CATVI: Conditional and Adaptively Truncated Variational Inference for Hierarchical Bayesian Nonparametric Models

Yirui Liu
London School of Economics

Xinghao Qiao
London School of Economics

Jessica Lam
JP Morgan Chase & Co.

Abstract

Current variational inference methods for hierarchical Bayesian nonparametric models can neither characterize the correlation structure among latent variables due to the mean-field setting, nor infer the true posterior dimension because of the universal truncation. To overcome these limitations, we propose the conditional and adaptively truncated variational inference method (CATVI) by maximizing the nonparametric evidence lower bound and integrating Monte Carlo into the variational inference framework. CATVI enjoys several advantages over traditional methods, including a smaller divergence between variational and true posteriors, reduced risk of underfitting or overfitting, and improved prediction accuracy. Empirical studies on three large datasets reveal that CATVI applied in Bayesian nonparametric topic models substantially outperforms competing models, providing lower perplexity and clearer topic-words clustering.

1 Introduction

Hierarchical Bayesian nonparametric (HBNP) models are widely used in bioinfomatics, language processing, computer version and network analysis (Sudderth and Jordan, 2009; Caron and Fox, 2017; Williamson, 2016; Yurochkin et al., 2019). A major benefit of HBNP models is their ability to relax the fixed dimension assumption in parametric models. For example, in natural language processing, hierarchical Dirichlet process (HDP) model (Teh et al., 2006) replaces the finite-dimensional Dirichlet distribution in latent Dirichlet allocation (LDA) with a countable-dimensional Dirichlet process (DP). This is done by regarding the number of topics as a random variable that can be inferred from the data, rather than as a parametric value (Blei et al., 2003).

However, it is much harder to implement HBNP models than their parametric counterparts. In particular, due to a HBNP model's infinite-dimensional nature, a finite-dimensional truncation is needed to approximate the posterior. Yet, the selection of the optimal truncation level poses several challenges. On one hand, the traditional Markov chain Monte Carlo (MCMC) methods (Teh et al., 2006) can produce an adaptive selection of the truncated dimension, but they are not computationally scalable especially for big data. On the other hand, standard variational inference methods (Teh et al., 2008; Wang et al., 2011; Hoffman et al., 2013; Roychowdhury and Kulis, 2015; Xu et al., 2019) can accelerate the computation, but they suffer from an universal selection of the truncation level by truncating the dimension of all latent variables to a prespecified value. Using a prespecified value is problematic, because a subjective selection of the fixed truncation level can make the model prone to overfitting or underfitting, leading to low predictive accuracy. These existing challenges in universal truncation contradict the motivation and advantages of using HBNP models.

In this paper, we propose a general framework, called conditional and adaptively truncated variational inference (CATVI), to infer HBNP models in the following steps. First, we convert the inference problem to an optimization task of maximizing our proposed nonparametric evidence lower bound based on finite partitions. Second, we introduce a conditional setting when factorizing variational distributions by conditioning variables in the middle layers on two adjacent layers. Third, to handle big data, we develop a stochastic variational inference framework (Blei et al., 2017) under our conditional setting. Finally, we obtain empirical distributions from Monte Carlo sampling of local latent variables, which are then used to update the variational parameters for the global latent vari-

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

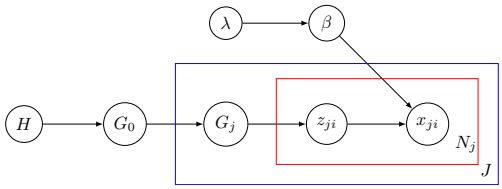

Figure 1: The HBNP models. The blue and red boxes correspond to J and N_j replicates, respectively.

ables. This enables us to truncate the dimension of the latent variational distributions to that of the empirical distribution.

Our proposed method benefits from both the inferential accuracy of Monte Carlo sampling and the computational efficiency of variational inference. First, our method rebuilds the correlation structure and hence attains a smaller Kullback–Leibler (KL) divergence between the variational distribution and the true posterior. Such procedure removes the unrealistic mean-field assumption, and searches for an optimal variational distribution over a wider family. Second, it adjusts the dimension of variation distributions, which converges to a stable level balancing the goodness-of-fit and model complexity. With these advantages, CATVI provides an adaptive selection of the truncated dimension, reducing the risk of overfitting or underfitting, while also enabling more accurate predictions without sacrificing the computational efficiency. Specific to the inference for the HDP model, CATVI enjoys several advantages over existing methods (Teh et al., 2006; Hoffman et al., 2013; Wang and Blei, 2012; Bryant and Sudderth, 2012), see Section 6 for our detailed discussion.

2 Background: HBNP Models

As a subclass of Bayesian nonparametric models, HBNP models extend the simplicity of using random measures (see Appendix A) as priors to the following hierarchical structure,

$$G_0|H \sim P(H), \quad \beta|\lambda \sim p(\beta|\lambda), \quad G_j|G_0 \sim R(G_0), \\ z_{ji}|G_j \sim G_j, \quad x_{ji}|z_{ji} \sim f(x_{ji}|\beta, z_{ji}), \tag{1}$$

for $j = 1, \ldots, J, i = 1, \ldots, N_j$, as illustrated in Figure 1. In the top layer, $G_1, \ldots, G_J$ are generated from a random measure R with common base measure G_0, while in the bottom layer, G_0 itself is a realization of random measure P with base measure H. To ensure exchangeability, $G_1, \ldots, G_J$ are assumed to be identical and independent given G_0. Each local latent variable z_{ji} is sampled from G_j independently. Finally, the global parameter β is assigned a prior $p(\beta|\lambda)$, and the observation x_{ji} is generated from a likelihood function f, parameterized by both global latent variable β and local latent variables z_{ji}.

In topic modelling, the HDP model (Teh et al., 2006) uses a DP for both P and R in (1) as,

$$G_0|H \sim \mathrm{DP}(\alpha H), \quad G_j|G_0 \sim \mathrm{DP}(\gamma G_0), \tag{2}$$

where α, γ are concentration parameters, and H, G_0 are normalized based measures (see Appendix A). Suppose a corpus has J documents, each document j has N_j words, and each word is chosen from a vocabulary with W terms. Specifically, $G_0 = \sum_{k=1}^{\infty} G_{0k}\delta_{\phi_k}$ is generated from the distribution $\mathrm{DP}(\alpha H)$, and for each document j, a topic proportion, defined as $G_j = \sum_{k=1}^{\infty} G_{jk}\delta_{\phi_k}$, is independently sampled from the distribution $\mathrm{DP}(\gamma G_0)$. For each topic k, the distribution of words over vocabulary is sampled from a W-dimensional Dirichlet distribution parameterized by η, $\beta_k = (\beta_{k,1}, \cdots, \beta_{k,W})^{\mathrm{T}} \sim \mathrm{Dir}(\eta)$. For each word i in document j, a topic assignment $z_{ji} = \phi_k$ is chosen from $z_{ji} \sim \mathrm{Multinomial}(G_j)$, where ϕ_k represents topic k. Finally, the observation x_{ji} is generated from the assigned topic and the corresponding within-topic word distribution, $x_{ji}|\{z_{ji} = \phi_k\} \sim \mathrm{Multinomial}(\beta_k)$.

The necessity to let G_0 be atomic can be shown in the HDP model. If $G_1, \ldots, G_J$ are sampled from a Dirichlet process with a diffuse base measure instead of an atomic G_0, $G_1, \ldots, G_J$ will not share any support almost surely, and thus none of the topics being shared across the documents. However, for a general HBNP model, as long as G_0 is atomic, it is not necessary to restrict the prior for G_0 to be a Dirichlet process or a probability random measure. For example, the ΓDP model, which has the following structure,

$$G_0|H \sim \Gamma\mathrm{P}(\alpha H), \quad G_j|G_0 \sim \mathrm{DP}(G_0), \tag{3}$$

allows for more flexibility by removing the constraint on the concentration parameter in the top layer. Other choices of the prior for G_0 include beta process, stable process and inverse Gaussian process (Ghosal and Van der Vaart, 2017).

To infer the HBNP models, we set up the theoretical foundations for nonparametric KL divergence and evidence lower bound, and then propose a novel methodology in the following Sections 3 and 4.

3 Nonparametric Evidence Lower Bound

3.1 KL Divergence between Random Measures

The object of variational inference is to minimize the KL divergence between the variational distribution

and the true posterior. For two infinite-dimensional random measures, their KL divergence is well defined even though an infinite-dimensional density function does not exist in a conventional sense. Given two random measures P and Q from $(\Theta, \mathcal{M})$ into $(\Omega, \mathcal{F})$, the Radon–Nikodym derivative dQ/dP exits if Q is absolutely continuous with respect to P. Their KL divergence is defined as

$$\mathcal{KL}(Q \parallel P) = \int_{\Theta} \log(dQ/dP) dQ,$$

which is intractable due to the infinite-dimensional integral (Matthews et al., 2016). We have developed a new approach to calculate it using the limit superior of the divergence between corresponding finite-dimensional induced measures, that is,

$$\mathcal{KL}(Q \parallel P) = \limsup_{\varOmega} \mathcal{KL}(q^{\varOmega} \parallel p^{\varOmega}), \tag{4}$$

where $p^{\varOmega}$ and $q^{\varOmega}$ are respectively induced measures from P and Q on a finite partition $\varOmega = (A_1, \ldots, A_n)$, such that $p^{\varOmega}(A_i) = P(A_i)$ and $q^{\varOmega}(A_i) = Q(A_i)$ for each $A_i \in \varOmega$. With an induced random variable $Z^{\varOmega}\colon \Theta \to \mathbb{R}^n$, we can also denote the induced measures by $p(Z^{\varOmega})$ and $q(Z^{\varOmega})$. The result in (4) is justified in Appendix C.1. We use the following two examples to illustrate (4).

