq d_{TV}(P_F, P_{F'}). \quad (\text{F.2})$$

Next, we parametrize the feature-allocation matrices in a way that is convenient for the analysis of total variation distance. Let  $J$  be the number of columns of  $F$ . Our parametrization involves  $d_{n,x}$ , for  $n \in [N]$  and  $x \in \mathbb{N}$ , and  $s_j$ , for  $j \in [J]$ :

1. For  $n = 1, 2, \dots, N$ :
  - (a) If  $n = 1$ , for each  $x \in \mathbb{N}$ ,  $d_{1,x}$  counts the number of columns  $j$  where  $F_{1,j} = x$ .
  - (b) For  $n \geq 2$ , for each  $x \in \mathbb{N}$ , let  $J_n = \{j : \forall i < n, F_{i,j} = 0\}$  i.e. no observation before  $n$  manifests the atom locations indexed by columns in  $J_n$ . For each  $x \in \mathbb{N}$ ,  $d_{n,x}$  counts the number of columns  $j \in J_n$  where  $F_{n,j} = x$ .
2. For  $j = 1, 2, \dots, J$ , let  $I_j = \min\{i : F_{i,j} > 0\}$  i.e. the first row to manifest the  $j$ th atom location. Let  $s_j = F_{I_j:N,j}$  i.e. the history of the  $j$ th atom location.

In words,  $d_{n,x}$  is the number of atom locations that is first instantiated by the individual  $n$  and each atom has size  $x$ , while  $s_j$  is the history of the  $j$ th atom location.  $\sum_{n=1}^N \sum_{x=1}^{\infty} d_{n,x}$  is exactly  $J$ , the number of columns. We use the short-hand  $d$  to refer to the collection of  $d_{n,x}$  and  $s$  the collection of  $s_j$ . There is a one-to-one mapping between  $(d, s)$  and the feature allocation matrix  $f$ . Let  $(D, S)$  be the distribution of  $d$  and  $s$  under the target model, while  $(D', S')$  is the distribution under the approximate model. We now aim to compare the joint distribution:

$$d_{TV}(P_F, P_{F'}) = d_{TV}(P_{D,S}, P_{D',S'}).$$

Because total variation distance is the infimum of difference probability over all couplings, to find an upper bound on  $d_{TV}(P_{D,S}, P_{D',S'})$ , it suffices to demonstrate a joint distribution such that  $\mathbb{P}((D, S) \neq (D', S'))$  is small. The rest of the proof is dedicated to that end. To start, we only assume that  $(D, S, D', S')$  is a proper coupling, in that marginally  $(D, S) \sim P_{D,S}$  and  $(D', S') \sim P_{D',S'}$ . As we progress, gradually more structure is added to the joint distribution  $(D, S, D', S')$  to control  $\mathbb{P}((D, S) \neq (D', S'))$ .

We first decompose  $\mathbb{P}((D, S) \neq (D', S'))$  into other probabilistic quantities which can be analyzed using Assumption 2. Define the *typical* set:

$$\mathcal{D}^* = \left\{ d : \sum_{n=1}^N \sum_{x=1}^{\infty} d_{n,x} \leq (\beta + 1) \max(C(K, C_1), C(N, C_1)) \right\}.$$

$d \in \mathcal{D}^*$  means that the feature-allocation matrix  $f$  has a bounded number of columns. The claim is that:

$$\mathbb{P}((D, S) \neq (D', S')) \leq \mathbb{P}(D \neq D') + \mathbb{P}(S \neq S' | D = D', D \in \mathcal{D}^*) + \mathbb{P}(D \notin \mathcal{D}^*). \quad (\text{F.3})$$

This is true from basic properties of probabilities and conditional probabilities:

$$\begin{aligned} & \mathbb{P}((D, S) \neq (D', S')) \\ &= \mathbb{P}(D \neq D') + \mathbb{P}(S \neq S', D = D') \\ &= \mathbb{P}(D \neq D') + \mathbb{P}(S \neq S', D = D', D \in \mathcal{D}^*) + \mathbb{P}(S \neq S', D = D', D \notin \mathcal{D}^*) \\ &\leq \mathbb{P}(D \neq D') + \mathbb{P}(S \neq S' | D = D', D \in \mathcal{D}^*) + \mathbb{P}(D \notin \mathcal{D}^*), \end{aligned}$$

The three ideas behind this upper bound are the following. First, because of the growth condition, we can analyze the atypical set probability  $\mathbb{P}(D \notin \mathcal{D}^*)$ . Second, because of the total variation between  $h_c$  and  $\tilde{h}_c$ , we can analyze  $\mathbb{P}(S \neq S' | D = D', D \in \mathcal{D}^*)$ . Finally,

we can analyze  $\mathbb{P}(D \neq D')$  because of the total variation between  $K\tilde{h}_c$  and  $M_{n,\cdot}$ . In what follows we carry out the program.

**Atypical set probability** The  $\mathbb{P}(D \notin \mathcal{D}^*)$  term in Eq. (F.3) is easiest to control. Under the target model Proposition C.1, the  $D_{i,x}$ 's are independent Poissons with mean  $M_{i,x}$ , so the sum  $\sum_{i=1}^N \sum_{x=1}^{\infty} D_{i,x}$  is itself a Poisson with mean  $M = \sum_{i=1}^N \sum_{x=1}^{\infty} M_{i,x}$ . Because of Lemma D.3, for any  $x > 0$ :

$$\mathbb{P}\left(\sum_{i=1}^N \sum_{x=1}^{\infty} D_{i,x} > M + x\right) \leq \exp\left(-\frac{x^2}{2(M+x)}\right).$$

For the event  $\mathbb{P}(D \notin \mathcal{D}^*)$ ,  $M + x = (\beta + 1) \max(C(K, C_1), C(N, C_1))$ ,  $M \leq C(N, C_1)$  due to Eq. (7), so that  $x \geq \beta \max(C(K, C_1), C(N, C_1))$ . Therefore:

$$\mathbb{P}(D \notin \mathcal{D}^*) \leq \exp\left(-\frac{\beta^2}{2(\beta+1)} \max(C(K, C_1), C(N, C_1))\right). \quad (\text{F.4})$$

**Difference between histories** To minimize the difference probability between the histories of atom sizes i.e. the  $\mathbb{P}(S \neq S' | D = D', D \in \mathcal{D}^*)$  term in Eq. (F.3), we will use Eq. (9). The claim is, there exists a coupling of  $S' | D'$  and  $S | D$  such that:

$$\mathbb{P}(S \neq S' | D = D', D \in \mathcal{D}^*) \leq \frac{(\beta+1) \max(C(K, C_1), C(N, C_1))}{K} C(N, C_1). \quad (\text{F.5})$$

Fix some  $d \in \mathcal{D}^*$  – since we are in the typical set, the number of columns in the feature-allocation matrix is at most  $(\beta+1) \max(C(K, C_1), C(N, C_1))$ . Conditioned on  $D = d$ , there is a finite number of history variables  $S$ , one for each atom location; similar for conditioning of  $S'$  on  $D' = d$ . For both the target and the approximate model, the density of the joint distribution factorizes:

$$\begin{aligned} \mathbb{P}(S = s | D = d) &= \prod_{j=1}^J \mathbb{P}(S_j = s_j | D = d) \\ \mathbb{P}(S' = s | D' = d) &= \prod_{j=1}^J \mathbb{P}(S'_j = s_j | D' = d), \end{aligned}$$

since in both marginal processes, the atom sizes for different atom locations are independent of each other. This means we can use Lemma D.8:

$$d_{TV}(P_{S|D=d}, P_{S'|D'=d}) \leq \sum_{j=1}^J d_{TV}(P_{S_j|D=d}, P_{S'_j|D'=d}).$$

We inspect each  $d_{TV}(P_{S_j|D=d}, P_{S'_j|D'=d})$ . Fixing  $d$  also fixes  $I_j$ , the first row to manifest the  $j$ th atom location. The history  $s_j$  is then a  $N - I_j + 1$  dimensional integer vector, whose  $t$ th entry is the atom size over the  $j$ th atom location of the  $t + I_j - 1$  row. Because of Eq. (9), we know that conditioned on the same partial history  $S_j(1 : (t-1)) = S'_j(1 : (t-1)) = s$ , the distributions  $S_j(t)$  and  $S'_j(t)$  are very similar. The conditional distribution  $S_j(t) | D = d, S_j(1 : (t-1)) = s$  is governed by  $h_c$  Proposition C.1 while  $S'_j(t) | D' = d, S'_j(1 : (t-1)) = s$  is governed by  $\tilde{h}_c$  Proposition C.2. Hence:

$$d_{TV}\left(P_{S_j(t)|D=d, S_j(1:(t-1))=s}, P_{S'_j(t)|D'=d, S'_j(1:(t-1))=s}\right) \leq 2 \frac{1}{K} \frac{C_1}{t + I_j - 2 + C_1},$$

for any partial history  $s$ . To use this conditional bound, we again leverage Lemma D.6 to compare the joint  $S_j = (S_j(1), S_j(2), \dots, S_j(N-I_j+1))$  with the joint  $S'_j = (S'_j(1), S'_j(2), \dots, S'_j(N-I_j+1))$ , peeling off one layer at a time.

