Bayesian Analysis (2024) **19**, Number 4, pp. 1187–1224

Independent Finite Approximations for Bayesian Nonparametric Inference*

Tin D. Nguyen[†], Jonathan Huggins[‡], Lorenzo Masoero[§], Lester Mackey[¶], Tamara Broderick[‖]

Abstract. Completely random measures (CRMs) and their normalizations (NCRMs) offer flexible models in Bayesian nonparametrics. But their infinite dimensionality presents challenges for inference. Two popular finite approximations are truncated finite approximations (TFAs) and independent finite approximations (IFAs). While the former have been well-studied, IFAs lack similarly general bounds on approximation error, and there has been no systematic comparison between the two options. In the present work, we propose a general recipe to construct practical finite-dimensional approximations for homogeneous CRMs and NCRMs, in the presence or absence of power laws. We call our construction the *automated independent finite approximation* (AIFA). Relative to TFAs, we show that AIFAs facilitate more straightforward derivations and use of parallel computing in approximate inference. We upper bound the approximation error of AIFAs for a wide class of common CRMs and NCRMs — and thereby develop guidelines for choosing the approximation level. Our lower bounds in key cases suggest that our upper bounds are tight. We prove that, for worst-case choices of observation likelihoods, TFAs are more efficient than AIFAs. Conversely, we find that in real-data experiments with standard likelihoods, AIFAs and TFAs perform similarly. Moreover, we demonstrate that AIFAs can be used for hyperparameter estimation even when other potential IFA options struggle or do not apply.

1 Introduction

Many data analysis problems can be seen as discovering a latent set of traits in a population — for example, recovering 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 number of observations. One might choose a prior for different data set sizes, but then

*Tin D. Nguyen, Jonathan Huggins, Lorenzo Masoero, and Tamara Broderick were supported in part by ONR grant N00014-17-1-2072, NSF grant CCF-2029016, ONR MURI grant N00014-11-1-0688, and a Google Faculty Research Award. Jonathan Huggins was also supported by the National Institute of General Medical Sciences of the National Institutes of Health under grant number R01GM144963 as part of the Joint NSF/NIGMS Mathematical Biology Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[†]LIDS, MIT, tdn@mit.edu
[‡]Department of Mathematics & Statistics, Boston University, huggins@bu.edu
[§]LIDS, MIT, lom@mit.edu
[¶]Microsoft Research New England, lmackey@microsoft.com
[‖]LIDS, MIT, tamarab@mit.edu

 https://doi.org/10.1214/23-BA1385

model construction potentially becomes inconvenient and unwieldy. A simpler approach is to choose a single prior that naturally yields different expected numbers of traits for different numbers of data points. In theory, Bayesian nonparametric (BNP) priors have exactly this desirable property due to a countable infinity of traits, so that there are always more traits to reveal through the accumulation of more data.

However, the infinite-dimensional parameter presents a practical challenge; namely, 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; James, 2017; Broderick, Wilson and Jordan, 2018) to circumvent the infinite representation; in particular, these models allow marginalization of the infinite collection of latent traits. These models will typically be part of a more complex generative model where the remaining components are all finite. Therefore, users can apply approximate inference schemes such as Gibbs sampling. However, these marginal forms typically limit the user to a constrained family of models; are not amenable to parallelization; would require substantial new development to use with modern inference engines like NIMBLE (de Valpine et al., 2017); and are not straightforward to use with variational Bayes.

An alternative approach 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 an approximation to the BNP prior. Therefore, its cardinality can be informed directly by the BNP prior and the size of the observed data. Any moderately complex model will necessitate approximate inference, such as Markov chain Monte Carlo (MCMC) or variational Bayes (VB). Therefore, as long as the error due to the finite-dimensional prior approximation is small compared to the error due to using approximate inference, inferential quality is not affected. Unlike marginal representations, probabilistic programming languages like NIMBLE (de Valpine et al., 2017) natively support such finite approximations.

Much of the previous work on finite approximations developed and analyzed truncations of series representations of the random measures underlying the nonparametric prior; we call these *truncated finite approximations* (TFAs) and refer to Campbell et al. (2019) for a thorough study. TFAs start from a sequential ordering of population traits in a random measure. The TFA retains a finite set of approximating traits; these match the population traits until a finite point and do not include terms beyond that (Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012; Roychowdhury and Kulis, 2015; Arbel and Prünster, 2017; Campbell et al., 2019). However, we show in Section 5 that the sequential nature of TFAs makes it difficult to derive update steps in an approximate inference algorithm (either MCMC or VB) and is not amenable to parallelization.

Here, we instead develop and analyze a general-purpose finite approximation consisting of independent and identically distributed (i.i.d.) representations of the traits together with their rates within the population; we call these *independent finite approximations* (IFAs). At the time of writing, we are aware of two alternative lines of work on generic constructions of finite approximations using i.i.d. random variables, namely Lijoi, Prünster and Rigon (2023) and Lee, Miscouridou and Caron (2022); Lee, James

and Choi (2016). Lijoi, Prünster and Rigon (2023) design approximations for clustering models, characterize the posterior predictive distribution, and derive tractable inference schemes. However, the authors have not developed their method for trait allocations, where data points can potentially belong to multiple traits and can potentially exhibit traits in different amounts. And in particular it would require additional development to perform inference in trait allocation models using their approximations.[1] Lee, Miscouridou and Caron (2022); Lee, James and Choi (2016) construct finite approximations through a novel augmentation scheme. However, Lee, Miscouridou and Caron (2022); Lee, James and Choi (2016) lack explicit constructions in important situations, such as exponential-family rate measures, because the functions involved in the augmentation are, in general, only implicitly defined. When the augmentation is implicit, there is not currently a way to evaluate (up to proportionality constant) the probability density of the finite-dimensional distribution; therefore standard Markov chain Monte Carlo and variational approaches for approximate inference are unavailable.

Our contributions. We propose a general-purpose construction for IFAs that subsumes a number of special cases that have already been successfully used in applications (Section 3.1). We call our construction the *automated independent finite approximation*, or AIFA. We show that AIFAs can handle a wide variety of models — including homogeneous completely random measures (CRMs) and normalized CRMs (NCRMs) (Section 3.3).[2] Our construction can handle (N)CRMs exhibiting power laws and has an especially convenient form for exponential family CRMs (Section 3.2). We show that our construction works for useful CRMs not previously seen in the BNP literature (Example 3.4). Unlike marginal representations, AIFAs do not require conditional conjugacy and can be used with VB. We show that, unlike TFAs, AIFAs facilitate straightforward derivations within approximate inference schemes such as MCMC or VB and are amenable to parallelization during inference (Section 5). In existing special cases, practitioners report similar predictive performance between AIFAs and TFAs (Kurihara, Welling and Teh, 2007) and that AIFAs are also simpler to use compared to TFAs (Fox et al., 2010; Johnson and Willsky, 2013). In contrast to the methods of Lee, Miscouridou and Caron (2022); Lee, James and Choi (2016), one can always evaluate the probability density (up to a proportionality constant) of AIFAs; furthermore, in Section 6.4, AIFAs accurately learn model hyperparameters by maximizing the marginal likelihood where the methods of Lee, Miscouridou and Caron (2022); Lee, James and Choi (2016) struggle.

In Section 4, we bound the error induced by approximating an exact infinite-dimensional prior with an AIFA. Our analysis provides interpretable error bounds with explicit dependence on the size of the approximation and the data cardinality; our bounds can be used to set the size of the approximation in practice. Our error bounds reveal that for the worst-case choice of observation likelihood, to approximate the target

[1] We also note that, without modification, their approximation is not suitable for use in statistical models where the unnormalized atom sizes of the CRM are bounded, as arise when modeling the frequencies (in $[0, 1]$) of traits. While model reparameterization may help, it requires (at least) additional steps.

[2] NCRMs are also called *normalized random measures with independent increments* (NRMIs) (Regazzini, Lijoi and Prünster, 2003; James, Lijoi and Prünster, 2009).

to a desired accuracy, it is necessary to use a large IFA model while a small TFA model would suffice. However, in practical experiments with standard observations likelihoods, we find that AIFAs and TFAs of equal sizes have similar performance. Likewise, we find that, when both apply, AIFAs and alternative IFAs (Lee, Miscouridou and Caron, 2022; Lee, James and Choi, 2016) exhibit similar predictive performance (Section 6.3). But AIFAs apply more broadly and are amenable to hyperparameter learning via optimizing the marginal likelihood, unlike Lee, Miscouridou and Caron (2022); Lee, James and Choi (2016) (Section 6.4). As a further illustration, we show that we are able to learn whether a model is over- or underdispersed, and by how much, using an AIFA approximating a novel BNP prior in Section 6.5.

2 Background

Our work will approximate nonparametric priors, so we first review construction of these priors from completely random measures (CRMs). Then we cover existing work on the construction of truncated and independent finite approximations for these CRM priors. For some space Ψ, let $\psi_i \in \Psi$ represent the i-th trait of interest, and let $\theta_i > 0$ represent the corresponding rate or frequency of this trait in the population. If the set of traits is finite, we let I equal its cardinality; if the set of traits is countably infinite, we let $I = \infty$. Collect the pairs of traits and frequencies in a measure Θ that places non-negative mass θ_i at location ψ_i: $\Theta := \sum_{i=1}^{I} \theta_i \delta_{\psi_i}$, where δ_{ψ_i} is a Dirac measure placing mass 1 at location ψ_i. To perform Bayesian inference, we need to choose a prior distribution on Θ and a likelihood for the observed data $Y_{1:N} := \{Y_n\}_{n=1}^{N}$ given Θ. Then, applying a disintegration, we can obtain the posterior on Θ given the observed data.

Homogeneous completely random measures. Many common BNP priors can be formulated as completely random measures (Kingman, 1967; Lijoi and Prünster, 2010).[3] CRMs are constructed from Poisson point processes,[4] which are straightforward to manipulate analytically (Kingman, 1992). Consider a Poisson point process on $\mathbb{R}_+ := [0, \infty)$ with rate measure $\nu(\mathrm{d}\theta)$ such that $\nu(\mathbb{R}_+) = \infty$ and $\int \min(1, \theta)\nu(\mathrm{d}\theta) < \infty$. Such a process generates a countably infinite set of rates $(\theta_i)_{i=1}^{\infty}$ with $\theta_i \in \mathbb{R}_+$ and $0 < \sum_{i=1}^{\infty} \theta_i < \infty$ almost surely. We assume throughout that $\psi_i \overset{\text{i.i.d.}}{\sim} H$ for some diffuse distribution H. The distribution H, called the ground measure, serves as a prior on the traits in the space Ψ. For example, consider a common topic model. Each trait ψ_i represents a latent topic, modeled as a probability vector in the simplex of vocabulary words. And θ_i represents the frequency with which the topic ψ_i appears across documents in a corpus. H is a Dirichlet distribution over the probability simplex, with dimension given by the number of words in the vocabulary.

By pairing the rates from the Poisson process with traits drawn from the ground measure, we obtain a completely random measure and use the shorthand $\mathrm{CRM}(H, \nu)$

[3] Conversely, some important priors, such as Pitman-Yor processes, are not CRMs or their normalizations and are outside the scope of the present paper (Pitman and Yor, 1997; Arbel, De Blasi and Prünster, 2019; Lijoi, Prünster and Rigon, 2020a).

[4] For brevity, we do not consider the fixed-location and deterministic components of a CRM (Kingman, 1967). When these are purely atomic, they can be added to our analysis without undue effort.

for its law: $\Theta = \sum_i \theta_i \delta_{\psi_i} \sim \mathrm{CRM}(H, \nu)$. Since the traits ψ_i and the rates θ_i are independent, the CRM is *homogeneous*. When the total mass $\Theta(\Psi)$ is strictly positive and finite, the corresponding *normalized CRM* (NCRM) is $\Xi := \Theta/\Theta(\Psi)$, which is a discrete probability measure: $\Xi = \sum_i \xi_i \delta_{\psi_i}$, where $\xi_i = \theta_i/(\sum_j \theta_j)$ (Regazzini, Lijoi and Prünster, 2003; James, Lijoi and Prünster, 2009).

The CRM prior on Θ is typically combined with a likelihood that generates trait counts for each data point. Let $\ell(\cdot \,|\, \theta)$ be a proper probability mass function on $\mathbb{N} \cup \{0\}$ for all θ in the support of ν. The process $X_n := \sum_i x_{ni} \delta_{\psi_i}$ collects the trait counts, where $x_{ni} \,|\, \Theta \sim \ell(\cdot \,|\, \theta_i)$ independently across atom index i and i.i.d. across data index n. We denote the distribution of X_n as $\mathrm{LP}(\ell, \Theta)$, which we call the *likelihood process*. Together, the prior on Θ and likelihood on X given Θ form a generative model for allocation of data points to traits; hence, this generative model is a special case of a *trait allocation model* (Campbell, Cai and Broderick, 2018). Analogously, when the trait counts are restricted to $\{0, 1\}$, this generative model represents a special case of a *feature allocation model*.

Since the trait counts are typically just a latent component in a full generative model specification, we define the observed data to be $Y_n \,|\, X_n \overset{\text{indep}}{\sim} f(\cdot \,|\, X_n)$ for a probability kernel $f(\mathrm{d}Y \,|\, X)$. Consider the topic modeling example: θ_i represents the rate of topic ψ_i in a document corpus; Θ captures the rates of all topics; X_n captures how many words in document n are generated from each topic; and Y_n gives the observed collection of words for that document.

Finite approximations. Since the set $\{\theta_i\}_{i=1}^{\infty}$ is countably infinite, it is not possible to simulate or perform posterior inference for every θ_i. One approximation scheme uses a *finite approximation* $\Theta_K := \sum_{i=1}^{K} \rho_i \delta_{\psi_i}$. The atom sizes $\{\rho_i\}_{i=1}^{K}$ are designed so that Θ_K is a good approximation of Θ in a suitable sense. Since it involves a finite number of parameters unlike Θ, Θ_K can be used directly in standard posterior approximation schemes such as Markov chain Monte Carlo or variational Bayes. But not using the full CRM Θ introduces approximation error.