Example 1 For Poisson processes $P = \mathrm{PP}(\Lambda + b\delta_{\phi})$ and $Q = \mathrm{PP}(\Lambda + a\delta_{\phi})$, where Λ is the intensity function defined on Ω, $a, b \in \mathbb{R}^+$, and δ_{ϕ} is a Dirac function at point $\phi \in \Omega$. Under partition $\varOmega = (\phi, \Omega/\phi)$, the limit superior in (4) is achieved, that is,

$$\mathcal{KL}(Q \parallel P) = \mathcal{KL}\big(\mathrm{Pois}(a) \parallel \mathrm{Pois}(b)\big),$$

where $\mathrm{Pois}(a)$ is the Poisson distribution with intensity a.

Example 2 For Dirichlet processes $P = \mathrm{DP}(\alpha H + \sum_{i=1}^{n} b_i \delta_{\phi_i})$ and $Q = \mathrm{DP}(\alpha H + \sum_{i=1}^{n} a_i \delta_{\phi_i})$, where H is the base measure, α is the concentration parameter, $a_i, b_i \in \mathbb{R}^+$ and $\phi_i \in \Omega$ for $i = 1, \ldots n$. Similarly, under the partition $\varOmega = \big(\phi_1, \ldots, \phi_n, \Omega/\{\phi_i\}_{i=1}^n\big)$,

$$\mathcal{KL}(Q \parallel P) =$$
$$\mathcal{KL}\big(\mathrm{Dir}(\alpha, a_1, \ldots, a_n) \parallel \mathrm{Dir}(\alpha, b_1, \ldots, b_n)\big),$$

where $\mathrm{Dir}(\alpha, a_1, \ldots, a_n)$ is the Dirichlet distribution with parameters $\alpha, a_1, \ldots, a_n$.

With the KL divergence between random measures represented under a finite partition, we can then define the nonparametric counterpart of evidence lower bound below.

3.2 Nonparametric Evidence Lower Bound

The parametric variational inference algorithm uses a finite-dimensional variational distribution to approximate the posterior by maximizing the evidence lower bound (Blei et al., 2017). In contrast, HBNP models uses a random measure for the variational distribution, due to the infinite dimensionality of latent variables. We propose a general inference framework for HBNP models by maximizing the nonparametric evidence lower bound (NPELBO), defined as the limit inferior of the parametric evidence lower bound under a finite partition, $\liminf_{\varOmega}(\mathrm{ELBO}^{\varOmega})$, that is

$$\liminf_{\varOmega} \big\{ \mathrm{E}_{q(Z^{\varOmega})} \log p(X, Z^{\varOmega}) - \mathrm{E}_{q(Z^{\varOmega})} \log q(Z^{\varOmega}) \big\}, \tag{5}$$

where $p(X, Z^{\varOmega})$ and $q(Z^{\varOmega})$ correspond to the induced measures from the joint distribution and the variational distribution on $\varOmega$, and where Z and X are the observations and latent variables, respectively. Moreover, given the KL divergence between random measures in (4), in Appendix C.2 we show that

$$\mathcal{KL}\big(Q(Z) \parallel P(Z|X)\big) + \mathrm{NPELBO} = \log p(X). \tag{6}$$

This demonstrates the equivalence between maximizing the NPELBO in (5) and minimizing the KL divergence between the variational distribution $Q(Z)$ and the true posterior $P(Z|X)$. The task of maximizing the NPELBO is general and can be applied broadly within Bayesian nonparametrics. To simplify notation, we will use $p(\cdot)$ and $q(\cdot)$ to denote the true and variational distributions, respectively, where the context is clear. To infer HBNP models, we aim to maximize the defined NPELBO, while truncating the dimension of variational distribution adaptively as follows.

4 Methodology

CATVI adopts the stochastic variational inference framework (Hoffman et al., 2013), where the computation is accelerated by selecting a small batch of data and updating variational parameters with an unbiased random gradient. We first build the foundation of conditional variational inference as follows.

4.1 Conditional Variational Inference

Conditional setting HBNP models in (1) contain global latent variable β, local latent variables $\boldsymbol{z}$, global prior G_0, local priors $\boldsymbol{G}_{[J]}$ and observations $\boldsymbol{x}$, where $\boldsymbol{z} = \{\boldsymbol{z}_j\}_{j=1}^{J}$, $\boldsymbol{z}_j = \{z_{ji}\}_{i=1}^{N_j}$, $\boldsymbol{x} = \{\boldsymbol{x}_j\}_{j=1}^{J}$, $\boldsymbol{x}_j = \{x_{ji}\}_{i=1}^{N_j}$ and $\boldsymbol{G}_{[J]} = \{G_j\}_{j=1}^{J}$. We aim to find the variational distribution to maximize the NPELBO. In contrast to traditional approaches under the mean-field setting, we factorize the variational distribution

as

$$q(\beta, \boldsymbol{z}, G_0, \boldsymbol{G}_{[J]}) = q(\beta)q(G_0)\prod_{j=1}^{J} q(G_j|G_0, \boldsymbol{z}_j)\prod_{i=1}^{N_j} q(z_{ji}), \tag{7}$$

in the sense of the probability law. Such conditional design facilitates the recovery of the dependence structure among G_0, $\boldsymbol{G}_{[J]}$ and $\boldsymbol{z}$.

Combing (5) and (7), we seek to maximize the following NPELBO:

$$\begin{aligned}\liminf_{\varrho} \Big\{ & \mathrm{E}_{q(\beta,\boldsymbol{z},G_0^\varrho,\boldsymbol{G}_{[J]}^\varrho)} \log p(\boldsymbol{x}, \beta, \boldsymbol{z}, G_0^\varrho, \boldsymbol{G}_{[J]}^\varrho) \\ & - \sum_{j=1}^{J} \mathrm{E}_{q(G_0^\varrho)} \mathrm{E}_{q(\boldsymbol{z}_j)} \mathrm{E}_{q(G_j^\varrho|G_0^\varrho,\boldsymbol{z}_j)} \log q(G_j^\varrho|G_0^\varrho, \boldsymbol{z}_j) \\ & - \mathcal{H}\big(q(G_0^\varrho)\big) - \mathcal{H}\big(q(\beta)\big) - \sum_{j=1}^{J}\sum_{i=1}^{N_j} \mathcal{H}\big(q(z_{ji})\big) \Big\},\end{aligned} \tag{8}$$

where the entropy $\mathcal{H}\big(q(\cdot)\big) = E_{q(\cdot)} \log q(\cdot)$ and Ω is a partition of the sample space Ω for G_0 and $\boldsymbol{G}_{[J]}$.

Conditional variational distribution To maximize the NPELBO in (8), we first compute the optimal variational distribution of G_j given G_0 and $\boldsymbol{z}_j$ for each j. As $p(\boldsymbol{x}, \beta, \boldsymbol{z}, G_0^\varrho, \boldsymbol{G}_{[J]}^\varrho) = p(G_0^\varrho, \boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \prod_{j=1}^{J} p(G_j^\varrho|G_0^\varrho, \boldsymbol{z}_j)$, the non-constant term in (8) with respect to $q(G_j|G_0, \boldsymbol{z}_j)$ is

$$\begin{aligned}\liminf_{\varrho} \Big\{ & \sum_{j=1}^{J} \mathrm{E}_{q(G_0^\varrho)} \mathrm{E}_{q(\boldsymbol{z}_j)} \mathrm{E}_{q(G_j^\varrho|G_0^\varrho,\boldsymbol{z}_j)} \log p(G_j^\varrho|G_0^\varrho, \boldsymbol{z}_j) \\ & - \log q(G_j^\varrho|G_0^\varrho, \boldsymbol{z}_j) \Big\}.\end{aligned}$$

Note that the above expression can be viewed as the negative of a KL divergence whose maximum is zero. Therefore, to enable the NPELBO to reach the maximum, the optimal conditional variational distribution for G_j should be $p(G_j|G_0, \boldsymbol{z}_j)$. Consequently, NPELBO in (8) does not contain any term related to $q(G_j|G_0, \boldsymbol{z}_j)$. In Appendix C.3, we derive NPELBO with respect to $q(G_0^\varrho)$ as

$$\begin{aligned}\liminf_{\varrho} \Big\{ & \sum_{j=1}^{J} \mathrm{E}_{q(G_0^\varrho)} \mathrm{E}_{q(\boldsymbol{z}_j)} \log \mathrm{E}_{p(G_j^\varrho|G_0^\varrho)} p(\boldsymbol{z}_j|G_j^\varrho) \\ & - \mathcal{KL}\big(q(G_0^\varrho) \| p(G_0^\varrho)\big) \Big\}\end{aligned} \tag{9}$$

up to a constant. It is important to note that $\mathrm{E}_{p(G_j^\varrho|G_0^\varrho)}$ is with respect to the prior $p(G_j^\varrho|G_0^\varrho)$ instead of the variational distribution $q(G_j^\varrho|G_0^\varrho)$, and hence this expectation can often be calculated analytically in HBNP models due to the conjugacy.

4.2 Empirical Distribution and Evidence Lower Bound

Within the conditional variational framework, for the task of adaptive truncation, CATVI integrates Monte Carlo sampling to variational inference by iterating the following steps till convergence, (i) using Monte Carlo sampling to get an empirical optimal variational distribution for local variables $\boldsymbol{z}$ and (ii) updating the variational distributions for global variables G_0 and β.

