# Generic finite approximations for practical Bayesian nonparametrics

**Jonathan Huggins**[1] **Lorenzo Masoero**[1] **Lester Mackey**[2] **Tamara Broderick**[1]
[1]CSAIL, MIT [2]Microsoft Research
jhuggins@mit.edu lom@mit.edu lmackey@microsoft.com tbroderick@csail.mit.edu

## 1 Introduction

Many data analysis problems can be seen as discovering a latent set of traits in a population. For instance, we might recover topics or themes from scientific papers, ancestral populations from genetic data, interest groups from social network data, or unique speakers across audio recordings of many meetings [3, 11, 29]. In all of these cases, we might reasonably expect the number of latent traits present in a data set to grow with the size of the data. One modeling option is to choose a different prior for different data set sizes, but constructing a new prior separately for every size of data set is unwieldy. A simpler option is to choose a single prior that naturally yields different expected numbers of traits for different numbers of data points. In theory, *Bayesian nonparametrics* (BNP) provides a rich set of priors with exactly this desirable property. BNP accomplishes this property by imagining a latent countable infinity of traits, so that there are always more traits to reveal in the data at larger data set sizes. This latent, infinite-dimensional parameter presents a major practical challenge, though. In what follows, we propose a simple approximation across a wide range of BNP models that is amenable to modern, efficient inference schemes and black-box code; fits easily within complex, potentially deep generative models; and admits straightforward parallelization.

A particular challenge of the infinite-dimensional parameter is that it is impossible to store an infinity of random variables in memory or learn the distribution over an infinite number of variables in finite time. Some authors have developed conjugate priors and likelihoods [27] to circumvent the infinite representation via marginalization and thereby perform exact Bayesian posterior inference [6, 17]. However, these priors and likelihoods are often just a single piece within a more complex generative model, which is no longer fully conjugate and therefore requires an approximate posterior inference scheme such as Markov Chain Monte Carlo (MCMC) or variational Bayes (VB). Some local steps in, e.g., an MCMC sampler can still take advantage of conditional conjugacy via special marginal forms such the Chinese restaurant process [38] or the Indian buffet process [12]; see Broderick et al. [6] and James [17] for general treatments. But using these marginal distributions rather than a full and explicit representation of the latent variables typically necessitates a Gibbs sampler, which can be slow to mix and may require special-purpose, model-specific sampling moves. To take advantage of black-box variational inference methods [23, 31], modern MCMC methods such as Metropolis-adjusted Langevin algorithm [33] or Hamiltonian Monte Carlo (HMC) [2, 26], or modern probabilistic programming systems such as Stan [8], a full trait representation is generally required.

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

Advances in Approximate Bayesian Inference Workshop at NIPS 2017

Previous work developed and analyzed general mechanisms for finite approximations based on truncating the random measures underlying BNP [7]. In the present work, we instead consider a finite approximation consisting of independent and identical representations of the traits together with their rates within the population. This approach has the potential to be simpler to incorporate in a complex model, to exhibit improved mixing, and to be amenable to parallelizing computation during inference. In fact, certain special cases of this approach have already been successfully used in applications, with practitioners reporting similar performance to the truncation approach but with faster mixing [11, 19, 24, 35]. In what follows, we start by reviewing the random processes underlying a broad toolbox for BNP in and rigorously define finite approximations in Section 2. We propose a broad mechanism for our i.i.d. finite approximation, which we call a *non-nested finite approximation* (NNFA), and relate it to existing work in Section 3. Finally, we demonstrate in preliminary experiments in Section 4 that NNFAs can provide good approximations for a full BNP model.

## 2 Background

Let $\psi_i$ represent the $i$th trait of interest; e.g., a topic in topic model. Let $\theta_i$ represent the rate, or frequency, of this trait in the population. We can collect the pairs of traits with their frequencies $(\psi_i, \theta_i)$ in a measure that places non-negative mass $\psi_i$ at location $\psi_i$: $\Theta := \sum_{i=1}^{I} \theta_i \delta_{\psi_i}$. $I$, the total number of traits, may be finite or, as in the nonparametric setting, countably infinite. To perform Bayesian inference, we need to choose a prior on $\Theta$, to choose a likelihood for the observed data $X_{1:M} := \{X_m\}_{m=1}^{M}$ given $\Theta$, and finally to apply Bayes Theorem to obtain the posterior on $\Theta$ given the observed data.