A *truncated finite approximation* (TFA; Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012; Roychowdhury and Kulis, 2015; Arbel and Prünster, 2017; Campbell et al., 2019) requires constructing an ordering on the set of rates from the Poisson process; let $(\theta_i)_{i=1}^{\infty}$ be the corresponding *sequence* of rates. The approximation uses $\rho_i = \theta_i$ for i up to some K; i.e. one keeps the first K rates in the sequence and ignores the remaining ones. We refer to the number of instantiated atoms K as the *approximation level*. Campbell et al. (2019) categorizes and analyzes TFAs. TFAs offer an attractive nested structure: to refine an existing truncation, it suffices to generate the additional terms in the sequence. However, the complex dependencies between the rates $(\theta_i)_{i=1}^{K}$ potentially make inference more challenging.

We instead develop a family of *independent finite approximations* (IFAs). An IFA is defined by a sequence of probability measures $\nu_1, \nu_2, \ldots$ such that at approximation level K, there are K atoms whose weights are given by $\rho_1, \ldots, \rho_K \overset{\text{i.i.d.}}{\sim} \nu_K$. The probability measures are chosen so that the sequence of approximations converges in distribution to the target CRM: $\Theta_K \overset{\mathcal{D}}{\to} \Theta$ as $K \to \infty$. For random measures, convergence in distribution

can also be characterized by convergence of integrals under the measures (Kallenberg, 2002, Lemma 12.1 and Theorem 16.16). The advantages and disadvantages of IFAs reverse those of TFAs: the atoms are now i.i.d., potentially making inference easier, but a completely new approximation must be constructed if K changes.

Next consider approximating an NCRM $\Xi = \sum_i \xi_i \delta_{\psi_i}$, where $\xi_i = \theta_i / (\sum_j \theta_j)$, with a finite approximation. A normalized TFA might be defined in one of two ways. In the first approach, the rates $\{\rho_i\}_{i=1}^K$ that target the CRM rates $\{\theta_i\}_{i=1}^\infty$ are normalized to form the NCRM approximation; i.e. the approximation has atom sizes $\rho_i / \sum_{j=1}^K \rho_j$ (Campbell et al., 2019). The second approach directly constructs an ordering over the sequence of normalized rates ξ_i and truncates this representation.[5] We construct normalized IFAs in a similar manner to the first TFA approach: the NCRM approximation has atom sizes $\rho_i / \sum_{j=1}^K \rho_j$ where $\{\rho_i\}_{i=1}^K$ are the IFA rates.

In the past, independent finite approximations have largely been developed on a case-by-case basis (Paisley and Carin, 2009; Broderick et al., 2015; Acharya, Ghosh and Zhou, 2015; Lee, James and Choi, 2016). Our goal is to provide a general-purpose mechanism. Lijoi, Prünster and Rigon (2023) and Lee, Miscouridou and Caron (2022) have also recently pursued a more general construction, but we believe there remains room for improvement. Lijoi, Prünster and Rigon (2023) focus on NCRMs for clustering; it is not immediately clear how to adapt this work for inference in trait allocation models. Also, Lijoi, Prünster and Rigon (2023, Theorem 1) employ infinitely divisible random variables. Since infinitely divisible distributions that are not Dirac measures cannot have bounded support, the approximate rates $\{\rho_i\}_{i=1}^K$ are not naturally compatible with the trait likelihood $\ell(\cdot \,|\, \theta)$ if the support of the rate measure ν is bounded. But the support of ν is often bounded in applications to trait allocation models; e.g., θ_i may represent a feature frequency, taking values in $[0, 1]$, and $\ell(\cdot \,|\, \theta)$ may take the form of a Bernoulli, binomial, or negative binomial distribution. Therefore, applications of the finite approximations of Lijoi, Prünster and Rigon (2023, Theorem 1) to these models may require some additional work. The construction in Lee, Miscouridou and Caron (2022, Proposition 3.2) yields $\{\rho_i\}_{i=1}^K$ that are compatible with $\ell(\cdot \,|\, \theta)$ and recovers important cases in the literature. However, outside these special cases, it is unknown if the i.i.d. distributions are tractable because the densities ν_K are not explicitly defined; see the discussion around Eq. (3) for more details.

Example 2.1 (Running example: beta process)**.** For concreteness, we consider the *(three-parameter) beta process*[6] (Teh and Görür, 2009; Broderick, Jordan and Pitman, 2012) as a running example of a CRM. The process $\mathrm{BP}(\gamma, \alpha, d)$ is defined by a mass parameter $\gamma > 0$, discount parameter $d \in [0, 1)$, and concentration parameter $\alpha > -d$. It has rate measure

$$\nu(\mathrm{d}\theta) = \gamma \frac{\Gamma(\alpha + 1)}{\Gamma(1 - d)\Gamma(\alpha + d)} \mathbf{1}\{0 \leq \theta \leq 1\} \theta^{-d-1} (1 - \theta)^{\alpha + d - 1} \mathrm{d}\theta. \tag{1}$$

[5] In this case, $\sum_{i=1}^K \xi_i < 1$. Therefore, setting the final atom size in the NCRM approximation to be $1 - \sum_{i=1}^K \xi_i$ ensures the approximation is a probability measure.

[6] Also known as the *stable beta process* (Teh and Görür, 2009).

The $d = 0$ case yields the standard beta process (Hjort, 1990; Thibaux and Jordan, 2007). The beta process is typically paired with the Bernoulli likelihood process with conditional distribution $\ell(x \mid \theta) = \theta^x(1-\theta)^{1-x}\mathbf{1}\{x \in \{0,1\}\}$. The resulting *beta–Bernoulli process* has been used in factor analysis models (Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012) and for dictionary learning (Zhou et al., 2009).

3 Automated independent finite approximations

In this section we introduce *automated independent finite approximations,* a practical construction of independent finite approximations (IFAs) for a broad class of CRMs. We highlight a useful special case of our construction for exponential family CRMs (Broderick, Wilson and Jordan, 2018) without power laws and apply our construction to approximate NCRMs. In all of these cases, we prove that as the approximation size increases, the distribution of the approximation converges (in some relevant sense) to that of the exact infinite-dimensional model.

3.1 Applying our approximation to CRMs

Formally, we define IFAs in terms of a fixed, diffuse probability measure H and a sequence of probability measures $\nu_1, \nu_2, \ldots$. The K-atom IFA Θ_K is

$$\Theta_K := \textstyle\sum_{i=1}^K \rho_i \delta_{\psi_i}, \qquad \rho_i \overset{\text{i.i.d.}}{\sim} \nu_K, \qquad \psi_i \overset{\text{i.i.d.}}{\sim} H,$$

which we write as $\Theta_K \sim \mathrm{IFA}_K(H, \nu_K)$. We consider CRM rate measures ν with densities that, near zero, are (roughly) proportional to θ^{-1-d}, where $d \in [0,1)$ is the *discount* parameter. We will propose a general construction for IFAs given a target random measure and prove that it converges to the target (Theorem 3.1). We first summarize our requirements for which CRMs we approximate in Assumption 1. We show in Nguyen et al. (2023, Section S1) that popular BNP priors satisfy Assumption 1; specifically, we check the beta, gamma (Ferguson and Klass, 1972; Kingman, 1975; Titsias, 2008), generalized gamma (Brix, 1999), beta prime (Broderick et al., 2015), and PG(α, ζ)-generalized gamma (James, 2013) processes.

Assumption 1. For $d \in [0,1)$ and $\eta \in V \subseteq \mathbb{R}^d$, we take $\Theta \sim \mathrm{CRM}(H, \nu(\cdot; \gamma, d, \eta))$ for

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

such that

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

Other than the discount d and mass γ, the rate measure ν potentially depends on additional hyperparameters η. The finiteness of the normalizer Z is necessary in defining finite-dimensional distributions whose densities are similar in form to ν. The conditions on the behaviors of $g(\theta)$ and $h(\theta;\eta)$ ensure that the overall rate measure's behavior near $\theta = 0$ is dominated by the θ^{-1-d} term. The support of the rate measure is implicitly determined by $h(\theta;\eta)$.

Given a CRM satisfying Assumption 1, we can construct a sequence of IFAs that converge in distribution to that CRM.

Theorem 3.1. *Suppose Assumption 1 holds. Let*

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

For $c := \gamma h(0;\eta)/Z(1-d,\eta)$*, let*

$$\nu_K(\mathrm{d}\theta) := \theta^{-1+cK^{-1}-dS_{1/K}(\theta-1/K)} g(\theta)^{cK^{-1}-d} h(\theta;\eta) Z_K^{-1} \mathrm{d}\theta$$

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

See Nguyen et al. (2023, Section S2) for a proof of Theorem 3.1. We choose the particular form of $S_b(\theta)$ in Eq. (2) for concreteness and convenience. But our theory still holds for a more general class of S_b forms, as we describe in more detail in the proof of Theorem 3.1.

Definition 3.2. We call the K-atom IFA resulting from Theorem 3.1 the *automated IFA* (AIFA_K).

Although the normalization constant Z_K is not always available analytically, numerical implementation remains straightforward. When Z_K is a quantity of interest, such as in Section 6.4, we estimate it using standard numerical integration schemes for a one-dimensional integral (Piessens et al., 2012; Virtanen et al., 2020). For other tasks, we need not access Z_K directly. In our experiments, we show that we can use either Markov chain Monte Carlo (Sections 6.1 and 6.5) or variational Bayes (Sections 6.2 and 6.3) with the unnormalized density.

To illustrate our construction, we next apply Theorem 3.1 to $\mathrm{BP}(\gamma, \alpha, d)$ from Example 2.1. In Nguyen et al. (2023, Section S1) we show how to construct AIFAs for the beta prime, gamma, generalized gamma, and $\mathrm{PG}(\alpha,\zeta)$-generalized gamma processes.

Example 3.1 (Beta process AIFA)**.** To apply Assumption 1, let $\eta = \alpha + d$, $V = \mathbb{R}_+$, $g(\theta) = 1$, $h(\theta;\eta) = (1-\theta)^{\eta-1}\mathbb{1}[\theta \le 1]$, and $Z(\xi,\eta)$ equal the beta function $B(\xi,\eta)$. Then the CRM rate measure ν in Assumption 1 corresponds to that of $\mathrm{BP}(\gamma,\alpha,d)$ from Example 2.1. Note that we make no additional restrictions on the hyperparameters γ, α, d beyond those in the original CRM (Example 2.1). Observe that h is continuous

and bounded on $[0, 1/2]$, and the normalization function $B(\xi, \eta)$ is finite for $\xi > 0, \eta \in V$; it follows that Assumption 1 holds. By Theorem 3.1, then, the AIFA density is

$$\frac{1}{Z_K}\theta^{-1+c/K-dS_{1/K}(\theta-1/K)}(1-\theta)^{\alpha+d-1}\mathbf{1}\{0 \le \theta \le 1\}d\theta,$$

where $c \coloneqq \gamma/B(\alpha+d, 1-d)$ and Z_K is the normalization constant. The density does not in general reduce to a beta distribution in θ due to the θ in the exponent.

Comparison to an alternative IFA construction. Lee, Miscouridou and Caron (2022, Proposition 3.2) verify the validity of a different IFA construction. Their construction requires two functions: (1) a bivariate function $\Lambda(\theta, t)$ such that for any $t > 0, \Delta(t) \coloneqq \int \Lambda(\theta, t)\nu(d\theta) < \infty$ and (2) a univariate function $f(n)$ such that $\Delta(f(n))$ is bounded from both above and below by n as $n \to \infty$. If these functions exist and

$$\widetilde{\nu}_K(d\theta) \coloneqq \frac{\Lambda(\theta, f(K))\nu(d\theta)}{\Delta(f(K))}, \tag{3}$$

Lee, Miscouridou and Caron (2022, Proposition 3.2) show that $\mathrm{IFA}_K(H, \widetilde{\nu}_K)$ converges in distribution to $\mathrm{CRM}(H, \nu)$ as $K \to \infty$. The usability of Eq. (3) in practice depends on the tractability of Λ and f. There are typically many tractable $\Lambda(\theta, t)$ (Lee, Miscouridou and Caron, 2022, Section 4). Proposition B.2 of Lee, Miscouridou and Caron (2022) lists tractable f for the important cases of the beta process and generalized gamma process with $d > 0$. However, the choice of f provided there for general power-law processes is not tractable because its evaluation requires computing complicated inverses in the asymptotic regime. Furthermore, for processes without power laws, no general recipe for f is known. In contrast, the AIFA construction in Theorem 3.1 always yields densities that can be evaluated up to proportionality constants.

Example 3.2 (Beta process: an IFA comparison)**.** We next compare our beta process AIFA to the two separate IFAs proposed by Lee, Miscouridou and Caron (2022) and Lee, James and Choi (2016) for disjoint subcases within the case $d > 0$. First consider the subcase where $\alpha = 0, d > 0$. Lee, James and Choi (2016) derive[7] what we call[8] the *BFRY IFA*. The IFA density, denoted $\nu_{\mathrm{BFRY}}(d\theta)$, is equal to

$$\frac{\gamma}{K}\frac{\theta^{-d-1}(1-\theta)^{d-1}}{B(d, 1-d)}\left[1 - \exp\left(-\left(\frac{K\Gamma(d)d}{\gamma}\right)^{1/d}\frac{\theta}{1-\theta}\right)\right]\mathbf{1}\{0 \le \theta \le 1\}d\theta. \tag{4}$$

Second, consider the subcase where $\alpha > 0, d > 0$, Lee, Miscouridou and Caron (2022, Section 4.5) derive another K-atom IFA, which we call[9] the *generalized Pareto*

[7] There is a typo in Lee, James and Choi (2016, Theorem 2, item (iii)): θ/K should be $(\theta/\Gamma(\alpha))/K$.

