Independent versus truncated finite approximations for Bayesian nonparametric inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

Bayesian nonparametric models based on completely random measures (CRMs) offer flexibility when the number of clusters or latent components in a data set is unknown. However, managing the infinite dimensionality of CRMs often leads to slow computation during inference. Practical inference typically relies on either integrating out the infinite-dimensional parameter or using a *finite approximation*: a truncated finite approximation (TFA) or an independent finite approximation (IFA). The atom weights of TFAs are constructed sequentially, while the atoms of IFAs are independent, which facilitates more convenient inference schemes. While the approximation error of TFA has been systematically addressed, there has not yet been a similar study of IFA. We quantify the approximation error between IFAs and two common target nonparametric priors (beta-Bernoulli process and Dirichlet process mixture model) and prove that, in the worst-case, TFAs provide more component-efficient approximations than IFAs. However, in experiments on image denoising and topic modeling tasks with real data, we find that the error of Bayesian approximation methods overwhelms any finite approximation error, and IFAs perform very similarly to TFAs.

1 Introduction

Many data analyses can be seen as discovering a latent set of traits in a population. For instance, we might recover topics or themes from scientific papers, ancestral populations from genetic data, interest groups from social network data, or unique speakers across audio recordings of many meetings [Palla et al., 2012, Blei et al., 2010, Fox et al., 2010]. In all of these cases, we might reasonably expect the number of latent traits present in a data set to grow with the size of the data. A powerful modelling option is to choose a single prior that naturally yields different expected numbers of traits for different numbers of data points. In theory, *Bayesian nonparametrics* (BNP) provides a rich set of priors with this desirable property thanks to a latent countable infinity of traits, so that there are always more traits to reveal through the accumulation of more data. This latent, infinite-dimensional parameter presents a major practical challenge, however. 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.

To apply BNP, a common technique is approximating the infinite-dimensional prior with a finite-dimensional prior that essentially replaces the infinite collection of random traits by a finite subset of "likely" traits. Unlike a fixed, finite-dimensional prior across all data set sizes, this finite-dimensional prior is seen as an approximation to the BNP prior and thereby its cardinality is informed directly by the BNP prior.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Finite approximations can be divided into two approaches. On the one hand, we call those based on truncations of the random measures underlying the nonparametric prior [Doshi-Velez et al., 2009, Paisley et al., 2012, Roychowdhury and Kulis, 2015, Campbell et al., 2019] *truncated finite approximations* (TFAs) and refer to Campbell et al. [2019] for a thorough study of constructions for TFAs. On the other hand, *independent finite approximations* (IFAs) consist of independent and identical (i.i.d.) representations of the traits together with their rates within the population [Kurihara et al., 2007, Saria et al., 2010, Fox et al., 2010, Johnson and Willsky, 2013]; we refer to Huggins et al. [2017] for a recent study of constructions for IFAs. The IFA approach has the potential to be simpler to incorporate in a complex hierarchical model, to exhibit improved mixing, and to be amenable to parallelizing computation during inference.

Our work aims to determine which approach, TFA or IFA, is the better one. We focus on two Bayesian nonparametric target models: the beta process [Hjort, 1990, Thibaux and Jordan, 2007, Teh and Görür, 2009, Broderick et al., 2012] and Dirichlet process mixture models [Ferguson, 1973, Sethuraman, 1994]. In general, between two approximations with equal accuracy, we prefer the approximation with fewer atoms since it will use fewer computational resources during inference. We show that, in the worst case, TFAs are more component-efficient than IFAs. However, experiments with image denoising and topic modeling tasks reveal that IFA and TFA have very similar performance across the number of instantiated components. Future work should analyze the average-case behavior of IFA, or the additional sources of error that come from approximate Bayesian inference.

In what follows, we first review the role of random measures in Bayesian nonparametrics and finite approximations. We then quantify the effect of replacing the infinite-dimensional priors with an IFA, providing interpretable error bounds with explicit dependence on the size of the approximation and the data cardinality. Finally, we confirm through experiments with image denoising and topic modeling that IFAs and TFAs perform similarly on applied problems.

2 Nonparametric models and finite approximations

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

Completely random measures. Most common BNP priors can be conveniently formulated as *completely random measures* (CRMs) or normalizations of CRMs. CRMs are constructed from Poisson point processes. 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 an infinite number of rates $(\theta_i)_{i=1}^{\infty}$, $\theta_i \in \mathbb{R}_+$, having an almost surely finite sum $\sum_{i=1}^{\infty} \theta_i < \infty$. We assume throughout that $\psi_i \in \Psi$ for some space Ψ and $\psi_i \overset{\text{i.i.d.}}{\sim} H$ for some diffuse distribution H. H serves as a prior on the trait values; for instance, in topic modeling, each topic is a probability vector in the simplex of vocabulary words, and it is typical to use $H = \mathsf{Dir}$. In general, the resulting measure Θ is a *completely random measure* (CRM) [Kingman, 1967]. As shorthand, we will write $\mathrm{CRM}(H, \nu)$ for the completely random measure generated as just described: $\Theta := \sum_i \theta_i \delta_{\psi_i} \sim \mathrm{CRM}(H, \nu)$. The corresponding *normalized CRM* (NCRM) is $\Xi := \Theta/\Theta(\Psi)$, which is a discrete probability measure. The set of atom locations of Ξ is the same as that of Θ, while the atom sizes are normalized $\Xi = \sum_i \xi_i \delta_{\psi_i}$ where $\xi_i = \theta_i / (\sum_j \theta_j)$.

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

A *truncated finite approximation* (TFA) [Doshi-Velez et al., 2009, Paisley et al., 2012, Roychowdhury and Kulis, 2015] requires constructing an ordering on the indices of the atoms sizes $(\theta_i)_{i=1}^{\infty}$ such that θ_i is a function of some auxiliary random variables $\xi_1, \ldots, \xi_i$; hence, θ_{i+1} reuses the same auxiliary randomness as θ_i as well as an additional random variable ξ_{i+1}. These approximations are nested; in general, the approximation quality increases with K, and to refine existing truncations, it suffices to generate the next terms in the sequence.

An *independent finite approximation* (IFA) involves choosing a sequence of probability measures $\nu_1, \nu_2, \ldots$ such that for any approximation level K, we choose $\theta_1, \ldots, \theta_K \overset{\text{i.i.d.}}{\sim} \nu_K$. The ν_K are chosen in such a way that $\Theta_K \overset{\mathcal{D}}{\Longrightarrow} \Theta$ as $K \to \infty$; that is, the IFAs converge in distribution to the CRM. The pros and cons of the IFA invert those of the TFA: the atoms are now i.i.d., potentially making inference easier, but a completely new approximation must be constructed if K changes.

For the normalized atom sizes $\xi_i = \theta_i / \sum_j \theta_j$, finite approximations also involve random measures with finite support $\Xi_K = \sum_{i=1}^{K} \xi_i \delta_{\psi_i}$. TFAs can be defined in one of two ways. In the first approach, the TFA corresponding to the CRM can be normalized to form the approximation of the NCRM [Campbell et al., 2019]. The second approach instead directly constructs an ordering over the sequence $(\xi_i)_{i=1}^{\infty}$ and truncates this representation [Ishwaran and James, 2001, Blei and Jordan, 2006]. We consider a single way to construct IFAs in the normalized case; we take the IFA approximation for the unnormalized CRM and normalize it to form the approximation of the corresponding NCRM.