Empirical distribution From the entire data $\boldsymbol{x}$, we randomly sample a subset $\{\boldsymbol{x}_s : \boldsymbol{x}_s \in \boldsymbol{x}\}_{s=1}^{S}$, where S is the batch size with $S \ll J$. Given a partition Ω in the current training iteration, we aim to update the parameters for $q(G_0^\varrho)$ conditional on $q(\beta)$ and $\{q(\boldsymbol{z}_s)\}_{s=1}^{S}$. While standard stochastic variational inference updates parameters analytically, we use Monte Carlo sampling to draw T_s samples for each $\boldsymbol{z}_s$ from $q(\boldsymbol{z}_s)$, thus constructing an empirical distribution,

$$\hat{q}(\boldsymbol{z}_s) = \frac{1}{T_s}\sum_{t=1}^{T_s} \delta_{\hat{\boldsymbol{z}}_{s,t}}, \quad \hat{\boldsymbol{z}}_{s,t} \sim q(\boldsymbol{z}_s).$$

Empirical evidence lower bound Using the empirical distribution $\hat{q}(\boldsymbol{z}_s)$, we obtain an empirical evidence lower bound with respect to $q(G_0^\varrho)$, $\widehat{\mathrm{ELBO}}^\varrho$, by replacing $q(\boldsymbol{z}_s)$ in (9) with $\hat{q}(\boldsymbol{z}_s)$, that is,

$$\begin{aligned}\sum_{s=1}^{S}\sum_{t=1}^{T_s} & \frac{J}{ST_s} \mathrm{E}_{q(G_0^\varrho)} \log \mathrm{E}_{p(G_s^\varrho|G_0^\varrho)} p(\hat{\boldsymbol{z}}_{s,t}|G_s^\varrho) \\ & - \mathcal{KL}\big(q(G_0^\varrho) \| p(G_0^\varrho)\big)\end{aligned} \tag{10}$$

up to a constant. It is obvious that $\mathrm{E}(\widehat{\mathrm{ELBO}}^\varrho) = \mathrm{ELBO}^\varrho$, thus satisfying the key condition for stochastic variational inference (Hoffman et al., 2013), that is, the random gradient is unbiased. Therefore, according to (10), we can use the random gradient generated from $\hat{\boldsymbol{z}}_s = \{\hat{\boldsymbol{z}}_{s,t}\}_{t=1}^{T_s}$ to update the parameters for $q(G_0^\varrho)$.

Resampling We next present the procedure to get the empirical distribution $\hat{q}(\boldsymbol{z}_s)$. As G_s is integrated out, the local latent variables $\{z_{si}\}_{i=1}^{N_s}$ can not be sampled independently when we use Monte Carlo sampling to draw $\hat{\boldsymbol{z}}_s$ given $q(G_0^\varrho)$ and $q(\beta)$. Therefore, we propose the following Gibbs sampling approach to get samples under optimal variational distributions. Conditional on $q(G_0^\varrho)$, $q(\beta)$ and samples $\hat{\boldsymbol{z}}_{s,i^-} = \{\hat{z}_{sl} : l = 1, \ldots, N_s, l \neq i\}$, it follows from (8) that the optimal variational distribution of $\log q(z_{si})$ is proportional to

$$\begin{aligned}& \mathrm{E}_{q(G_0^\varrho)} \mathrm{E}_{p(G_j^\varrho|G_0^\varrho)} p(z_{si}, \hat{\boldsymbol{z}}_{s,i^-}|G_s^\varrho) \\ & + \mathrm{E}_{q(\beta)} \log p(x_{si}|z_{si}, \beta).\end{aligned} \tag{11}$$

Then we sample $\hat{z}_{si} \sim q(z_{si})$ for each i iteratively, which constructs a Markov chain. Noting that $q(z_{si})$ is

often multinomial, sampling from its logarithm is commonly used. After the convergence, we can resample $\hat{\boldsymbol{z}}_{s,1}, \ldots, \hat{\boldsymbol{z}}_{s,T_s}$ from the stable Markov chain to update the parameters of $q(G_0^o)$ according to (10). Similarly, we derive the empirical evidence lower bound with respect to $q(\beta)$ in Appendix B.1 and can update the parameters for $q(\beta)$ using $\hat{\boldsymbol{z}}_{s,1}, \ldots, \hat{\boldsymbol{z}}_{s,T_s}$ correspondingly.

4.3 Adaptive Truncation

Finally, we seek to obtain the finite partition Ω that could reach the limit inferior in NPELBO. Rather than having Ω fixed on an universal truncation level, we enable the dimension of Ω to gradually adjust to a stable level. This partition or truncation is dependent on data-fitting and embedded within the optimization process, providing another key advantage of using a Monte Carlo sampling scheme in the stochastic variational inference framework.

Partition refinement According to the structure of HBNP models, $\hat{z}_{si}$ are sampled from the atomic support of G_0, $\{\phi_k\}_{k=1}^{\infty}$. Without loss of generality, we assume that the current partition Ω consists of atomic elements $\phi_1, \ldots, \phi_K \in \{\phi_k\}_{k=1}^{\infty}$ and their complement $\phi_0 = \Omega / \{\phi_1, \ldots, \phi_K\}$. Under this partition, $q(z_{si} \in \phi_0)$ is positive given (11), and hence $\hat{z}_{si}$ can be sampled within ϕ_0, that is, $\hat{z}_{si}$ is a new sample, distinct from $\phi_1, \ldots, \phi_K$. If this happens, we draw a new ϕ_{K+1} and refine the partition as $\big(\phi_0, \phi_1, \ldots, \phi_K, \phi_{K+1}\big)$, where ϕ_0 is updated as $\Omega / \{\phi_1, \ldots, \phi_K, \phi_{K+1}\}$.

Remark: The partition refinement procedure reaches the limit inferior of empirical evidence lower bound as follows. Since there is no sampling within ϕ_0 after each update, to minimize the KL divergence, the posterior should be proportional to the prior on ϕ_0, $q\big(G_0(\phi_0)\big) \propto p\big(G_0(\phi_0)\big)$. Moreover, if we further partition ϕ_0 into $\phi_0^1 \cup \phi_0^2$, the KL divergence stays the same. Thus, $\mathrm{E}(\widehat{\mathrm{ELBO}}^{\,\Omega}) = \mathrm{ELBO}^{\Omega} = \mathrm{NPELBO}$. See Appendix C.5 for a justification.

We summarize the CATVI algorithm in Algorithm 1.

Algorithm 1: CATVI Algorithm

```
Initialize the partition Ω, the parameters for
q(G_0), q(β) and set up the step-size {ρ_τ}_{τ≥1}.
repeat
  Randomly select x_1, ..., x_S from the dataset.
  for s ∈ {1, ..., S} do
    repeat
      for i ∈ {1, ..., N_s} do
        Sample ẑ_si | q(G_0), q(β), ẑ_{s,i^-} .
        if Sampling a new ẑ_si then
          Refine the partition Ω.
        end if
      end for
    until convergence
    Resample ẑ_s = {ẑ_{s,t}}_{t=1}^{T_s}.
  end for
  Update parameters for q(G_0) and q(β) given
  samples {ẑ_s}_{s=1}^{S} using the step-size ρ_τ .
until convergence
```

5 Applications in the Topic Models

5.1 CATVI for the HDP Model

We apply the proposed CATVI method to the HDP model. Specifically, we factorize the variational distributions in the conditional setting and specify the variational family as follows. First, the variational distribution of G_s for each s is given by $q(G_s | G_0, \boldsymbol{z}_s) = \mathrm{DP}\big(\sum_{k=1}^{\infty} n_{sk}\delta_{\phi_k} + G_0\big)$, where $n_{sk} = \sum_{i=1}^{N_s} I(z_{si} = \phi_k)$ and $I(\cdot)$ is the indicator function. Second, $q(\beta_k)$ for each topic k is set as a W-dimensional Dirichlet distribution, $q(\beta_k) = \mathrm{Dirichlet}(\lambda_k)$, where $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{kW})^{\mathrm{T}}$ is the parameter of vocabulary distribution for topic k. The variational distribution for topics without any observation remains the same as the prior, hence $q(\beta_0) = \mathrm{Dirichlet}(\eta)$. Third, we specify the variational family for G_0 using spike and slab distributions (Andersen et al., 2017) as $q(G_0) = \sum_{k=1}^{K} m_k \delta_{\phi_k} + m_0 \mathrm{DP}(\alpha H)$, such that $\sum_{k=0}^{K} m_k = 1$. Finally, following (11) we use Monte Carlo sampling to obtain samples $\{\hat{\boldsymbol{z}}_s\}_{s=1}^{S}$, avoiding the need to parametrize their variational distributions.

As different samples in $\{\hat{\boldsymbol{z}}_s\}_{s=1}^{S}$ are used to represent different topic clusters in topic modelling, their exact values in the sample space do not contain any statistical information. We can then simply index the topics from 1 to K and denote the different clusters by distinct points $\phi_1, \ldots, \phi_K$ in Ω, and cluster 0 is the topic without any observation. Given samples $\{\hat{\boldsymbol{z}}_s\}_{s=1}^{S}$, we define the number of topics with observations by $K = \sum_{k=0}^{\infty} I(\sum_{s=1}^{S} \sum_{t=1}^{T_s} \hat{n}_{sk,t} > 0)$, where $\hat{n}_{sk,t} = \sum_{i=1}^{N_s} I(\hat{z}_{si,t} = \phi_k)$. In Appendix C.4, we rely on (10) to derive the empirical evidence lower bound with respect to $q(G_0)$,

$$\begin{aligned} &\alpha \log m_0 - \sum_{k=0}^{K} \log m_k \\ &+ \sum_{s=1}^{S} \sum_{k=1}^{K} \sum_{t=1}^{T_s} \frac{J}{S T_s} \log \frac{\Gamma(\gamma m_k + \hat{n}_{sk,t})}{\Gamma(\gamma m_k)} \end{aligned} \tag{12}$$

up to a constant. According to Algorithm 1, we repeatedly select documents of a batch size S, sample

$\{\hat{\boldsymbol{z}}_s\}_{s=1}^S$, and update parameters for G_0 and β iteratively until the empirical evidence lower bound converges to its maximum. During Gibbs sampling, once a document is sampled in cluster 0, we add a new cluster $K+1$, thus partitioning Ω to be $(K+1)$-dimensional, with K single points $\{\phi_k\}_{k=1}^K$ and one complement set $\phi_0 = \Omega/\{\phi_k\}_{k=1}^K$. During the training, this procedure is repeated until Ω is optimized. See Appendix B.2 for detailed steps on updating the variational parameters and refining the partition.