$$\begin{aligned} & d_{TV}(P_{S_j|D=d}, P_{S'_j|D'=d}) \\ & \leq \sum_{t=1}^{N-I_j+1} \max_s d_{TV} \left( P_{S_j(t)|D=d, S_j(1:(t-1))=s}, P_{S'_j(t)|D'=d, S'_j(1:(t-1))=s} \right) \\ & \leq \sum_{t=1}^{N-I_j+1} 2 \frac{1}{K} \frac{C_1}{t + I_j - 2 + C_1} \\ & \leq 2 \frac{C(N, C_1)}{K}. \end{aligned}$$

Multiplying the right hand side by  $(\beta+1) \max(C(K, C_1), C(N, C_1))$ , the upper bound on  $J$ , we arrive at the same upper bound for the total variation between  $P_{S|D=d}$  and  $P_{S'|D'=d}$  in Eq. (F.5). Furthermore, our analysis of the total variation can be back-tracked to construct the coupling between the conditional distributions  $S|D=s$  and  $S'|D'=d$  which attains that small probability of difference. Since the choice of conditioning  $d \in \mathcal{D}^*$  was arbitrary, we have actually shown Eq. (F.5).

**Difference between new atom sizes** Finally, to control the difference probability for the distribution over new atom sizes i.e. the  $\mathbb{P}(D \neq D')$  term in Eq. (F.3), we will utilize Eqs. (8) and (10). For each  $n$ , define the short-hand  $d_{1:n}$  to refer to the collection  $d_{i,x}$  for  $i \in [n]$ ,  $x \in \mathbb{N}$ , and the typical sets:

$$\mathcal{D}_n^* = \left\{ d_{1:n} : \sum_{i=1}^n \sum_{x=1}^{\infty} d_{i,x} \leq (\beta+1) \max(C(K, C_1), C(N, C_1)) \right\}.$$

The type of expansion performed in Eq. (F.3) can be done once here to see that:

$$\begin{aligned} & \mathbb{P}(D \neq D') \\ & = \mathbb{P}((D_{1:(N-1)}, D_N) \neq (D'_{1:(N-1)}, D'_N)) \\ & \leq \mathbb{P}(D_{1:(N-1)} \neq D'_{1:(N-1)}) + \mathbb{P}(D_N \neq D'_N | D_{1:(N-1)} = D'_{1:(N-1)}, D_{1:(N-1)} \in \mathcal{D}_{n-1}^*) + \mathbb{P}(D_{1:(N-1)} \notin \mathcal{D}_{n-1}^*). \end{aligned}$$

Apply the expansion once more to  $\mathbb{P}(D_{1:(N-1)} \neq D'_{1:(N-1)})$ , then to  $\mathbb{P}(D_{1:(N-2)} \neq D'_{1:(N-2)})$ . If we define:

$$B_j = \mathbb{P}(D_j \neq D'_j | D_{1:(j-1)} = D'_{1:(j-1)}, D_{1:(j-1)} \in \mathcal{D}_{j-1}^*),$$

with the special case  $B_1$  simply being  $\mathbb{P}(D_1 \neq D'_1)$ , then:

$$\mathbb{P}(D \neq D') \leq \sum_{j=1}^N B_j + \sum_{j=2}^N \mathbb{P}(D_{1:(j-1)} \notin \mathcal{D}_{j-1}^*). \quad (\text{F.6})$$

The second summation in Eq. (F.6), comprising of only atypical probabilities, is easier to control. For any  $j$ , since  $\sum_{i=1}^{j-1} \sum_{x=1}^{\infty} D_{i,x} \leq \sum_{i=1}^N \sum_{x=1}^{\infty} D_{i,x}$ ,  $\mathbb{P}(D_{1:(j-1)} \notin \mathcal{D}_{j-1}^*) \leq \mathbb{P}(D \notin \mathcal{D}^*)$ , so a generous upper bound for the contribution of all the atypical probabilities

including the first one from Eq. (F.4) is:

$$\begin{aligned} \mathbb{P}(D \notin \mathcal{D}^*) + \sum_{j=2}^N \mathbb{P}(D_{1:(j-1)} \notin \mathcal{D}_{j-1}^*) \\ \leq \exp\left(-\left(\frac{\beta^2}{2(\beta+1)} \max(C(K, C_1), C(N, C_1)) - \ln N\right)\right). \end{aligned}$$

By Lemma D.9,  $\max(C(K, C_1), C(N, C_1)) \geq C_1(\max(\ln N, \ln K) - C_1(\psi(C_1) + 1))$ . Since we have set  $\beta$  so that  $\frac{\beta^2}{\beta+1}C_1 = 2$ , we have:

$$\frac{\beta^2}{2(\beta+1)} \max(C(K, C_1), C(N, C_1)) - \ln N \geq \ln K - \text{constant}.$$

meaning the overall atypical probabilities is at most:

$$\mathbb{P}(D \notin \mathcal{D}^*) + \sum_{j=2}^N \mathbb{P}(D_{1:(j-1)} \notin \mathcal{D}_{j-1}^*) \leq \frac{\text{constant}}{K}. \quad (\text{F.7})$$

As for the first summation in Eq. (F.6), we look at the individual  $B_j$ 's. For any fixed  $d_{1:(j-1)} \in \mathcal{D}_{j-1}^*$ , we claim that there exists a coupling between the conditionals  $D_j|D_{1:(j-1)} = d_{1:(j-1)}$  and  $D'_j|D'_{1:(j-1)} = d_{1:(j-1)}$  such that  $\mathbb{P}(D_j \neq D'_j|D_{1:(j-1)} = D'_{1:(j-1)} = d_{1:(j-1)})$  is at most:

$$\frac{\text{constant}}{K} \frac{1}{(j-1+C_1)^2} + \text{constant} \frac{(\ln N + \ln K)}{K} \frac{1}{j-1+C_1}. \quad (\text{F.8})$$

Because the upper bound hold for arbitrary values  $d_{1:(j-1)}$ , the coupling actually ensures that, as long as  $D_{1:(j-1)} = D'_{1:(j-1)}$  for some value in  $\mathcal{D}_{j-1}^*$ , the probability of difference between  $D_j$  and  $D'_j$  is small i.e.  $B_j$  is at most the right hand side.

Such a coupling exists because the total variation between the two distributions  $P_{D_j|D_{1:(j-1)}=d_{1:(j-1)}}$  and  $P_{D'_j|D'_{1:(j-1)}=d_{1:(j-1)}}$  is small. In particular, there exists a distribution  $U = \{U_x\}_{x=1}^\infty$  of independent Poisson random variables, such that both the total variation between  $P_{D_j|D_{1:(j-1)}=d_{1:(j-1)}}$  and  $P_U$  and the total variation between  $P_{D'_j|D'_{1:(j-1)}=d_{1:(j-1)}}$  and  $P_U$  is small – we then use triangle inequality to bound the original total variation. Here, each  $U_x$  has mean:

$$\mathbb{E}(U_x) = \left(K - \sum_{i=1}^{j-1} \sum_{y=1}^\infty d_{i,y}\right) \tilde{h}_c(x|x_{1:(j-1)} = 0).$$

On the one hand, conditioned on  $D'_{1:(j-1)} = d_{1:(j-1)}$ ,  $D'_j = \{D'_{j,x}\}_{x=1}^\infty$  is the joint distribution of types of successes of type  $x$ , where there are  $K - \sum_{i=1}^{j-1} \sum_{x=1}^\infty d_{i,x}$  independent trials and types  $x$  success has probability  $\tilde{h}_c(x|x_{1:(j-1)} = 0)$  by Proposition C.2. Because of Lemma D.4 and Eq. (8):

$$\begin{aligned} d_{TV}\left(P_{D'_j|D'_{1:(j-1)}=d_{1:(j-1)}}, P_U\right) &\leq \left(K - \sum_{i=1}^{j-1} \sum_{y=1}^\infty d_{i,y}\right) \left(\sum_{x=1}^\infty \tilde{h}_c(x|x_{1:(j-1)} = 0)\right)^2 \\ &\leq K \left(\frac{1}{K} \frac{C_1}{j-1+C_1}\right)^2 \\ &\leq \frac{C_1^2}{K} \frac{1}{(j-1+C_1)^2}. \end{aligned} \quad (\text{F.9})$$