Beta-Bernoulli model. The first model under focus is the beta process [Hjort, 1990, Thibaux and Jordan, 2007]. We denote its distribution as $\mathrm{BP}(\gamma, \alpha)$, with scale parameter $\alpha > 0$, mass parameter $\gamma > 0$, and rate measure $\nu(\mathrm{d}\theta) = \gamma\alpha \mathbb{1}[\theta \leq 1]\theta^{-1}(1-\theta)^{\alpha-1}\mathrm{d}\theta$. The beta process prior on Θ is combined with a Bernoulli likelihood that generates trait counts for each data point. A collection of conditionally independent observations $X_{1:N}$ given Θ are distributed according to the *likelihood process* $\mathrm{LP}(l, \Theta)$ — i.e., $X_n := \sum_i x_{ni}\delta_{\psi_i} \overset{\text{i.i.d.}}{\sim} \mathrm{LP}(\mathrm{Ber}, \Theta)$ — if $x_{ni} \sim \mathrm{Ber}(\cdot \mid \theta_i)$ independently across i and i.i.d. across n. Since the trait counts are typically latent in a full generative model specification, define the observed data $Y_n \mid X_n \overset{\text{indep}}{\sim} f(\cdot \mid X_n)$ for a probability kernel f. The target nonparametric model can thus be summarized as

$$\Theta \sim \mathrm{BP}(\gamma, \alpha; H), \quad X_n|\Theta \overset{\text{i.i.d.}}{\sim} \mathrm{LP}(\mathrm{Ber}; \Theta), \quad Y_n|X_n \overset{indep}{\sim} f(\cdot \mid X_n), \quad n = 1, 2, \ldots, N. \tag{1}$$

Dirichlet process mixture model. The second model under focus is the Dirichlet process [DP] [Ferguson, 1973, Sethuraman, 1994] — which is the normalization of a non-power law gamma process. The Dirichlet process is one of the most widely used nonparametric priors. The gamma process CRM has rate measure $\nu(\mathrm{d}\theta) = \gamma\lambda\theta^{-1}e^{-\lambda\theta}\mathrm{d}\theta$. We denote its distribution as $\Gamma\mathrm{P}(\gamma, \lambda)$. The normalization of $\Gamma\mathrm{P}(\gamma, 1)$ is a Dirichlet process with mass parameter γ [Kingman, 1975, Ferguson, 1973]. We consider Dirichlet process mixture models [Antoniak, 1974] with latent clusters X_n mapping to observations Y_n through the observational likelihood f:

$$\Theta \sim \mathrm{DP}(\alpha; H), \quad X_n|\Theta \overset{\text{i.i.d.}}{\sim} \Theta, \quad Y_n|X_n \overset{indep}{\sim} f(\cdot \mid X_n), \quad n = 1, 2, \ldots, N. \tag{2}$$

3 Theoretical error bounds

In this section, we derive novel upper and lower bounds on the approximation error incurred by IFA as a function of the approximation level K and data cardinality N. The upper bound holds for any

observational likelihood f mapping from latents to observations, while the lower bound holds for "bad" f. Juxtaposed with upper bounds for TFA, they reveal that in the worst case, for the same K, the IFA error is much larger than TFA error.

3.1 Beta-Bernoulli model

We first discuss how approximation error is defined. Let FA_K be some finite approximation at level K. After replacing the beta process with FA_K, we have the following generative process

$$\Theta_K \sim \mathrm{FA}_K, \quad Z_n|\Theta_K \overset{\text{i.i.d.}}{\sim} \mathrm{LP}(\mathrm{Ber};\Theta_K), \quad W_n|Z_n \overset{indep}{\sim} f(.|Z_n), \quad n = 1, 2, \ldots, N. \tag{3}$$

Let $P_{N,\infty}$ be the distribution of the observations $Y_{1:N}$ in Equation (1), and $P_{N,K}$ be the distribution of the observations $W_{1:N}$ in Equation (3). We define *approximation error* to be the total variation distance $d_{TV}(P_{N,K}, P_{N,\infty})$ between the two observational process [Ishwaran and Zarepour, 2002, Doshi-Velez et al., 2009, Paisley et al., 2012, Campbell et al., 2019]. Recall that total variation distance is the supremum difference in probability mass over measurable sets.

Paisley and Carin [2009], Huggins et al. [2017] constructed IFAs where each ν_K is a proper beta distribution:

$$\mathrm{IFA}_K := \sum_{i=1}^{K} \xi_{K,i}\delta_{\psi_{K,i}} \text{ where } \xi_{K,i} \overset{i.i.d.}{\sim} \mathsf{Beta}(\gamma\alpha/K, \alpha) \text{ and } \psi_{K,i} \overset{i.i.d.}{\sim} H. \tag{4}$$

We now quantify the approximation error for IFA_K. Let $P_{N,K}$ be the observational process using IFA_K in Equation (4). Theorem 3.1 upper bounds the approximation error.

Theorem 3.1 (Upper bound for beta-Bernoulli)**.** *There exist positive constants C', C'', C''' depending only on γ and α such that*

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

Theorem 3.1 states that the IFA approximation error grows as $O(\ln^2 N)$ with fixed K and decreases as $O\left(\frac{\ln K}{K}\right)$ for fixed N. For fixed K, we expect that the error increases as N increases. In particular, as the data set size N increases, we expect to see increasingly smaller components represented in the data. To capture these components, we require finite approximations of increasingly larger sizes. For fixed N, the error goes to zero at least as fast as $O\left(\frac{\ln K}{K}\right)$.

The $1/K$ dependence in the upper bound in Theorem 3.1 is *tight* (modulo logarithmic factors).

Theorem 3.2 (Lower bound for beta-Bernoulli)**.** *There exists an observational likelihood f, independent of K and N, such that for any N,*

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

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

While Theorem 3.1 implies that an IFA with $K = O\left(\mathrm{poly}(\ln N)/\epsilon\right)$ atoms suffices in approximating the target model to less than ϵ error, Theorem 3.2 implies that an IFA with $K = \Omega\left(1/\epsilon\right)$ atoms is *necessary* in the worst case. This dependence on ϵ means that IFAs are worse than TFAs in theory. For example, consider Bondesson approximations [Bondesson, 1982] of $\mathrm{BP}(\gamma, \alpha; H)$ for $\alpha > 1$

$$\mathrm{TFA}_K := \sum_{k=1}^{K} \theta_k \delta_{\psi_k} \text{ where } \theta_k = V_k \exp(-\Gamma_k/\gamma\alpha), V_k \overset{iid}{\sim} \mathsf{Beta}(1, \alpha-1) \text{ and } \psi_k \overset{iid}{\sim} H. \tag{5}$$

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

Proposition 3.3. *[Campbell et al., 2019, Appendix A.1] Let $Q_{N,K}$ be the distribution of the observational process using TFA_K. Then:*

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

Proposition 3.3 implies that a TFA with $K = O\left(\ln\left(N/\epsilon\right)\right)$ atoms suffices in approximating the target model to less than ϵ error. Modulo log factors, comparing the necessary $\frac{1}{\epsilon}$ level for IFA and the sufficient $\ln\left(\frac{1}{\epsilon}\right)$ level for TFA, we conclude that the necessary size for IFA is exponentially larger than the sufficient size for TFA, in the worst case.

3.2 Dirichlet process mixture model

Approximation error is defined analogously to the previous section. After replacing the Dirichlet process with some finite approximation FA_K, we have the following generative process:

$$\Theta_K \sim \mathrm{FA}_K, \quad Z_n|\Theta_K \overset{\text{i.i.d.}}{\sim} \Theta_K, \quad W_n|Z_n \overset{indep}{\sim} f(.|Z_n), \quad n = 1, 2, \ldots, N. \tag{6}$$

Let $P_{N,\infty}$ be the distribution of the observations $Y_{1:N}$ in Equation (2), and $P_{N,K}$ be the distribution of the observations $W_{1:N}$ in Equation (6). *Approximation error* remains the total variation distance $d_{TV}(P_{N,K}, P_{N,\infty})$.

Acharya et al. [2015], Huggins et al. [2017] constructed IFAs targeting gamma process $\Gamma\mathrm{P}(\alpha, 1; H)$ where each ν_K is a proper gamma distribution:

$$\mathrm{IFA}_K \coloneqq \sum_{i=1}^{K} \xi_{K,i}\delta_{\psi_{K,i}} \text{ where } \xi_{K,i} \overset{i.i.d.}{\sim} \mathrm{Gamma}(\alpha/K, 1) \text{ and } \psi_{K,i} \overset{i.i.d.}{\sim} H.$$

Because the normalization of independent gamma random variables is a Dirichlet random variable, the normalization of IFA_K is equal in distribution to

$$\mathrm{FSD}_K \coloneqq \sum_{i=1}^{K} p_{K,i}\delta_{\psi_{K,i}} \text{ where } \psi_{K,i} \overset{\text{i.i.d.}}{\sim} H \text{ and } \{p_{K,i}\}_{i=1}^{K} \sim \mathsf{Dir}\left(\frac{\alpha}{K}\mathbf{1}_K\right). \tag{7}$$

We now quantify the approximation error for FSD_K.