5.2 CATVI for Generic HBNP Models

The CATVI algorithm can also be applied to a general class of HBNP models, where the global prior G_0 is generated from a completely random measure. In these models, the concentration parameter for any G_j is not fixed, and G_0 is not restricted to be a probability measure. The corresponding inference algorithm is similar to that of the HDP model, but requires a new parameter μ to approximate $G_0(\Omega)$. We choose the variational family for the global prior G_0 as $q(G_0) = \mu\big(\sum_{k=1}^K m_k\delta_{\phi_k} + m_0\widetilde{N}(\alpha H)\big)$, where $\widetilde{N}$ is the normalization of the corresponding completely random measure and $\sum_{k=0}^K m_k = 1$. We provide the corresponding empirical evidence lower bounds and algorithms to infer more general HBNP models, including ΓDP model, in Appendix B.3.

6 Relationship to Relevant Work

In this section, we discuss several advantages of CATVI compared with traditional methods (Hoffman et al., 2013; Wang et al., 2011; Wang and Blei, 2012), although these are specific to the inference for the HDP model. First, CATVI replaces the unrealistic mean-field assumption with the conditional setting to capture the correlation structure among latent variables. Second, CATVI approximates the posterior group-wisely instead of updating the stick-breaking parameters sequentially, and hence avoids the gradient vanishing problem. By contrast, Hoffman et al. (2013) and Wang et al. (2011) perform inference separately over each atomic location and weight of G_0 using the stick-breaking representation $G_{0K} = g_{0K}\prod_{k=1}^{K-1}(1-g_{0k})$, where g_{0k}s are the representation parameters. However, this may cause the gradient vanishing problem of G_{0K} if k is large, because $\prod_{k=1}^{K-1}(1-g_{0k})$ is close to zero. Third, these traditional methods universally truncate the dimension of G_0 to a fixed level, contradicting the motivation and advantages of using HBNP models. Finally, CATVI is guaranteed to maximize the NPELBO. By comparison, Wang and Blei (2012) update parameters using the locally collapsed Gibbs sampling, but their work leads to an approximation that fails to maximize the ELBO, especially when the variance of distributions is large.

From a computational perspective, CATVI inherits the fast speed of stochastic variational inference, while other methods that truncate the dimension in a truly nonparametric way are very slow, such as the split-merge variational inference (Bryant and Sudderth, 2012) and the pure Gibbs sampling (Teh et al., 2006). To check the split-merge criterion, the split-merge variational inference requires calculating the likelihood before and after a split or merge, which is computationally infeasible in practice. Moreover, the pure Gibbs sampling is not scalable as well. As pure Gibbs sampling does not have batch selection, the Markov chains would converge very slowly when the sample size is large. As a result, these methods cannot be used to handle big data.

7 Experiments

7.1 Datasets and Architectures

We apply the CATVI algorithm to three large datasets, *arXiv*, *NYT* and *Wiki*, and compare the performance of CATVI with the online variational inference (OVI) (Wang et al., 2011), the memorized online variational inference (MOVI) (Hughes and Sudderth, 2013), and the split-merge variational inference (SMVI) (Bryant and Sudderth, 2012).

arXiv The corpus contains descriptive metadata of articles on *arXiv* up to September 1, 2019, resulting in 1.03M documents and 44M words from a vocabulary of 7,500 terms.

NYT The corpus contains all articles published by *New York Times* from January 1987 to June 2007 (Sandhaus, 2008), resulting in 1.56M documents and 176M words from a vocabulary of 7,600 terms.

Wiki The corpus contains entries from all English *Wikipedia* websites on January 1, 2019, resulting in 4.03M documents and 423M words from a vocabulary of 8,000 terms.

For the preprocessing, stemming and lemmatization are used to clean the raw text, and then words with too high or too low frequency, as well as common stop words, are filtered out.

To evaluate the performance of CATVI, we set aside a test set of 10,000 documents for each dataset and calculate the predictive perplexity as

$$\text{perplexity} = \exp\left\{-\frac{\sum_{j\in \mathrm{D}_{\text{test}}} \log p(\boldsymbol{x}_j^{\text{test}}|\boldsymbol{x}_j^{\text{train}}, \mathrm{D}_{\text{train}})}{\sum_{j\in \mathrm{D}_{\text{test}}} |\boldsymbol{x}_j^{\text{test}}|}\right\},$$

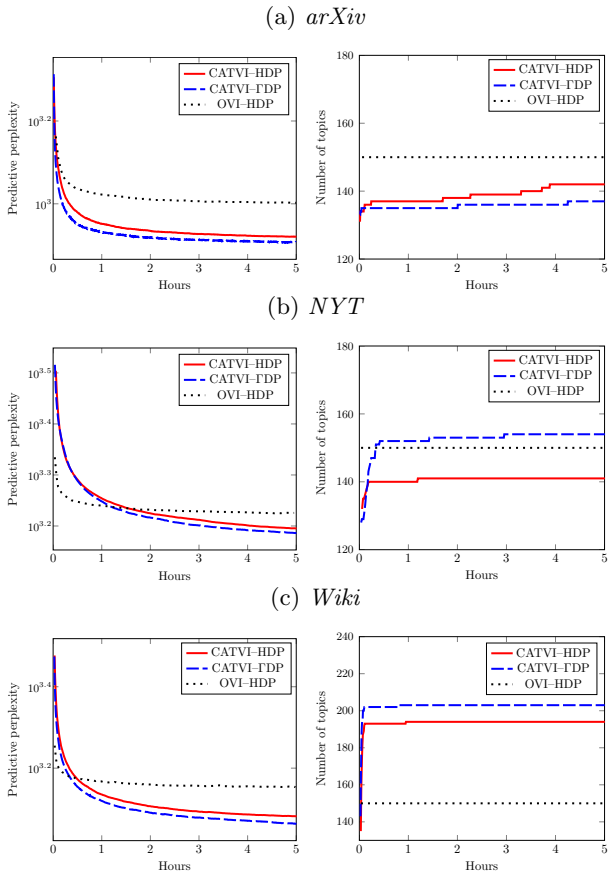

Figure 2: Left column: plots for the perplexity vs the running time up to 5 hours, Right column: plots for the number of topics vs the running time.

where $\mathrm{D}_{\text{train}}$ and D_{test} represent the training and test data, respectively, $\boldsymbol{x}_j^{\text{train}}$ and $\boldsymbol{x}_j^{\text{test}}$ are the training and test words in test document j, respectively, and $|\boldsymbol{x}_j^{\text{test}}|$ is the number of words in $\boldsymbol{x}_j^{\text{test}}$ (Ranganath and Blei, 2018). The perplexity measures the uncertainty of fitted models, where a lower perplexity will result in a better language model with higher predictive likelihood. Since the perplexity can not be computed exactly, the standard routine uses $\mathrm{D}_{\text{train}}$ to compute the variational distribution for β and G_0, then obtains the variational distribution for G_j based on G_0 and $\boldsymbol{x}_j^{\text{test}}$, and then approximates the likelihood by $p(\boldsymbol{x}_j^{\text{test}}|\boldsymbol{x}_j^{\text{train}}) = \prod_{w\in\boldsymbol{x}_j^{\text{test}}} \sum_{k=0}^{K} \overline{G}_{jk}\overline{\beta}_{kw}$, where $\overline{G}_{jk}$ and $\overline{\beta}_{kw}$ are the variational expectations of G_{jk} and β_{kw}, respectively (Blei et al., 2003). Experiments are run with the three datasets above using both the HDP and ΓDP models. For the HDP model, we set the hyperparameters as $\alpha = \gamma = \eta = 5$, where α and γ are the concentration parameters for G_0 and G_j respectively, and η is the hyperparameter for the prior on the distribution of words. The initial number of topics is set to be 100. The parameters are then optimized using stochastic gradient descent, with a batch size of 256 and a linear decaying learning rate adopted in Hoffman et al. (2010). For the ΓDP model, we use the same settings but discard γ. In the experiments, we remove clusters with fewer than 1 document during the training.

Table 1: A summary of predictive perplexity results.

Model	Method	*arXiv*	*NYT*	*Wiki*
HDP	MOVI	1901	2921	1876
HDP	SMVI	1917	2866	1877
HDP	OVI	1005	1681	1422
HDP	CATVI	832	1569	1207
ΓDP	CATVI	808	1536	1157

7.2 Empirical Results

Predictive perplexity The top row of Figure 2 plots the predictive perplexity as a function of running time for the three comparison methods using the three datasets. As MOVI and SMVI provide significantly higher perplexities, we do not plot their results in Figure 2. Table 1 reports numerical summaries for all comparison methods. Several conclusions can be drawn here. First, on all three datasets, CATVI uniformly outperforms competing methods. The improvement is highly significant especially for *arXiv* and *Wiki*. For *NYT*, there is moderate improvement, likely due to the long length of documents in this corpus. Second, for each dataset, the ΓDP model attains a lower perplexity than the HDP model, consistent with the fact that the ΓDP model removes a restriction of the HDP model and hence is more flexible. Third, CATVI is empirically shown to be computationally efficient, reaching the lowest perplexity within the same training time. Although it involves Monte Carlo sampling, the perplexity converges fast. This is because the convergence of local Markov chains to assign words to topics is accelerated by a clear topic-words clustering as the global variational distributions approach to the optimal.