**Completely random measures.** Most common BNP priors can be conveniently formulated as (normalizations of) *completely random measures* (CRMs). CRMs are constructed from a Poisson process, which is straightforward to manipulate both analytically and algorithmically. 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 $\Psi$ and $\psi_i \overset{\text{i.i.d.}}{\sim} H$ for some distribution $H$. $H$ serves as a prior on the trait values. The resulting measure $\Theta$ in this case is a *completely random measure* (CRM) [21]. 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.[1]

The CRM prior on $\Theta$ is typically combined with a likelihood that generates trait counts for each data point. Let $h(\cdot \mid \theta)$ be a proper probability mass function on $\mathbb{N} \cup \{0\}$ for all $\theta$ in the support of $\nu$. Then a collection of conditionally independent observations $Z_{1:M}$ given $\Theta$ are distributed according to the *likelihood process* $\mathrm{LP}(h, \Theta)$, i.e. $Z_m := \sum_i z_{mi} \delta_{\psi_i} \overset{\text{i.i.d.}}{\sim} \mathrm{LP}(h, \Theta)$, if $z_{mi} \sim h(\cdot \mid \theta_i)$ independently across $i$ and i.i.d. across $m$. Since the trait counts are typically latent in a full generative model specification, define the observed data $X_m \mid Z_m \overset{\text{indep}}{\sim} f(\cdot \mid Z_m)$ for a conditional density $f$ with respect to a measure $\mu$ on some space.

**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* $\Theta_n := \sum_{i=1}^{n} \theta_i \delta_{\psi_i}$. Since it involves a finite number of parameters, $\Theta_n$ can be used for efficient posterior inference, including with black-box MCMC and VB algorithms—but some approximation error is introduced by not using the full CRM $\Theta$.

A *truncated finite approximation* (TFA) requires constructing an ordering on the sequence $(\theta_i)_{i=1}^{\infty}$ such that $\theta_i$ is a function of some auxiliary random variables $\xi_1, \ldots, \xi_i$; hence, $\theta_{i+1}$ reuses the same auxiliary randomness as $\theta_i$, plus uses an additional random variable $\xi_{i+1}$. Thus, the value of $\theta_{i+1}$ implicitly depends on the values of $\theta_1, \ldots, \theta_i$. Truncated finite approximations are attractive because the approximation level $n$ does not need to be chosen ahead of time. On the other hand the complex dependences between the atoms $\theta_1, \theta_2, \ldots$ potentially make inference more challenging.

[1] The possible fixed-location and deterministic components of an (N)CRM [21] are not considered here for brevity; these components can be added (assuming they are purely atomic) and our analysis modified without undue effort.

We here instead pursue what we call a *non-nested finite approximation* (NNFA), which involves choosing a sequence of probability measures $\nu_1, \nu_2, \ldots$ such that for any approximation level $n$, we choose $\theta_1, \ldots, \theta_n \overset{\text{i.i.d.}}{\sim} \nu_n$. The $\nu_i$ are chosen in such a way that $\Theta_n \overset{\mathcal{D}}{\Longrightarrow} \Theta$ — that is, the NNFAs converge in distribution to the CRM. The pros and cons of the NNFA 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 $n$ changes.

**The gamma process.** For concreteness, we consider the *gamma process* [4, 10, 16, 22, 39] as a running example of a CRM. It is widely used in applications including image modeling [39], text and music [1], and disease progression [34]. We denote its distribution as $\Gamma\mathrm{P}(\gamma, \lambda, d)$, with discount parameter $d \in [0, 1)$, scale parameter $\lambda > 0$, mass parameter $\gamma > 0$, and rate measure $\nu(\mathrm{d}\theta) = \gamma \frac{\lambda^{1-d}}{\Gamma(1-d)} \theta^{-d-1} e^{-\lambda\theta} \mathrm{d}\theta$. The normalized gamma process [15, 18, 25, 30, 32] is a popular prior over probability measures, and in the case of $d = 0$ yields the Dirichlet process [9, 36]. Appendix A provides additional example applications of our main result for three other CRMs: the beta process [5, 37], the beta prime process [6], and a novel generalized gamma process.