Theorem 3.4 (Upper bound for DP mixture model)**.** *For some constants C_1, C_2, C_3 that depend only on α,*

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

Theorem 3.4 is similar to Theorem 3.1. The $O(\ln^2 N)$ growth of the bound for fixed N can likely be reduced to $O(\ln N)$, the inherent growth rate of DP mixture models [Arratia et al., 2003, Section 5.2]. The $O\left(\frac{\ln K}{K}\right)$ rate of decrease to zero is tight because of a $\frac{1}{K}$ lower bound on the approximation error.

Theorem 3.5 ($1/K$ lower bound)**.** *There exists an observational likelihood $f(.)$, independent of K, N, such that for any $N \geq 2$,*

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

While Theorem 3.4 implies that the normalized IFA_K with $K = O\left(\mathrm{poly}(\ln N)/\epsilon\right)$ atoms suffices in approximating the DP mixture model to less than ϵ error, Theorem 3.5 implies that a normalized IFA with $K = \Omega\left(1/\epsilon\right)$ atoms is *necessary* in the worst case. This worst-case behavior is analogous to Theorem 3.2 for DP-based models.

The $\frac{1}{\epsilon}$ dependence means that IFAs are worse than TFAs in theory. It is known that small TFA models are already excellent approximations of the DP. For example, consider truncated stick-breaking approximation of $\mathrm{DP}(\alpha; H)$ [Sethuraman, 1994]:

$$\mathrm{TSB}_K \coloneqq \sum_{k=1}^{K} \xi_k \delta_{\psi_k} \text{ where } \xi_i = v_i \prod_{j=1}^{i-1}(1 - v_j) \text{ with } v_i \overset{\text{i.i.d.}}{\sim} \mathsf{Beta}(1, \alpha) \text{ and } \psi_k \overset{\text{i.i.d.}}{\sim} H. \tag{8}$$

The following result gives a bound on the error of the truncated stick-breaking approximation:

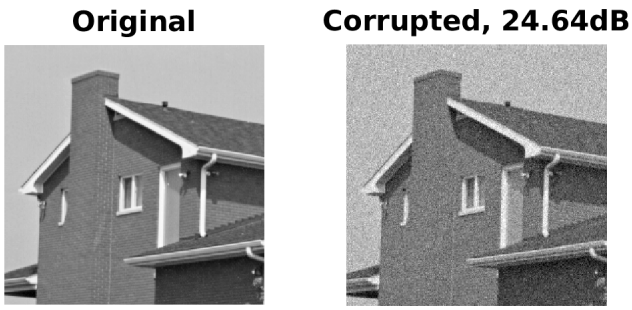


Figure 1: Original versus corrupted images. The number plotted on top of the noisy image is peak signal-to-noise-ratio, or PSNR, with respect to the noiseless image.

Proposition 3.6. *[Ishwaran and James, 2001, Theorem 2] Let $Q_{N,K}$ be the distribution of the observations under* TSB_K*. Then*

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

Proposition 3.6 implies that a TFA with $K = O\left(\ln\left(N/\epsilon\right)\right)$ atoms suffices in approximating the DP mixture model to less than ϵ error. Modulo log factors, comparing the necessary $\frac{1}{\epsilon}$ level for IFA and the sufficient $\ln\left(\frac{1}{\epsilon}\right)$ level for TFA, we conclude that the necessary size for normalized IFA is exponentially larger than the sufficient size for TFA, in the worst case.

4 Performance in applications

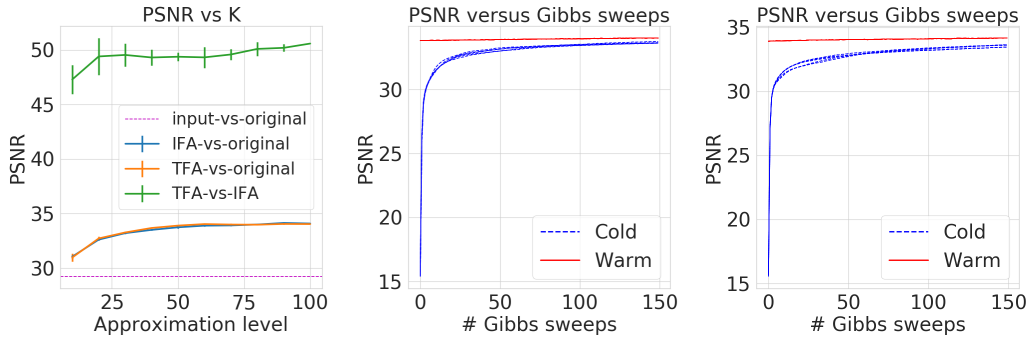
We compare the practical performance of IFAs and TFAs on two real-data examples: an image denoising application using the beta-Bernoulli model and topic modeling using a hierarchical modification of DP mixtures. Existing empirical work (e.g., Doshi-Velez et al. [2009, Table 1,2] and Kurihara et al. [2007, Figure 4]) suggests two patterns: that the approximations improve in performance as the number of instantiated atoms K increase, and for the same K, normalized IFA and TFA have similar performance. Our experiments confirm and expand upon these previous findings. The worst-case behaviors discussed in the previous section are perhaps too pessimistic, since the observational likelihoods f that trigger the lower bounds are different from usage in common probabilistic models.

4.1 Image denoising

Image denoising through dictionary learning is an application where finite approximations of BNP model — in particular beta-Bernoulli — have proven useful [Zhou et al., 2009]. The goal is recovering the original noiseless image (left of Figure 1) from a corrupted one (right of Figure 1). To do so, the input image is deconstructed into small contiguous patches and we postulate that each patch is a combination of underlying *basis elements*. By estimating the coefficients expressing the combination, one can denoise the individual patches and ultimately the overall image. Posterior inference using the beta-Bernoulli process allows simultaneous estimation of both basis elements and basis assignments, and automatically deals with the cumbersome problem of calibrating the number of basis elements. Better denoised images have high peak signal-to-noise-ratio, or PSNR [Hore and Ziou, 2010], with respect to the noiseless image.

We use a sequential[1] Gibbs sampler, which traverses the posterior over latent variables following a fixed scheme. The final denoised image is a weighted average of the candidate images encountered

[1] We introduce patches (i.e. the observed data) in epochs. The sampler only modifies the latent variables of the current epoch's observations.



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

Figure 2: Image denoising results on the house image.

during the sampler run. We initialize the latent variables at random, as well as in the simulation of the Gibbs conditionals. For a 256×256 image like the right panel of Figure 1, the number of extracted patches, N, is about $60k$.

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

Figure 2b uses the output of inference with IFA model as initial values for inference with TFA; similarly Figure 2c uses the output of TFA for inference with IFA. For both kinds of approximation, $K = 60$. Rather than randomly initializing the latent variables at the beginning of the Gibbs sampler of one model (i.e., cold start), we can use the last configuration of latent variables visited in the other model as the initial state of the Gibbs sampler (i.e., warm start). To isolate the effect of the initial conditions, all the patches are available from the start as opposed to being gradually introduced. For both kinds of approximation, the Gibbs sampler initialized at the warm start visits candidate images that basically have the same PSNR as the starting configuration. The early iterates of 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. Each trace of PSNR of cold-start Gibbs corresponds to a random seed in intialization and simulation of the conditionals, while each trace of warm-start PSNR corresponds to a different final state of the alternative model's training. The variation across warm starts is tiny; the variation across cold starts is larger but still very small. In all, the modes of TFA posterior are good initializations for inference with the IFA model, and vice-versa.

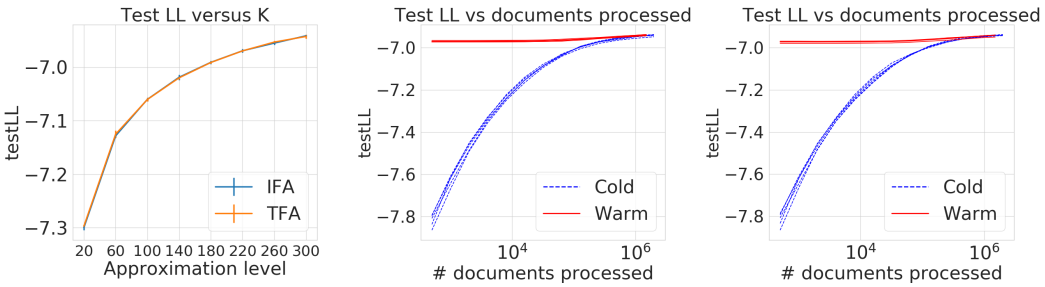
4.2 Topic modelling

Finally, we compare the performance of normalized IFA (i.e., FSD_K, Equation (7)) and TFA (i.e., TSB_K, Equation (8)) when used in DP-based model. In this section, we provide evidence of the same trends in the modified HDP — a hierarchical extension of the Dirichlet process mixture model — when analyzing Wikipedia documents.