Number of topics The bottom row of Figure 2 plots the number of topics during the training process. For OVI, the number of topics remains constant at the prespecified value, while for CATVI, this value first increases steeply and then converges to a stable level. For example, the number of topics in *Wiki* sharply increase from 100 to around 190 for the HDP model and around 200 for ΓDP model. The sharp increase is driven by the data complexity, while the stable level is achieved due to the dimension penalty effect from the priors in HBNP models. Although the estimation of the number of topics is not consistent, CATVI can

provide some useful information about topics in data. For instance, the data from the *arXiv* corpus in these experiments are limited to abstracts of scientific articles, and thus it has the smallest number of topics. By contrast, *NYT* is a compilation of all new articles covering a wider range of areas, and hence consists of more topics. Similarly, *Wiki* has the largest number of topics as it contains almost every aspect of an encyclopedia. It is important to note that we do not need to set a fixed number of topics before the inference. Instead, CATVI starts from an initial value, for example 100 in our experiments, then automatically converges to a stable optimal number of topics.

Topic-words clustering CATVI is shown to reveal much better linguistic results. To compare CATVI with OVI for the HDP model, we report the top 12 words in the top 10 topics with biggest weights for both methods on *arXiv* and *Wiki* in Tables 2a and 2b, respectively. We observe a few apparent patterns. First, the topic-word clusters from CATVI hardly contains replicated topics, whereas those from OVI results have similar word components, such as those shown in blue in columns 1-6 in the bottom part of Table 2a. An ideal topic-word clustering should allocate these words into just one topic. However, the prespecified number of topics is fixed at 150 in OVI, which is larger than the ground truth, resulting in generating replicated topics. By contrast, the topic-word clustering by CATVI does not have such redundancy. It is apparent that our top 10 topics are mostly distinct. Second, CATVI leads to much clearer topic-word clustering. For both datasets, our results indicate that all of our detected words within any column are highly relevant and should intuitively be grouped into one cluster with clear linguistic meaning. For example, column 7 of Table 2b for CATVI presents several words all related to military, but words in the same column for OVI seem to be a mixture of several loosely connected topics including 'human, character, reveal, episode, comic, voice', 'human, earth' and 'human, kill, attack, fight, battle, doctor'. This mixture of topics makes the topic-word clustering in this column ambiguous. Furthermore, CATVI identifies a topic about popular English given names in column 5 of Table 2b. Although these given names are not shown in a single document, CATVI can successfully discover that they belong to one topic, while OVI fails. This is because CATVI does not force the topics to merge together if the prespecified number of topics is not large enough, thus reducing the noise in the clusters.

We also perform sensitivity analysis of CATVI using *arXiv* under the HDP model as an example. The left and right panels of Figure 3 in Appendix E respectively plot the results as the batch size varies from 128 to 1024 and the initial number of topics varies from 60 to 140. We observe that the performance is not sensitive to the change of these hyperparameters. Moreover, the best results are obtained for the case with a smaller batch size and a larger initial number of topics.

Table 2: Top 12 words in top 10 topics.

(a) *arXiv*

	1	2	3	4	5	6	7	8	9	10
CATVI	galaxy	group	network	neutrino	star	gaug	prove	algorithm	collision	test
	cluster	algebra	learn	higg	dwarf	string	bound	optim	product	error
	redshift	construct	train	matter	survey	brane	theorem	converge	decay	samples
	samples	finit	neural	dark	object	symmetry	finit	solve	hadron	statist
	luminos	lie	image	decay	binari	dimension	class	linear	jet	uncertainties
	formation	categories	deep	standard	variable	couple	position	approximate	transverse	fit
	survey	prove	dataset	couple	cluster	action	inequalities	gradient	gev	systematic
	agnes	class	feature	mix	stellar	conform	dimension	minim	lhc	accuracy
	lar	complex	task	boson	period	construct	converge	matrix	cross	correct
	star	map	object	lepton	photometr	correspond	continual	iter	section	procedure
	populated	invariable	convolut	violate	distanc	dual	regular	spars	quark	improve
	host	manifold	detect	symmetry	samples	background	compact	constraint	collid	bias
OVI	galaxy	galaxy	star	xray	emiss	galaxy	higg	star	emiss	radio
	cluster	redshift	cluster	emiss	star	line	neutrino	planet	gammaray	emiss
	halo	source	abundance	source	region	emiss	decay	period	source	galaxy
	star	survey	galaxy	accret	line	gas	dark	orbit	grb	source
	stellar	samples	stellar	kev	gas	star	boson	dwarf	xray	xray
	formation	cluster	metal	line	dust	redshift	matter	binari	ray	jet
	velocity	luminos	age	variable	disk	absorption	standard	detect	detect	line
	dark	agnes	populated	spectral	molecular	samples	couple	stellar	burst	region
	gas	xray	ngc	star	cloud	quasar	gev	transit	flux	cluster
	matter	radio	dwarf	flux	detect	luminos	particle	variable	radio	detect
	profile	star	samples	spectrum	formation	region	mix	light	spectrum	gas
	disk	optic	giant	detect	galaxy	detect	symmetry	companion	jet	star

(b) *Wiki*

	1	2	3	4	5	6	7	8	9	10
CATVI	tell	increase	band	human	james	claim	armies	process	polit	album
	tried	effect	album	natur	robert	issue	battle	model	parti	chart
	want	case	guitar	tradition	charles	announce	attack	inform	union	song
	friend	process	vocal	term	david	critic	troop	effect	communist	track
	leave	measure	track	idea	thomas	controversi	command	problem	movement	video
	ask	caus	rock	view	richard	proposal	soldier	experience	independence	billboard
	feel	require	drum	word	michael	agreement	military	test	social	label
	decide	rate	bass	theorie	frank	polit	fight	example	republic	peak
	turn	example	song	philosophies	peter	allegation	tanks	research	leader	week
	good	reduce	tour	culture	andrew	statement	brigade	specific	worker	digitated
	away	possibilities	studio	believe	brown	agree	german	individual	socialist	hot
	believe	occur	label	conception	henry	minister	capture	object	liberal	remix
OVI	album	episode	actor	album	album	episode	character	novel	animal	ship
	band	tell	movi	song	song	televis	kill	character	episode	navies
	song	character	character	band	chart	drama	human	love	character	class
	track	kill	critic	tour	video	actor	earth	poem	voice	boat
	guitar	friend	cast	love	track	comedies	attack	london	movi	naval
	vocal	leave	review	blue	billboard	actress	reveal	king	video	command
	rock	tried	televis	artist	love	movi	episode	tell	air	vessel
	tour	relationship	episode	rock	label	theatre	comic	fiction	dvd	submarine
	chart	need	scene	track	version	voice	fight	narrated	televis	gun
	studio	love	theatre	label	week	uncredit	doctor	friend	ray	fleet
	bass	reveal	picture	chart	peak	nominal	battle	mother	blu	sail
	drum	mother	love	singer	remix	cast	voice	critic	song	destroy

8 Discussion

CATVI can also be applied to other HBNP models including, for example, hierarchical Pitman–Yor process model (Teh and Jordan, 2010) and hierarchical beta process model (Thibaux and Jordan, 2007). CATVI will provide more advantages in these applications, because the hierarchical Pitman–Yor process, with heavy tail behavior, and the hierarchical beta process, with sparse structure, may suffer more from the universal truncation.

References

Andersen, M. R., Vehtari, A., Winther, O., and Hansen, L. K. (2017). Bayesian inference for spatio-temporal spike-and-slab priors. *Journal of Machine Learning Research*, 18(1):5076–5133.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bryant, M. and Sudderth, E. B. (2012). Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 25*, pages 2699–2707.

Caron, F. and Fox, E. B. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B*, 79(5):1295–1366.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge.

Hoffman, M., Bach, F. R., and Blei, D. M. (2010). Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, pages 856–864.

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

Hughes, M. C. and Sudderth, E. (2013). Memoized online variational inference for Dirichlet process mixture models. In *Advances in Neural Information Processing Systems 26*, pages 1133–1141.

Kingman, J. F. C. (1993). *Poisson Processes*. Clarendon Press, Oxford.

Matthews, A. G. d. G., Hensman, J., Turner, R., and Ghahramani, Z. (2016). On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*.

Ranganath, R. and Blei, D. M. (2018). Correlated random measures. *Journal of the American statistical Association*, 113(521):417–430.

Roychowdhury, A. and Kulis, B. (2015). Gamma processes, stick-breaking, and variational inference. In *Proceedings of the 8th International Conference on Artificial Intelligence and Statistics*.

Sandhaus, E. (2008). *The New York Times annotated corpus*. Linguistic Data Consortium, Philadelphia.

Sudderth, E. B. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman–Yor processes. In *Advances in Neural Information Processing Systems 21*, pages 1585–1592.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, pages 158–207. Cambridge University Press.

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

Teh, Y. W., Kurihara, K., and Welling, M. (2008). Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems 20*.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, volume 2, pages 564–571.

Wang, C. and Blei, D. M. (2012). Truncation-free online variational inference for Bayesian nonparametric models. In *Advances in Neural Information Processing Systems 25*.

Wang, C., Paisley, J., and Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*.

Williamson, S. A. (2016). Nonparametric network models for link prediction. *Journal of Machine Learning Research*, 17(202):1–21.

Xu, K., Srivastava, A., and Sutton, C. (2019). Variational Russian Roulette for deep Bayesian nonparametrics. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6963–6972.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7252–7261.

Acknowledgments

Opinions expressed in this paper are those of the authors, and do not necessarily reflect the view of JP Morgan.