On the other hand, conditioned on  $D_{1:(j-1)}$ ,  $D_j = \{D_{j,x}\}_{x=1}^\infty$  consists of independent Poissons, where the mean of  $D_{j,x}$  is  $M_{j,x}$  by Proposition C.1. We recursively apply Lemma D.8 and Lemma D.5:

$$\begin{aligned}
& d_{TV}(P_U, P_{D_j}) \\
& \leq \sum_{x=1}^\infty d_{TV}(P_{U_x}, P_{D_{j,x}}) \\
& \leq \sum_{x=1}^\infty \left| M_{j,x} - \left( K - \sum_{i=1}^{j-1} \sum_{y=1}^\infty d_{i,y} \right) \tilde{h}_c(x|x_{1:(j-1)} = 0) \right| \\
& \leq \sum_{x=1}^\infty \left( |M_{j,x} - K \tilde{h}_c(x|x_{1:(j-1)} = 0)| + \sum_{i=1}^{j-1} \sum_{y=1}^\infty d_{i,y} \tilde{h}_c(x|x_{1:(j-1)} = 0) \right) \\
& \leq \sum_{x=1}^\infty |M_{j,x} - K \tilde{h}_c(x|x_{1:(j-1)} = 0)| + \left( \sum_{i=1}^{j-1} \sum_{y=1}^\infty d_{i,y} \right) \left( \sum_{x=1}^\infty \tilde{h}_c(x|x_{1:(j-1)} = 0) \right). \quad (\text{F.10})
\end{aligned}$$

The first term is upper bounded by Eq. (10). Regarding the second term, since we are in the typical set,  $\sum_{i=1}^{j-1} \sum_{y=1}^\infty d_{i,y}$  is upper bounded. Therefore the overall bound on the second term is:

$$(\beta + 1) \max(C(K, C_1), C(N, C_1)) \frac{1}{K} \frac{C_1}{j-1 + C_1}.$$

Combining the two bounds give the bound on  $d_{TV}(P_U, P_{D_j})$ :

$$\begin{aligned}
& \frac{1}{K} \frac{C_4 \ln j + C_5}{j-1 + C_1} + (\beta + 1) \max(C(K, C_1), C(N, C_1)) \frac{1}{K} \frac{C_1}{j-1 + C_1} \\
& \leq \text{constant} \frac{(\ln N + \ln K)}{K} \frac{1}{j-1 + C_1}. \quad (\text{F.11})
\end{aligned}$$

Combining Eqs. (F.9) and (F.11) gives the upper bound in Eq. (F.8). The summation of the right hand side of Eq. (F.8) across  $j$  leads to:

$$\sum_{j=1}^N B_j \leq \frac{\text{constant}}{K} + \text{constant} \frac{(\ln N + \ln K) \ln N}{K}. \quad (\text{F.12})$$

In all, because of Eqs. (F.7) and (F.12), we can couple  $D$  and  $D'$  such that  $\mathbb{P}(D \neq D')$  is at most:

$$\frac{\text{constant}}{K} + \text{constant} \frac{(\ln N + \ln K) \ln N}{K}. \quad (\text{F.13})$$

Aggregating the results from Eqs. (F.4), (F.5) and (F.13), we are done.  $\square$

## F.2. Lower bound

*Proof of Theorem 4.3.* First we mention which probability kernel  $f$  results in the large total variation distance: the pathological  $f$  is the Dirac measure i.e.,  $f(\cdot | X) := \delta_X(\cdot)$ . With this conditional likelihood  $X_n = Y_n$  and  $Z_n = W_n$ , meaning:

$$d_{TV}(P_{N,\infty}, P_{N,K}) = d_{TV}(P_{X_{1:N}}, P_{Z_{1:N}}).$$

Now we discuss why the total variation is lower bounded by the function of  $N$ . Let  $\mathcal{A}$  be the event that there are at least  $\frac{1}{2}C(N, \alpha)$  unique atom locations in among the latent states:

$$\mathcal{A} := \left\{ x_{1:N} : \#\text{unique atom locations} \geq \frac{1}{2}C(N, \alpha) \right\}.$$

The probabilities assigned to this event by the approximate and the target models are very different from each other. On the one hand, since  $K < \frac{\gamma C(N, \alpha)}{2}$ , under  $\text{IFA}_K$ ,  $\mathcal{A}$  has measure zero:

$$\mathbb{P}_{Z_{1:N}}(\mathcal{A}) = 0. \quad (\text{F.14})$$

On the other hand, under beta-Bernoulli, the number of unique atom locations drawn is a Poisson random variable with mean exactly  $\gamma C(N, \alpha)$  – see Proposition C.1 and Example 4.2. The complement of  $\mathcal{A}$  is a lower tail event. By Lemma D.3 with  $\lambda = \gamma C(N, \alpha)$  and  $x = \frac{1}{2}\gamma C(N, \alpha)$ :

$$\mathbb{P}_{X_{1:N}}(\mathcal{A}) \geq 1 - \exp\left(-\frac{\gamma C(N, \alpha)}{8}\right). \quad (\text{F.15})$$

Because of Lemma D.9, we can lower bound  $C(N, \alpha)$  by a multiple of  $\ln N$ :

$$\exp\left(-\frac{\gamma C(N, \alpha)}{8}\right) \leq \exp\left(-\frac{\gamma \alpha \ln N}{8} + \frac{\alpha \gamma (\psi(\alpha) + 1)}{8}\right) = \frac{\text{constant}}{N^{\gamma \alpha / 8}}.$$

We now combine Eqs. (F.14) and (F.15) and recall that total variation is the maximum over probability discrepancies.  $\square$

The proof of Theorem 4.4 relies on the ability to compute a lower bound on the total variation distance between a Binomial distribution and a Poisson distribution.

**Proposition F.1** (Lower bound on total variation between Binomial and Poisson). *For all  $K$ , it is true that:*

$$d_{TV}\left(\text{Poisson}(\gamma), \text{Binom}\left(K, \frac{\gamma/K}{\gamma/K+1}\right)\right) \geq C(\gamma)K\left(\frac{\gamma/K}{\gamma/K+1}\right)^2,$$

where:

$$C(\gamma) = \frac{1}{8} \frac{1}{\gamma + \exp(-1)(\gamma + 1) \max(12\gamma^2, 48\gamma, 28)}.$$

*Proof of Proposition F.1.* We adapt the proof of (Barbour and Hall, 1984, Theorem 2) to our setting. The  $\text{Poisson}(\gamma)$  distribution satisfies the functional equality:

$$\mathbb{E}[\gamma y(Z+1) - Zy(Z)] = 0, \quad (\text{F.16})$$

where  $y$  is any real-valued function and  $Z \sim \text{Poisson}(\gamma)$ .

Denote  $\gamma_K = \frac{\gamma}{\gamma/K+1}$ . For  $m \in \mathbb{N}$ , let

$$x(m) = m \exp\left(-\frac{m^2}{\gamma_K \theta}\right),$$

where  $\theta$  is a constant which will be specified later.  $x(m)$  serves as a test function to lower bound the total variation distance between  $\text{Poisson}(\gamma)$  and  $\text{Binom}(K, \gamma_K/K)$ . Let  $X_i \sim$

$\text{Ber}(\frac{\gamma_K}{K})$ , independently across  $i$  from 1 to  $K$ , and  $W = \sum_{i=1}^K$ . Then  $W \sim \text{Binomial}(K, \gamma_K/K)$ . The following identity is adapted from (Barbour and Hall, 1984, Equation 2.1):

$$\mathbb{E}[\gamma_K x(W+1) - Wx(W)] = \left(\frac{\gamma_K}{K}\right)^2 \sum_{i=1}^K \mathbb{E}[x(W_i+2) - x(W_i+1)]. \quad (\text{F.17})$$

where  $W_i = W - X_i$ .