For both IFA and TFA, we use stochastic variational inference with mean-field factorization [Hoffman et al., 2013] to approximate the posterior over the latent topics based on training documents. The training corpus is nearly one million documents from Wikipedia. There is randomness in the initial values of the variational parameters, as well as in the order that data minibatches are processed. The quality of inferred topics is measured by the predictive log-likelihood on a set of $10k$ held-out documents.

In Figure 3a, the quality of the inferred topics improves as the approximation level grows; furthermore, the quality is very similar across the two types of approximation. The improvement with K is largest for small K; the slope plateaus for large K. For a given approximation level, the quality of TFA topics and that of normalized IFA are almost the same. The error bars reflect variation across both the random initialization and the ordering of data minibatches processed by stochastic variational inference.

Figure 3b uses the output of inference with (normalized) IFA model as initial values for inference with TFA; similarly Figure 3c uses the output of TFA for inference with (normalized) IFA. The number of topics is fixed to be $K = 300$. Rather than randomly initializing the variational parameters at the start of variational inference of one model (i.e., cold start), we can use the variational parameters at the end of the other model's training as the initialization (i.e., warm start). For both kinds of approximation, the test log-likelihood basically stays the same for warm-start training iterates, hinting that such initialization is part of an attractive region. The early iterates of cold start are noticeably lower in quality compared to the warm iterates; however at the end of training, the test log-likelihoods are nearly the same. Each trace of cold start corresponds to a different initialization and ordering of data batches processed. Each trace of warm start corresponds to a different output of the other model's training and a different ordering of data batches processed. The variation across either cold starts or warm starts is small. In all, the modes of TFA posterior are good initializations for inference with the IFA model, and vice-versa.



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

Figure 3: Topic modeling results on Wikipedia documents.

5 Conclusion

Our analysis of independent finite approximations reveals that in the worst case, for the same number of atoms instantiated, an independent-based approximation has larger error than a truncation-based approximation. However, we have not observed the worst case in our experiments, suggesting that either the error bounds can be tightened for relevant observational likelihoods f or that additional sources of error, such as those from approximate inference, dominate approximation error made by the finite approximations.

References

A. Acharya, J. Ghosh, and M. Zhou. Nonparametric Bayesian factor analysis for dynamic count matrices. In *AISTATS*, 2015.

J. A. Adell and A. Lekuona. Sharp estimates in signed Poisson approximation of Poisson mixtures. *Bernoulli*, 11(1):47–65, 2005.

D. Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198, 1985.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

R. Arratia, A. D. Barbour, and S. Tavaré. *Logarithmic combinatorial structures: a probabilistic approach*, volume 1. European Mathematical Society, 2003.

A. D. Barbour and P. Hall. On the rate of Poisson convergence. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 95, pages 473–480. Cambridge University Press, 1984.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Polya urn schemes. *Ann. Statist.*, 1(2): 353–355, 03 1973. doi: 10.1214/aos/1176342372.

D. M. Blei and M. I. Jordan. Variational Inference for Dirichlet Process Mixtures . *Bayesian Analysis*, 1(1):121–144, 2006.

D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, Jan. 2010.

L. Bondesson. On simulation from infinitely divisible distributions. *Advances in Applied Probability*, 1982.

T. Broderick, M. I. Jordan, and J. Pitman. Beta processes, stick-breaking and power laws. *Bayesian analysis*, 7(2):439–476, 2012.

T. Campbell, J. H. Huggins, J. P. How, and T. Broderick. Truncated random measures. *Bernoulli*, 25 (2):1256–1288, 05 2019. doi: 10.3150/18-BEJ1020.

F. Doshi-Velez, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process. In *Artificial Intelligence and Statistics*, pages 137–144, 2009.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

E. B. Fox, E. Sudderth, M. I. Jordan, and A. S. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. *The Annals of Applied Statistics*, 5(2A):1020–1056, Nov. 2010.

L. Gordon. A stochastic approach to the gamma function. *The American Mathematical Monthly*, 101 (9):858–865, 1994. ISSN 00029890, 19300972.

T. L. Griffiths and Z. Ghahramani. The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. *the Annals of Statistics*, 18(3):1259–1294, 1990.

M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864, 2010.

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

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

J. Huggins, L. Masoero, L. Mackey, and T. Broderick. Generic finite approximations for practical Bayesian nonparametrics. In *NeurIPS 2017 Workshop on Advances in Approximate Bayesian Inference*, 2017.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, Mar. 2001.

H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

M. J. Johnson and A. S. Willsky. Bayesian Nonparametric Hidden Semi-Markov Models. *Journal of Machine Learning Research*, 14:673–701, 2013.

J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

J. F. C. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society B*, 37(1): 1–22, 1975.

K. Kurihara, M. Welling, and Y. W. Teh. Collapsed Variational Dirichlet Process Mixture Models. In *International Joint Conference on Artificial Intelligence*, pages 2796–2801, 2007.

G. Last and M. Penrose. *Lectures on the Poisson Process*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2017.

L. Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.*, 10(4): 1181–1197, 1960.

N. Madras and D. Sezer. Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli*, 16(3):882–908, 08 2010. doi: 10.3150/09-BEJ238.

J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 777–784, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553474.

J. Paisley, D. M. Blei, and M. I. Jordan. Stick-breaking beta processes and the Poisson process. In *Artificial Intelligence and Statistics*, pages 850–858, 2012.

K. Palla, D. A. Knowles, and Z. Ghahramani. An Infinite Latent Attribute Model for Network Data. In *International Conference on Machine Learning*. University of Cambridge, 2012.

J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.

D. Pollard. *A User's Guide to Measure Theoretic Probability*, volume 8. Cambridge University Press, 2001.

A. Roychowdhury and B. Kulis. Gamma processes, stick-breaking, and variational inference. In *Artificial Intelligence and Statistics*, pages 800–808, 2015.

S. Saria, D. Koller, and A. Penn. Learning individual and population level traits from clinical temporal data. In *NeurIPS Predictive Models in Personalized Medicine*, 2010.

J. Sethuraman. A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

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

R. Thibaux and M. I. Jordan. Hierarchical Beta Processes and the Indian Buffet Process. In *International Conference on Artificial Intelligence and Statistics*, 2007.

C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

M. Zhou, H. Chen, L. Ren, G. Sapiro, L. Carin, and J. W. Paisley. Non-parametric Bayesian dictionary learning for sparse image representations. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2295–2303. Curran Associates, Inc., 2009.

A Proofs

A.1 Technical tools

Lemma A.1 (Order of growth of harmonic-like sums)**.**

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

where ψ is the digamma function.

Proof of Lemma A.1. It is well-known (for instance `https://en.wikipedia.org/wiki/Chinese_restaurant_process`) that:

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

[Gordon, 1994, Theorem 5] says that

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

□

Lemma A.2 (Modified upper tail Chernoff bound)**.** *Let $X = \sum_{i=1}^{n} X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let μ be an upper bound on $E(X) = \sum_{i=1}^{n} p_i$. Then for all $\delta > 0$:*

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

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

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

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

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

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

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

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

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

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

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

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

$$(u+1)\psi(u) - 1 \geq 0. \tag{9}$$

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

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

Since $g''(u) \geq 0$, $g'(u)$ is monotone increasing. Since $g'(0) = 1$, $g'(u) > 0$ for $u \geq 0$, hence $g(u)$ is monotone increasing. Because $g(0) = 0$, we conclude that $g(u) \geq 0$ for $u > 0$ and Eq. (9) holds. Plugging in $u = x/\lambda$:

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

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

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

$$\mathbb{P}(X \leq \lambda - x) \leq \exp\left(-\frac{x^2}{2\lambda}\psi\left(-\frac{x}{\lambda}\right)\right). \tag{10}$$

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

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

Therefore:

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

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

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

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

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

which is exactly the right hand side of Eq. (10). Hence to demonstrate the lower tail bound, it suffices to show that:

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

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

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

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

Lemma A.4 (Chain rule)**.** *Suppose (X_1, Y_1) and (X_2, Y_2) are two distributions, over $\mathcal{A} \times \mathcal{B}$, that have densities w.r.t a common measure. Then:*

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

Proof of Lemma A.4. Because both P_{X_1,Y_1} and P_{X_2,Y_2} have densities, total variation distance is half of L_1 distance between the densities:

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

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

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

Regarding the second term:

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

Sum of the first and second upper bound give the total variation chain rule. □

Lemma A.5 (Propagation rule)**.** *Suppose (X_1, Y_1) and (X_2, Y_2) are two distributions over $\mathcal{A} \times \mathcal{B}$. Suppose the conditional $Y_2|X_2 = a$ is the same as the conditional $Y_1|X_1 = a$, which we just denote as $Y|X = a$. Then:*

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

Proof of Lemma A.5. It is well-known that total variation between P_U and P_V is the infimum of $\mathbb{P}(U \neq V)$ over all couplings (U, V) where $U \sim P_U$ and $V \sim P_V$ ([Madras and Sezer, 2010, Equation 13]). For any joint distribution of (X_1, Y_1, X_2, Y_2) where marginally $(X_1, Y_1) \sim P_{X_1,Y_1}$ and $(X_2, Y_2) \sim P_{X_2,Y_2}$, (Y_1, Y_2) is a coupling where $Y_1 \sim P_{Y_1}$ and $Y_2 \sim P_{Y_2}$. Therefore:

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

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

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

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

Lemma A.6 (Product rule)**.** *$Z_1 = (X_1, Y_1)$ and $Z_2 = (X_2, Y_2)$ are two distributions over $\mathcal{A} \times \mathcal{B}$. Suppose P_{X_1,Y_1} factorizes into $P_{X_1}P_{Y_1}$ and similarly $P_{X_2,Y_2} = P_{X_2}P_{Y_2}$. Then:*

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

Proof of Lemma A.6. Consider any (X_1, X_2) that is a coupling of P_{X_1} and P_{X_2}, and any (Y_1, Y_2) that is a coupling of P_{Y_1} and P_{Y_2}. Because of the factorization structure between the $X's$ and the Y', we can construct (X_1', X_2', Y_1', Y_2') such that $(X_1', X_2') \stackrel{D}{=} (X_1, X_2)$, $(Y_1', Y_2') \stackrel{D}{=} (Y_1, Y_2)$, $(X_1', Y_1') \sim P_{X_1,Y_1}$, $(X_2', Y_2') \sim P_{X_2,Y_2}$. By union bound:

$$\mathbb{P}((X_1', Y_1') \neq (X_2', Y_2')) \leq \mathbb{P}(X_1' \neq X_2') + \mathbb{P}(Y_1' \neq Y_2')$$

Because $\inf_{\text{coupling } P_{Z_1}, P_{Z_2}} \mathbb{P}(Z_1 \neq Z_2) \leq \mathbb{P}((X_1', Y_1') \neq (X_2', Y_2'))$, we have:

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

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

Lemma A.7 (Total variation between Poissons [Adell and Lekuona, 2005, Corrollary 3.1])**.** *Let P_1 be the Poisson distribution with mean s, P_2 the Poisson distribution with mean t. Then:*

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

Proposition A.8 (Lower bound on total variation between Binomial and Poisson)**.** *For all K, it is true that:*

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

where:

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

Proof of Proposition A.8. We adapt the proof of [Barbour and Hall, 1984, Theorem 2] to our setting. The $\text{Poisson}(\gamma)$ distribution satisfies the functional equality:

$$\mathbb{E}[\gamma y(Z+1) - Zy(Z)] = 0, \tag{11}$$

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

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

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

where θ is a constant which will be specified later. $x(m)$ serves as a test function to lower bound the total variation distance between $\text{Poisson}(\gamma)$ and $\text{Binom}\left(K, \gamma_K/K\right)$. Let $X_i \sim \text{Ber}(\frac{\gamma_K}{K})$, independently across i from 1 to K, and $W = \sum_{i=1}^K$. Then $W \sim \text{Binomial}\left(K, \gamma_K/K\right)$. The following identity is adapted from [Barbour and Hall, 1984, Equation 2.1]:

$$\mathbb{E}[\gamma_K x(W+1) - Wx(W)] = \left(\frac{\gamma_K}{K}\right)^2 \sum_{i=1}^K \mathbb{E}[x(W_i+2) - x(W_i+1)]. \tag{12}$$

where $W_i = W - X_i$.

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

$$\mathbb{E}[x(W_i+2) - x(W_i+1)] \geq 1 - \frac{3\gamma_K^2 + 12\gamma_K + 7}{\theta\gamma_K}. \tag{13}$$

Consider the derivative of $x(m)$:

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

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

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

Taking expectations, noting that $\mathbb{E}(W_i) \leq \gamma_K$ and $\mathbb{E}(W_i^2) = \text{Var}(W_i) + [\mathbb{E}(W_i)]^2 \leq \sum_{j=1}^{K} \frac{\gamma_K}{K} + (\gamma_K)^2 = \gamma_K^2 + \gamma_K$ we have proven Eq. (13).

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

$$\mathbb{E}[\gamma x(W+1) - Wx(W)] \geq \mathbb{E}[\gamma_K x(W+1) - Wx(W)]. \tag{14}$$

Combining Eq. (12), Eq. (13) and Eq. (14) we have:

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

Recalling Eq. (11), for any coupling (W, Z) such that $W \sim \text{Binom}\left(K, \frac{\gamma/K}{\gamma/K+1}\right)$ and $Z \sim \text{Poisson}(\gamma)$:

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

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

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

Taking expectations on both sides, we conclude that

$$2d_{TV}(P_W, P_Z) \times \sup_m |\gamma x(m+1) - mx(m)| \geq \frac{\gamma_K^2}{K}\left(1 - \frac{3\gamma_K^2 + 12\gamma_K + 7}{\theta\gamma_K}\right) \tag{15}$$

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

$$\begin{aligned}
|\gamma x(m+1) - mx(m)| &= |\gamma(x(m+1) - x(m)) + (\gamma - m)x(m)| \\
&\leq \gamma + (m+\gamma)m \exp\left(-\frac{m^2}{\gamma_K\theta}\right) \\
&\leq \gamma + (\gamma+1)m^2 \exp\left(-\frac{m^2}{\gamma_K\theta}\right) \\
&\leq \gamma + \theta\gamma_K(\gamma+1)\exp(-1).
\end{aligned} \tag{16}$$

where the last inequality owes to the easy-to-verify $x\exp(-x) \leq \exp(-1)$. Combining Eq. (16) and Eq. (15) we have that:

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

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

Lemma A.9 (Multinomial-Poisson approximation)**.** *Let $\{p_i\}_{i=1}^{\infty}$, $p_i \geq 0$, $\sum_{i=1}^{\infty} p_i < 1$. Suppose there are n independent trials: in each trial, success of type i has probability p_i. Let $X = \{X_i\}_{i=1}^{\infty}$ be the number of type i successes after n trial. Let $Y = \{Y_i\}_{i=1}^{\infty}$ be independent Poisson random variables, where Y_i has mean np_i. Then:*

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

Proof of Lemma A.9. First we remark that both X and Y can be sampled in two-steps.

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

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

Observing that the conditional $X|N_1 = n$ is the same as $Y|N_2 = n$, we use propagation rule Lemma A.5:

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

Total variation between N_1 and N_2 is just the classic Binomial-Poisson approximation Le Cam [1960].

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

□

A.2 Beta-Bernoulli model

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

The marginal representation of beta-Bernoulli model is the well-known Indian buffet process.

Proposition A.10 (Beta-Bernoulli marginal process [Griffiths and Ghahramani, 2011])**.** *For any n, $X_n | X_{n-1}, \ldots, X_1$ is a random measure with finite support.*

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

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

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

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

The marginal representation of IFA of beta-Bernoulli model is as follows.