## 3 Constructing non-nested finite approximations

Finite approximations are a powerful and convenient approach to obtaining practical approximate inference schemes for BNP models. It remains to demonstrate how we can construct an accurate finite approximation in practice. Campbell et al. [7] recently provided a thorough study of constructions for truncated approximations, but we are unaware of any general-purpose results on constructing non-nested approximations, which will be our focus in this section. Specifically, our main result shows how to construct NNFAs that converge in distribution to CRMs with rate measures of a particular form. As an important special case, if the CRM is an exponential family CRM [6] and the "discount" parameter $d = 0$, then the NNFA is constructed from random variables in the same exponential family, a connection which is useful for approximate inference algorithms. We leave it for future work to obtain error guarantees on these finite approximations, which are available for TFAs (see Campbell et al. [7] and citations therein).

Formally, NNFAs take the following form. For probability measures $H$ and $\nu_n$, write $\Theta_n \sim \mathrm{NNFA}_n(H, \nu_n)$ if

$$\Theta_n = \textstyle\sum_{i=1}^n \theta_{n,i} \delta_{\psi_{n,i}} \qquad \theta_{n,i} \overset{\text{indep}}{\sim} \nu_n \qquad \psi_{n,i} \overset{\text{i.i.d.}}{\sim} H.$$

We consider CRMs with rate measures $\nu$ with densities that, near zero, are (essentially) proportional to $\theta^{-1-d}$, where $d \in [0, 1)$ is the "discount" parameter. (This family of CRMs includes the most popular BNP priors.) We will define a sequence of NNFAs that converge in distribution to such a CRM. Our NNFA construction requires the following definition.

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

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

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

Our main result now shows how to construct NNFAs that provably converge to our family of CRMs.

**Theorem 3.2.** *For $d \in [0, 1)$ and $\eta \in E \subseteq \mathbb{R}^d$, let $\Theta \sim \mathrm{CRM}(H, \nu(\cdot; d, \eta))$, where*

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

*Assume that:*

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

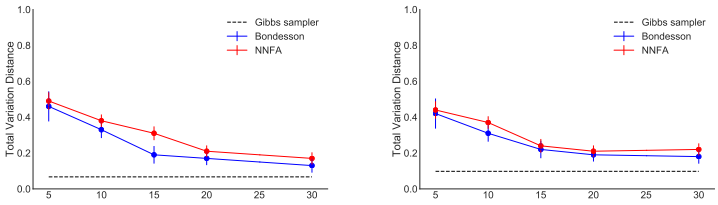


Figure 1: Comparison of the performance of the two approximate models on the total variation distance between the approximate and true distributions on number of clusters for varying approximation levels $n$ ($x$-axis). Plots refer to different dataset of sizes $N = 500$ (left) and $N = 1000$ (right) in $\mathbb{R}^4$.

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

$$\nu_n(\mathrm{d}\theta) := \theta^{-1+cn^{-1}-dS_{b_n}(\theta-an^{-1})}g(\theta)^{cn^{-1}-d}h(\theta;\eta)Z_n^{-1}\mathrm{d}\theta$$

*be a family of probability densities, where $Z_n$ is chosen such that $\int\nu_n(\mathrm{d}\theta) = 1$. If $\Theta_n \sim \mathrm{NNFA}_n(H,\nu_n)$, then $\Theta_n \overset{\mathcal{D}}{\Longrightarrow} \Theta$.*

The NNFAs obtained from Theorem 3.2 are particularly simple when the discount parameter $d = 0$.