[8] Devroye and James (2014) introduce the acronym BFRY to denote a distribution named for the authors Bertoin et al. (2006). We here use "BFRY IFA" to denote what Lee, James and Choi (2016) call the "BFRY process" and thereby emphasize that this process forms an IFA.

[9] We use the term "generalized Pareto" because Lee, Miscouridou and Caron (2022, Section 4.5) use generalized Pareto variates to define $\Lambda(\theta, t)$ from Eq. (3).

IFA (GenPar IFA). The IFA density, denoted $\nu_{\text{GenPar}}(d\theta)$, is equal to

$$\frac{\gamma}{K}\frac{\theta^{-d-1}(1-\theta)^{\alpha+d-1}}{B(1-d,\alpha+d)}\left(1-\frac{1}{\left(\theta\left[\left(1+\frac{Kd}{\gamma\alpha}\right)^{\frac{1}{d}}-1\right]+1\right)^{\alpha}}\right)\mathbf{1}\{0\le\theta\le 1\}d\theta. \tag{5}$$

Since the BFRY IFA and GenPar IFA apply to disjoint hyperparameter regimes, they are not directly comparable. Since our AIFA applies to the whole domain $\alpha \ge -d$, we can separately compare it to each of these alternative IFAs; we also highlight that the AIFA still applies when $\alpha \in (-d, 0)$, a case not covered by either the BFRY IFA or GenPar IFA.

We find in Section 6.3 that the AIFA and BFRY IFA have comparable predictive performance; the AIFA and GenPar IFA also have comparable predictive performance. But in Section 6.4, we show that the AIFA is much more reliable than the BFRY IFA or the GenPar IFA for estimating the discount (d) hyperparameter by maximizing the marginal likelihood. Conversely, sampling from a BFRY IFA or GenPar IFA prior is easier than sampling from an AIFA prior since the BFRY and GenPar IFA priors are formed from standard distributions.

3.2 Applying our approximation to exponential family CRMs

Exponential family CRMs with $d = 0$ comprise a widely used special case of CRMs. In what follows, we show how Theorem 3.1 simplifies in this special case.

In common BNP models, the relationship between the likelihood $\ell(\cdot \mid \theta)$ and the CRM prior is closely related to finite-dimensional exponential family conjugacy (Broderick, Wilson and Jordan, 2018, Section 4). In particular, the likelihood has an exponential family form,

$$\ell(x \mid \theta) := \kappa(x)\theta^{\phi(x)}\exp\left(\langle\mu(\theta), t(x)\rangle - A(\theta)\right). \tag{6}$$

Here $x \in \mathbb{N}\cup\{0\}$, $\kappa(x) \in \mathbb{R}$ is the base density, $\phi(x) \in \mathbb{R}$ and $t(x) \in \mathbb{R}^{D'}$ (for some D') form the vector of sufficient statistics $(t(x), \phi(x))^T$, $A(\theta) \in \mathbb{R}$ is the log partition function, $\mu(\theta) \in \mathbb{R}^{D'}$ and $\ln\theta$ form the vector of natural parameters $(\mu(\theta), \ln\theta)^T$, and $\langle\mu(\theta), t(x)\rangle$ denotes the standard Euclidean inner product. The rate measure nearly matches the form of the conjugate prior, but behaves like θ^{-1} near 0:

$$\nu(\mathrm{d}\theta) := \gamma'\theta^{-1}\exp\left\{\left\langle\begin{pmatrix}\psi\\ \lambda\end{pmatrix}, \begin{pmatrix}\mu(\theta)\\ -A(\theta)\end{pmatrix}\right\rangle\right\}\mathbf{1}\{\theta\in U\}\mathrm{d}\theta, \tag{7}$$

where $\gamma' > 0$, $\lambda > 0$, $\psi \in \mathbb{R}^{D'}$ and $U \subseteq \mathbb{R}_+$ is the support of ν. Eq. (7) leads to the suggestive terminology of *exponential family* CRMs. The θ^{-1} dependence near 0 means that these models lack power-law behavior. Models that can be cast in this form include the standard beta process with Bernoulli or negative binomial likelihood (Zhou et al., 2012; Broderick et al., 2015) and the gamma process with Poisson likelihood (Acharya, Ghosh and Zhou, 2015; Roychowdhury and Kulis, 2015). We refer to these models as, respectively, the beta–Bernoulli, beta–negative binomial, and gamma–Poisson processes.

We now specialize Assumption 1 and Theorem 3.1 to exponential family CRMs in Assumption 2 and Corollary 3.3, respectively.

Assumption 2. Let ν be of the form in Eq. (7) and assume that

1. For any $\xi > -1$, for any $\eta = (\psi, \lambda)^T$ where $\lambda > 0$, the normalizer defined as
$$Z(\xi, \eta) := \int_U \theta^\xi \exp\left\{\left\langle \eta, \begin{pmatrix} \mu(\theta) \\ -A(\theta) \end{pmatrix} \right\rangle\right\} d\theta \tag{8}$$
is finite, and
2. there exists $\epsilon > 0$ such that, for any $\eta = (\psi, \lambda)^T$ where $\lambda > 0$, the map
$$\varsigma : \theta \mapsto \exp\left\{\left\langle \eta, \begin{pmatrix} \mu(\theta) \\ -A(\theta) \end{pmatrix} \right\rangle\right\} \mathbf{1}\{\theta \in U\}$$
is a continuous and bounded function of θ on $[0, \epsilon]$.

Corollary 3.3. *Suppose Assumption 2 holds. For* $c := \gamma'\varsigma(0)$*, let*
$$\nu_K(\theta) := \frac{\theta^{c/K-1}\varsigma(\theta)}{Z\left(c/K - 1, \eta\right)}. \tag{9}$$
If $\Theta_K \sim \mathrm{IFA}_K(H, \nu_K)$*, then* $\Theta_K \xrightarrow{\mathcal{D}} \Theta$.

The density in Eq. (9) is almost the same as the rate measure of Eq. (7), except the θ^{-1} term has become $\theta^{c/K-1}$. As a result, Eq. (9) is a proper exponential-family distribution. In Nguyen et al. (2023, Section S1), we detail the corresponding $d = 0$ special cases of the AIFA for beta prime, gamma, generalized gamma, and PG(α,ζ)-generalized gamma processes. We cover the beta process case next.

Example 3.3 (Beta process AIFA for $d = 0$)**.** Corollary 3.3 is sufficient to recover known IFA results for $\mathrm{BP}(\gamma, \alpha, 0)$; when $d = 0$, the AIFA from Example 3.1 simplifies to $\nu_K = \mathsf{Beta}\left(\gamma\alpha/K, \alpha\right)$. Doshi-Velez et al. (2009) approximates $\mathrm{BP}(\gamma, 1, 0)$ with $\nu_K = \mathsf{Beta}\left(\gamma/K, 1\right)$. For $\mathrm{BP}(\gamma, \alpha, 0)$, Griffiths and Ghahramani (2011) set $\nu_K = \mathsf{Beta}\left(\gamma\alpha/K, \alpha\right)$, and Paisley and Carin (2009) use $\nu_K = \mathsf{Beta}\left(\gamma\alpha/K, \alpha(1 - 1/K)\right)$. The difference between $\mathsf{Beta}\left(\gamma\alpha/K, \alpha\right)$ and $\mathsf{Beta}\left(\gamma\alpha/K, \alpha(1 - 1/K)\right)$ is negligible for moderately large K.

We can also use Corollary 3.3 to create a new finite approximation for a nonparametric process so far not explored in the Bayesian nonparametric literature.

Example 3.4 (CMP likelihood and extended gamma process)**.** The *CMP likelihood*[10] (Shmueli et al., 2005) is given by
$$\ell(x \mid \theta) = \frac{\theta^x}{(x!)^\tau} \frac{1}{Z_\tau(\theta)}, \quad \text{where } Z_\tau(\theta) := \sum_{y=0}^{\infty} \frac{\theta^y}{(y!)^\tau}. \tag{10}$$

[10] CMP stands for Conway-Maxwell-Poisson.

The conjugate CRM prior, which we call an *extended gamma* (or *Xgamma*) *process*, has four hyperparameters: mass γ, concentration c, maximum T, and shape τ:

$$\nu(d\theta) = \gamma\theta^{-1}Z_\tau^{-c}(\theta)1\{0 \le \theta \le T\}d\theta. \tag{11}$$

Unlike existing BNP models, the model in Eqs. (10) and (11), which we call *Xgamma–CMP process*, is able to capture different dispersion regimes. For $\tau < 1$, the variance of the counts from $\ell(x \mid \theta)$ is larger than the mean of the counts, corresponding to overdispersion. For $\tau > 1$, the variance of the counts from $\ell(x \mid \theta)$ is smaller than the mean of the counts, corresponding to underdispersion. As we show in Section 6.5, the latent shape τ can be inferred using observed data. Zhou et al. (2012); Broderick et al. (2015) provide BNP trait allocation models that handle overdispersion. Canale and Dunson (2011) provide a BNP model that handles both underdispersion and overdispersion, but for clustering rather than traits. We are not aware of trait allocation models that handle underdispersion, or any trait allocation models that handle both underdispersion and overdispersion. Following the approach of Broderick, Wilson and Jordan (2018), in Nguyen et al. (2023, Section S4) we show that as long as $\gamma > 0$, $c > 0$, $T \ge 1$, and $\tau > 0$, the total mass of the rate measure is infinite and the number of active traits is almost surely finite. Under these conditions, we show in Nguyen et al. (2023, Section S1) that Corollary 3.3 applies to the CRM in Eq. (11), and we construct the resulting AIFA.

3.3 Normalized independent finite approximations

Given that AIFAs are approximations that converge to the corresponding target CRM, it is natural to ask if normalizations of AIFAs converge to the corresponding normalization of the target CRM, i.e., the corresponding NCRM. Our next result shows that normalized AIFAs indeed converge, in the sense that the exchangeable partition probability functions, or EPPFs (Pitman, 1995), converge. Given a random sample of size N from an NCRM Ξ, the EPPF gives the probability of the induced partition from such a sample. In particular, consider the model $\Xi \sim \text{NCRM}, X_n \mid \Xi \overset{\text{i.i.d.}}{\sim} \Xi$ for $1 \le n \le N$.[11] Grouping the indices n with the same value of X_n induces a partition over the set $\{1, 2, \ldots, N\}$. Let b represent the number of distinct values in the set $\{X_n\}_{n=1}^N$, so $b \le N$. Let n_i be the number of indices n with X_n equal to the i-th distinct value of X_n, for some ordering of the values. So $\sum_{i=1}^b n_i = N$ and $\forall i, n_i \ge 1$. With this notation in hand, we can write the EPPF, which gives the probability of the induced partition under the model, as a symmetric function $p(n_1, n_2, \ldots, n_b)$ that depends only on the counts n_i. Similarly, we let $p_K(n_1, n_2, \ldots, n_b)$ be the EPPF for the normalized AIFA_K. Note that $p_K(n_1, n_2, \ldots, n_b) = 0$ when $K < b$ since the normalized AIFA_K at approximation level K generates at most K blocks.

Theorem 3.4. *Suppose Assumption 1 holds. Take any positive integers $N, b, \{n_i\}_{i=1}^b$ such that $b \le N$, $n_i \ge 1$, and $\sum_{i=1}^b n_i = N$. Let p be the EPPF of the NCRM $\Xi :=$*

[11] We reuse the X_n notation from the CRM description, even though X_n now is a scalar, because the role of the draws from Ξ is the same as that of the draws from Θ.

$\Theta/\Theta(\Psi)$. *If* Θ_K *is the AIFA for* Θ *at approximation level* K*, and* p_K *is the EPPF for the corresponding NCRM approximation* $\Theta_K/\Theta_K(\Psi)$*, then*

$$\lim_{K\to\infty} p_K(n_1, n_2, \ldots, n_b) = p(n_1, n_2, \ldots, n_b).$$

See Nguyen et al. (2023, Section S2.3) for the proof. 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 sampling from the normalized AIFA_K converges to that induced by sampling from the target NCRM, for any finite sample size N.

4 Non-asymptotic error bounds

Theorems 3.1 and 3.4 justify the use of our proposed AIFA construction in the limit $K \to \infty$ but do not provide guidance on how to choose the approximation level K when N observations are available. In Section 4.1, we quantify the error introduced by replacing an exponential family CRM with the AIFA. In Section 4.2, we quantify the error introduced by replacing a Dirichlet process (DP) (Ferguson, 1973; Sethuraman, 1994) with the corresponding normalized AIFA. We derive error bounds that are simple to manipulate and yield recommendations for the appropriate K for a given N and a desired accuracy level.

4.1 Bounds when approximating an exponential family CRM

Recall from Section 2 that the CRM prior Θ is typically paired with a likelihood process LP, which manifests features X_n, and a probability kernel f relating active features to observations Y_n. The target nonparametric model can be summarized as

$$\begin{aligned} \Theta &\sim \text{CRM}(H, \nu), \\ X_n \mid \Theta &\overset{\text{i.i.d.}}{\sim} \text{LP}(\ell, \Theta), \quad n = 1, 2, \ldots, N, \\ Y_n \mid X_n &\overset{\text{indep}}{\sim} f(\cdot \mid X_n), \quad n = 1, 2, \ldots, N. \end{aligned} \tag{12}$$

The approximating model, with ν_K as in Theorem 3.1 (or Corollary 3.3), is

$$\begin{aligned} \Theta_K &\sim \text{AIFA}_K(H, \nu_K), \\ Z_n \mid \Theta_K &\overset{\text{i.i.d.}}{\sim} \text{LP}(\ell, \Theta_K), \quad n = 1, 2, \ldots, N, \\ W_n \mid Z_n &\overset{\text{indep}}{\sim} f(\cdot \mid Z_n), \quad n = 1, 2, \ldots, N. \end{aligned} \tag{13}$$

Active traits in the approximate model are collected in Z_n and observations are W_n. Let $P_{N,\infty}$ be the marginal distribution of the observations $Y_{1:N}$ and $P_{N,K}$ be the marginal distribution of the observations $W_{1:N}$. The *approximation error* we analyze is the total variation distance $d_{\text{TV}}(P_{N,K}, P_{N,\infty}) := \sup_{0\le g\le 1} |\int g \mathrm{d}P_{N,K} - \int g \mathrm{d}P_{N,\infty}|$ between the two observational processes, one using the CRM and the other one using the approximate AIFA_K as the prior. Total variation is a standard choice of error when analyzing

CRM approximations (Ishwaran and Zarepour, 2002; Doshi-Velez et al., 2009; Paisley, Blei and Jordan, 2012; Campbell et al., 2019). Small total variation distance implies small differences in expectations of bounded functions.