We first argue that the right hand side is not too small i.e. for any  $i$ :

$$\mathbb{E}[x(W_i+2) - x(W_i+1)] \geq 1 - \frac{3\gamma_K^2 + 12\gamma_K + 7}{\theta\gamma_K}. \quad (\text{F.18})$$

Consider the derivative of  $x(m)$ :

$$\frac{d}{dm}x(m) = \exp\left(-\frac{m^2}{\gamma_K\theta}\right) \left(1 - \frac{2m^2}{\gamma_K\theta}\right) \geq 1 - \frac{3m^2}{\theta\gamma_K}.$$

because of the easy-to-verify inequality  $e^{-x}(1-2x) \geq 1-3x$  for  $x \geq 0$ . This means:

$$x(W_i+2) - x(W_i+1) \geq \int_{W_i+1}^{W_i+2} \left(1 - \frac{3m^2}{\theta\gamma_K}\right) dm = 1 - \frac{1}{\theta\gamma_K}(3W_i^2 + 9W_i + 7).$$

Taking expectations, noting that  $\mathbb{E}(W_i) \leq \gamma_K$  and  $\mathbb{E}(W_i^2) = \text{Var}(W_i) + [\mathbb{E}(W_i)]^2 \leq \sum_{j=1}^K \frac{\gamma_K}{K} + (\gamma_K)^2 = \gamma_K^2 + \gamma_K$  we have proven Eq. (F.18).

Now, because of positivity of  $x$ , and that  $\gamma \geq \gamma_K$ , we trivially have:

$$\mathbb{E}[\gamma x(W+1) - Wx(W)] \geq \mathbb{E}[\gamma_K x(W+1) - Wx(W)]. \quad (\text{F.19})$$

Combining Eq. (F.17), Eq. (F.18) and Eq. (F.19) we have:

$$\mathbb{E}[\gamma x(W+1) - Wx(W)] \geq K \left(\frac{\gamma_K}{K}\right)^2 \left(1 - \frac{3\gamma_K^2 + 12\gamma_K + 7}{\theta\gamma_K}\right).$$

Recalling Eq. (F.16), for any coupling  $(W, Z)$  such that  $W \sim \text{Binom}\left(K, \frac{\gamma/K}{\gamma/K+1}\right)$  and  $Z \sim \text{Poisson}(\gamma)$ :

$$\mathbb{E}[\gamma(x(W+1) - x(Z+1)) + Zx(Z) - Wx(W)] \geq \frac{\gamma_K^2}{K} \left(1 - \frac{3\gamma_K^2 + 12\gamma_K + 7}{\theta\gamma_K}\right).$$

Suppose  $(W, Z)$  is the maximal coupling attaining the total variation distance between  $P_W$  and  $P_Z$  i.e.  $\mathbb{P}(W \neq Z) = d_{TV}(P_Y, P_Z)$ . Clearly:

$$\begin{aligned} & \gamma(x(W+1) - x(Z+1)) + Zx(Z) - Wx(W) \\ & \leq \mathbf{1}\{W \neq Z\} \sup_{m_1, m_2} |(\gamma x(m_1+1) - m_1 x(m_1)) - (\gamma x(m_2+1) - m_2 x(m_2))| \\ & \leq 2\mathbf{1}\{W \neq Z\} \sup_m |(\gamma x(m+1) - mx(m))|. \end{aligned}$$

Taking expectations on both sides, we conclude that

$$2d_{TV}(P_W, P_Z) \times \sup_m |\gamma x(m+1) - mx(m)| \geq \frac{\gamma_K^2}{K} \left(1 - \frac{3\gamma_K^2 + 12\gamma_K + 7}{\theta\gamma_K}\right) \quad (\text{F.20})$$

It remains to upper bound  $\sup_m |\gamma x(m+1) - mx(m)|$ . Recall that the derivative of  $x$  is  $\exp\left(-\frac{m^2}{\gamma_K \theta}\right) \left(1 - \frac{2m^2}{\gamma_K \theta}\right)$ , taking values in  $[-2e^{-3/2}, 1]$ . This means for any  $m$ ,  $-2e^{-3/2} \leq x(m+1) - x(m) \leq 1$ . Hence:

$$\begin{aligned} |\gamma x(m+1) - mx(m)| &= |\gamma(x(m+1) - x(m)) + (\gamma - m)x(m)| \\ &\leq \gamma + (m + \gamma)m \exp\left(-\frac{m^2}{\gamma_K \theta}\right) \\ &\leq \gamma + (\gamma + 1)m^2 \exp\left(-\frac{m^2}{\gamma_K \theta}\right) \\ &\leq \gamma + \theta \gamma_K (\gamma + 1) \exp(-1). \end{aligned} \quad (\text{F.21})$$

where the last inequality owes to the easy-to-verify  $x \exp(-x) \leq \exp(-1)$ . Combining Eq. (F.21) and Eq. (F.20) we have that:

$$d_{TV} \left( \text{Binomial} \left( K, \frac{\gamma/K}{\gamma/K + 1} \right), \text{Poisson}(\gamma) \right) \geq \frac{1}{2} \frac{1 - \frac{3\gamma_K^2 + 12\gamma_K + 7}{\theta \gamma_K}}{\gamma + (\gamma + 1)\theta \gamma_K \exp(-1)} K \left( \frac{\gamma_K}{K} \right)^2.$$

Finally, we calibrate  $\theta$ . By selecting  $\theta = \max\left(12\gamma_K, \frac{28}{\gamma_K}, 48\right)$  we have that the numerator of the unwieldy fraction is at least  $\frac{1}{4}$  and its denominator is at most  $\gamma + \exp(-1)(\gamma + 1) \max(12\gamma^2, 48\gamma, 28)$ , because  $\gamma_K < \gamma$ . This completes the proof.  $\square$

*Proof of Theorem 4.4.* First we mention which probability kernel  $f$  results in the large total variation distance. For any discrete measure  $\sum_{i=1}^M \delta_{\psi_i}$ ,  $f$  is the Dirac measure sitting on  $M$ , the number of atoms.

$$f(\cdot | \sum_{i=1}^M \delta_{\psi_i}) := \delta_M(\cdot). \quad (\text{F.22})$$

Now we show that under such  $f$ , the total variation distance is lower bounded. First, observe that:

$$d_{TV}(P_{Y_{1:N}}, P_{W_{1:N}}) \geq d_{TV}(P_{Y_1}, P_{W_1}). \quad (\text{F.23})$$

Truly, suppose  $(Y_{1:N}, W_{1:N})$  is any coupling of  $P_{Y_{1:N}}, P_{W_{1:N}}$ . Elementarily we have  $P(Y_{1:N} \neq W_{1:N}) \geq P(Y_1 \neq W_1)$ . Taking the infimum over couplings to attain the total variation distance, we have shown Eq. (F.23). Hence it suffices to show:

$$d_{TV}(P_{Y_1}, P_{W_1}) \geq C(\gamma) \frac{\gamma^2}{K} \frac{1}{(1 + \gamma/K)^2}.$$

Recall the generative process defining  $P_{Y_1}$  and  $P_{W_1}$ .  $Y_1$  is an observation from the target Beta-Bernoulli model, so by Proposition C.1

$$N_T \sim \text{Poisson}(\gamma), \quad \psi_k \stackrel{iid}{\sim} H, \quad X_1 = \sum_{i=1}^{N_T} \delta_{\psi_k}, \quad Y_1 \sim f(\cdot | X_1).$$

$W_1$  is an observation from the approximate model, so by Proposition C.2

$$N_A \sim \text{Binom} \left( K, \frac{\gamma/K}{1 + \gamma/K} \right), \quad \phi_k \stackrel{iid}{\sim} H, \quad Z_1 = \sum_{i=1}^{N_A} \delta_{\phi_k}, \quad W_1 \sim f(\cdot | Z_1).$$

Because of the choice of  $f$ ,  $Y_1 = N_T$  and  $W_1 = N_A$ . Hence, by Proposition F.1:

$$\begin{aligned} d_{TV}(P_{Y_1}, P_{W_1}) &= d_{TV}(P_{N_T}, P_{N_A}) \\ &\geq C(\gamma) \frac{\gamma^2}{K} \frac{1}{(1 + \gamma/K)^2}. \end{aligned}$$

□

## Appendix G: Proofs of DP bounds

Our technique to analyze the error made by  $\text{FSD}_K$  follows a similar vein to the technique in Appendix F. We compare the joint distribution of the latents  $X_{1:N}$  and  $Z_{1:N}$  (with the underlying  $\Theta$  or  $\Theta_K$  marginalized out) using the conditional distributions  $X_n | X_{1:(n-1)}$  and  $Z_n | Z_{1:(n-1)}$ . Before going into the proofs, we give the form of the conditionals.

The conditional  $X_{1:N} | X_{1:(n-1)}$  is the well-known Blackwell-MacQueen prediction rule.

**Proposition G.1.** *Blackwell and MacQueen (1973)* For  $n = 1$ ,  $X_1 \sim H$ . For  $n \geq 2$ :

$$X_n | X_{n-1}, X_{n-2}, \dots, X_1 \sim \frac{\alpha}{n-1+\alpha} H + \sum_j \frac{n_j}{n-1+\alpha} \delta_{\psi_j}.$$

where  $\{\psi_j\}$  is the set of unique values among  $X_{n-1}, X_{n-2}, \dots, X_1$  and  $n_j$  is the cardinality of the set  $\{i : 1 \leq i \leq n-1, X_i = \psi_j\}$ .