Supplement to "CATVI: Conditional and Adaptively Truncated Variational Inference for Hierarchical Bayesian Nonparametric Models"

Yirui Liu, Xinghao Qiao and Jessica Lam

This supplementary material contains a short review of completely random measures in Appendix A, CATVI algorithm and its applications to the HDP model and the ΓDP model in Appendix B, technical proofs and derivations in Appendix C, computational complexity analysis and code in Appendix D and sensitivity analysis results in Appendix E.

A A Short Review of Completely Random Measures

Suppose that $(\Omega, \mathcal{F})$ is a Polish sample space, Θ is the set of all bounded measures on $(\Omega, \mathcal{F})$ and $\mathcal{M}$ is a σ-algebra on Θ. A random measure G on $(\Omega, \mathcal{F})$ is a transition kernel from $(\Theta, \mathcal{M})$ into $(\Omega, \mathcal{F})$ such that (i) $G \mapsto G(A)$ is $\mathcal{M}$-measurable for any $A \in \mathcal{F}$ and (ii) $A \mapsto G(A)$ is a measure for any realization of G (Ghosal and Van der Vaart, 2017). For example, a Dirichlet process P with base measure P_0 satisfies

$$\big(P(A_1), \ldots, P(A_n)\big) \sim \text{Dirichlet}\big(P_0(A_1), \ldots, P_0(A_n)\big)$$

for any partition $\Omega = (A_1, \ldots, A_n)$ of Ω, that is, a finite number of measurable, nonempty and disjoint sets such that $\bigcup_{i=1}^n A_i = \Omega$. The Dirichlet process is denoted by $P \sim \text{DP}(P_0)$ or $P \sim \text{DP}(\alpha H)$ with concentration parameter $\alpha = P_0(\Omega)$ and center measure $H = \alpha^{-1} P_0$. Moreover, a random measure is called a completely random measure (Kingman, 1993) if it also satisfies the condition that (iii) $P(A_i)$ is independent of $P(A_j)$ for any disjoint subsets A_i and A_j in Ω. Completely random measures and their normalizations (Ghosal and Van der Vaart, 2017), for example, the Gamma process and Dirichlet process, respectively, are commonly used as priors for infinite-dimensional latent variables in HBNP models, because their realizations are atomic measures with countable-dimensional supports.

A completely random measure (Kingman, 1993) is characterized by its Laplace transform,

$$\text{E}\big[e^{-tP(A)}\big] = \exp\Big\{ -\int_A \int_{(0,\infty]} (1 - e^{-t\pi}) v^c(dx, ds) \Big\},$$

where A is any measurable subset of Ω and $v^c(dx, ds)$ is called the Lévy measure. If $v^c(dx, ds) = \kappa(dx)v(ds)$, where $\kappa(\cdot)$ and $v(\cdot)$ are measures on Ω and $(0, \infty]$, respectively, the completely random measure is homogeneous (Ghosal and Van der Vaart, 2017). In such a case, we call $v(\cdot)$ the weight intensity measure. We can view completely random measure as a Poisson process on the product space $\Omega \times (0, \infty]$ using its Lévy measure as the mean measure.

B CATVI Algorithm

B.1 Empirical ELBO for $q(\beta)$

To maximize the NPELBO, we iterate the following three steps: (i) randomly select a small batch from the entire data, (ii) sample $\{\tilde{\boldsymbol{z}}_s\}_{s=1}^S$ by Monte Carlo method, and (iii) update $q(G_0^a)$ and $q(\beta)$ in the stochastic variational inference framework.

In an analogy to (9), the NPELBO with respect to $q(\beta)$ is

$$\liminf_{\Omega} \Big\{ \sum_{j=1}^{J} \mathrm{E}_{q(\beta)} \mathrm{E}_{q(\boldsymbol{z}_j)} \log p(\boldsymbol{x}_j | \boldsymbol{z}_j, \beta) - \mathcal{KL}\big(q(\beta) \| p(\beta)\big) \Big\}, \tag{A.1}$$

and the empirical evidence lower bound with respect to $q(\beta)$, $\widehat{\mathrm{ELBO}}^{\Omega}$ is,

$$\sum_{s=1}^{S} \sum_{t=1}^{T_s} \frac{J}{ST_s} \mathrm{E}_{q(\beta)} \log p(\boldsymbol{x}_s | \hat{\boldsymbol{z}}_{s,t}, \beta) - \mathcal{KL}\big(q(\beta) | p(\beta)\big) \tag{A.2}$$

up to a constant and then we can update its parameter with the corresponding random gradient in a similar way. We summarize the details of CATVI algorithm in Algorithm 1.

B.2 CATVI for the HDP Model

We repeatedly select documents of a batch size and update parameters iteratively according to the following three steps, until the NPELBO attains its maximum.

Inference for G_0. There is no closed-form expression for the parameters $\{m_k\}_{k=0}^K$ to attain the maximum in (12). Moreover, the standard gradient descent algorithm fails in this case, because $\{m_k\}_{k=0}^K$ may easily exceed the simplex during the updating procedure. Instead, given the parameters $\{m_k^{(\tau)}\}_{k=0}^K$ in the τ-th iteration, we first define

$$m_k^* \propto \begin{cases} JS^{-1}\gamma \sum_{s=1}^{S} \big\{ T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m_k^{(\tau)} + \hat{n}_{sk,t}) - \Phi(\gamma m_k^{(\tau)}) \big\} m_k^{(\tau)} - 1 & k = 1, \ldots, K, \\ \alpha - 1 & k = 0, \end{cases} \tag{B.1}$$

where $\Phi(\cdot)$ denotes the log-gamma function, such that $\sum_{k=0}^K m_k^* = 1$, and then we update the parameters by $m^{(\tau+1)} = (1-\rho_t) m^{(\tau)} + \rho_\tau m_k^*$, where ρ_t is the step size defined in Algorithm 1. This updating algorithm is consistent to the gradient descent after the inverse logit transformation. See Appendix C.6 for a justification. In the process of updating, the condition $\sum_{k=0}^K m_k^* = 1$ always holds, and hence we eliminate the risk of exceeding the simplex.

Inference for β. By (A.2), we update the parameters for $q(\beta)$ using samples $\{\hat{\boldsymbol{z}}_s\}_{s=1}^S$. We define λ_{kw}^* for topic k and word w as,

$$\lambda_{kw}^* = \eta + \sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{i=1}^{N_s} \frac{J}{ST_S} I(\hat{z}_{si,t} = \phi_k, x_{si} = w), \tag{B.2}$$

and update the parameter λ_k by $\lambda_k^{(\tau+1)} = (1-\rho_t)\lambda_k^{(\tau)} + \rho_\tau \lambda_k^*$ for each k, where $\lambda_k^* = (\lambda_{k1}^*, \ldots, \lambda_{kW}^*)^{\mathrm{T}}$.

Sampling for $\boldsymbol{z}$. According to (11) we sample $\hat{z}_{si}$ conditional on $q(G_0)$ and $\hat{\boldsymbol{z}}_{si^-}$ by

$$q(z_{si} = \phi_k) \ \propto \ \begin{cases} (\gamma m_k + \hat{n}_{s,i^-}^k) \exp\big(\Phi(\lambda_{kx_{si}}) - \Phi(\sum_{w=1}^W \lambda_{kw})\big) & k = 1, \ldots, K, \\ \gamma m_0 \exp\big(\Phi(\eta) - \Phi(W\eta)\big) & k = 0, \end{cases} \tag{B.3}$$

to construct the Markov chain, where $\hat{n}_{s,i^-}^k = \sum_{1 \le l \le N_s, l \ne i} I(\hat{z}_{sl} = \phi_k)$. Whenever the sampled $\hat{z}_{si}$ is in ϕ_0, meaning $\hat{z}_{si}$ forms a new point not belonging to $\{\phi_1, \ldots, \phi_K\}$, we need to update the partition and add a new topic indicated by ϕ_{K+1}. Otherwise the partition dimension remains the same. Iterating the sampling scheme till convergence, we obtain the samples $\{\hat{z}_{si,t}\}_{1 \le s \le S, 1 \le i \le N_s, 1 \le t \le T_s}$ and corresponding $\{\hat{n}_{sk,t}\}_{1 \le s \le S, 1 \le k \le K, 1 \le t \le T_s}$ for the selected chunk.

B.3 CATVI for the ΓDP Model

ΓDP releases the constraint of fixed concentration parameter γ in HDP. Therefore, the CATVI algorithm for ΓDP inherits the steps in (B.1) and (B.3), except that a parameter μ replaces the concentration parameter γ in both formulas.

We derive the empirical evidence lower bound in Appendix C.7 with respect to $q(G_0)$ as,

$$\sum_{k=1}^{K}\log v(\mu m_k)+\log u(\mu m_0)+\sum_{s=1}^{S}\frac{J}{S}\log\frac{\Gamma(\mu)}{\Gamma(\mu+N_s)}+\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{t=1}^{T_s}\frac{J}{ST_s}\log\frac{\Gamma(\mu m_k+\hat{n}_{sk,t})}{\Gamma(\mu m_k)}+K\log\mu \tag{B.4}$$

up to a constant, where $v(\cdot)$ is the weight intensity measure (see Appendix A) for the completely random measure, and $u(\cdot)$ is the density function for $G_0(\Omega)$ that can be derived using its Laplace transform. Therefore, we can update $\{m_k\}_{k=0}^K$ in the same way as the HDP model.