**Corollary 3.3.** *Under the conditions of Theorem 3.2, if $d = 0$, then $\nu_n(d\theta) = \theta^{-1+cn^{-1}}g(\theta)^{cn^{-1}}h(\theta;\eta)/Z(cn^{-1},\eta)d\theta$. In particular, if $\nu$ takes the form of an (improper) exponential family and $d = 0$, then $\nu_n$ will belong to the same (but proper) exponential family.*

Corollary 3.3 is sufficient to recover all known NNFA results, including those for the Dirichlet process (i.e., the normalized gamma process) [13, 14] and the beta process [28]. We next apply Theorem 3.2 to the gamma process, with additional examples deferred to Appendix A.

**Example 3.1** (Gamma process)**.** Taking $E = \mathbb{R}_+$, $g(\theta) = 1$, $h(\theta;\eta) = e^{-\eta\theta}$, and $Z(\xi,\eta) = \Gamma(\xi)\eta^{-\xi}$ in Theorem 3.2 yields the gamma process, with $p(\theta;\xi,\eta) = \mathsf{Gam}(\theta;\xi,\eta)$. Since $h(\theta;\eta)$ is continuous and bounded on $[0,1]$, the hypotheses of Theorem 3.2 hold. In the case of $d = 0$, $c = \gamma\eta$ and $\nu_n(\theta) = \mathsf{Gam}(\theta;\gamma\eta/n,\eta)$.

## 4 Experiments

As a proof-of-concept, we consider a Dirichlet process mixture model with Gaussian observations with known isotropic covariance:

$$\Xi \sim \mathrm{DP}(\alpha,H), \qquad X_m \,|\, \Xi \sim \textstyle\sum_{i=1}^{\infty}\xi_i\mathcal{N}(\psi_i,\sigma I), \qquad m \geq 1,$$

where the base measure is $H = \mathcal{N}(0,\sigma_0^2 I)$. Instead of directly approximating the Dirichlet process we approximate a gamma process and use the fact that if $\Theta \sim \Gamma\mathrm{P}(\gamma,\alpha,0)$ then $\Theta/\Theta(\Psi) \sim \mathrm{DP}(\alpha)$. We compare a TFA called the Bondesson representation [7] to the NNFA obtained in Example 3.1. We use a standard CRP Gibbs sampler to obtain an approximation to the ground truth posterior.

We generated several datasets from a modified Pitman-Yor process with concentration parameter 2, and discount parameter 0.25. We implemented the truncated and non-nested approximations in Stan [8] using HMC and compared the accuracy of the approximate posteriors to a non-finite approximation of the "exact" posterior. (Our metrics and experiments are described in detail in Appendix B.) Comparison of a co-clustering metric and test log-likelihood indicated that both the Bondesson and NNFA approximations performed very well across component levels considered. Fig. 1 shows the total variation distance between the finite and non-finite approximations of the posterior marginal over number of clusters as the approximation level $n$ varies. While the noise is sufficiently high across data sets in this experiment as to be inconclusive, it seems plausible that the Bondesson approximation is actually yielding the best performance. We expect that NNFA will demonstrate better performance in more complex modeling scenarios, in comparisons that more directly measure mixing quality, and in cases where parallelism is of interest. We plan to investigate these cases in future work.

**Acknowledgments**

JHH, LM, and TB are supported in part by ONR grant N00014-17-1-2072, ONR MURI grant N00014-11-1-0688, and a Google Faculty Research Award.

## References

[1] A. Acharya, J. Ghosh, and M. Zhou. Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices. In *AISTATS*, 2015.

[2] M. Betancourt. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv.org*, Jan. 2017.

[3] 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.

[4] A. Brix. Generalized gamma measures and shot-noise cox processes. *Advances in Applied Probability*, 31:929–953, 1999.

[5] T. Broderick, M. I. Jordan, and J. Pitman. Beta Processes, Stick-Breaking and Power Laws. *Bayesian Analysis*, 7(2):439–476, 2012.

[6] T. Broderick, A. C. Wilson, and M. I. Jordan. Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*, 2017.

[7] T. Campbell, J. H. Huggins, J. P. How, and T. Broderick. Truncated random measures. *arXiv.org*, March 2016.

[8] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017.

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

[10] T. S. Ferguson and M. J. Klass. A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics*, 43(5), 1972.

[11] 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.