The conditionals  $Z_n | Z_{1:(n-1)}$  are related to the Blackwell-MacQueen prediction rule.

**Proposition G.2.** *Pitman (1996)* For  $n = 1$ ,  $Z_1 \sim H$ . For  $n \geq 2$ , let  $\{\psi_j\}_{j=1}^{J_n}$  be the set of unique values among  $Z_{n-1}, Z_{n-2}, \dots, Z_1$  and  $n_j$  is the cardinality of the set  $\{i : 1 \leq i \leq n-1, Z_i = \psi_j\}$ . If  $J_n < K$ :

$$Z_n | Z_{n-1}, Z_{n-2}, \dots, Z_1 \sim \frac{(K - J_n)\alpha/K}{n-1+\alpha} H + \sum_{j=1}^{J_n} \frac{n_j + \alpha/K}{n-1+\alpha} \delta_{\psi_j},$$

Otherwise, if  $J_n = K$ , there is zero probability of drawing a fresh component from  $H$  i.e.  $Z_n$  comes only from  $\{\psi_j\}_{j=1}^{J_n}$ :

$$Z_n | Z_{n-1}, Z_{n-2}, \dots, Z_1 \sim \sum_{j=1}^{J_n} \frac{n_j + \alpha/K}{n-1+\alpha} \delta_{\psi_j},$$

$J_n \leq K$  is an invariant of these of prediction rules: once  $J_n = K$ , all subsequent  $J_m$  for  $m \geq n$  is also equal to  $K$ .

### G.1. Upper bounds

*Proof of Theorem 5.1.* First, because of Lemma D.7, it suffices to show that  $d_{TV}(P_{X_{1:N}}, P_{Z_{1:N}})$  is small, since the conditional distributions of the observations given the latent variables are the same across target and approximate models.

To show that  $d_{TV}(P_{X_{1:N}}, P_{Z_{1:N}})$  is small, we will construct a coupling of  $X_{1:N}$  and  $Z_{1:N}$  such that for any  $n \geq 1$ :

$$\mathbb{P}(X_n \neq Z_n | X_{1:(n-1)} = Z_{1:(n-1)}) \leq 2 \frac{\alpha}{K} \frac{J_n}{n-1+\alpha}, \quad (\text{G.1})$$

where  $J_n$  is the number of unique atom locations among  $X_{1:(n-1)}$ . Such a coupling exists because the total variation distance between the prediction rules  $X_n | X_{1:(n-1)}$  and  $Z_n | Z_{1:(n-1)}$  is small: as total variation is the minimum difference probability, there exists a coupling that achieves the total variation distance. Consider any measurable set  $A$ . If  $J_n < K$ , the probability of  $A$  under the two rules are respectively:

$$\begin{aligned} & \frac{\alpha(1 - J_n/K)}{n-1+\alpha} H(A) + \sum_{j=1}^{J_n} \frac{n_j + \alpha/K}{n-1+\alpha} \delta_{\psi_j}(A) \\ & \frac{\alpha}{n-1+\alpha} H(A) + \sum_{j=1}^{J_n} \frac{n_j}{n-1+\alpha} \delta_{\psi_j}(A) \end{aligned}$$

meaning the absolute difference in probability mass is:

$$\begin{aligned} \left| \frac{\alpha}{K} \frac{J_n H(A)}{n-1+\alpha} - \frac{\alpha}{K} \sum_{j=1}^{J_n} \frac{\delta_j(A)}{n-1+\alpha} \right| & \leq \left| \frac{\alpha}{K} \frac{J_n H(A)}{n-1+\alpha} \right| + \left| \frac{\alpha}{K} \sum_{j=1}^{J_n} \frac{\delta_j(A)}{n-1+\alpha} \right| \\ & \leq \frac{\alpha}{K} \frac{J_n}{n-1+\alpha} + \frac{\alpha}{K} \frac{J_n}{n-1+\alpha} \\ & = 2 \frac{\alpha}{K} \frac{J_n}{n-1+\alpha}. \end{aligned}$$

The same upper bound holds for the case  $J_n = K$ . The couplings for different  $n$  are naturally glued together because of the recursive nature of the conditional distributions.

We now show that for the coupling satisfying Eq. (G.1), the overall probability of difference  $\mathbb{P}(X_{1:N} \neq Z_{1:N})$  is small. Define the short hand:

$$C(N, \alpha) := \sum_{n=1}^N \frac{\alpha}{n-1+\alpha}.$$

The definition of the typical set depends on the relative deviation  $\delta$ , which we calibrate at the end of the proof. Define the *typical* set:

$$\mathcal{D}_n := \{x_{1:(n-1)} : J_n \leq (1 + \delta) \max(C(N-1, \alpha), C(K, \alpha))\}.$$

In other words, the number of unique values among the  $x_{1:(n-1)}$  is small. The following decomposition is used to investigate the difference probability on the typical set:

$$\begin{aligned} \mathbb{P}(X_{1:N} \neq Z_{1:N}) & = \mathbb{P}((X_{1:(N-1)}, X_N) \neq (Z_{1:(N-1)}, Z_N)) \\ & = \mathbb{P}(X_{1:(N-1)} \neq Z_{1:(N-1)}) + \mathbb{P}(X_N \neq Z_N, X_{1:(N-1)} = Z_{1:(N-1)}) \quad (\text{G.2}) \end{aligned}$$

The second term can be further expanded:

$$\begin{aligned} & \mathbb{P}(X_N \neq Z_N, X_{1:(N-1)} = Z_{1:(N-1)}, X_{1:(N-1)} \in \mathcal{D}_N) \\ & + \mathbb{P}(X_N \neq Z_N, X_{1:(N-1)} = Z_{1:(N-1)}, X_{1:(N-1)} \notin \mathcal{D}_N) \end{aligned}$$

The former term is at most:

$$\mathbb{P}(X_N \neq Z_N | X_{1:(N-1)} = Z_{1:(N-1)}, X_{1:(N-1)} \in \mathcal{D}_N),$$

while the latter term is at most:

$$\mathbb{P}(X_{1:(N-1)} \notin \mathcal{D}_N).$$

To recap, we can bound  $\mathbb{P}(X_{1:N} \neq Z_{1:N})$  by bounding three quantities:

1. The difference probability of a shorter process  $\mathbb{P}(X_{1:(N-1)} \neq Z_{1:(N-1)})$ .
2. The difference probability of the prediction rule on typical sets  $\mathbb{P}(X_N \neq Z_N | X_{1:(N-1)} = Z_{1:(N-1)}, X_{1:(N-1)} \in \mathcal{D}_N)$ .
3. The probability of the atypical set  $\mathbb{P}(X_{1:(N-1)} \notin \mathcal{D}_N)$ .

By recursively applying the expansion initiated in Eq. (G.2) to  $\mathbb{P}(X_{1:(N-1)} \neq Z_{1:(N-1)})$ , we actually only need to bound difference probability of the different prediction rules on typical sets and the atypical set probabilities.

Regarding difference probability of the different prediction rules, being in the typical set allows us to control  $J_n$  in Eq. (G.1). Summation across  $n = 1$  through  $N$  gives the overall bound of:

$$2 \frac{\alpha}{K} (1 + \delta) \max(C(N-1, \alpha), C(K, \alpha)) C(N, \alpha) \leq \text{constant} \frac{\ln N (\ln N + \ln K)}{K}. \quad (\text{G.3})$$

Regarding the atypical set probabilities, because  $J_{n-1}$  is stochastically dominated by  $J_n$  i.e., the number of unique values at time  $n$  is at least the number at time  $n-1$ , all the atypical set probabilities are upper bounded by the last one i.e.  $\mathbb{P}(X_{1:(N-1)} \notin \mathcal{D}_N)$ .  $J_{N-1}$  is the sum of independent Poisson trials, with an overall mean equaling exactly  $C(N-1, \alpha)$ . Therefore, the atypical event has small probability because of Lemma D.1:

$$\begin{aligned} & \mathbb{P}(J_{N-1} > (1 + \delta) \max(C(N-1, \alpha), C(K, \alpha))) \\ & \leq \exp\left(-\frac{\delta^2}{2 + \delta} \max(C(N-1, \alpha), C(K, \alpha))\right). \end{aligned}$$

Even accounting for all  $N$  atypical events, the total probability is small:

$$\exp\left(-\left(\frac{\delta^2}{2 + \delta} \max(C(N-1, \alpha), C(K, \alpha)) - \ln(N-1)\right)\right)$$

By Lemma D.9,  $\max(C(N-1, \alpha), C(K-1, \alpha)) \geq \alpha \max(\ln(N-1), \ln K - \alpha(\psi(\alpha) + 1))$ . Therefore, if we set  $\delta$  such that  $\frac{\delta^2}{2+\delta}\alpha = 2$ , we have:

$$\frac{\delta^2}{2 + \delta} \max(C(N-1, \alpha), C(K-1, \alpha)) - \ln(N-1) \geq \ln K - \text{constant}$$

meaning the overall atypical probabilities is at most:

$$\frac{\text{constant}}{K}. \quad (\text{G.4})$$

The overall total variation bound combines Eqs. (G.3) and (G.4).  $\square$

*Proof of Corollary 5.2.* The main idea is reducing to the Dirichlet process mixture model situation. This can be done in two steps.