Conditions. In our analysis, we focus on exponential family CRMs and conjugate likelihood processes. We will suppose Assumption 2 holds. Our analysis guarantees that $d_{\mathrm{TV}}(P_{N,K}, P_{N,\infty})$ is small whenever a conjugate exponential family CRM–likelihood pair and the corresponding AIFA model satisfy certain conditions, beyond those already stated in Assumption 2. In the proof of the error bound, these conditions serve as intermediate results that ultimately lead to small approximation error. Because we can verify the conditions for common models, we have error bounds in the most prevalent use cases of CRMs. To express these conditions, we use the *marginal process* representation of the target and the approximate model, i.e., the series of conditional distributions of $X_n \,|\, X_{1:(n-1)}$ (or $Z_n \,|\, Z_{1:(n-1)}$) with Θ (or Θ_K) integrated out. Corollary 6.2 of Broderick, Wilson and Jordan (2018) guarantees that the marginal $X_n \,|\, X_{1:(n-1)}$ is a random measure with finite support and with a convenient form. Since we will use this form to write our conditions (Condition 1 below), we first review the requisite notation — and establish analogous notation for $Z_n \,|\, Z_{1:(n-1)}$.

We start by defining h and M to describe the conditional distribution $X_n \,|\, X_{1:(n-1)}$. Let K_{n-1} be the number of unique atom locations in $X_1, X_2, \ldots, X_{n-1}$, and let $\{\zeta_i\}_{i=1}^{K_{n-1}}$ be the collection of unique atom locations in $X_1, X_2, \ldots, X_{n-1}$. Fix an atom location ζ_j (the choice of j does not matter). For m with $1 \leq m \leq n$, let x_m be the atom size of X_m at atom location ζ_j; x_m may be zero if there is no atom at ζ_j in X_m. The distribution of x_n depends *only* on the $x_{1:(n-1)}$ values, which are the atom sizes of previous measures X_m at ζ_j. We use $h(x \mid x_{1:(n-1)})$ to denote the probability mass function (p.m.f.) of x_n at value x. Furthermore, X_n has a finite number of new atoms, which can be grouped together by atom size. Consider any potential atom size $x \in \mathbb{N}$. Define $p_{n,x}$ to be the number of atoms of size x. Regardless of atom size, each atom location is a fresh draw from the ground measure H and $p_{n,x}$ is Poisson-distributed; we use $M_{n,x}$ to denote the mean of $p_{n,x}$.

Next, we define $\widetilde{h}$, which governs the conditional distribution of $Z_n \,|\, Z_{1:(n-1)}$. Let 0_{n-1} be the zero vector with $n-1$ components. Although $h(x \mid x_{1:(n-1)})$ is defined only for count vectors $x_{1:(n-1)}$ that are not identically zero, we will see that $\widetilde{h}(x \mid 0_{n-1})$ is well-defined. In particular, let $\{\zeta_i\}_{i=1}^{K_{n-1}}$ be the union of atom locations in $Z_1, Z_2, \ldots, Z_{n-1}$. Fix an atom location ζ_j. For $1 \leq m \leq n$, let x_m be the atom size of Z_m at atom location ζ_j. We write the p.m.f. of x_n at x as $\widetilde{h}(x \mid x_{1:(n-1)})$. In addition, Z_n also has a maximum of $K - K_{n-1}$ new atoms with locations disjoint from $\{\zeta_i\}_{i=1}^{K_{n-1}}$, and the distribution of atom sizes is governed by $\widetilde{h}(x \mid 0_{n-1})$. Note that we reuse the x_n and ζ_j notation from $X_n \,|\, X_{1:(n-1)}$ without risk of confusion, since x_n and ζ_j are dummy variables whose meanings are clear given the context of h or $\widetilde{h}$.

In Nguyen et al. (2023, Section S3), we describe the marginal processes in more detail and give formulas for h, $\widetilde{h}$, and $M_{n,x}$ in terms of the functions that parametrize Eqs. (6) and (7) and the normalizer Eq. (8). For the beta–Bernoulli process with $d = 0$, the functions have particularly convenient forms.

Example 4.1. For the beta–Bernoulli model with $d = 0$, we have

$$h(x \mid x_{1:(n-1)}) = \frac{\sum_{i=1}^{n-1} x_i}{\alpha - 1 + n}\mathbf{1}\{x = 1\} + \frac{\alpha + \sum_{i=1}^{n-1}(1 - x_i)}{\alpha - 1 + n}\mathbf{1}\{x = 0\}.$$

$$\widetilde{h}(x \mid x_{1:(n-1)}) = \frac{\sum_{i=1}^{n-1} x_i + \gamma\alpha/K}{\alpha - 1 + n + \gamma\alpha/K}\mathbf{1}\{x = 1\} + \frac{\alpha + \sum_{i=1}^{n-1}(1 - x_i)}{\alpha - 1 + n + \gamma\alpha/K}\mathbf{1}\{x = 0\},$$

$$M_{n,1} = \frac{\gamma\alpha}{\alpha - 1 + n}, \quad M_{n,x} = 0 \text{ for } x > 1.$$

We now formulate conditions on h, $\widetilde{h}$, and $M_{n,x}$ that will yield small $d_{\mathrm{TV}}(P_{N,K}, P_{N,\infty})$.

Condition 1. There exist constants $\{C_i\}_{i=1}^5$ such that

1. for all $n \in \mathbb{N}$,
$$\sum_{x=1}^{\infty} M_{n,x} \leq \frac{C_1}{n - 1 + C_1}; \tag{14}$$
2. for all $n \in \mathbb{N}$,
$$\sum_{x=1}^{\infty} \widetilde{h}(x \mid x_{1:(n-1)} = 0_{n-1}) \leq \frac{1}{K}\frac{C_1}{n - 1 + C_1}; \tag{15}$$
3. for any $n \in \mathbb{N}$, for any $\{x_i\}_{i=1}^{n-1} \neq 0_{n-1}$,
$$\sum_{x=0}^{\infty} \left| h(x \mid x_{1:(n-1)}) - \widetilde{h}(x \mid x_{1:(n-1)}) \right| \leq \frac{1}{K}\frac{C_1}{n - 1 + C_1}; \text{ and} \tag{16}$$
4. for all $n \in \mathbb{N}$, for any $K \geq C_2(\ln n + C_3)$,
$$\sum_{x=1}^{\infty} \left| M_{n,x} - K\widetilde{h}(x \mid x_{1:(n-1)} = 0_{n-1}) \right| \leq \frac{1}{K}\frac{C_4 \ln n + C_5}{n - 1 + C_1}. \tag{17}$$

Note that the conditions depend only on the functions governing the exponential family CRM prior and its conjugate likelihood process — and not on the observation likelihood f. Eq. (14) constrains the growth rate of the target model since $\sum_{n=1}^{N}\sum_{x=1}^{\infty} M_{n,x}$ is the expected number of components for data cardinality N. Because each $\sum_{x=1}^{\infty} M_{n,x}$ is at most $O(1/n)$, the total number of components after N samples is $O(\ln N)$. Similarly, Eq. (15) constrains the growth rate of the approximate model. The third condition (Eq. (16)) ensures that $\widetilde{h}$ is a good approximation of h in total variation distance and that there is also a reduction in the error as n increases. Finally, Eq. (17) implies that $K\widetilde{h}(x \mid 0_{n-1})$ is an accurate approximation of $M_{n,x}$, and there is also a reduction in the error as n increases.

We show that Condition 1 holds for the most commonly used non-power-law CRM models; see Example 4.2 for the case of the beta–Bernoulli model with discount $d = 0$

and Nguyen et al. (2023, Section S6) for the beta–negative binomial and gamma–Poisson models with $d = 0$. As we detail next, we believe Condition 1 is also reasonable beyond these common models. The $O(1/n)$ quantity in Eq. (14) is the typical expected number of new features after observing n observations in non-power-law BNP models. Eqs. (15), (16) and (17) are likely to hold when $\widetilde{h}$ is a small perturbation of h and $K\widetilde{h}$ is a small perturbation of $M_{n,x}$. For instance, in Example 4.1, the functional form of $\widetilde{h}$ is very similar to that of h, except that $\widetilde{h}$ has the additional $\gamma\alpha/K$ factor in both numerator and denominator. The functional form of $K\widetilde{h}$ is very similar to that of $M_{n,x}$, except that $K\widetilde{h}$ has an additional $\gamma\alpha/K$ factor in the denominator.

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

$$\widetilde{h}(1 \mid x_{1:(n-1)} = 0_{n-1}) = \frac{\gamma\alpha/K}{\alpha - 1 + n + \gamma\alpha/K} \leq \frac{1}{K}\frac{\gamma\alpha}{n-1+\alpha}.$$

Since both h and $\widetilde{h}$ are supported on $\{0, 1\}$, Eq. (16) becomes

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

And because $M_{n,x} = 0 = \widetilde{h}(x \mid \cdot\,)$ for $x > 1$, Eq. (17) becomes

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

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

Upper bound. We now make use of Condition 1 to derive an upper bound on the approximation error induced by AIFAs.

Theorem 4.1 (Upper bound for exponential family CRMs)**.** *Recall that $P_{N,\infty}$ is the distribution of $Y_{1:N}$ from Eq. (12) while $P_{N,K}$ is the distribution of $W_{1:N}$ from Eq. (13). If Assumption 2 and Condition 1 hold, then there exist positive constants C', C'', C''', C'''' depending only on $\{C_i\}_{i=1}^5$ such that*

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

See Nguyen et al. (2023, Section S7.1) for explicit values of the constants as well as the proof. Theorem 4.1 states that the AIFA approximation error grows as $O(\ln^2 N)$ with fixed K, and decreases as $O(\ln K/K)$ for fixed N. The bound accords with our intuition that, for fixed K, the error should increase as N increases: with more data, the

expected 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. It is likely that the $O(\ln^2 N)$ factor can be improved to $O(\ln N)$ due to $O(\ln N)$ being the natural growth rate; 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, we expect that error should decrease as K increases and the approximation thus has greater capacity. This behavior also matches Theorem 3.1, which guarantees that sufficiently large finite models have small error.

We highlight that Theorem 4.1 provides upper bounds both (i) for approximations that were already known in the literature but where bounds were not already known, as in the case of the beta–negative binomial process, and (ii) for processes and approximations not previously studied in the literature in any form.

Lower bounds. From the upper bound in Theorem 4.1, we know how to set a sufficient number of atoms for accurate approximations: for the total variation to be less than some ϵ, we solve for the smallest K such that the right hand side of Theorem 4.1 is smaller than ϵ. We now derive lower bounds on the AIFA approximation error to characterize a *necessary* number of atoms for accurate approximations, by looking at worst-case observational likelihoods f. In particular, Theorem 4.1 implies that an AIFA with $K = O(\text{poly}(\ln N)/\epsilon)$ atoms suffices in approximating the target model to less than ϵ error. In Theorem 4.2 below, we establish that K must grow at least at a $\ln N$ rate in the worst case. In Theorem 4.3 below, we establish that the $1/\epsilon$ term is necessary. To the best of our knowledge, Theorems 4.2 and 4.3 are the *first* lower bounds on IFA approximation error for any process.

Our lower bounds apply to the beta–Bernoulli process with $d = 0$. Recall that $P_{N,\infty}$ is the distribution of $Y_{1:N}$ from Eq. (12) while $P_{N,K}$ is the distribution of $W_{1:N}$ from Eq. (13). In what follows, $P^{\text{BP}}_{N,\infty}$ refers to the marginal distribution of the observations that arises when we use the prior $\text{BP}(\gamma, \alpha, 0)$. Analogously, $P^{\text{BP}}_{N,K}$ is the observational distribution that arises when we use the AIFA_K approximation in Example 3.1. The observational likelihood f will be clear from context. The worst-case observational likelihoods f are pathological. We leave to future work to lower bound the approximation error when more common likelihoods f, such as Gaussian or Dirichlet, are used.

For the first result, it will be useful to define the *growth function* for any $N \in \mathbb{N}$, $\alpha > 0$:

$$C(N, \alpha) := \sum_{n=1}^{N} \frac{\alpha}{n - 1 + \alpha}. \tag{18}$$

$C(N, \alpha)$ satisfies $\lim_{N\to\infty} C(N, \alpha)/(\alpha \ln N) = 1$; this asymptotic equivalence is a corollary of Nguyen et al. (2023, Lemma S5.10) or Theorem 2.3 from Korwar and Hollander (1972). Our next result shows that our AIFA approximation can be poor if the approximation level K is too small compared to the growth function $C(N, \alpha)$.