Similar to (B.1), we update $\{m_k\}_{k=0}^K$ according to

$$m_k^*\propto\begin{cases}JS^{-1}\mu\sum_{s=1}^{S}\left\{T_s^{-1}\sum_{t=1}^{T_s}\Phi(\mu m_k^{(\tau)}+\hat{n}_{sk,t})-\Phi(\mu m_k^{(\tau)})\right\}m_k^{(\tau)}-1 & k=1,\ldots,K,\\ \alpha-1 & k=0,\end{cases} \tag{B.5}$$

and $m^{(\tau+1)}=(1-\rho_t)m^{(\tau)}+\rho_\tau m_k^*$. Moreover, in an analogy to (B.3), the probability to sample $\hat{z}_{si}$ is defined as

$$q(z_{si}=\phi_k)\ \propto\ \begin{cases}(\mu m_k+\hat{n}_{s,i^-}^k)\exp\left(\Phi(\lambda_{kx_{si}})-\Phi(\sum_{w=1}^{W}\lambda_{kw})\right) & k=1,\ldots,K,\\ \mu m_0\exp\left(\Phi(\eta)-\Phi(W\eta)\right) & k=0.\end{cases} \tag{B.6}$$

Finally, we apply the gradient ascent to update μ. In Appendix C.7, we derive the gradient of empirical evidence lower bound with respect to μ as

$$g'(\mu)=\frac{\alpha-1}{\mu}-1+\frac{J}{S}\sum_{s=1}^{S}\Big\{\Phi(\mu)-\Phi(\mu+N_s)+\sum_{k=1}^{K}\frac{1}{T_s}\sum_{t=1}^{T_s}m_k\big(\Phi(\mu m_k+\hat{n}_{sk,t})-\Phi(\mu m_k)\big)\Big\}, \tag{B.7}$$

and then update μ by $\mu^{(\tau+1)}=\mu^{(\tau)}+\rho_\tau g'(\mu^{(\tau)})$.

C Technical Proofs and Derivations

C.1 Proof for (4)

By definition of induced measure, $q^\varnothing(d\Theta)=Q(d\Theta)$ for any $\mathcal{M}$-measurable $d\Theta$, we have

$$\int_\Theta\log\frac{dq^\varnothing}{dp^\varnothing}dq^\varnothing=\int_\Theta\log\frac{dq^\varnothing}{dp^\varnothing}dQ.$$

It follows from $\limsup_\varnothing dq^\varnothing/dp^\varnothing=dQ/dP$ and the monotone convergence theorem that

$$\limsup_\varnothing\int_\Theta\log\frac{dq^\varnothing}{dp^\varnothing}dQ=\int_\Theta\log\frac{dQ}{dP}dQ.$$

Combining the above equations yields (4). Furthermore, suppose there exists a sequence of partition $\{\Omega_i\}_{i\geq1}$ such that $\limsup\Omega_i=\Omega$, we have

$$\limsup_{\Omega_i}\int_\Theta\log\frac{dq^{\varnothing_i}}{dp^{\varnothing_i}}dq^{\varnothing_i}=\limsup_{\Omega_i}\int_\Theta\log\frac{dq^{\varnothing_i}}{dp^{\varnothing_i}}dQ=\int_\Theta\log\frac{dq^\varnothing}{dp^\varnothing}dQ=\int_\Theta\log\frac{dq^\varnothing}{dp^\varnothing}dq^\varnothing.$$

Hence $\limsup_{\Omega_i}\mathrm{KL}(q^{\varnothing_i}\|p^{\varnothing_i})=\mathrm{KL}(q^\varnothing\parallel p^\varnothing)$, which will be used in Appendix C.4.

C.2 Proof for (6)

By $p(X,Z)=p(Z|X)p(X)$, we have

$$\int\log\frac{p(X,Z^\varnothing)}{q(Z^\varnothing)}q(dZ^\varnothing)=\log p(X)+\int\log\frac{p(Z^\varnothing|X)}{q(Z^\varnothing)}q(dZ^\varnothing).$$

Taking the limit inferior on both sides, we have

$$\liminf_\varnothing\int\log\frac{p(X,Z^\varnothing)}{q(Z^\varnothing)}q(dZ^\varnothing)=\log p(X)-\limsup_\varnothing\left\{-\int\log\frac{p(Z^\varnothing|X)}{q(Z^\varnothing)}q(dZ^\varnothing)\right\}.$$

Combing the above equation with the definition of NPELBO in (5) and the KL divergence in (4) yields (6).

C.3 Derivation for (9)

By $p(G_0^\varrho, \{\boldsymbol{z}_j\}_{j=1}^J) = \int\cdots\int p\big(G_0^\varrho, \{G_j\}_{j=1}^J, \{\boldsymbol{z}_j\}_{j=1}^J\big) dG_1 dG_2 \cdots dG_J$ and the hierarchical generative structure, the evidence lower bound under partition Ω with respect to $q(G_0^\varrho)$ equals,

$$\begin{aligned}
&\mathrm{ELBO}^\varrho\\
=\;& \mathrm{E}_{q(G_0^\varrho)}\mathrm{E}_{q(\{\boldsymbol{z}_j\}_{j=1}^J)} \log p(G_0^\varrho, \{\boldsymbol{z}_j\}_{j=1}^J) - \mathrm{E}_{q(G_0^\varrho)} \log q(G_0^\varrho) + \text{constant}\\
=\;& \mathrm{E}_{q(G_0^\varrho)}\mathrm{E}_{q(\{\boldsymbol{z}_j\}_{j=1}^J)} \log q(G_0^\varrho) \prod_{j=1}^J \int p(G_j^\varrho|G_0^\varrho) p(\boldsymbol{z}_j|G_j^\varrho) dG_j - E_{q(G_0^\varrho)} \log q(G_0^\varrho) + \text{constant}\\
=\;& \sum_{j=1}^J \mathrm{E}_{q(G_0^\varrho)}\mathrm{E}_{q(\boldsymbol{z}_j)} \log \mathrm{E}_{p(G_j^\varrho|G_0^\varrho)} p(\boldsymbol{z}_j|G_j^\varrho) + \mathrm{E}_{q(G_0^\varrho)} \log p(G_0^\varrho) - \mathrm{E}_{q(G_0^\varrho)} \log q(G_0^\varrho) + \text{constant}.
\end{aligned}$$

Furthermore, based on the equation above, (8) can be expressed as $\mathrm{NPELBO} = \liminf_\varrho \mathrm{ELBO}^\varrho$.

C.4 Derivation for (12)

By the formula of moments for Dirichlet-distributed random variables, we obtain

$$\mathrm{E}_{p(G_s^\varrho|G_0^\varrho)} p(\hat{\boldsymbol{z}}_{s,t}|G_s^\varrho) = \frac{\Gamma(\gamma)}{\Gamma(\gamma+N_s)} \prod_{k=1}^K \frac{\Gamma(\gamma G_{0k} + \hat{n}_{sk,t})}{\Gamma(\gamma G_{0k})}.$$

Based on the points $\{\phi_k\}_{k=1}^K$ defined in Section 5.1, we propose a sequence of partition $\{\Omega_c : \Omega_c = \bigcup_{k=0}^K \Omega_{ck}\}_{c\geq 1}$ to approach Ω, where $\Omega_{ck} = (\phi_k - c^{-1}, \phi_k + c^{-1}]$ for $k = 1, \ldots, K$ and Ω_{c0} is the corresponding complement. Under Ω_c, $q(G_0^{\varrho_c}) = \mathrm{d}_{K+1}\big(m_0^{-1}(G_0^{\varrho_c} - M^{\varrho_c})\big)$ and $p(G_0^{\varrho_c}) = \mathrm{d}_{K+1}(G_0^{\varrho_c})$, where $\mathrm{d}_{K+1}(\cdot)$ denotes the density function for $(K+1)$-dimensional Dirichlet distribution, $M = \sum_{k=1}^K m_k \delta_{\phi_k}$ and M^{ϱ_c} is the corresponding induced random variable. By (10), the empirical evidence lower bound under Ω_c is

$$\begin{aligned}
&E_{q(G_0^{\varrho_c})}\Big\{ \sum_{k=1}^K (\alpha H_k^{\varrho_c} - 1) \log \frac{m_0 G_{0k}}{(G_{0k} - m_k)} + (\alpha H_0^{\varrho_c} - 1) \log m_0\\
&+ \sum_{s=1}^S \sum_{k=1}^K \sum_{t=1}^{T_s} \frac{J}{ST_s} \log \frac{\Gamma(\gamma G_{0k} + \hat{n}_{sk,t})}{\Gamma(\gamma G_{0k})} \Big\} + \text{constant},
\end{aligned}$$

where $H_k^{\varrho_c} = H(\Omega_{ck})$. Since $(G_{0k} - m_k)/m_0 \sim \mathrm{Beta}(H_k^{\varrho_c})$ under $q(G_0^{\varrho_c})$, the term $E_{q(G_0^{\varrho_c})}(\alpha H_k^{\varrho_c} - 1) \log m_0 (G_{0k} - m_k)^{-1}$ is constant with respect to parameters $\{m_k\}_{k=0}^K$. Taking lim sup on both sides of the above equation with $\limsup_{\varrho_c} E_{q(G_0^{\varrho_c})}(\log G_{0k}) = \log m_k$, $\limsup_{\varrho_c} H_k^{\varrho_c} = 0$ for $k > 0$ and $\limsup_{\varrho_c} H_0^{\varrho_c} = 1$, we obtain equation (12).