First, the conditional distribution of the observations  $W|H_{1:D}$  of the target model is the same as the conditional distribution  $Z|F_{1:D}$  of the approximate model if  $H_{1:D} = F_{1:D}$ . Hence to control the total variation between  $P_W$  and  $P_Z$  it suffices to control the total variation between  $P_{H_{1:D}}$  and  $P_{F_{1:D}}$  because of Lemma D.7. Second, the distance between  $P_{H_{1:D}}$  and  $P_{F_{1:D}}$  can be upper bounded by the distance between the atom locations that define  $H_{1:D}$  and  $F_{1:D}$ . Recall the construction of the  $F_d$  in terms of atom locations  $\phi_{d,j}$  and stick-breaking weights  $\gamma_{d,j}$ :

$$\begin{aligned} G_K &\sim \text{FSD}_K(\omega, H) \\ \phi_{dj} | G_K &\stackrel{iid}{\sim} G_K(\cdot) && \text{across } d, j \\ \gamma_{dj} &\stackrel{iid}{\sim} \text{Beta}(1, \alpha) && \text{across } d, j \text{ (except } \gamma_{dT} = 1) \\ F_d | \phi_{d,\cdot}, \gamma_{d,\cdot} &= \sum_{i=1}^T \left( \gamma_{di} \prod_{j<i} (1 - \gamma_{dj}) \right) \delta_{\phi_{dj}}. \end{aligned}$$

Similarly  $H_d$  is also constructed in terms of atom locations  $\lambda_{d,j}$  and stick-breaking weights  $\eta_{d,j}$ :

$$\begin{aligned} G &\sim \text{DP}(\omega, H) \\ \lambda_{dj} | G &\stackrel{iid}{\sim} G(\cdot) && \text{across } d, j \\ \eta_{dj} &\stackrel{iid}{\sim} \text{Beta}(1, \alpha) && \text{across } d, j \text{ (except } \eta_{dT} = 1) \\ H_d | \lambda_{d,\cdot}, \eta_{d,\cdot} &= \sum_{i=1}^T \left( \eta_{di} \prod_{j<i} (1 - \eta_{dj}) \right) \delta_{\lambda_{dj}}. \end{aligned}$$

Let  $\Lambda = \{\lambda_{dj}\}_{d,j}$  and  $\Phi = \{\phi_{dj}\}_{d,j}$ . It is apparent that the conditional distribution  $H_{1:D}|\Lambda$  is the same as the conditional distribution  $F_{1:D}|\Phi$  if  $\Lambda = \Phi$ . Therefore, we only need to control total variation between  $P_\Lambda$  and  $P_\Phi$ , again by Lemma D.7.

Because of Theorem 5.1, we already know how to compare  $P_\Lambda$  and  $P_\Phi$ . On the one hand, since  $\lambda_{dj}$  are conditionally iid given  $G$  across  $d, j$ , the joint distribution of  $\lambda_{dj}$  is from a DPMM (probability kernel  $f$  being Dirac  $f(\cdot|x) = \delta_x(\cdot)$ ) where the underlying DP has concentration  $\omega$ . On the other hand, since  $\phi_{dj}$  are conditionally iid given  $G_K$  across  $d, j$ , the joint distribution  $\phi_{dj}$  comes from the finite mixture with  $\text{FSD}_K$ . Each observational has cardinality  $DT$ . Therefore:

$$d_{TV}(P_\Lambda, P_\Phi) \leq \frac{C_1 + C_2 \ln^2(DT) + C_3 \ln(DT) \ln K}{K},$$

where the constants  $C_i$  only depend on  $\omega$ . □

## G.2. Lower bounds

*Proof of Theorem 5.3.* First we mention which probability kernel  $f$  results in the large total variation distance: the pathological  $f$  is the Dirac measure i.e.,  $f(\cdot|x) = \delta_x(\cdot)$ . With this

conditional likelihood  $X_n = Y_n$  and  $Z_n = W_n$ , meaning:

$$d_{TV}(P_{N,\infty}, P_{N,K}) = d_{TV}(P_{X_{1:N}}, P_{Z_{1:N}}).$$

Now we discuss why the total variation is lower bounded by the function of  $N$ . Let  $\mathcal{A}$  be the event that there are at least  $\frac{1}{2}C(N, \alpha)$  unique components in among the latent states:

$$\mathcal{A} := \left\{ x_{1:N} : \#\text{unique values} \geq \frac{1}{2}C(N, \alpha) \right\}.$$

The probabilities assigned to this event by the approximate and the target models are very different from each other. On the one hand, since  $K < \frac{C(N, \alpha)}{2}$ , under  $\text{FSD}_K$ ,  $\mathcal{A}$  has measure zero:

$$\mathbb{P}_{Z_{1:N}}(\mathcal{A}) = 0. \quad (\text{G.5})$$

On the other hand, under DP, the number of unique atoms drawn is the sum of Poisson trials with expectation exactly  $C(N, \alpha)$ . The complement of  $\mathcal{A}$  is a lower tail event. Hence by Lemma D.2 with  $\delta = 1/2, \mu = C(N, \alpha)$ , we have:

$$\mathbb{P}_{X_{1:N}}(\mathcal{A}) \geq 1 - \exp\left(-\frac{C(N, \alpha)}{8}\right) \quad (\text{G.6})$$

Because of Lemma D.9, we can lower bound  $C(N, \alpha)$  by a multiple of  $\ln N$ :

$$\exp\left(-\frac{C(N, \alpha)}{8}\right) \leq \exp\left(-\frac{\alpha \ln N}{8} + \frac{\alpha(\psi(\alpha) + 1)}{8}\right) = \frac{\text{constant}}{N^{\alpha/8}}.$$

We now combine Eqs. (G.5) and (G.6) and recall that total variation is the maximum over probability discrepancies.  $\square$

*Proof of Theorem 5.4.* First we mention which probability kernel  $f$  results in the large total variation distance: the pathological  $f$  is the Dirac measure i.e.,  $f(\cdot | x) = \delta_x(\cdot)$ .

Now we show that under such  $f$ , the total variation distance is lower bounded. Observe that it suffices to understand the total variation between  $P_{Y_1, Y_2}$  and  $P_{W_1, W_2}$ . Truly, suppose  $(Y_{1:N}, W_{1:N})$  is any coupling of  $P_{Y_{1:N}}$  and  $P_{W_{1:N}}$ . Elementarily we have  $P(Y_{1:N} \neq W_{1:N}) \geq P((Y_1, Y_2) \neq (W_1, W_2))$ . Taking the infimum, we have:

$$d_{TV}(P_{N,\infty}, P_{N,K}) \geq d_{TV}(P_{Y_1, Y_2}, P_{W_1, W_2}).$$

Since  $f$  is Dirac,  $X_n = Y_n$  and  $Z_n = W_n$  and we have:

$$d_{TV}(P_{Y_1, Y_2}, P_{W_1, W_2}) = d_{TV}(P_{X_1, X_2}, P_{Z_1, Z_2}).$$

Now, let  $(X_1, X_2), (Z_1, Z_2)$  be any coupling of  $P_{X_1, X_2}$  and  $P_{Z_1, Z_2}$ . We have:

$$\begin{aligned} \mathbb{P}((X_1, X_2) \neq (Z_1, Z_2)) &= \mathbb{P}(X_2 \neq Z_2 | X_1 = Z_1) + \mathbb{P}(X_1 \neq Z_1) \mathbb{P}(X_2 = Z_2 | X_1 = Z_1) \\ &\geq \mathbb{P}(X_2 \neq Z_1 | X_1 = Z_2). \end{aligned}$$