Theorem 4.2 ($\ln N$ is necessary)**.** *For the beta–Bernoulli process model with $d = 0$, there exists an observation likelihood f, independent of K and N, such that for any N,*

if $K \leq 0.5\gamma C(N, \alpha)$*, then*

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

where C *is a constant depending only on* γ *and* α*.*

See Nguyen et al. (2023, Section S7.2) for the proof. The intuition is that, with high probability, the number of features that manifest in the target $X_{1:N}$ is greater than $0.5\gamma C(N, \alpha)$. However, the finite model $Z_{1:N}$ has fewer than $0.5\gamma C(N, \alpha)$ components. Hence, there is an event where the target and approximation assign drastically different probability masses. Theorem 4.2 implies that as N grows, if the approximation level K fails to surpass the $0.5\gamma C(N, \alpha)$ threshold, then the total variation between the approximate and the target model remains bounded from zero; in fact, the error tends to one.

We next show that the $1/K$ factor in the upper bound from Theorem 4.1 is *tight* (up to logarithmic factors).

Theorem 4.3 (Lower bound of $1/K$)**.** *For the beta–Bernoulli process model with* $d = 0$*, there exists an observation likelihood* f*, independent of* K *and* N*, such that for any* N*,*

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

where C *is a constant depending only on* γ*.*

See Nguyen et al. (2023, Section S7.2) for the proof. The intuition is that, under the pathological likelihood f, analyzing the AIFA approximation error is the same as analyzing the binomial–Poisson approximation error (Le Cam, 1960). We then show that $1/K$ is a lower bound using the techniques from Barbour and Hall (1984). Theorem 4.3 implies that an AIFA with $K = \Omega\,(1/\epsilon)$ atoms is necessary in the worst case.

Our lower bounds (which apply specifically to the beta–Bernoulli process) are much less general than our upper bounds. However, as a practical matter, generality in the lower bounds is not so crucial due to the different roles played by upper and lower bounds. Upper bounds give control over the approximation error; this control is what is needed to trust the approximation and to set the approximation level. Whether or not we have access to lower bounds, general-purpose upper bounds give us this control. Lower bounds, on the other hand, serve as a helpful check that the upper bounds are not too loose — and reassure us that we are not inefficiently using too many atoms in a too-large approximation. From that standpoint, the need for general-purpose lower bounds is not as pressing.

The dependence on the accuracy level in the $d = 0$ beta–Bernoulli process is worse for AIFAs than for TFAs. For example, consider the Bondesson approximation (Bondesson, 1982; Campbell et al., 2019) of $\mathrm{BP}(\gamma, \alpha, 0)$; we will see next that this approximation is a TFA with excellent error bounds.

Example 4.3 (Bondesson approximation (Bondesson, 1982))**.** Fix $\alpha \geq 1$, let $E_l \overset{\text{i.i.d.}}{\sim} \mathsf{Exp}(1)$, and let $\Gamma_k := \sum_{l=1}^{k} E_l$. The K-atom Bondesson approximation of $\mathrm{BP}(\gamma, \alpha, 0)$ is a TFA $\sum_{k=1}^{K} \theta_k \delta_{\psi_k}$, where $\theta_k := V_k \exp(-\Gamma_k/\gamma\alpha), V_k \overset{\text{i.i.d.}}{\sim} \mathsf{Beta}(1, \alpha - 1)$, and $\psi_k \overset{\text{i.i.d.}}{\sim} H$.

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

Proposition 4.4. *(Campbell et al., 2019, Appendix A.1) For $\gamma > 0, \alpha \geq 1$, let Θ_K be distributed according to a level-K Bondesson approximation of $\mathrm{BP}(\gamma, \alpha, 0)$, $R_n \,|\, \Theta_K \overset{i.i.d.}{\sim} \mathrm{LP}(\ell; \Theta_K), T_n \,|\, R_n \overset{indep}{\sim} f(\cdot \,|\, R_n)$ with N observations. Let $Q_{N,K}$ be the distribution of the observations $T_{1:N}$. Then: $d_{TV}\left(P^{BP}_{N,\infty}, Q_{N,K}\right) \leq N\gamma \left(\frac{\gamma\alpha}{1+\gamma\alpha}\right)^K$.*

Proposition 4.4 implies that a TFA with $K = O\left(\ln\{N/\epsilon\}\right)$ atoms suffices in approximating the target model to less than ϵ error. Up to log factors in N, comparing the necessary $1/\epsilon$ level for an AIFA and the sufficient $\ln(1/\epsilon)$ level for a TFA, we conclude that the necessary size for an AIFA is exponentially larger than the sufficient size for a TFA, in the worst-case observational likelihood f.

4.2 Approximating a (hierarchical) Dirichlet process

So far we have analyzed AIFA error for CRM-based models. In this section, we analyze the error that arises from using a normalized AIFA as an approximation for an NCRM; here, we focus on a Dirichlet process — i.e., a normalized gamma process without power-law behavior. We first consider a generative model with the same number of layers as in previous sections. But we also consider a more complex generative model, with an additional layer — as is common in, e.g., text analysis. Indeed, one of the strengths of Bayesian modeling is the flexibility facilitated by hierarchical modeling, and a goal of probabilistic programming is to provide fast, automated inference for these more complex models.

Dirichlet process. The Dirichlet process is one of the most widely used nonparametric priors and arises as a normalized gamma process. The generalized gamma process CRM is characterized by the rate measure $\nu(\mathrm{d}\theta) = \gamma \frac{\lambda^{1-d}}{\Gamma(1-d)} \theta^{-d-1} e^{-\lambda\theta} \mathrm{d}\theta$. We denote its distribution as $\Gamma\mathrm{P}(\gamma, \lambda, d)$. A normalized draw from $\Gamma\mathrm{P}(\gamma, 1, 0)$ is Dirichlet-process distributed with mass parameter γ (Kingman, 1975; Ferguson, 1973). By Corollary 3.3, $\mathrm{IFA}_K(H, \nu_K)$ with $\nu_K = \mathsf{Gam}(\gamma/K, 1)$ converges to $\Gamma\mathrm{P}(\gamma, 1, 0)$. Because the normalization of independent gamma random variables is a Dirichlet random variable, a normalized draw from $\mathrm{IFA}_K(H, \nu_K)$ is equal in distribution to $\sum_{i=1}^{K} p_i \delta_{\psi_i}$ where $\psi_i \overset{\text{i.i.d.}}{\sim} H$ and $\{p_i\}_{i=1}^{K} \sim \mathsf{Dir}(\{\gamma/K\}\mathbf{1}_K)$. We call this distribution the *finite symmetric Dirichlet* (FSD), and denote it as $\mathrm{FSD}_K(\gamma, H)$.[12]

In the simplest use case, the Dirichlet process is used as the de Finetti measure for observations X_n; i.e., $\Xi \sim \mathrm{DP}, X_n \mid \Xi \overset{\text{i.i.d.}}{\sim} \Xi$ for $1 \leq n \leq N$. In Nguyen et al. (2023, Section S8), we state error bounds when FSD_K replaces the Dirichlet process

[12]The name "finite symmetric Dirichlet" comes from Kurihara, Welling and Teh (2007). See Ishwaran and James (2001, Section 2.2) for other names this distribution has had in the literature.

as the mixing measure that are analogous to the results in Section 4.1. The upper bound is similar to Theorem 4.1 in that the error grows as $O(\ln^2 N)$ with fixed K, and decreases as $O(\ln K/K)$ for fixed N. The lower bounds, which are the analogues of Theorems 4.2 and 4.3, state that $K = \Omega(\ln N)$ is necessary for accurate approximations, and that truncation-based approximations are better than FSD_K, in the worst case. In comparison to existing results (Ishwaran and Zarepour, 2000, 2002), Theorem 1 of Ishwaran and Zarepour (2000) does not bound the distance between observational processes, so it is not directly comparable to our error bound. We improve upon Theorem 4 of Ishwaran and Zarepour (2002), whose upper bound on the FSD approximation error lacks an explicit dependence on K or N. So, unlike our bounds, that bound cannot be inverted to determine a sufficient approximation level K.

Hierarchical Dirichlet process. In modern applications such as text analysis, practitioners use additional hierarchical levels to capture group structure in observed data. In text, we might have D documents with N words in each. More, generally, we might have D groups (each indexed by d) with N observations (each indexed by n) each. We target the influential model of Wang, Paisley and Blei (2011); Hoffman et al. (2013), which is a variant of the hierarchical Dirichlet process (HDP; Teh et al., 2006) and which we refer to as the *modified HDP*. In the HDP, G is a population measure with $G \sim \mathrm{DP}(\omega, H)$. The measure for the d-th subpopulation is $G_d \,|\, G \sim \mathrm{DP}(\alpha, G)$; the concentrations ω and α are potentially different from each other. The modified HDP is defined in terms of the *truncated stick-breaking (TSB) approximation*:

Definition 4.5 (Stick-breaking approximation (Sethuraman, 1994))**.** For $i = 1, 2, \ldots, K-1$, let $v_i \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha)$. Set $v_K = 1$. Let $\xi_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$. Let $\psi_k \overset{\text{i.i.d.}}{\sim} H$, and $\Xi_K = \sum_{k=1}^{K} \xi_k \delta_{\psi_k}$. We denote the distribution of Ξ_K as $\mathrm{TSB}_K(\alpha, H)$.

In the modified HDP, the sub-population measure is distributed as $G_d \,|\, G \sim \mathrm{TSB}_T(\alpha, G)$. Wang, Paisley and Blei (2011) and Hoffman et al. (2013) set T to be small so that inference in the modified HDP is more efficient than in the HDP, since the number of parameters per group is greatly reduced. From a modeling standpoint, small T is a reasonable assumption since documents typically manifest a small number of topics from the corpus, with the total number depending on the document length and independent of corpus size. For completeness, the generative process of the modified HDP is

$$\begin{aligned} G &\sim \mathrm{DP}(\omega, H), \\ H_d \,|\, G &\overset{\text{i.i.d.}}{\sim} \mathrm{TSB}_T(\alpha, G) && \text{across } d, \\ \beta_{dn} \,|\, H_d &\overset{\text{indep}}{\sim} H_d(\cdot) && \text{across } d, n \\ W_{dn} \,|\, \beta_{dn} &\overset{\text{indep}}{\sim} f(\cdot \,|\, \beta_{dn}) && \text{across } d, n. \end{aligned} \tag{19}$$

H_d contains at most T distinct atom locations, all shared with the base measure G.

The finite approximation we consider replaces the population-level Dirichlet process

with FSD_K, keeping the other conditionals intact:[13]

$$\begin{aligned} G_K &\sim \mathrm{FSD}_K(\omega, H), \\ F_d \mid G_K &\overset{\text{i.i.d.}}{\sim} \mathrm{TSB}_T(\alpha, G_K) && \text{across } d, \\ \psi_{dn} \mid F_d &\overset{\text{indep}}{\sim} F_d(\cdot) && \text{across } d, n, \\ Z_{dn} \mid \psi_{dn} &\overset{\text{indep}}{\sim} f(\cdot \mid \psi_{dn}) && \text{across } d, n. \end{aligned} \tag{20}$$

Our contribution is analyzing the error of Eq. (20).

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 bound on the total variation distance between $P_{(N,D),\infty}$ and $P_{(N,D),K}$.

Theorem 4.6 (Upper bound for modified HDP)**.** *For some constants C', C'', C''', C'''' that depend only on ω,*

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

See Nguyen et al. (2023, Section S9) for explicit values of the constants as well as the theorem's proof. For fixed K, Theorem 4.6 is independent of N, the number of observations in each group, but scales with the number of groups D like $O(\mathrm{poly}(\ln D))$. For fixed D, the approximation error decreases to zero at rate no slower that $O(\ln K/K)$. The $O(\ln(DT))$ factor is related to the expected logarithmic growth rate of Dirichlet process mixture models (Arratia, Barbour and Tavaré, 2003, Section 5.2) in the following way. Since there are D groups, each manifesting at most T distinct atom locations from an underlying Dirichlet process prior, the situation is akin to generating DT samples from a common Dirichlet process prior. Hence, the expected number of unique samples is $O(\ln(DT))$. Similar to Theorem 4.1, we speculate that the $O(\ln^2(DT))$ factor can be improved to $O(\ln(DT))$. For error bounds of truncation-based approximations of hierarchical processes, such as the HDP, we refer to Lijoi, Prünster and Rigon (2020b, Theorem 1).

5 Conceptual benefits of finite approximations

Though approximation error lends itself more readily to analysis, ease-of-use considerations are often at the forefront of users' choice of finite approximation in practice. Therefore, we next compare AIFAs to TFAs in this dimension. We see that AIFAs offer more straightforward updates in approximate inference algorithms and easier implementation of parallelism.

[13] Our construction in Eq. (20) is slightly different from Eqs. 5.5 and 5.6 in Fox et al. (2010). Our document-level process F_d contains at most T topics from the underlying corpus; by contrast, the Fox et al. (2010) document-level process contains as many topics as the corpus-level process. However, the novelty of Eq. (20) is incidental since the replacement of the population-level DP with the FSD in the modified HDP is analogous to the DP case.

To reduce notation in this section, we let a term without subscripts represent the collection of all subscripted terms: $\rho := (\rho_k)_{k=1}^K$ denotes the collection of atom sizes, $\psi := (\psi_k)_{k=1}^K$ denotes the collection of atom locations, $x := (x_{n,k})_{k=1,n=1}^{K,N}$ denotes the latent trait counts of each observation,[14] and $y := (y_n)_{n=1}^N$ denotes the observed data. We use a dot to collect terms across the corresponding subscript: $x_{\cdot,k} := (x_{n,k})_{n=1}^N$ denotes trait counts across observations of the k-th trait. We next consider algorithms to approximate the posterior distribution $\mathbb{P}(\rho, \psi, x \,|\, y)$ of the finite approximation.