C.5 Justification for Section 4.3

In this section, we show that the empirical evidence lower bound achieves the limit inferior in NPELBO. With the partition $\Omega = (\phi_0, \phi_1, \ldots, \phi_K)$ defined in Section 4.3, there is no sampling within ϕ_0, and hence we have

$$q\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big) \propto p\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big) p\big(\boldsymbol{x} \mid G_0(\phi_1), \cdots, G_0(\phi_K)\big).$$

As the likelihood part $p\big(\boldsymbol{x}|G_0(\phi_1), \cdots, G_0(\phi_K)\big)$ does not contain $G(\phi_0)$, by integrating both sides with respect to $G_0(\phi_1), \cdots, G_0(\phi_K)$, we can get $q\big(G_0(\phi_0)\big) \propto p\big(G_0(\phi_0)\big)$. Moreover, the KL divergence between the variational distribution and true posterior is

$$\begin{aligned}
&\mathcal{KL}\big(q(G_0^\varrho) \parallel p(G_0^\varrho \mid \boldsymbol{x})\big)\\
=\;& \int \log \frac{q\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big)}{p\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big) p\big(\boldsymbol{x} \mid G_0(\phi_1), \cdots, G_0(\phi_K)\big)} dq\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big)\\
=\;& -\log \mathcal{N},
\end{aligned}$$

because

$$q\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big) = p\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big) p\big(\boldsymbol{x} \mid G_0(\phi_1), \cdots, G_0(\phi_K)\big)/N,$$

where N is the normalization constant,

$$\begin{aligned}
\mathcal{N} &= \int \cdots \int p\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big) p\big(\boldsymbol{x} \mid G_0(\phi_1), \cdots, G_0(\phi_K)\big) dG_0(\phi_1) dG_0(\phi_1) \cdots dG_0(\phi_K) \\
&= \int p\big(\boldsymbol{x} \mid G_0(\phi_1), \cdots, G_0(\phi_K)\big) dp\big(G_0(\phi_0), G_0(\phi_1), \cdots, G_0(\phi_K)\big) \\
&= \int p\big(\boldsymbol{x} \mid G_0(\phi_1), \cdots, G_0(\phi_K)\big) dp\big(G_0(\phi_1), \cdots, G_0(\phi_K)\big).
\end{aligned}$$

It is obvious that $\mathcal{N}$ is independent of $p\big(G_0(\phi_0)\big)$. Therefore, if we partition ϕ_0 into $\phi_0^1 \cup \phi_0^2$, the normalization constant $\mathcal{N}$ will not change, that is, the KL divergence under Ω and $\Omega' = (\phi_0^1, \phi_0^2, \phi_1, \ldots, \phi_K)$ are the same. Consequently, the partition Ω enables the limit superior of KL divergence to be reached. By (6), the limit inferior of NPELBO is also attained.

C.6 Derivation for (B.1)

Consider the Lagrange multiplier of constrained optimization,

$$L' = -\sum_{k=1}^{K} \log m_k + (\alpha - 1) \log m_0 + \sum_{s=1}^{S} \sum_{k=1}^{K} \frac{J}{S T_s} \sum_{t=1}^{T_s} \log \frac{\Gamma(\gamma m_k + \hat{n}_{sk,t})}{\Gamma(\gamma m_k)} - \lambda\Big(\sum_{k=0}^{K} m_k - 1\Big),$$

its first order conditions satisfy,

$$\begin{cases} JS^{-1}\gamma \sum_{s=1}^{S} \big\{T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m_k + \hat{n}_{sk,t}) - \Phi(\gamma m_k)\big\} m_k - 1 = m_k \lambda, & k = 1, \ldots, K, \\ \alpha - 1 = m_0 \lambda, & k = 0. \end{cases}$$

Dividing λ on both sides of the above equations, the definition of $\{m_k^*\}_{k=0}^K$ in (B.1) follows.

We next show that this updating is consistent with the gradient descent after the inverse logit transformation, that is, transforming $\{m_k\}_{k=0}^K$ by $m_k = e^{\theta_k} / \sum_{l=0}^K e^{\theta_l}$ to remove the constraint of $\sum_{k=0}^K m_k = 1$. By $\partial m_k / \partial \theta_k = m_k - m_k^2$, $\partial m_l / \partial \theta_k = -m_k m_l$ for $l \neq k$, and the chain rule, we have

$$\frac{\partial L}{\partial \theta_k} = \begin{cases} JS^{-1}\gamma \sum_{s=1}^{S} \big\{T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m_k + \hat{n}_{sk,t}) - \Phi(\gamma m_k)\big\} m_k - 1 - \Lambda m_k & k = 1, \ldots, K, \\ \alpha - 1 - \Lambda m_k & k = 0, \end{cases}$$

where L denotes $\widehat{\mathrm{ELBO}}^{\Omega}$ in (12) and

$$\Lambda = \alpha - 1 + \sum_{k=1}^{K} \Big[JS^{-1}\gamma \sum_{s=1}^{S} \big\{T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m_k + \hat{n}_{sk,t}) - \Phi(\gamma m_k)\big\} m_k - 1 \Big].$$

As $\partial L / \partial \theta_k = \Lambda(m_k^* - m_k)$, $(m_k^* - m_k)$ represents the gradient with respect to θ_k after the inverse logit transformation.

C.7 Derivation for (B.4)

For the HBNP model, we use an unnormalized random measure as the prior of G_0. Given moments for Dirichlet-distributed random variables, we obtain

$$\log \mathrm{E}_{p(G_s^{\Omega} | G_0^{\Omega})} p(\hat{\boldsymbol{z}}_{s,t} | G_s^{\Omega}) = \log \frac{\Gamma(\sum_{k=0}^K G_{0k})}{\Gamma(\sum_{k=0}^K G_{0k} + N_s)} \prod_{k=1}^{K} \frac{\Gamma(G_{0k} + \hat{n}_{sk,t})}{\Gamma(G_{0k})},$$

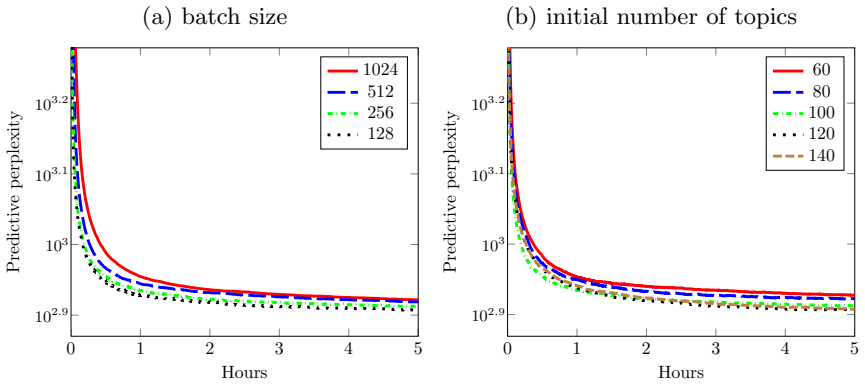

Figure 3: Plots for the perplexity vs running time for different batch sizes and initial numbers of topics.

In analogy to Appendix C.4, the empirical evidence lower bound under Ω_c is

$$K\log\mu + E_{q(G_0^{\Omega_c})}\Big\{\sum_{k=1}^{K}\log p(G_{0k}^{\Omega_c}) + \log p(G_{00}^{\Omega_c}) + \frac{J}{S}\sum_{s=1}^{S}\sum_{k=1}^{K}T_s^{-1}\sum_{t=1}^{T_s}\log\frac{\Gamma(\gamma G_{0k}^{\Omega_c}+\hat{n}_{sk,t})}{\Gamma(\gamma G_{0k}^{\Omega_c})}\Big\},$$

up to a constant, where $K\log\mu$ comes from the Jacob matrix from $G_0, G_1, \ldots, G_K$ to $\mu, m_1, \ldots, m_K$. As the partition converges to single points and the corresponding complement, $\limsup_{\Omega_c} p(G_{0k}^{\Omega_c}) = v(G_{0k}^{\Omega_c})$ and $\limsup_{\Omega_c} p(G_{00}^{\Omega_c}) = u(G_{00}^{\Omega_c})$. Therefore, we can obtain (B.4) by $\limsup_{\Omega_c} G_{0k} = \mu m_k$ for $k \neq 0$ and $\limsup_{\Omega_c} G_{00} = \mu m_0$. Specially, for the ΓDP model, $\widehat{\text{ELBO}}^{\Omega}$ takes the form of

$$\mu - \sum_{k=1}^{K}\log m_k + (\alpha-1)\log\mu m_0 + \frac{J}{S}\sum_{s=1}^{S}\Big\{\log\frac{\Gamma(\mu)}{\Gamma(\mu+N_s)} + \sum_{k=1}^{K}\frac{1}{T_s}\sum_{t=1}^{T_s}\log\frac{\Gamma(\mu m_k+\hat{n}_{sk,t})}{\Gamma(\mu m_k)}\Big\},$$

up to a constant and (B.7) is also attained.

D Computational Complexity Analysis, Data and Code

For CATVI, updating the global variables takes linear time, and the Monte Carlo step iteratively samples each z_{ji} from K possible topics. Therefore, the computational complexity of Algorithm 1 is dominated by $O(K+\overline{T_s}K\overline{N_S})$, where $\overline{N_S}$ is the average number of words in a document, and $\overline{T_s}$ is the average of T_s defined in Algorithm 1. To implement this algorithm, we conduct our experiments on a c5d.4xlarge instance on the AWS EC2 platform, with 16 vCPUs and 32 GB RAM. It takes at most 5 hours to run all numerical experiments.

Python code for CATVI is available at

`https://anonymous.4open.science/r/46d58511-4743-43cc-8f6a-36ccca5661d8`

We obtain the *arXiv* and *Wiki* data from public open resources `https://arxiv.org/help/bulk_data` and `https://dumps.wikimedia.org`, respectively. The *NYT* data are from Sandhaus (2008). For the comparison methods, we implement OVI using the Python package 'gensim.models.hdpmodel' under GNU Lesser general public license v2.1. Moreover, we implement MOVI and SMVI using the Python package 'bnpy' under 3-clause BSD license, which is available at `https://github.com/bnpy/bnpy`.

E Sensitivity Analysis Results of CATVI

Figure 3a plots the sensitivity analysis with respect to the batch size varying from 128 to 1024. Figure 3b plots the sensitivity analysis with respect to the initial number of topics varying from 60 to 140. We observe that the performance is not sensitive to the change of these hyperparameters.