We now investigate how small  $\mathbb{P}(X_2 \neq Z_2 | X_1 = Z_2)$  can be. In the conditioning  $X_1 = Z_1$ , let the common atom be  $\psi_1$ . The prediction rule  $X_2 | X_1 = \psi_1$  puts mass  $\frac{1}{1+\alpha}$  on  $\psi_1$  while the

prediction rule  $Z_2|Z_1 = \psi_1$  puts mass  $\frac{1+\alpha/K}{1+\alpha}$ . This means that the total variation distance between the two prediction rules is at least:

$$\frac{1 + \alpha/K}{1 + \alpha} - \frac{1}{1 + \alpha} = \frac{\alpha}{1 + \alpha} \frac{1}{K}.$$

Since the minimum difference probability is at least the total variation distance, we conclude that for any coupling  $(X_1, X_2), (Z_1, Z_2)$

$$\mathbb{P}(X_2 \neq Z_2 | X_1 = Z_1) \geq \frac{\alpha}{1 + \alpha} \frac{1}{K}.$$

Hence we have a lower bound on  $\mathbb{P}((X_1, X_2) \neq (Z_1, Z_2))$  itself. As the coupling was arbitrary, we take the infimum to attain the lower bound on total variation.  $\square$

## Appendix H: Experimental setup

### H.1. Image denoising

The experiments in this section aim to isolate the effect of TFA versus IFA, by fitting different approximations of the beta-Bernoulli model to denoise<sup>4</sup> an image. We give a description of our models and their hyper-parameter settings. Each patch  $x_i$  is flattened into a vector in  $\mathbb{R}^n$ . Let  $\mathbf{I}_n$  be the  $n \times n$  identity matrix, and similarly for  $\mathbf{I}_K$ . The base measure generating the basis elements is the same:

$$\psi_k \stackrel{iid}{\sim} \mathcal{N}(0, n^{-1}\mathbf{I}_n) \quad k = 1, 2, \dots, K$$

The observational likelihood conditioned on feature-allocation matrix  $F \in \{0, 1\}^{N \times K}$  and basis elements  $\{\psi_k\}_{k=1}^K$  is the same for both models.

$$\begin{aligned} \gamma_w &\sim \text{Gamma}(10^{-6}, 10^{-6}) \\ \gamma_e &\sim \text{Gamma}(10^{-6}, 10^{-6}) \\ w_i &\stackrel{iid}{\sim} \mathcal{N}(0, \gamma_w^{-1}\mathbf{I}_K) \quad i = 1, 2, \dots, N \\ \epsilon_i &\stackrel{iid}{\sim} \mathcal{N}(0, \gamma_e^{-1}\mathbf{I}_n) \quad i = 1, 2, \dots, N \\ x_i &= \sum_{k=1}^K F_{i,k} w_{i,k} \psi_k + \epsilon_i \quad i = 1, 2, \dots, N \end{aligned} \tag{H.1}$$

where we are using the shape-rate parametrization of the gamma. Finally, how the feature-allocation matrix  $F$  is generated is the sole difference between TFA and IFA. The underlying beta process being approximated has rate measure  $\nu(\theta) = \theta^{-1} \mathbf{1}\{\theta \leq 1\}$ .

- TFA:

$$\begin{aligned} v_k &\stackrel{iid}{\sim} \text{Beta}(1, 1) \\ \pi_k &= \prod_{i=1}^k v_i, \quad k = 1, 2, \dots, K \\ F_{i,k} | \pi_k &\stackrel{indep}{\sim} \text{Ber}(\pi_k) \quad i = 1, 2, \dots, N \end{aligned}$$

<sup>4</sup>The posterior over (trait, frequency) and per-observation allocation is traversed for a certain number of steps using a Gibbs sampler. Each visited dictionary and assignment is used to compute each patch's mean value: the candidate output pixel value is the mean over patches covering that pixel. We aggregate the output images across Gibbs steps by a weighted averaging mechanism.

- IFA:

$$\pi_k \stackrel{iid}{\sim} \text{Beta}\left(\frac{1}{K}, 1\right) \quad k = 1, 2, \dots, K$$

$$F_{i,k} | \pi_k \stackrel{indep}{\sim} \text{Ber}(\pi_k) \quad i = 1, 2, \dots, N$$

In Eq. (H.1), we are enriching the basic feature-allocation structure by introducing weights  $w_{i,k}$  which allow an observation to manifest a non-integer (and potentially negative) scaled version of the basis element. Following (Zhou et al., 2009), we are *uninformative* about the noise precisions by choosing Gamma( $10^{-6}$ ,  $10^{-6}$ ). Regarding the choice of hyper-parameters for the underlying beta process, (Zhou et al., 2009) suggests that the performance of the denoising routine is insensitive to the choice of  $\gamma$  and  $\alpha$ : we picked  $\gamma, \alpha = 1$  for computational convenience, especially since for the beta process for  $\alpha = 1$  admits the simple stick-breaking construction.

## H.2. Topic modelling

Nearly 1m random wikipedia documents were downloaded and processed following (Hoffman, Bach and Blei, 2010).

IFA:

$$G_0 \sim \text{FSD}_K(\omega, \text{Dir}(\eta \mathbf{1}_V))$$

$$G_d \sim \text{T-DP}_T(\alpha, G_0) \quad \text{independently across } d = 1, 2, \dots, D$$

$$\beta_{dn} | G_d \sim G_d(\cdot) \quad \text{independently across } n = 1, 2, \dots, N_d$$

$$w_{dn} | \beta_{dn} \sim \text{Categorical}(\beta_{dn}) \quad \text{independently across } n = 1, 2, \dots, N_d$$

TFA:

$$G_0 \sim \text{T-DP}_K(\omega, \text{Dir}(\eta \mathbf{1}_V))$$

$$G_d \sim \text{T-DP}_T(\alpha, G_0) \quad \text{independently across } d = 1, 2, \dots, D$$

$$\beta_{dn} | G_d \sim G_d(\cdot) \quad \text{independently across } n = 1, 2, \dots, N_d$$

$$w_{dn} | \beta_{dn} \sim \text{Categorical}(\beta_{dn}) \quad \text{independently across } n = 1, 2, \dots, N_d$$

Hyper-parameter settings follow (Wang, Paisley and Blei, 2011) in that  $\eta = 0.01, \alpha = 1.0, \omega = 1.0, T = 20$ .

We approximate the posterior in each model using stochastic variational inference (Hoffman et al., 2013). Both models have nice conditional conjugacies that allow the use of exponential family variational distributions and closed-form expectation equations. Batch size is 500, learning rate parametrized by  $\rho_t = (t + \tau)^{-\kappa}$  where by default  $\tau = 1.0$  and  $\kappa = 0.9$ .

We discuss how held-out log-likelihood is computed. Each held-out document  $d'$  is separated into two parts  $w_{ho}$  and  $w_{obs}$ <sup>5</sup>, with no common words between the two. In our

<sup>5</sup>How each document is separated into these two parts can have an impact on the range of test log-likelihood values encountered. For instance, if the first (in order of appearance in the document)  $x\%$  of words were the observed words and the last  $(100 - x)\%$  words were unseen, then the test log-likelihood is low, presumably since predicting future words using only past words and without any filtering is challenging. Randomly assigning words to be observed and unseen gives better test log-likelihood.

experiments, we set 75% of words to be observed, the remaining 25% unseen. The predictive distribution of each word  $w_{new}$  in the  $w_{ho}$  is exactly equal to:

$$p(w_{new}|\mathcal{D}, w_{obs}) = \int_{\theta_{d'}, \beta} p(w_{new}|\theta_{d'}, \beta)p(\theta_{d'}, \beta|\mathcal{D}, w_{obs})d\theta_{d'}d\beta.$$

This is an intractable computation as the posterior  $p(\theta_{d'}, \beta|\mathcal{D}, w_{obs})$  is not analytical. We approximate it with a factorized distribution:

$$p(\theta_{d'}, \beta|\mathcal{D}, w_{obs}) \approx q(\beta|\mathcal{D})q(\theta_{d'}),$$

where  $q(\beta|\mathcal{D})$  is fixed to be the variational approximation found during training and  $q(\theta_{d'})$  minimizes the KL between the variational distribution and the posterior. Operationally, we do an E-step for the document  $d'$  based on the variational distribution of  $\beta$  and the observed words  $w_{obs}$ , and discard the distribution over  $z_{d', \cdot}$ , the per-word topic assignments because of the mean-field assumption. Using those approximations, the predictive approximation is approximately:

$$p(w_{new}|\mathcal{D}, w_{obs}) \approx \tilde{p}(w_{new}|\mathcal{D}, w_{obs}) = \sum_{k=1}^K \mathbb{E}_q(\theta_{d'}(k))\mathbb{E}_q(\beta_k(w_{new})),$$

and the final number we report for document  $d'$  is:

$$\frac{1}{|w_{ho}|} \sum_{w \in w_{ho}} \log \tilde{p}(w|\mathcal{D}, w_{obs}).$$

## Appendix I: Additional experiments

### I.1. Plane

The results for the plane image are Figs. I.1, I.2 and I.3.