Proposition A.11 (IFA marginal process)**.** *For any n, $Z_n | Z_{n-1}, \ldots, Z_1$ is a random measure with finite support.*

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

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

2. *$K - K_{n-1}$ atom locations are generated iid from H. Z_n has $p_{n,x}$ atoms whose size is exactly x (for $x \in \mathbb{N} \cup \{0\}$) over these $K - K_{n-1}$ atom locations (the $p_{n,0}$ atoms whose atom size is 0 can be interpreted as not present in Z_n). The joint distribution of $p_{n,x}$ is a Multinomial with $K - K_{n-1}$ trials, with success of type x having probability $\widetilde{h}_c(x|x_{1:(n-1)} = 0)$.*

Let $C_1 = \alpha \max(\gamma, 1)$, $C_2 = C_3 = C_4 = 0$ and $C_5 = \gamma^2\alpha$. It is straightforward to verify that the functions h_c, $\widetilde{h}_c$ and $M_{n,x}$ in Propositions A.10 and A.11 satisfy the following inequalities:

1. For all $n \in \mathbb{N}$,
$$\sum_{x=1}^{\infty} M_{n,x} \le \frac{C_1}{n-1+C_1}. \tag{17}$$
2. For all $n \in \mathbb{N}$,
$$\sum_{x=1}^{\infty} h(x|x_{1:(n-1)} = 0) \le \frac{1}{K}\frac{C_1}{n-1+C_1}. \tag{18}$$
3. For any $n \in \mathbb{N}$, for any $\{x_i\}_{i=1}^{n-1}$,
$$\sum_{x=0}^{\infty} \left| h_c(x|x_{1:(n-1)}) - \widetilde{h}_c(x|x_{1:(n-1)}) \right| \le \frac{1}{K}\frac{C_1}{n-1+C_1}. \tag{19}$$
4. For all $n \in \mathbb{N}$, for any $K \ge C_2(\ln n + C_3)$,
$$\sum_{x=1}^{\infty} \left| M_{n,x} - K\widetilde{h}_c(x|x_{1:(n-1)} = 0) \right| \le \frac{1}{K}\frac{C_4 \ln n + C_5}{n-1+C_1}. \tag{20}$$

Proof of Theorem 3.1. Let β be the smallest positive constant where $\beta^2 C_1/(1+\beta) \ge 2$. We will focus on the case where the approximation level K is essentially $\Omega(\ln N)$:

$$K \ge \max\left((\beta+1)\max(C(K, C_1), C(N, C_1)), C_2(\ln N + C_3)\right). \tag{21}$$

To see why it is sufficient, observe that the upper bound in Theorem 3.1 naturally holds for K smaller than $\ln N$. Total variation distance is always upper bounded by 1; if $K = o(\ln N)$, then by selecting reasonable constants C', C'', C''', we can make the right hand side at least 1, and satisfy the inequality. In the sequel, we will only consider the situation in Eq. (21).

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

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

The marginal process that described the atom sizes of $X_n|X_{n-1}, X_{n-2}, \ldots, X_1$ in Proposition A.10 is also the description of how the rows of F are generated. The joint distribution $X_1, X_2, \ldots, X_n$ can be two-step sampled. First, the feature-allocation matrix F is sampled. Then, the atom locations are drawn iid from the base measure H: each column of F is assigned an atom location, and the

latent measure X_i has atom size $F_{i,j}$ on the jth atom location. A similar two-step sampling generates $Z_1, Z_2, \ldots, Z_n$, the latent measures under the approximate model: the distribution over the feature-allocation matrix F' follows Proposition A.11 instead of Proposition A.10, but conditioned on the feature-allocation matrix, the process generating atom locations and constructing latent measures is exactly the same. In other words, this implies that the conditional distributions $Y_{1:N}|F = f$ and $W_{1:N}|F' = f$ are the same, since both models have the same the observational likelihood f given the latent measures 1 through N. Denote P_F to be the distribution of the feature-allocation matrix under the target model, and $P_{F'}$ the distribution of the feature-allocation matrix under the approximate model. Lemma A.5 implies that:

$$d_{TV}(P_{W_{1:N}}, P_{Y_{1:N}}) \leq d_{TV}(P_F, P_{F'}). \tag{22}$$

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

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

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

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

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

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

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

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

$$\mathbb{P}((D, S) \neq (D', S')) \leq \mathbb{P}(D \neq D') + \mathbb{P}(S \neq S' | D = D', D \in \mathcal{D}^*) + \mathbb{P}(D \notin \mathcal{D}^*). \tag{23}$$

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

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

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

Atypical set probability The $\mathbb{P}(D \notin \mathcal{D}^*)$ term in Eq. (23) is easiest to control. Under the target model Proposition A.10, the $D_{i,x}$'s are independent Poissons with mean $M_{i,x}$, so the sum $\sum_{i=1}^{N}\sum_{x=1}^{\infty} D_{i,x}$ is itself a Poisson with mean $M = \sum_{i=1}^{N}\sum_{x=1}^{\infty} M_{i,x}$. Because of Lemma A.3, for any $x > 0$:

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

For the event $\mathbb{P}(D \notin \mathcal{D}^*)$, $M + x = (\beta+1)\max(C(K, C_1), C(N, C_1))$, $M \leq C(N, C_1)$ due to Eq. (17), so that $x \geq \beta \max(C(K, C_1), C(N, C_1))$. Therefore:

$$\mathbb{P}(D \notin \mathcal{D}^*) \leq \exp\left(-\frac{\beta^2}{2(\beta+1)}\max(C(K, C_1), C(N, C_1))\right). \tag{24}$$

Difference between histories To minimize the difference probability between the histories of atom sizes i.e. the $\mathbb{P}(S \neq S'|D = D', D \in \mathcal{D}^*)$ term in Eq. (23), we will use Eq. (19). The claim is, there exists a coupling of $S'|D'$ and $S|D$ such that:

$$\mathbb{P}(S \neq S'|D = D', D \in \mathcal{D}^*) \leq \frac{(\beta+1)\max(C(K, C_1), C(N, C_1))}{K} C(N, C_1). \tag{25}$$

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

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

since in both marginal processes, the atom sizes for different atom locations are independent of each other. This means we can use Lemma A.6:

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

We inspect each $d_{TV}(P_{S_j|D=d}, P_{S'_j|D'=d})$. Fixing d also fixes I_j, the first row to manifest the jth atom location. The history s_j is then a $N - I_j + 1$ dimensional integer vector, whose tth entry is the atom size over the $j$the atom location of the $t + I_j - 1$ row. Because of Eq. (19), we know that conditioned on the same partial history $S_j(1:(t-1)) = S'_j(1:(t-1)) = s$, the distributions $S_j(t)$ and $S'_j(t)$ are very similar. The conditional distribution $S_j(t)|D = d, S_j(1:(t-1)) = s$ is governed by h_c Proposition A.10 while $S'_j(t)|D' = d, S'_j(1:(t-1)) = s$ is governed by $\widetilde{h}_c$ Proposition A.11. Hence:

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

for any partial history s. To use this conditional bound, we again leverage Lemma A.4 to compare the joint $S_j = (S_j(1), S_j(2), \ldots, S_j(N - I_j + 1))$ with the joint $S'_j = (S'_j(1), S'_j(2), \ldots, S'_j(N -$

$I_j + 1)$), peeling off one layer at a time.

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

Multiplying the right hand side by $(\beta+1)\max(C(K,C_1), C(N,C_1))$, the upper bound on J, we arrive at the same upper bound for the total variation between $P_{S|D=d}$ and $P_{S'|D'=d}$ in Eq. (25). Furthermore, our analysis of the total variation can be back-tracked to construct the coupling between the conditional distributions $S|D=s$ and $S'|D'=d$ which attains that small probability of difference. Since the choice of conditioning $d \in \mathcal{D}^*$ was arbitrary, we have actually shown Eq. (25).