Gibbs sampling. When all latent parameters are continuous, Hamiltonian Monte Carlo methods are increasingly standard for performing Markov chain Monte Carlo (MCMC) posterior approximation (Hoffman and Gelman, 2014; Carpenter et al., 2017). However, due to the discreteness of the trait counts x, successful MCMC algorithms for CRMs or their approximations 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 $\mathbb{P}(\rho \,|\, x, \psi, y), \mathbb{P}(x \,|\, \psi, \rho, y)$, and $\mathbb{P}(\psi \,|\, x, \rho, y)$ are easy to simulate from.[15]

Different finite approximations with the same number of atoms K change only $\mathbb{P}(\rho)$ in the generative model. So, of the conditionals, we expect only $\mathbb{P}(\rho \,|\, x, \psi, y)$ to differ across finite approximations. We next show in Proposition 5.1 that the form of $\mathbb{P}(\rho \,|\, x, \psi, y)$ is particularly tractable for AIFAs. Then we will discuss how Gibbs derivations are substantially more involved for TFAs.

Proposition 5.1 (Conditional conjugacy of AIFA)**.** *Suppose the likelihood is an exponential family (Eq. (6)) and the AIFA prior ν_K is as in Corollary 3.3. Then the complete conditional of the atom sizes factorizes across atoms as:*

$$\mathbb{P}(\rho \mid x, \psi, y) = \prod_{k=1}^{K} \mathbb{P}(\rho_k \mid x_{\cdot,k}).$$

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

$$\mathbf{1}\{\rho \in U\}\rho^{c/K+\sum_{n=1}^N \phi(x_{n,k})-1} \exp\left(\langle \psi + \sum_{n=1}^{N} t(x_{n,k}), \mu(\rho)\rangle + (\lambda + N)[-A(\rho)]\right). \tag{21}$$

See Nguyen et al. (2023, Section S10.2) for the proof of Proposition 5.1. For common models — such as beta–Bernoulli, gamma–Poisson, and beta–negative binomial — we

[14] The usage of x in this section is different from the usage in the remaining sections: in Eq. (6), x is a single observation from the likelihood process.

[15] Because of the factorization $\mathbb{P}(x \,|\, \psi, \rho, y) = \prod_{n=1}^N \mathbb{P}(x_{n,\cdot} \,|\, \psi, \rho, y_n)$, Gibbs sampling over the finite approximation can be an appealing technique even when Gibbs sampling over the marginal process is not. In particular, the wall-time of a Gibbs iteration for the finite approximation can be small by drawing $\mathbb{P}(x_{n,\cdot} \,|\, \psi, \rho, y_n)$ in parallel. Meanwhile, any iteration to update the trait counts with the marginal process representation needs to sequentially process the data points, prohibiting speed up through parallelism.

see that the complete conditionals over AIFA atom sizes are in forms that are well known and easy to simulate.

There are many different types of TFAs, but typical TFA Gibbs updates pose additional challenges. Even when $\mathbb{P}(\rho)$ is easy to sample from, $\mathbb{P}(\rho \,|\, x)$ can be intractable, as we see in the following example.

Example 5.1 (Stick-breaking approximation (Broderick, Jordan and Pitman, 2012; Paisley, Carin and Blei, 2011))**.** Consider the TFA for $\mathrm{BP}(\gamma, \alpha, 0)$ given by

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

where $C_i \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\gamma)$, $V_{i,j}^{(l)} \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha)$ and $\psi_{i,j} \overset{\text{i.i.d.}}{\sim} H$. One can sample the atom sizes $V_{i,j}^{(i)} \prod_{l=1}^{i-1}(1 - V_{i,j}^{(l)})$. But there is no tractable way to sample from the conditional distribution $\mathbb{P}(\rho \,|\, x)$ because of the dependence on C_i as well as the entangled form of each ρ. Strategies to make sampling more tractable include introducing auxiliary round indicator variables r_k and marginalizing out the stick-breaking proportions (Broderick, Jordan and Pitman, 2012). However, the final model still contains one Gibbs conditional that is difficult to sample from (Broderick, Jordan and Pitman, 2012, Equation 37).

Other superposition-based approximations, like decoupled Bondesson or power-law (Campbell et al., 2019), present similar challenges due to the number of atoms per round variables C_i and the dependence among the atom sizes.

Mean-field variational inference (MFVI). Analogous to Hamiltonian Monte Carlo for MCMC, black-box variational methods are increasingly used for variational inference when the latent parameters are continuous (Ranganath, Gerrish and Blei, 2014; Kingma and Welling, 2014; Rezende, Mohamed and Wierstra, 2014; Burda, Grosse and Salakhutdinov, 2016; Kucukelbir et al., 2017; Bingham et al., 2018). Mean-field coordinate ascent updates (Wainwright and Jordan, 2008, Section 6.3) remain popular for cases with discrete variables, including the present trait counts x.[16]

MFVI posits a factorized distribution q to approximate the exact posterior. In our case, we approximate $\mathbb{P}(\rho, \psi, x \,|\, y)$ with $q(\rho, \psi, x) = q_\rho(\rho) q_\psi(\psi) q_x(x)$. We focus on $q_\rho(\rho)$. For fixed $q_\psi(\psi)$ and $q_x(x)$, the optimal q_ρ^* minimizes the (reverse) Kullback-Leibler divergence between the posterior and $q_\rho^* q_\psi q_x$:

$$q_\rho^* := \operatorname*{argmin}_{q_\rho} \mathrm{KL}\left(q_\rho(\cdot) q_\psi(\cdot) q_x(\cdot) \,||\, \mathbb{P}(\cdot, \cdot, \cdot \,|\, y)\right). \tag{22}$$

Our next result shows that q_ρ^* takes a convenient form when using AIFAs.

[16]When discrete latent variables are present, black-box variational methods typically utilize enumeration strategies to marginalize out the discrete variables. There exists a tradeoff between user time and wall time. The user time is small since there is no need to derive update equations, but the wall time can be large depending on the enumeration strategy.

Corollary 5.2 (AIFA optimal distribution is in exponential family)**.** *Suppose the likelihood is an exponential family (Eq. (6)) and the AIFA prior ν_K is as in Corollary 3.3. Then, the density of q_ρ^* is given by*

$$q_\rho^*(\rho) = \prod_k \widetilde{p}_k(\rho_k), \tag{23}$$

where each $\widetilde{p}_k$ has density at ρ_k proportional to

$$\mathbf{1}\{\rho_k \in U\}\rho_k^{c/K+\sum_n \mathbb{E}_{x_{n,k}\sim q_x}\phi(x_{n,k})-1} \exp\left\langle \begin{bmatrix} \psi + \sum_n \mathbb{E}_{x_{n,k}\sim q_x} t(x_{n,k}) \\ \lambda + N \end{bmatrix}, \begin{bmatrix} \mu(\rho_k) \\ -A(\rho_k) \end{bmatrix} \right\rangle \tag{24}$$

where $x_{n,k} \sim q_x$ denotes the marginal distribution of $x_{n,k}$ under $q_x(x)$.

That is, when using the AIFA, the optimal q_ρ^* factorizes across the K atoms, and each distribution is in the conjugate exponential family for the likelihood $\ell(x_{n,k} \mid \rho_k)$. Typically users will report summary statistics like means or variances of the variational approximations q_ρ^*. These are typically straightforward from the exponential family form.

The TFA case is much more complex and requires both more steps in the inference scheme as well as additional approximations. See Nguyen et al. (2023, Section S10.1) for two illustrative examples.

Parallelization. We end with a brief discussion on parallelization. In both Proposition 5.1 and Corollary 5.2, the update distribution for ρ factorizes across the K atoms. Hence, AIFA updates can be done in parallel across atoms, yielding speed-ups in wall-clock time, with the gains being greatest when there are many instantiated atoms. For TFAs, due to the complicating coupling among the atom rates, there is no such benefit from parallelization.

6 Empirical evaluation

In our experiments, we compare our AIFA constructions to TFAs and to other IFA constructions (Lee, James and Choi, 2016; Lee, Miscouridou and Caron, 2022) on a variety of synthetic and real-data examples. Even though our theory suggests better performance of TFAs than AIFAs for worst-case likelihoods, we find comparable performance of TFAs and AIFAs in predictive tasks (Sections 6.1 and 6.2). Likewise, we find comparable performance of AIFAs and alternative IFAs in predictive tasks (Section 6.3). However, we find that AIFAs can be used to learn model hyperparameters where alternative IFA approximations fail (Section 6.4). And we show that AIFAs can be used to learn model hyperparameters for new models, not previously explored in the BNP literature (Section 6.5).

In relation to prior studies, existing empirical work has compared IFAs and TFAs only for simpler models and smaller data sets (e.g., Doshi-Velez et al. (2009, Table 1,2) and Kurihara, Welling and Teh (2007, Figure 4)). Our comparison is grounded in models

with more levels and analyzes datasets of much larger sizes. For instance, in our topic modeling application, we analyze nearly 1 million documents, while the comparison in Kurihara, Welling and Teh (2007) utilizes only 200 synthetic data points.

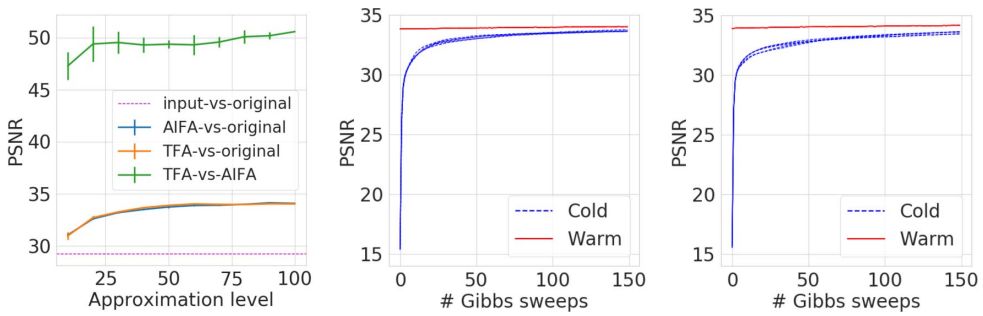
6.1 Image denoising with the beta–Bernoulli process

Our first experiments show comparable performance of the AIFA and TFA at an image denoising task with a CRM-based target model. We use MCMC for image denoising through dictionary learning because it is an application where finite approximations of BNP models — in particular the beta–Bernoulli process with $d = 0$ — have proven useful (Zhou et al., 2009). The observation likelihood in this dictionary learning model is not one of the worst cases in Section 4.1. We find that the performance of AIFAs and TFAs is comparable across K, and the posterior modes across TFA and AIFA models are similar to each other.

The goal of image denoising is to recover the original, noiseless image (e.g., Figure 1a) from a corrupted one (e.g., Figure 1b). The input image is first decomposed into small contiguous patches. The model assumes that each patch is a combination of latent *basis elements*. By estimating the coefficients expressing the combination, one can denoise the individual patches and ultimately the overall image. The beta–Bernoulli process allows simultaneous estimation of both basis elements and basis assignments. The number of extracted patches depends on both the patch size and the input image size. So even on the same input image, the analysis might process a varying number of "observations." The nonparametric nature of the beta–Bernoulli process sidesteps the cumbersome problem of calibrating the number of basis elements for these different data set sizes, which can be large even for a relatively small image; for a 256×256 image like Figure 1b, the number of extracted patches, N, is about 60,000. We quantify denoising quality by computing the peak signal-to-noise ratio (PNSR) between the original and the denoised image (Hore and Ziou, 2010). The higher the PNSR, the more similar the images.

(a) Original (b) Input, 24.64 dB (c) AIFA, 33.81 dB (d) TFA, 34.03 dB

Figure 1: AIFA and TFA denoised images have comparable quality. **(a)** The noiseless image. **(b)** The corrupted image. **(c,d)** Sample denoised images from finite models with $K = 60$. We report PSNR (in dB) with respect to the noiseless image.



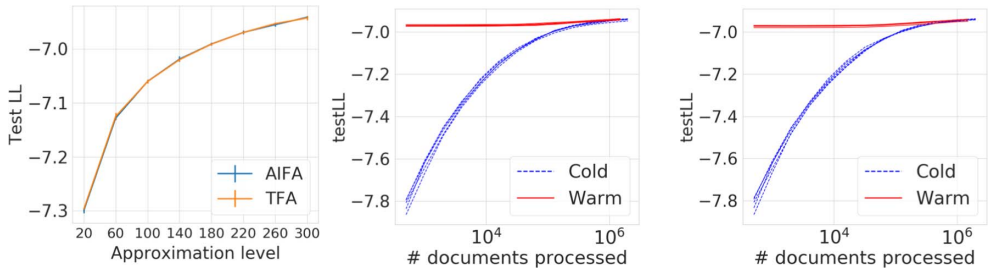
(a) Performance across K (b) TFA training (c) AIFA training

Figure 2: **(a)** Peak signal-to-noise ratio (PNSR) as a function of approximation level K. Error bars depict 1-standard-deviation ranges across 5 trials. **(b,c)** How PSNR evolves during inference across 10 trials, with 5 each starting from respectively cold or warm starts.

We use Gibbs sampling to approximate the posterior distributions. To ensure stability and accuracy of the sampler, patches (i.e., observations) are gradually introduced in epochs, and the sampler modifies only the latent variables of the current epoch's observations. See Nguyen et al. (2023, Section S11.1) for more details about the finite approximations, the hyperparameter settings, and the inference algorithm.

Figures 1c and 1d visually summarize the results of posterior inference for a particular image. We report experiments with other images in Nguyen et al. (2023, Section S12.1). Our results across all images indicate that the AIFA and TFA perform similarly, and both approximations perform much better than the baseline (i.e., the noisy input image). Figure 2 quantitatively confirms these qualitative findings; Figure 2a shows that, for approximation levels we considered, the PSNR between either the TFA or AIFA output image and the original image are always very similar and substantially higher (between 30 and 35) than the PSNR between the original and corrupted image (below 30). In fact, each TFA denoised image is more similar to the AIFA denoised image than to the original image; the PSNR between the TFA and AIFA outputs is about 50. We also see from Figure 2a that the quality of denoised images improves with increasing K. The improvement with K is largest for small K, and plateaus for larger values of K.