[12] T. L. Griffiths and Z. Ghahramani. Infinite Latent Feature models and the Indian Buffet Process. In *Advances in Neural Information Processing Systems*, 2005.

[13] H. Ishwaran and L. F. James. Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information. *Journal of Computational and Graphical Statistics*, 11(3):508–532, Sept. 2002.

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

[15] L. F. James. Poisson Process Partition Calculus with applications to Exchangeable models and Bayesian Nonparametrics. *arXiv.org*, May 2002.

[16] L. F. James. Stick-breaking PG($\alpha$,$\zeta$)-Generalized Gamma Processes. *arXiv.org*, Aug. 2013.

[17] L. F. James. Bayesian Poisson calculus for latent feature modeling via generalized Indian Buffet Process priors. *The Annals of Statistics*, 45(5):2016–2045, Oct. 2017.

[18] L. F. James, A. Lijoi, and I. Prünster. Posterior Analysis for Normalized Random Measures with Independent Increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

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

[20] O. Kallenberg. *Foundations of modern probability*. Springer, New York, 2nd edition, 2002.

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

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

[23] A. Kucukelbir, R. Ranganath, A. Gelman, and D. M. Blei. Automatic Variational Inference in Stan. In *Advances in Neural Information Processing Systems*, June 2015.

[24] 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.

[25] A. Lijoi and I. Prünster. Models beyond the dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. Walker, editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge University Press, 2010.

[26] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011.

[27] P. Orbanz. Conjugate Projective Limits. *arXiv.org*, Dec. 2010.

[28] J. Paisley and M. I. Jordan. A constructive definition of the beta process. *arXiv.org*, Apr. 2016.

[29] 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.

[30] J. Pitman. Poisson-kingman partitions. *Lecture Notes-Monograph Series*, 2003.

[31] R. Ranganath, S. Gerrish, and D. M. Blei. Black Box Variational Inference. In *International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

[32] E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585, 2003.

[33] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, Nov. 1996.

[34] A. Saeedi and A. Bouchard-Côté. Priors over Recurrent Continuous Time Processes. In *Advances in Neural Information Processing Systems*, pages 2052–2060, 2011.

[35] S. Saria, D. Koller, and A. Penn. Learning individual and population level traits from clinical temporal data. Technical report, 2010.

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

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

[38] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, Dec. 2006.

[39] M. Titsias. The infinite gamma-poisson feature model. In *Advances in Neural Information Processing Systems*, 2008.

## A Additional Examples

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

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

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

Since $h$ is continuous and bounded on $[0, 1/2]$, Theorem 3.2 applies. In the case of $d = 0$, $c = \gamma\eta$ and

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

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

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

Since $g$ is continuous, $g(0) = 1$, $1 \leq g(\theta) \leq 1 + \theta$, and $h(\theta; \eta)$ is continuous and bounded on $[0, 1]$, Theorem 3.2 applies. In the case of $d = 0$, $c = \gamma\eta$ and

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

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

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

which is the rate measure for the gamma process $\Gamma\mathrm{P}(\gamma, \eta, d)$. Since $h(\theta; \eta)$ is continuous and bounded on $[0, 1]$, Theorem 3.2 applies. In the case of $d = 0$, $c = \frac{\gamma\eta_1\eta_2}{\Gamma(\eta_2^{-1})}$ and

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

## B Experiments and Metrics for Checking Approximate Model Quality

The datasets for which we report the results in figure Fig. 1 were generated from the following modified Pitman Yor process with concentration parameter $\alpha = 2$, discount parameter $\theta = 0.25$, and base measure $H = \mathcal{N}(0, I)$. We first drew all the labels $\{z_i\}_{i=1}^M$ for the $M$ points in the dataset, sequentially, according to the Pitman Yor scheme. We then drew $C$ means $\mu_i \sim H$, where $C$ is the number of distinct labels. To make sure that the clusters were somewhat separated, we forced the means to satisfy $d(\mu_i, \mu_j) > 0.25$ for all $i \neq j$, where $d$ is the Euclidean distance, by rejecting those that were too close to the ones already instantiated. Last, we sampled independently $X_m \sim N(\mu_{z_m}, I)$ for $m = 1, \ldots, M$.