Difference between new atom sizes Finally, to control the difference probability for the distribution over new atom sizes i.e. the $\mathbb{P}(D \neq D')$ term in Eq. (23), we will utilize Eqs. (18) and (20). For each n, define the short-hand $d_{1:n}$ to refer to the collection $d_{i,x}$ for $i \in [n]$, $x \in \mathbb{N}$, and the typical sets:

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

The type of expansion performed in Eq. (23) can be done once here to see that:

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

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

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

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

$$\mathbb{P}(D \neq D') \leq \sum_{j=1}^{N} B_j + \sum_{j=2}^{N} \mathbb{P}(D_{1:(j-1)} \notin \mathcal{D}^*_{j-1}). \tag{26}$$

The second summation in Eq. (26), comprising of only atypical probabilities, is easier to control. For any j, since $\sum_{i=1}^{j-1}\sum_{x=1}^{\infty} D_{i,x} \leq \sum_{i=1}^{N}\sum_{x=1}^{\infty} D_{i,x}$, $\mathbb{P}(D_{1:(j-1)} \notin \mathcal{D}^*_{j-1}) \leq \mathbb{P}(D \notin \mathcal{D}^*)$, so a generous upper bound for the contribution of all the atypical probabilities including the first one from Eq. (24) is:

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

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

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

meaning the overall atypical probabilities is at most:

$$\mathbb{P}(D \notin \mathcal{D}^*) + \sum_{j=2}^{N} \mathbb{P}(D_{1:(j-1)} \notin \mathcal{D}^*_{j-1}) \leq \frac{\text{constant}}{K}. \tag{27}$$

As for the first summation in Eq. (26), we look at the individual B_j's. For any fixed $d_{1:(j-1)} \in \mathcal{D}^*_{j-1}$, we claim that there exists a coupling between the conditionals $D_j|D_{1:(j-1)} = d_{1:(j-1)}$ and $D'_j|D'_{1:(j-1)} = d_{1:(j-1)}$ such that $\mathbb{P}(D_j \neq D'_j|D_{1:(j-1)} = D'_{1:(j-1)} = d_{1:(j-1)})$ is at most:

$$\frac{\text{constant}}{K}\frac{1}{(j-1+C_1)^2} + \text{constant}\frac{(\ln N + \ln K)}{K}\frac{1}{j-1+C_1}. \tag{28}$$

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

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

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

On the one hand, conditioned on $D'_{1:(j-1)} = d_{1:(j-1)}$, $D'_j = \{D'_{j,x}\}_{x=1}^{\infty}$ is the joint distribution of types of successes of type x, where there are $K - \sum_{i=1}^{j-1}\sum_{x=1}^{\infty} d_{i,x}$ independent trials and types x success has probability $\widetilde{h}_c(x|x_{1:(j-1)} = 0)$ by Proposition A.11. Because of Lemma A.9 and Eq. (18):

$$\begin{aligned} d_{TV}\left(P_{D'_j|D'_{1:(j-1)}=d_{1:(j-1)}}, P_U\right) &\leq \left(K - \sum_{i=1}^{j-1}\sum_{y=1}^{\infty} d_{i,y}\right)\left(\sum_{x=1}^{\infty}\widetilde{h}_c(x|x_{1:(j-1)} = 0)\right)^2 \\ &\leq K\left(\frac{1}{K}\frac{C_1}{j-1+C_1}\right)^2 \\ &\leq \frac{C_1^2}{K}\frac{1}{(j-1+C_1)^2}. \end{aligned} \tag{29}$$

On the other hand, conditioned on $D_{1:(j-1)}$, $D_j = \{D_{j,x}\}_{x=1}^{\infty}$ consists of independent Poissons, where the mean of $D_{j,x}$ is $M_{j,x}$ by Proposition A.10. We recursively apply Lemma A.6 and Lemma A.7:

$$\begin{aligned} &d_{TV}(P_U, P_{D_j}) \\ &\leq \sum_{x=1}^{\infty} d_{TV}(P_{U_x}, P_{D_{j,x}}) \\ &\leq \sum_{x=1}^{\infty}\left|M_{j,x} - \left(K - \sum_{i=1}^{j-1}\sum_{y=1}^{\infty} d_{i,y}\right)\widetilde{h}_c(x|x_{1:(j-1)} = 0)\right| \\ &\leq \sum_{x=1}^{\infty}\left(|M_{j,x} - K\widetilde{h}_c(x|x_{1:(j-1)} = 0)| + \sum_{i=1}^{j-1}\sum_{y=1}^{\infty} d_{i,y}\widetilde{h}_c(x|x_{1:(j-1)} = 0)\right) \\ &\leq \sum_{x=1}^{\infty}|M_{j,x} - K\widetilde{h}_c(x|x_{1:(j-1)} = 0)| + \left(\sum_{i=1}^{j-1}\sum_{y=1}^{\infty} d_{i,y}\right)\left(\sum_{x=1}^{\infty}\widetilde{h}_c(x|x_{1:(j-1)} = 0)\right). \end{aligned} \tag{30}$$

The first term is upper bounded by Eq. (20). Regarding the second term, since we are in the typical set, $\sum_{i=1}^{j-1}\sum_{y=1}^{\infty} d_{i,y}$ is upper bounded. Therefore the overall bound on the second term is:

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

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

$$\frac{1}{K}\frac{C_4 \ln j + C_5}{j-1+C_1} + (\beta+1)\max(C(K,C_1), C(N,C_1))\frac{1}{K}\frac{C_1}{j-1+C_1}$$
$$\leq \text{constant}\frac{(\ln N + \ln K)}{K}\frac{1}{j-1+C_1}. \tag{31}$$

Combining Eqs. (29) and (31) gives the upper bound in Eq. (28). The summation of the right hand side of Eq. (28) across j leads to:

$$\sum_{j=1}^{N} B_j \leq \frac{\text{constant}}{K} + \text{constant}\frac{(\ln N + \ln K)\ln N}{K}. \tag{32}$$

In all, because of Eqs. (27) and (32), we can couple D and D' such that $\mathbb{P}(D \neq D')$ is at most:

$$\frac{\text{constant}}{K} + \text{constant}\frac{(\ln N + \ln K)\ln N}{K}. \tag{33}$$

Aggregating the results from Eqs. (24), (25) and (33), we are done.

□

Proof of Theorem 3.2. First we mention which probability kernel f results in the large total variation distance. For any discrete measure $\sum_{i=1}^{M}\delta_{\psi_i}$, f is the Dirac measure sitting on M, the number of atoms.

$$f(.|\sum_{i=1}^{M}\delta_{\psi_i}) := \delta_M(.). \tag{34}$$

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

$$d_{TV}(P_{Y_{1:N}}, P_{W_{1:N}}) \geq d_{TV}(P_{Y_1}, P_{W_1}). \tag{35}$$

Truly, suppose $(Y_{1:N}, W_{1:N})$ is any coupling of $P_{Y_{1:N}}, P_{W_{1:N}}$. Elementarily we have $P(Y_{1:N} \neq W_{1:N}) \geq P(Y_1 \neq W_1)$. Taking the infimum over couplings to attain the total variation distance, we have shown Eq. (35). Hence it suffices to show:

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

Recall the generative process defining P_{Y_1} and P_{W_1}. Y_1 is an observation from the target Beta-Bernoulli model, so by Proposition A.10

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

W_1 is an observation from the approximate model, so by Proposition A.11

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

Because of the choice of f, $Y_1 = N_T$ and $W_1 = N_A$. Hence, by Proposition A.8:

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

□

A.3 Dirichlet process mixture model

Our technique to analyze the error made by FSD_K follows a similar vein to the technique in Appendix A.2. We compare the joint distribution of the latents $X_{1:N}$ and $Z_{1:N}$ (with the underlying Θ or Θ_K marginalized out) using the conditional distributions $X_n \mid X_{1:(n-1)}$ and $Z_n \mid Z_{1:(n-1)}$. Before going into the proofs, we give the form of the conditionals.

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

Proposition A.12. *Blackwell and MacQueen [1973] For $n = 1$, $X_1 \sim H$. For $n \geq 2$:*

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

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

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

Proposition A.13. *Pitman [1996] For $n = 1$, $Z_1 \sim H$. For $n \geq 2$, let $\{\psi_j\}_{j=1}^{J_n}$ be the set of unique values among $Z_{n-1}, Z_{n-2}, \ldots, Z_1$ and n_j is the cardinality of the set $\{i : 1 \leq i \leq n-1, Z_i = \psi_j\}$. If $J_n < K$:*

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

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

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

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

Proof of Theorem 3.4. First, because of Lemma A.5, it suffices to show that $d_{TV}(P_{X_{1:N}}, P_{Z_{1:N}})$ is small, since the conditional distributions of the observations given the latent variables are the same across target and approximate models.