In addition to randomly initializing the latent variables at the beginning of the Gibbs sampler of one model ("cold start"), we can use the last configuration of latent variables visited in the other model as the initial state of the Gibbs sampler ("warm start"). In Figure 2b, the warm-start curve uses the output of inference with the AIFA as an initial value for inference with the TFA; similarly, the warm-start curve of Figure 2c uses the output with the TFA to initialize inference with the AIFA. For both approximations, $K = 60$. At the end of training, all latent variables for all patches have been assigned, so for the warm start experiment, we make all patches available from the start instead of gradually introducing patches. For both approximations, the Gibbs sampler initialized at the warm start visits candidate images that essentially have the same PSNR as the



(a) Performance across K (b) TFA training (c) AIFA training

Figure 3: **(a)** Test log-likelihood (testLL) as a function of approximation level K. Error bars show 1 standard deviation across 5 trials. **(b,c)** TestLL change during inference.

starting configuration; the PSNR values never deviate from the initial PSNR by more than 1%. The early iterates of the 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.[17] Each PSNR trace corresponds to a different set of initial values and simulation of the conditionals. The variation across the 5 warm-start trials is small; the variation across the 5 cold-start trials is larger but still quite small. In all, the modes of TFA posterior are good initializations for inference with the AIFA model, and vice versa.

6.2 Topic modelling with the modified hierarchical Dirichlet process

We next compare the performance of normalized AIFAs (namely, FSD_K) and TFAs (namely, TSB_K) in a DP-based model with additional hierarchy: the modified HDP from Section 4.2. As in Section 6.1, we find that the approximations perform similarly.

We use the modified HDP for topic modeling. We apply stochastic variational inference with mean-field factorization (Hoffman et al., 2013) to approximate the posterior over the latent topics. The training corpus consists of nearly one million documents from Wikipedia. We measure the quality of inferred topics via predictive log-likelihood on a set of 10,000 held-out documents. See Nguyen et al. (2023, Section S11.2) for complete experimental details.

Figure 3a shows that, as expected, the quality of the inferred topics improves as the approximation level grows. For a given approximation level, the quality of the topics learned using the TFA and the normalized AIFA are almost the same.

The warm start in this case corresponds to using variational parameters at the end of the other model's training. Figure 3b uses the outputs of inference with the normalized AIFA approximation as initial values for inference with the normalized TFA; similarly Figure 3c uses the TFA to initialize inference with the AIFA. We fix the number of topics to $K = 300$ and run 5 trials each with the cold start and warm start, respectively.

[17]Because the warm start represents the end of the training from the cold start with gradually introduced patches, the gap in final PSNR is due to the gradual patch introduction.

For both approximations, the test log-likelihood stays nearly the same for warm-start training iterates; the test log-likelihood for the iterates never deviate more than 0.5% from the initial value. The early iterates after the 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 corresponds to a different set of initial values and ordering of data batches processed. The variation across either cold starts or warm starts is small. So, in sum, the modes of the TFA posterior are good initializations for inference with the AIFA model, and vice versa.

6.3 Comparing predictions across independent finite approximations

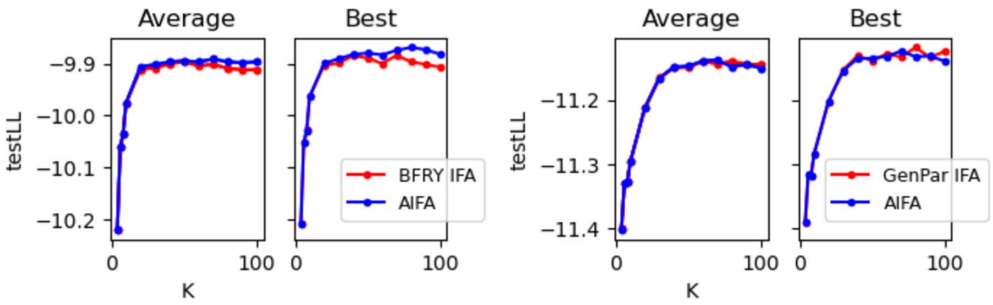
We next show that AIFAs have comparable predictive performance with other IFAs, namely the BFRY IFA and GenPar IFA. We consider a linear–Gaussian factor analysis model with the power-law beta–Bernoulli process (Griffiths and Ghahramani, 2011), where the AIFA, BFRY IFA, or GenPar IFA can be used directly.

Recall that the BFRY IFA applies only when the concentration hyperparameter is zero, and the GenPar IFA applies only when the concentration parameter is positive. We consider it a strength of the AIFA that it applies to both cases (and the negative range of the concentration hyperparameter) simultaneously. Nonetheless, we here generate two separate synthetic datasets: one to compare the BFRY IFA with the AIFA and one to compare the GenPar IFA with the AIFA. In each case, we generate 2,000 data points from the full CRM model with a discount of $d = 0.6$. We use 1,500 for training and report predictive log-likelihood on the 500 held-out data points. For posterior approximation, we use automatic differentiation variational inference as implemented in Pyro (Bingham et al., 2018). To isolate the effect of the approximation type, we use "ideal" initialization conditions: we initialize the variational parameters using the latent features, assignments, and variances that generated the training set. See Nguyen et al. (2023, Section S11.3) for more details about the BRFY IFA, GenPar IFA, and the approximate inference scheme. Figure 4a shows that across approximation levels K, the predictive performances of the AIFA and BFRY IFA are similar. Likewise, Figure 4b shows that the predictive performance of the AIFA and GenPar IFA are similar.

6.4 Discount estimation

We next show that AIFAs can reliably recover the beta process discount hyperparameter d, which governs the power law growth in the number of features. By contrast, we show that the BFRY IFA or GenPar IFA struggle at this task. In Nguyen et al. (2023, Section S12.3), we show that the AIFA can also reliably estimate the mass and concentration hyperparameters.

We generate a synthetic dataset so that the ground truth hyperparameter values are known. The data takes the form of a binary matrix X, with N rows and $\tilde{K}$ columns. We generate X from an Indian buffet process prior; recall that the Indian buffet process is the marginal process of a beta process CRM paired with Bernoulli likelihood. To learn the hyperparameter values with an AIFA, we maximize the marginal likelihood of



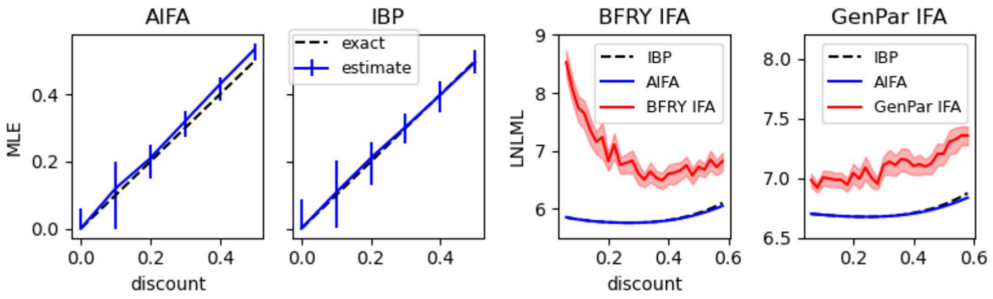
(a) BFRY IFA versus AIFA (b) GenPar IFA vs AIFA

Figure 4: **(a)** The left panel shows the average predictive log-likelihood of the AIFA (blue) and BFRY IFA (red) as a function of the approximation level K; the average is across 10 trials with different random seeds for the stochastic optimizer. The right panel shows highest predictive log-likelihood across the same 10 trials. **(b)** The panels are analogous to **(a)**, except the GenPar IFA is in red.

the observed matrix X implied by the AIFA. In particular, we compute the marginal likelihood by integrating the Bernoulli likelihood $\mathbb{P}(x_{n,k} \mid \theta_k)$ over θ_k distributed as the K-atom AIFA ν_K. To quantify the variability of the estimation procedure, we generate 50 feature matrices and compute the maximum likelihood estimate for each of these 50 trials. See Nguyen et al. (2023, Section S11.4) for more experimental details.

Figure 5a shows that we can use an AIFA to estimate the underlying discount for a variety of ground-truth discounts. Since the estimates and error bars are similar whether we use the AIFA (left) or full nonparametric process (right), we conclude that using the AIFA yields comparable inference to using the full process.

In theory, the marginal likelihood of the BFRY IFA can also be used to estimate the discount, but in practice we find that this approach is not straightforward and can yield unreliable estimates. At the time of writing, such an experiment had not yet been attempted; Lee, James and Choi (2016) focus on clustering models and do not discuss strategies to estimate any hyperparameter in a feature allocation model with a BFRY IFA. We are not aware of a closed-form formula for the marginal likelihood. Default schemes to numerically integrate $\mathbb{P}(0 \mid \theta_k)$ against the BFRY prior for θ_k fail because of overflow issues. $(K\Gamma(d)d/\gamma)^{1/d}$ is typically very large, especially for small d. Due to finite precision, $1 - \exp\left(-(Kd/\gamma)^{1/d}\frac{\theta}{1-\theta}\right)$ evaluates to 1 on the quadrature grid used by numerical integrators (Piessens et al., 2012). In this case, Eq. (4) behaves as θ^{-d-1} near 0, and thus the integral over θ diverges. To create the left panel of Figure 5b, we view the marginal likelihood as an expectation and construct Monte Carlo estimates; we draw 10^5 BFRY samples to estimate the marginal likelihood, and we take the estimate's logarithm as an approximation to the log marginal likelihood (red line). To quantify the uncertainty, we draw 100 batches of 10^5 samples (light red region). Even for this large number of Monte Carlo samples, the estimated log marginal likelihood curve is too noisy to be useful for hyperparameter estimation. By comparison, we can compute the log



(a) Maximum likelihood estimates (b) Log negative log marginal likelihood

Figure 5: **(a)** We estimate the discount by maximizing the marginal likelihood of the AIFA (left) or the full process (right). The solid blue line is the median of the estimated discounts, while the lower and upper bounds of the error bars are the 20% and 80% quantiles. The black dashed line is the ideal value of the estimated discount, equal to the ground-truth discount. **(b)** In each panel, the solid red line is the average log of negative log marginal likelihood (LNLML) across batches. The light red region depicts two standard errors in either direction from the mean.

marginal likelihood analytically for the IBP (dashed black line); it is much smoother and features a clear minimum. Moreover, we can compute the AIFA log marginal likelihood via numerical integration (solid blue line); it is also very smooth and features a clear minimum.

We again consider the BFRY IFA and GenPar IFA separately and generate separate simulated data for each case due to their disjoint assumptions; we generate data with concentration $\alpha = 0$ for the BFRY IFA and with $\alpha > 0$ for the GenPar IFA. An experiment to recover a discount hyperparameter with the GenPar IFA, analogous to the experiment above with the BFRY IFA, has also not previously been attempted. There is no analytical formula for the GenPar IFA marginal likelihood, and we again encounter overflow when trying numerical integration. Therefore, we resort to Monte Carlo; we find that estimates of the log marginal likelihood are too noisy for practical use in recovering the discount (the right panel of Figure 5b).

6.5 Dispersion estimation

Finally, we show that the AIFA can straightforwardly be adapted to estimate hyperparameters in other BNP processes, not just the beta process. In particular we show that AIFAs can be used to learn the dispersion parameter τ in the novel Xgamma–CMP process that we introduced in Example 3.4. We consider a well-known application of BNP trait-allocation models to matrix-factorization–based topic modeling (Roychowdhury and Kulis, 2015). The observed data is a count matrix X, with N rows, representing documents, and V columns, representing vocabulary words. We adjust the model of Roychowdhury and Kulis (2015) to use the Xgamma–CMP process of Example 3.4 instead of a gamma–Poisson process. The added flexibility of τ allows modeling trait

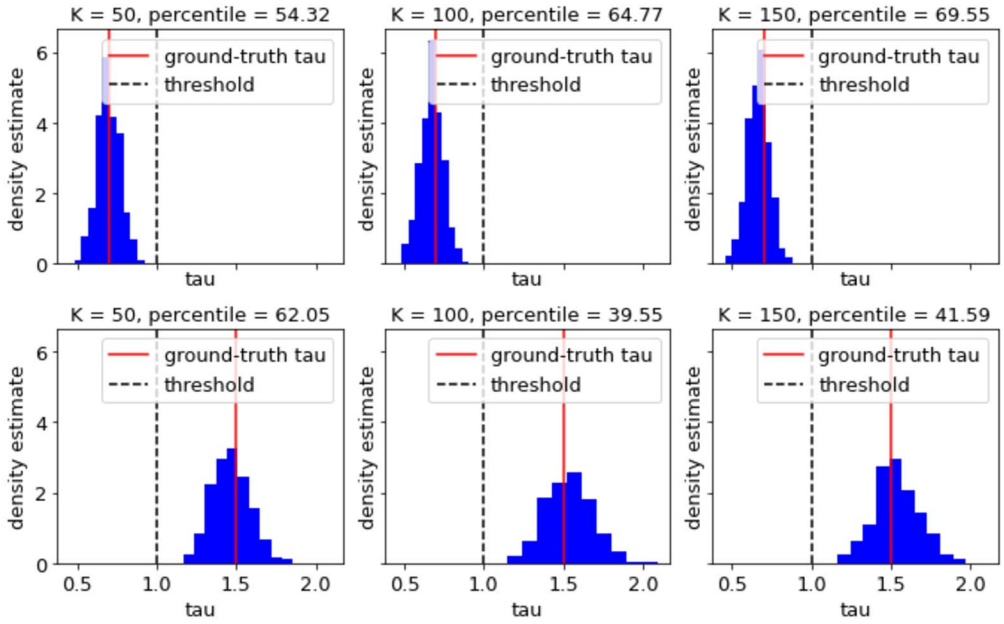


Figure 6: Blue histograms show posterior density estimates for τ from MCMC draws. The ground-truth τ (solid red line) is 0.7 in the overdispersed case (upper row) and 1.5 in the underdispersed case (lower row). The threshold $\tau = 1$ (dashed black line) marks the transition from overdispersion ($\tau < 1.0$) to underdispersion ($\tau > 1.0$). The percentile in each panel's title is the percentile where the ground truth τ falls in the posterior draws. The approximation size K of the AIFA increases in the plots from left to right.

count distributions that are over- or under-dispersed, which cannot be done with the gamma-Poisson process.