Let $\mathcal{D} = \{X_m\}_{m=1}^M$ denote the observed data.

**Test log-likelihood.** Let $\mathcal{D}' = \{X'_j\}_{j=1}^J$ be a held out test set. The test log-likelihood is $\mathrm{TLL}(\mathcal{D}' \,|\, \mathcal{D}) := \sum_{j=1}^J \log \mathbb{E}[p(X'_j \,|\, \Theta) \,|\, \mathcal{D}]$, where the expectation is with respect to either the true or the approximate model.

**Co-clustering error.** In the mixture model case let $Z_m$ denote the cluster assignment of $X_m$. Calculate the matrix $\mathrm{CCP}_{m\ell}(\mathcal{D}) = \mathbb{P}[Z_m = Z_\ell \,|\, \mathcal{D}]$ under the true and approximate model. Then for a matrix norm $\|\cdot\|$, the co-clustering error is $\mathcal{E}_{CC}(\mathcal{D}, \|\cdot\|) = \|\mathrm{CCP}_{true}(\mathcal{D}) - \mathrm{CCP}_{approx}(\mathcal{D})\|$.

**Distance between the distributions over the number of clusters.** Let $p_{\mathcal{D}}(k) = \mathbb{P}[\#\{z_1, \ldots, z_M\} = k \,|\, \mathcal{D}]$ be the probability that the data is in $k$ clusters. We calculate the total variation distance between this distribution under true and approximate model: $d_{TV}(p_{\mathcal{D},true}, p_{\mathcal{D},approx}) = \frac{1}{2}\sum_{k=1}^\infty |p_{\mathcal{D},true}(k) - p_{\mathcal{D},approx}(k)|$.

[2] `https://en.wikipedia.org/wiki/Generalized_gamma_distribution`

## C Proofs

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

**Lemma C.1** ([20, Lemmas 12.1 and 12.2])**.** *Let $\Theta$ be a random measure and $\Theta_1, \Theta_2, \ldots$ a sequence of random measures. If for all measurable sets $A$ and $t > 0$,*

$$\lim_{n\to\infty} \mathbb{E}[e^{-t\Theta_n(A)}] = \mathbb{E}[e^{-t\Theta(A)}],$$

*then $\Theta_n \overset{\mathcal{D}}{\Longrightarrow} \Theta$.*

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

**Proposition C.2.** *Let $\Theta \sim \mathrm{CRM}(H, \nu)$ and for $n = 1, 2, \ldots$, let $\Theta_n \sim \mathrm{NNFA}_n(H, \nu_n)$ where $\nu$ is a measure and $\nu_1, \nu_2, \ldots$ are probability measures on $\mathbb{R}_+$, all absolutely continuous with respect to Lebesgue measure. If $\|\mu(1, n\nu_n) - \mu(1, \nu)\|_1 \to 0$, then $\Theta_n \overset{\mathcal{D}}{\Longrightarrow} \Theta$.*

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

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

We have

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

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

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

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

$$\lim_{n\to\infty} \mathbb{E}[e^{-t\Theta_n(A)}] = \lim_{n\to\infty} \mathbb{E}[e^{-t\theta_{n,1}\mathbb{1}(\psi_{n,1}\in A)}]^n,$$

so result follows from Lemma C.1. □

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

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

*then*

$$\|\mu - \mu_n\|_1 \to 0.$$

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

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

The conclusion follows by dominated convergence. □

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

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

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

$$f_n(\theta) = nZ_n^{-1}(1 - e^{-\theta})\theta^{-1+cn^{-1}+d-dS_{b_n}(\theta-an^{-1})} g(\theta)^{-1+cn^{-1}},$$