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

$$\mathbb{P}(X_n \neq Z_n | X_{1:(n-1)} = Z_{1:(n-1)}) \leq 2 \frac{\alpha}{K} \frac{J_n}{n-1+\alpha}, \tag{36}$$

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

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

$$\frac{\alpha}{n-1+\alpha} H(A) + \sum_{j=1}^{J_n} \frac{n_j}{n-1+\alpha} \delta_{\psi_j}(A)$$

meaning the absolute difference in probability mass is:

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

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

We now show that for the coupling satisfying Eq. (36), the overall probability of difference $\mathbb{P}(X_{1:N} \neq Z_{1:N})$ is small. Define the short hand:

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

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

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

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

$$\begin{aligned}\mathbb{P}(X_{1:N} \neq Z_{1:N}) &= \mathbb{P}((X_{1:(N-1)}, X_N) \neq (Z_{1:(N-1)}, Z_N)) \\ &= \mathbb{P}(X_{1:(N-1)} \neq Z_{1:(N-1)}) + \mathbb{P}(X_N \neq Z_N, X_{1:(N-1)} = Z_{1:(N-1)})\end{aligned} \tag{37}$$

The second term can be further expanded:

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

The former term is at most:

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

while the latter term is at most:

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

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

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

By recursively applying the expansion initiated in Eq. (37) to $\mathbb{P}(X_{1:(N-1)} \neq Z_{1:(N-1)})$, we actually only need to bound difference probability of the different prediction rules on typical sets and the atypical set probabilities.

Regarding difference probability of the different prediction rules, being in the typical set allows us to control J_n in Eq. (36). Summation across $n = 1$ through N gives the overall bound of:

$$2\frac{\alpha}{K}(1+\delta)\max(C(N-1,\alpha), C(K,\alpha))C(N,\alpha) \leq \text{constant}\frac{\ln N(\ln N + \ln K)}{K}. \tag{38}$$

Regarding the atypical set probabilities, because J_{n-1} is stochastically dominated by J_n i.e., the number of unique values at time n is at least the number at time $n-1$, all the atypical set probabilities

are upper bounded by the last one i.e. $\mathbb{P}(X_{1:(N-1)} \notin \mathcal{D}_N)$. J_{N-1} is the sum of independent Poisson trials, with an overall mean equaling exactly $C(N-1,\alpha)$. Therefore, the atypical event has small probability because of Lemma A.2:

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

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

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

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

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

meaning the overall atypical probabilities is at most:

$$\frac{\text{constant}}{K}. \tag{39}$$

The overall total variation bound combines Eqs. (38) and (39). □

Proof of Theorem 3.5. First we mention which probability kernel f results in the large total variation distance: the pathological f is the Dirac measure i.e., $f(\cdot \mid x) = \delta_x(.)$.

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

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

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

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

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

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

We now investigate how small $\mathbb{P}(X_2 \neq Z_2 | X_1 = Z_2)$ can be. In the conditioning $X_1 = Z_1$, let the common atom be ψ_1. The prediction rule $X_2|X_1 = \psi_1$ puts mass $\frac{1}{1+\alpha}$ on ψ_1 while the prediction rule $Z_2|Z_1 = \psi_1$ puts mass $\frac{1+\alpha/K}{1+\alpha}$. This means that the total variation distance between the two prediction rules is at least:

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

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

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

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

□

B Experimental setup

B.1 Image denoising

The experiments in this section aim to isolate the effect of TFA versus IFA, by fitting different approximations of the beta-Bernoulli model to denoise[2] an image. We give a description of our models and their hyper-parameter settings. Each patch x_i is flattened into a vector in $\mathbb{R}^n$. Let $\mathbf{I}_n$ be the $n \times n$ identity matrix, and similarly for $\mathbf{I}_K$. The base measure generating the basis elements is the same:

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

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

$$\begin{aligned}
\gamma_w &\sim \mathrm{Gamma}(10^{-6}, 10^{-6}) \\
\gamma_e &\sim \mathrm{Gamma}(10^{-6}, 10^{-6}) \\
w_i &\overset{iid}{\sim} \mathcal{N}(0, \gamma_w^{-1}\mathbf{I}_K) \quad i = 1, 2, \ldots, N \\
\epsilon_i &\overset{iid}{\sim} \mathcal{N}(0, \gamma_e^{-1}\mathbf{I}_n) \quad i = 1, 2, \ldots, N \\
x_i &= \sum_{k=1}^{K} F_{i,k} w_{i,k} \psi_k + \epsilon_i \quad i = 1, 2, \ldots, N
\end{aligned} \tag{40}$$

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

- TFA:

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

- IFA:

$$\begin{aligned}
\pi_k &\overset{iid}{\sim} \mathsf{Beta}\left(\frac{1}{K}, 1\right) \quad k = 1, 2, \ldots, K \\
F_{i,k}|\pi_k &\overset{indep}{\sim} \mathrm{Ber}(\pi_k) \quad i = 1, 2, \ldots, N
\end{aligned}$$

In Eq. (40), we are enriching the basic feature-allocation structure by introducing weights $w_{i,k}$ which allow an observation to manifest a non-integer (and potentially negative) scaled version of the basis element. Following [Zhou et al., 2009], we are *uninformative* about the noise precisions by choosing $\mathrm{Gamma}(10^{-6}, 10^{-6})$. Regarding the choice of hyper-parameters for the underlying beta process, [Zhou et al., 2009] suggests that the performance of the denoising routine is insensitive to the choice of γ and α: we picked $\gamma, \alpha = 1$ for computational convenience, especially since for the beta process for $\alpha = 1$ admits the simple stick-breaking construction.

B.2 Topic modelling

Nearly $1m$ random wikipedia documents were downloaded and processed following [Hoffman et al., 2010].

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

IFA:

$$
\begin{aligned}
G_0 &\sim \mathrm{FSD}_K(\omega, \mathsf{Dir}(\eta \mathbf{1}_V)) \\
G_d &\sim \text{T-DP}_T(\alpha, G_0) && \text{independently across } d = 1, 2, \ldots, D \\
\beta_{dn}|G_d &\sim G_d(.) && \text{independently across } n = 1, 2, \ldots, N_d \\
w_{dn}|\beta_{dn} &\sim \mathsf{Categorical}(\beta_{dn}) && \text{independently across } n = 1, 2, \ldots, N_d
\end{aligned}
$$

TFA:

$$
\begin{aligned}
G_0 &\sim \text{T-DP}_K(\omega, \mathsf{Dir}(\eta \mathbf{1}_V)) \\
G_d &\sim \text{T-DP}_T(\alpha, G_0) && \text{independently across } d = 1, 2, \ldots, D \\
\beta_{dn}|G_d &\sim G_d(.) && \text{independently across } n = 1, 2, \ldots, N_d \\
w_{dn}|\beta_{dn} &\sim \mathsf{Categorical}(\beta_{dn}) && \text{independently across } n = 1, 2, \ldots, N_d
\end{aligned}
$$

Hyper-parameter settings follow [Wang et al., 2011] in that $\eta = 0.01, \alpha = 1.0, \omega = 1.0, T = 20$.

We approximate the posterior in each model using stochastic variational inference [Hoffman et al., 2013]. Both models have nice conditional conjugacies that allow the use of exponential family variational distributions and closed-form expectation equations. Batch size is 500, learning rate parametrized by $\rho_t = (t + \tau)^{-\kappa}$ where by default $\tau = 1.0$ and $\kappa = 0.9$. The learning rate for warm-start training is slightly different from that for cold start, to reflect the fact that many batches of data had been processed leading up to the warm-start variational parameters.

We discuss how held-out log-likelihood is computed. Each held-out document d' is separated into two parts w_{ho} and w_{obs}[3], with no common words between the two. In our experiments, we set 75% of words to be observed, the remaining 25% unseen. The predictive distribution of each word w_{new} in the w_{ho} is exactly equal to:

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

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

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

where $q(\beta|\mathcal{D})$ is fixed to be the variational approximation found during training and $q(\theta_{d'})$ minimizes the KL between the variational distribution and the posterior. Operationally, we do an E-step for the document d' based on the variational distribution of β and the observed words w_{obs}, and discard the distribution over $z_{d',.}$, the per-word topic assignments because of the mean-field assumption. Using those approximations, the predictive approximation is approximately:

$$p(w_{new}|\mathcal{D}, w_{obs}) \approx \widetilde{p}(w_{new}|\mathcal{D}, w_{obs}) = \sum_{k=1}^{K} \mathbb{E}_q(\theta_{d'}(k)) \mathbb{E}_q(\beta_k(w_{new})),$$

and the final number we report for document d' is:

$$\frac{1}{|w_{ho}|} \sum_{w \in w_{ho}} \log \widetilde{p}(w|\mathcal{D}, w_{obs}).$$

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