To have a notion of ground truth, we generate synthetic data (with $N = 600$) from a large AIFA (with $K = 500$) of the Xgamma–CMP process, which is a good approximation of the BNP limit.[18] In each set of experiments, the data are overdispersed ($\tau < 1$) or underdispersed ($\tau > 1$). In this case, we take a Bayesian approach to estimating τ, and put a uniform prior on $\tau \in (0, 100]$ since τ must be strictly positive. For smaller values of K ($K = 50$ to $K = 150$), we approximate the posterior for the K-atom AIFA using Gibbs sampling. See Nguyen et al. (2023, Section S11.5) for more details about the experimental setup.

Figure 6 shows that the posterior approximation agrees with the ground truth on the dispersion type (over or under) in each case. We also see from the figures that the 95% credible intervals contain the ground-truth τ value in each case.

[18]For the chosen number of documents N, let the number of traits with positive count be $\widehat{K}$. There is no noticeable difference in the distribution of $\widehat{K}$ between $K = 500$ and $K > 500$. The rates of the inactive (zero count) traits are smaller than $1/N$.

7 Discussion

We have provided a general construction of automated independent finite approximations (AIFAs) for completely random measures and their normalizations. Our construction provides novel finite approximations not previously seen in the literature. For processes without power-law behavior, we provide approximation error bounds; our bounds show that we can ensure accurate approximation by setting the number of atoms K to be (1) logarithmic in the number of observations N and (2) inverse to the error tolerance ϵ. We have discussed how the independence and automatic construction of AIFA atom sizes lead to convenient inference schemes. A natural competitor for AIFAs is a truncated finite approximation (TFA). We show that, for the worst case choice of observational likelihood and the same K, AIFAs can incur larger error than the corresponding TFAs. However, in our experiments, we find that the two methods have essentially the same performance in practice. Meanwhile, AIFAs are overall easier to work with than TFAs, whose coupled atoms complicate the development of inference schemes. Future work might extend our error bound analysis to conjugate exponential family CRMs with power-law behavior. An obstacle to upper bounds for the positive-discount case is the verification of the clauses in Condition 1. In the positive-discount case, the functions h and $M_{n,x}$, which describe the marginal representation of the nonparametric process, take forms that are straightforwardly amenable to analysis. But the function $\widetilde{h}$, which describes the finite approximations, is complex. In general, $\widetilde{h}$ is equal to the ratio of two normalization constants of different AIFAs. The normalization constants can be computed numerically. However, to make theoretical statements such as the clauses in Condition 1, we need to prove their smoothness properties. Another direction is to tighten the error upper bound by focusing on specific, commonly-used observational likelihoods — in contrast to the worst-case analysis we provide here. Finally, more work is required to directly compare the size of error in the finite approximation to the size of error due to approximate inference algorithms such as Markov chain Monte Carlo or variational inference.

Supplementary Material

Supplementary Material: Independent finite approximations for Bayesian nonparametric inference (DOI: 10.1214/23-BA1385SUPP; .pdf). The supplementary materials contain detailed proofs and more details on the experiments.

References

Acharya, A., Ghosh, J. and Zhou, M. (2015). Nonparametric Bayesian factor analysis for dynamic count matrices. In *International Conference on Artificial Intelligence and Statistics*. 1192, 1196

Arbel, J., De Blasi, P. and Prünster, I. (2019). Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis* **14** 1201–1219. MR4136558. doi: https://doi.org/10.1214/18-BA1127. 1190

Arbel, J. and Prünster, I. (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing* **27** 3–17. MR4136558. doi: https://doi.org/10.1214/18-BA1127. 1188, 1191

Arratia, R., Barbour, A. D. and Tavaré, S. (2003). *Logarithmic combinatorial structures: a probabilistic approach* **1**. European Mathematical Society. MR2032426. doi: https://doi.org/10.4171/000. 1207

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. MR0755837. doi: https://doi.org/10.1017/S0305004100061806. 1204

Bertoin, J., Fujita, T., Roynette, B. and Yor, M. (2006). On a particular class of self-decomposable random variables: the durations of Bessel excursions straddling independent exponential times. *Probability and Mathematical Statistics* **26** 315–366. MR2325310. 1195

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P. and Goodman, N. D. (2018). Pyro: deep universal probabilistic programming. *Journal of Machine Learning Research*. 1209, 1214

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. MR2606082. doi: https://doi.org/10.1145/1667053.1667056. 1187

Bondesson, L. (1982). On Simulation from Infinitely Divisible Distributions. *Advances in Applied Probability* **14** 855–869. MR0677560. doi: https://doi.org/10.2307/1427027. 1204, 1205

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* **31** 929–953. MR1747450. doi: https://doi.org/10.1239/aap/1029955251. 1193

Broderick, T., Jordan, M. I. and Pitman, J. (2012). Beta processes, stick-breaking and power laws. *Bayesian analysis* **7** 439–476. MR2934958. doi: https://doi.org/10.1214/12-BA715. 1192, 1209

Broderick, T., Wilson, A. C. and Jordan, M. I. (2018). Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli* **24** 3181–3221. MR3788171. doi: https://doi.org/10.3150/16-BEJ855. 1188, 1193, 1196, 1198, 1200

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. 1192, 1193, 1196, 1198

Burda, Y., Grosse, R. B. and Salakhutdinov, R. (2016). Importance Weighted Autoencoders. In *International Conference on Learning Representations*. 1209

Campbell, T., Cai, D. and Broderick, T. (2018). Exchangeable trait allocations. *Electronic Journal of Statistics* **12** 2290–2322. MR3832093. doi: https://doi.org/10.1214/18-EJS1455. 1191

Campbell, T., Huggins, J. H., How, J. P. and Broderick, T. (2019). Truncated random measures. *Bernoulli* **25** 1256–1288. MR3920372. doi: https://doi.org/10.3150/18-bej1020. 1188, 1191, 1192, 1200, 1204, 1205, 1209

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association* **106** 1528–1539. MR2896854. doi: https://doi.org/10.1198/jasa.2011.tm10552. 1198

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** 1–32. 1208

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D. and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* **26** 403–413. MR3640196. doi: https://doi.org/10.1080/10618600.2016.1172487. 1188

Devroye, L. and James, L. (2014). On simulation and properties of the stable law. *Statistical methods & applications* **23** 307–343. MR3233961. doi: https://doi.org/10.1007/s10260-014-0260-0. 1195

Doshi-Velez, F., Miller, K. T., Van Gael, J. and Teh, Y. W. (2009). Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*. 1188, 1191, 1193, 1197, 1200, 1210

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230. MR0350949. 1199, 1205

Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics* **43** 1634–1643. MR0373022. doi: https://doi.org/10.1214/aoms/1177692395. 1193

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. MR2840185. doi: https://doi.org/10.1214/10-AOAS395. 1187, 1189, 1207

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. 1208

Griffiths, T. L. and Ghahramani, Z. (2011). The Indian buffet process: an introduction and review. *Journal of Machine Learning Research* **12** 1185–1224. MR2804598. 1197, 1214

Hjort, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* **18** 1259–1294. MR1062708. doi: https://doi.org/10.1214/aos/1176347749. 1193

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. MR3214779. 1208

Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research* **14** 1303–1347. MR3081926. 1206, 1213

Hore, A. and Ziou, D. (2010). Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition* 2366–2369. IEEE. 1211

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96** 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 1205

Ishwaran, H. and Zarepour, M. (2000). Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-Parameter Process Hierarchical Models. *Biometrika* **87** 371–390. MR1782485. doi: https://doi.org/10.1093/biomet/87.2.371. 1206

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics* **30** 269–283. MR1926065. doi: https://doi.org/10.2307/3315951. 1200, 1206

James, L. F. (2013). Stick-breaking PG(α,ζ)-generalized gamma processes. *Available at arXiv:1308.6570v3*. 1193

James, L. F. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian Buffet Process priors. *The Annals of Statistics* **45** 2016–2045. MR3718160. doi: https://doi.org/10.1214/16-AOS1517. 1188

James, L. F., Lijoi, A. and Prünster, I. (2009). Posterior Analysis for Normalized Random Measures with Independent Increments. *Scandinavian Journal of Statistics* **36** 76–97. MR2508332. doi: https://doi.org/10.1111/j.1467-9469.2008.00609.x. 1189, 1191

Johnson, M. J. and Willsky, A. S. (2013). Bayesian nonparametric hidden semi-Markov models. *Journal of Machine Learning Research* **14** 673–701. MR3033344. 1189

Kallenberg, O. (2002). *Foundations of modern probability*, 2nd ed. Springer, New York. MR1876169. doi: https://doi.org/10.1007/978-1-4757-4015-8. 1192

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*. 1209

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics* **21** 59–78. MR0210185. 1190

KINGMAN, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society B* **37** 1–22. MR0368264. 1193, 1205

KINGMAN, J. (1992). *Poisson Processes* **3**. Clarendon Press. MR1207584. 1190

KORWAR, R. M. and HOLLANDER, M. (1972). Contributions to the theory of Dirichlet processes. *The Annals of Probability* **1** 705–711. MR2622286. 1203

KUCUKELBIR, A., TRAN, D., RANGANATH, R., GELMAN, A. and BLEI, D. M. (2017). Automatic Differentiation Variational Inference. *Journal of Machine Learning Research* **18** 1–45. MR3634881. 1209

KURIHARA, K., WELLING, M. and TEH, Y. W. (2007). Collapsed variational Dirichlet process mixture models. In *International Joint Conference on Artificial Intelligence*. 1189, 1205, 1210, 1211

LE CAM, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* **10** 1181–1197. MR0142174. 1204

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*. 1188, 1189, 1190, 1192, 1195, 1210, 1215

LEE, J., MISCOURIDOU, X. and CARON, F. (2022). A unified construction for series representations and finite approximations of completely random measures. *Bernoulli*. MR4580911. doi: https://doi.org/10.3150/22-bej1536. 1188, 1189, 1190, 1192, 1195, 1210

LIJOI, A., PRÜNSTER, I. and RIGON, T. (2020a). The Pitman–Yor multinomial process for mixture modelling. *Biometrika* **107** 891–906. MR4186494. doi: https://doi.org/10.1093/biomet/asaa030. 1190

LIJOI, A., PRÜNSTER, I. and RIGON, T. (2020b). Sampling Hierarchies of Discrete Random Structures. *Statistics and Computing* **30** 1591–1607. MR4156338. doi: https://doi.org/10.1007/s11222-020-09961-7. 1207

LIJOI, A. and PRÜNSTER, I. (2010). *Models beyond the Dirichlet process*. In *Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics* 80–136. Cambridge University Press. MR2730661. 1190

LIJOI, A., PRÜNSTER, I. and RIGON, T. (2023). Finite-dimensional Discrete Random Structures and Bayesian Clustering. *Journal of the American Statistical Association* **0** 1–13. 1188, 1189, 1192

NGUYEN, T. D., HUGGINS, J., MASOERO, L., MACKEY, L. and BRODERICK, T. (2023). Supplement to "Independent Finite Approximations for Bayesian Nonparametric Inference". *Bayesian Analysis*. doi: https://doi.org/10.1214/23-BA1385SUPP. 1193, 1194, 1197, 1198, 1199, 1200, 1202, 1203, 1204, 1205, 1207, 1208, 1210, 1212, 1213, 1214, 1215, 1217

ORBANZ, P. (2010). Conjugate projective limits. *Available at arXiv:1012.0363v2*. 1188

Paisley, J., Blei, D. M. and Jordan, M. I. (2012). Stick-breaking beta processes and the Poisson process. In *International Conference on Artificial Intelligence and Statistics*. 1188, 1191, 1193, 1200

Paisley, J. and Carin, L. (2009). Nonparametric factor analysis with beta process priors. In *International Conference on Machine Learning*. 1192, 1197

Paisley, J., Carin, L. and Blei, D. (2011). Variational inference for stick-breaking beta process priors. In *International Conference on Machine Learning*. 1209

Palla, K., Knowles, D. A. and Ghahramani, Z. (2012). An infinite latent attribute model for network data. In *International Conference on Machine Learning*. 1187

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W. and Kahaner, D. K. (2012). *QUADPACK: a subroutine package for automatic integration* **1**. Springer Science & Business Media. MR0712135. doi: https://doi.org/10.1007/978-3-642-61786-7. 1194, 1215

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102** 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 1198

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 1190

Ranganath, R., Gerrish, S. and Blei, D. M. (2014). Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*. 1209

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics* **31** 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 1189, 1191

Rezende, D. J., Mohamed, S. and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*. 1209

Roychowdhury, A. and Kulis, B. (2015). Gamma processes, stick-breaking, and variational inference. In *International Conference on Artificial Intelligence and Statistics*. 1188, 1191, 1196, 1216

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4** 639–650. MR1309433. 1199, 1206

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54** 127–142. MR2134602. doi: https://doi.org/10.1111/j.1467-9876.2005.00474.x. 1197

Teh, Y. W. and Görür, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*. 1192

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. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 1206

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*. 1193

Titsias, M. (2008). The infinite gamma-Poisson feature model. In *Advances in Neural Information Processing Systems*. 1193

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. and SciPy 1. 0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17** 261–272. 1194

Wainwright, M. J. and Jordan, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning* **1** 1–305. 1209

Wang, C., Paisley, J. and Blei, D. (2011). Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics*. 1206

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*. 1193, 1211

Zhou, M., Hannah, L., Dunson, D. and Carin, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *International Conference on Artificial Intelligence and Statistics*. 1196, 1198