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

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

For the remainder of the proof we will consider $n$ large enough that $an^{-1} + 2b_n$ and $cn^{-1}$ are less than $\kappa$. The normalizing constant $Z_n$ can be written as

$$\begin{aligned}
Z_n &= \int_0^{a/n} (\theta g(\theta))^{-1+cn^{-1}} \pi'(\mathrm{d}\theta) \\
&\quad + \int_{a/n}^\kappa \theta^{-1+cn^{-1}-dS_{b_n}(\theta-an^{-1})} g(\theta)^{-1+cn^{-1}} \pi'(\mathrm{d}\theta) \\
&\quad + \int_\kappa^\infty (\theta g(\theta))^{-1+cn^{-1}-d} \pi'(\mathrm{d}\theta).
\end{aligned}$$

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

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

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

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

For the third term,

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

Hence, $Z_n = \frac{n}{\gamma} + o(n)$ and $nZ_n^{-1} = \gamma(1+e_n)$, where $e_n = o(1)$.

Next, we have

$$\begin{aligned}&\sup_{\theta\in[\kappa,\infty)} |f(\theta) - f_n(\theta)|\\ &= \sup_{\theta\in[\kappa,\infty)} (1-e^{-\theta})(\theta g(\theta))^{-1}|\gamma - nZ_n^{-1}(\theta g(\theta))^{cn^{-1}}|\\ &\le \sup_{\theta\in[\kappa,\infty)} \gamma(\theta g(\theta))^{-1}|1-(1+e_n)(\theta g(\theta))^{cn^{-1}}|\\ &\le \gamma \sup_{\theta\in[\kappa,\infty)} (\theta g(\theta))^{-1}|1-(\theta g(\theta))^{cn^{-1}}|\\ &\quad + \gamma e_n \sup_{\theta\in[\kappa,\infty)} (\theta g(\theta))^{-1+cn^{-1}}.\end{aligned} \tag{C.1}$$

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

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

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

$$\begin{aligned}&\gamma \sup_{\theta\in[\kappa,\infty)} (\theta g(\theta))^{-1}|1-(\theta g(\theta))^{cn^{-1}}|\\ &\le \gamma \sup_{\psi\in[\tilde{\kappa},\infty)} \psi^{-1}|1-\psi^{cn^{-1}}|\\ &\le \gamma \sup_{\psi\in[\tilde{\kappa},1]} \psi^{-1}|1-\psi^{cn^{-1}}| + \gamma \sup_{\psi\in[1,\infty)} \psi^{-1}|1-\psi^{cn^{-1}}|\\ &\le \gamma\tilde{\kappa}^{-1} \sup_{\psi\in[\tilde{\kappa},1]} |1-\psi^{cn^{-1}}| + \gamma\left(\frac{n-c}{n}\right)^{nc^{-1}}\left|1-\frac{n}{n-c}\right|\\ &\le \gamma\tilde{\kappa}^{-1}(1-\tilde{\kappa}^{cn^{-1}}) + O(1)\times\frac{c}{n-c}\\ &= \gamma\tilde{\kappa}^{-1}\times o(1) + O(n^{-1})\\ &\to 0.\end{aligned}$$

Similarly, For the second term in Eq. (C.1) we have

$$\begin{aligned}\gamma e_n \sup_{\theta\in[\kappa,\infty)} (\theta g(\theta))^{-1+cn^{-1}} &\le \gamma e_n \sup_{\psi\in[\tilde{\kappa},\infty)} \psi^{-1+cn^{-1}}\\ &\le \gamma\tilde{\kappa}^{-1}e_n\\ &\to 0.\end{aligned}$$

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

$$\begin{aligned}&\int_0^\kappa |f'(\theta) - f_n'(\theta)|\mathrm{d}\theta = \int_0^\kappa (\theta g(\theta))^{-d}|f(\theta) - f_n(\theta)|\mathrm{d}\theta\\ &= \int_0^\kappa (1-e^{-\theta})(\theta g(\theta))^{-1-d}|\gamma - nZ_n^{-1}\theta^{cn^{-1}+d-dS_{b_n}(\theta-an^{-1})}g(\theta)^{cn^{-1}}|\mathrm{d}\theta\end{aligned}$$

$$\leq \gamma[c^*(1+\kappa)]^{1+d} \int_0^\kappa \theta^{-d} |1 - (1+e'_n)\theta^{cn^{-1}+d-dS_{b_n}(\theta-an^{-1})}| \mathrm{d}\theta$$
$$\leq \gamma \int_0^\kappa \theta^{-d} |1 - \theta^{cn^{-1}+d-dS_{b_n}(\theta-an^{-1})}| \mathrm{d}\theta + \gamma e'_n \int_0^\kappa \theta^{cn^{-1}+d-dS_{b_n}(\theta-an^{-1})} \mathrm{d}\theta. \quad \text{(C.2)}$$