### I.2. Truck

The results for the truck image are Figs. I.4, I.5 and I.6.

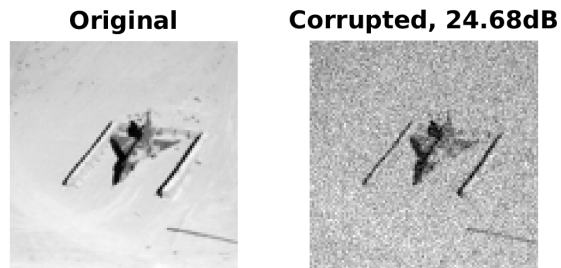


Fig I.1: Original versus corrupted image for plane.

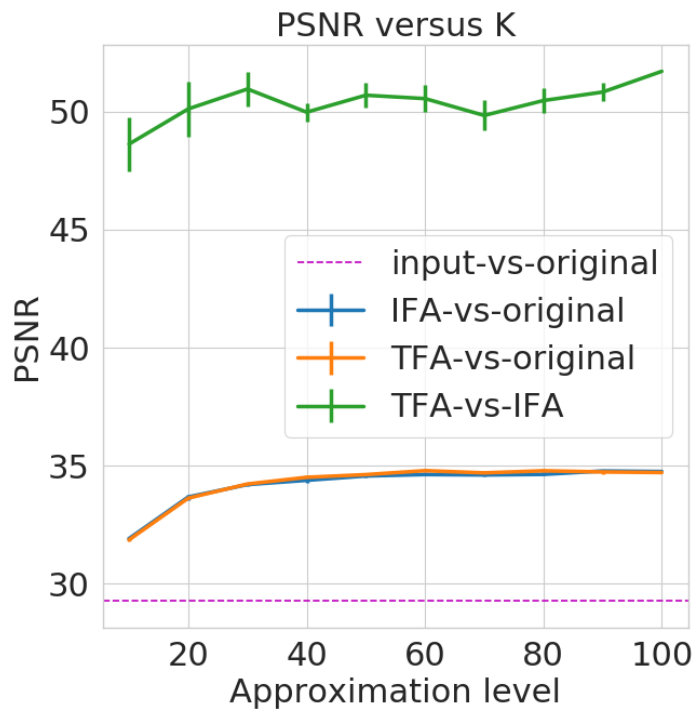


Fig I.2: PSNR versus approximation level for plane

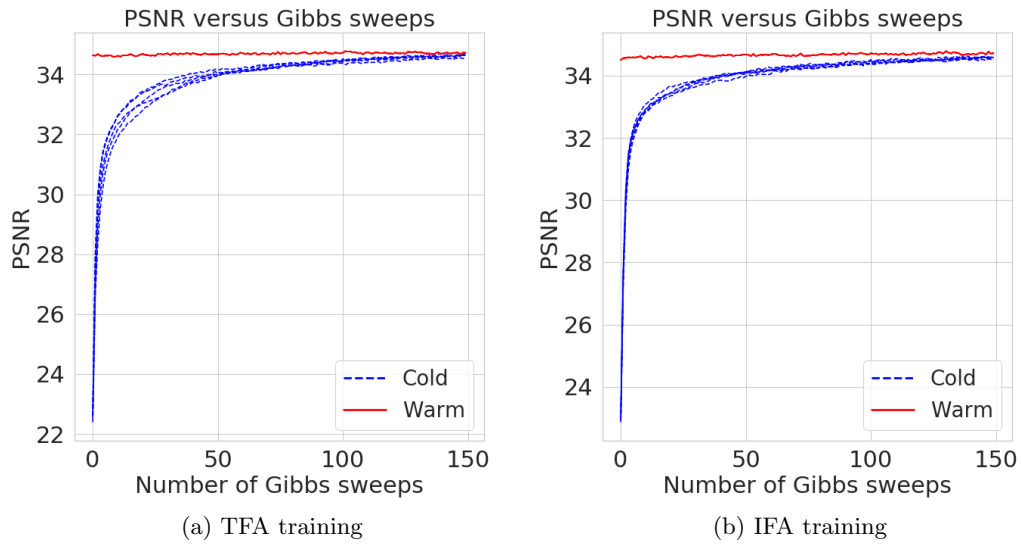


Fig I.3: The output of one model is a good initialization for the training of the other one. Here  $K = 60$ .

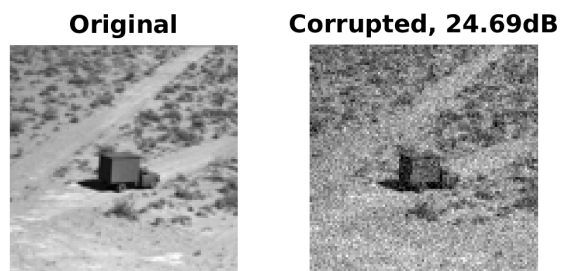


Fig I.4: Original versus corrupted images for truck.

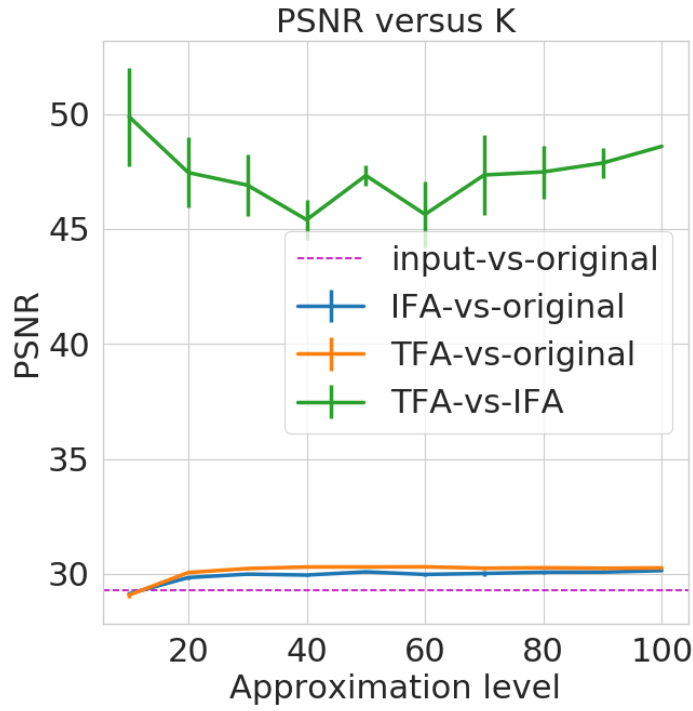
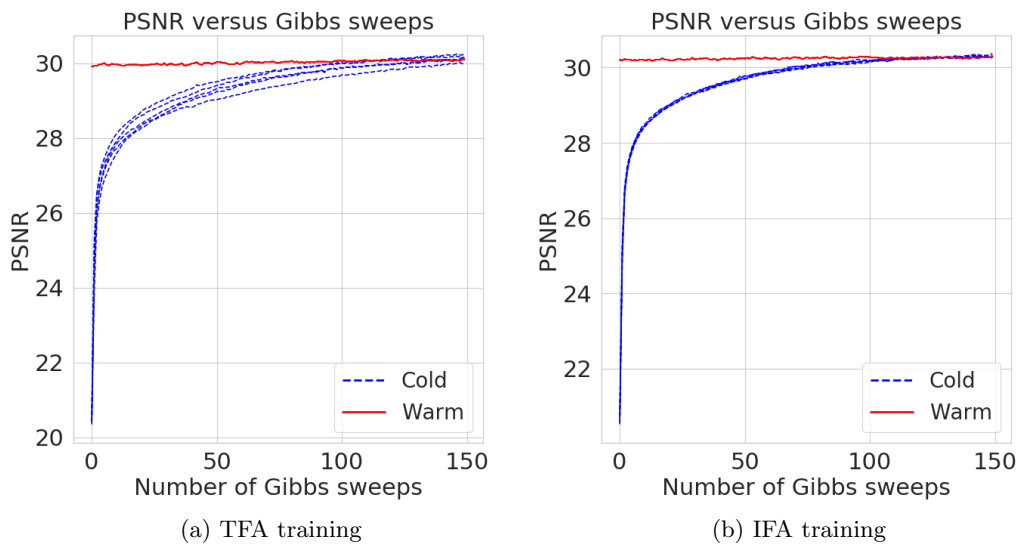


Fig I.5: PSNR versus approximation level for truck.



(a) TFA training

(b) IFA training

Fig I.6: The output of one model is a good initialization for the training of the other one. Here  $K = 60$ .