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

$$\begin{aligned}
\int_0^{an^{-1}} \theta^{-d} |1 - \theta^{d+cn^{-1}}| \mathrm{d}\theta &\leq \int_0^{an^{-1}} \theta^{-d} + \theta^{cn^{-1}} \mathrm{d}\theta \\
&= \frac{\theta^{1-d}}{1-d} + \frac{n}{c+n} \theta^{1+cn^{-1}} \Big|_0^{an^{-1}} \\
&= \frac{1}{1-d} (an^{-1})^{1-d} + \frac{n}{c+n} (an^{-1})^{1+cn^{-1}} \\
&\to 0.
\end{aligned}$$

The second part is equal to

$$\begin{aligned}
\int_{an^{-1}}^{an^{-1}+b_n} \theta^{-d} |1 - \theta^{cn^{-1}+d-dS_{b_n}(\theta-an^{-1})}| \mathrm{d}\theta &\leq \int_{an^{-1}}^{an^{-1}+b_n} \theta^{-d} + \theta^{cn^{-1}-d} \mathrm{d}\theta \\
&\leq 2 \int_{an^{-1}}^{an^{-1}+b_n} \theta^{-d} \mathrm{d}\theta \\
&= \frac{2}{1-d} \theta^{1-d} \Big|_{an^{-1}}^{an^{-1}+b_n} \\
&= \frac{2}{1-d} \left( (an^{-1}+b_n)^{1-d} - (an^{-1})^{1-d} \right) \\
&\to 0.
\end{aligned}$$

The third part is equal to

$$\begin{aligned}
\int_{an^{-1}+b_n}^{\kappa-b_n} \theta^{-d} |1 - \theta^{cn^{-1}}| \mathrm{d}\theta &= \int_{an^{-1}+b_n}^{\kappa-b_n} \theta^{-d} - \theta^{cn^{-1}-d} \mathrm{d}\theta \\
&= \frac{1}{1-d} \theta^{1-d} - \frac{n}{c+n(1-d)} \theta^{1-d+cn^{-1}} \Big|_{an^{-1}+b_n}^{\kappa-b_n} \\
&= \frac{(\kappa-b_n)^{1-d}}{1-d} - \frac{n}{c+n(1-d)} (\kappa-b_n)^{1-d+cn^{-1}} \\
&\quad - \frac{(an^{-1}+b_n)^{1-d}}{1-d} + \frac{n}{c+n} (an^{-1}+b_n)^{1-d+cn^{-1}} \\
&\to 0.
\end{aligned}$$

The fourth part is equal to

$$\begin{aligned}
\int_{\kappa-b_n}^{\kappa} \theta^{-d} |1 - \theta^{cn^{-1}}| \mathrm{d}\theta &\leq \int_{\kappa-b_n}^{\kappa} \theta^{-d} + \theta^{cn^{-1}-d} \mathrm{d}\theta \\
&\to 0
\end{aligned}$$

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

$$\gamma e'_n \int_0^\kappa \theta^{cn^{-1}-dS_{b_n}(\theta-an^{-1})} \mathrm{d}\theta \leq \gamma e'_n \int_0^\kappa \theta^{-d} \mathrm{d}\theta = \gamma e'_n \frac{\kappa^{1-d}}{1-d} = o(n).$$

Since $\sup_{\theta \in [0,\kappa]} \pi'(\theta) < \infty$ by the boundedness of $g$ and $h$ and $\pi$ is a probability density by construction, conclude using Lemma C.3 that $\|\mu - \mu_n\|_1 \to 0$. It then follows from Lemma C.1 that $\Theta_n \overset{\mathcal{D}}{\Longrightarrow} \Theta$. □