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

Yirui Liu  Xinghao Qiao  Jessica Lam
London School of Economics  London School of Economics  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 bioinformatics, language processing, computer vision 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 a 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-
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 avoids the risk of overfitting or underfitting, while also enabling more accurate predictions without sacrificing the computational efficiency. 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,

\[ G_0 | H \sim P(H), \quad \beta | \lambda \sim p(\beta | \lambda), \quad G_j | G_0 \sim \mathcal{R}(G_0), \]

\[ z_{ji} | G_j \sim G_j, \quad x_{ji} | z_{ji} \sim f(x_{ji} | \beta, z_{ji}), \]

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 \( \beta \) 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 \( \beta \) 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,

\[ G_0 | H \sim \text{DP}(\alpha H), \quad G_j | G_0 \sim \text{DP}(\gamma G_0), \]

where \( \alpha, \gamma \) are concentration parameters, and \( H, G_0 \) are normalized base 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 \( \text{DP}(\alpha H) \), 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 \( \text{DP}(\gamma G_0) \). For each topic \( k \), the distribution of words over vocabulary is sampled from a \( W \)-dimensional Dirichlet distribution parameterized by \( \eta \), \( \beta_k = (\beta_{k,1}, \ldots, \beta_{k,W})^T \sim \text{Dir}(\eta) \). For each word \( i \) in document \( j \), a topic assignment \( z_{ji} = \phi_k \) is chosen from \( z_{ji} \sim \text{Multinomial}(G_j) \), where \( \phi_k \) represents topic \( k \). Finally, the observation \( x_{ji} \) is generated from the assigned topic and the corresponding within-topic word distribution, \( x_{ji} | \{z_{ji} = \phi_k\} \sim \text{Multinomial}(\beta_k) \).

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 GDP model, which has the following structure,

\[ G_0 | H \sim \Gamma \text{P}(\alpha H), \quad G_j | G_0 \sim \text{DP}(G_0), \]

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.
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 use 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 (NP-ELBO), defined as the limit inferior of the parametric evidence lower bound under a finite partition, \(\liminf_a(\text{ELBO})\), that is

\[
\liminf_a \left\{ E_{q(Z')} \log p(X,Z') - E_{q(Z')} \log q(Z') \right\},
\]

where \(p(X,Z')\) and \(q(Z')\) correspond to the induced measures from the joint distribution and the variational distribution on \(\Omega\), 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

\[
\mathcal{KL}(Q(\Omega) \| P(Z|X)) + \text{NP-ELBO} = \log p(X).
\]

This demonstrates the equivalence between maximizing the NP-ELBO 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 NP-ELBO 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 NP-ELBO, 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 \(\beta\), local latent variables \(z\), global prior \(G_0\), local priors \(G_{[j]}\) and observations \(x\), where \(z = \{z_j\}_{j=1}^J\), \(z_j = \{z_{j,i}\}_{i=1}^{N_j}\), \(x = \{x_j\}_{j=1}^J\), \(x_j = \{x_{j,i}\}_{i=1}^{N_j}\) and \(G_{[j]} = \{G_j\}_{j=1}^N\). We aim to find the variational distribution to maximize the NP-ELBO. In contrast to traditional approaches under the mean-field setting, we factorize the variational distribution...
\begin{equation}
q(\beta, z, G_0, G_{[j]}) = q(\beta)q(G_0) \prod_{j=1}^{J} q(G_j|G_0, z_j) \prod_{i=1}^{N_j} q(z_{ji}).
\tag{7}
\end{equation}

in the sense of the probability law. Such conditional design facilitates the recovery of the dependence structure among \(G_0, G_{[j]}\) and \(z\).

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

\[
\begin{aligned}
\liminf_{n} & \left\{ E_q(\beta, z, G_0, G_{[j]}) \log p(x, \beta, z, G_0, G_{[j]}) \\
& - \sum_{j=1}^{J} E_q(G_0) E_q(z_j) \log q(G_j|G_0, z_j) \\
& - \mathcal{H}(q(G_0)) - \mathcal{H}(q(\beta)) - \sum_{j=1}^{J} \sum_{i=1}^{N_j} \mathcal{H}(q(z_{ji})) \right\},
\end{aligned}
\tag{8}
\]

where the entropy \(\mathcal{H}(q(\cdot)) = E_q(\cdot) \log q(\cdot)\) and \(\Omega\) is a partition of the sample space \(\Omega\) for \(G_0\) and \(G_{[j]}\).

**Conditional variational distribution** To maximize the NPELBO in (6), we first compute the optimal variational distribution of \(G_j\) given \(G_0\) and \(z_j\) for each \(j\). As \(p(x, \beta, z, G_0, G_{[j]}) = p(G_0, z)p(x|z) \prod_{j=1}^{J} p(G_j|G_0, z_j)\), the non-constant term in (8) with respect to \(q(G_j|G_0, z_j)\) is

\[
\liminf_{n} \left\{ \sum_{j=1}^{J} E_q(G_0) E_q(z_j) \log p(G_j|G_0, z_j) \\
- \log q(G_j|G_0, z_j) \right\}.
\tag{9}
\]

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 \(p(G_j|G_0, z_j)\). Consequently, NPELBO in (6) does not contain any term related to \(q(G_j|G_0, z_j)\). In Appendix C.3 we derive NPELBO with respect to \(q(G_0)\) as

\[
\liminf_{n} \left\{ \sum_{j=1}^{J} E_q(G_0) E_q(z_j) \log p(G_j|G_0) p(z_j) \\
- \mathcal{KL}(q(G_0)||p(G_0)) \right\}
\tag{10}
\]

up to a constant. It is important to note that \(E_p(G_j|G_0)\) is with respect to the prior \(p(G_j|G_0)\) instead of the variational distribution \(q(G_j|G_0)\), 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 \(z\) and (ii) updating the variational distributions for global variables \(G_0\) and \(\beta\).

#### Empirical distribution

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

\[
\hat{q}(z_s) = \frac{1}{T_s} \sum_{t=1}^{T_s} \delta_{\hat{z}_{s,t}}, \quad \hat{z}_{s,t} \sim q(z_s).
\]

#### Empirical evidence lower bound

Using the empirical distribution \(\hat{q}(z_s)\), we obtain an empirical evidence lower bound with respect to \(q(G_0)\), ELBO*, by replacing \(q(z_s)\) in (9) with \(\hat{q}(z_s)\), that is,

\[
\sum_{s=1}^{S} \sum_{t=1}^{T_s} \frac{J}{ST_s} E_q(G_0) \log E_p(G_j|G_0) p(\hat{z}_{s,t}|G_0) \\
- \mathcal{KL}(q(G_0)||p(G_0))
\tag{10}
\]

up to a constant. It is obvious that \(E(\text{ELBO}^*) = \text{ELBO}^*\), 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 \(\hat{z}_s = \{\hat{z}_{s,t}\}_{t=1}^{T_s}\) to update the parameters for \(q(G_0)\).

#### Resampling

We next present the procedure to get the empirical distribution \(\hat{q}(z_s)\). As \(G_0\) 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 \(\hat{z}_s\) given \(q(G_0)\) and \(q(\beta)\). Therefore, we propose the following Gibbs sampling approach to get samples under optimal variational distributions. Conditional on \(q(G_0)\), \(q(\beta)\) and samples \(\hat{z}_{s,i-} = \{\hat{z}_{si} : i = 1, \ldots, N_s, i \neq i\}\), it follows from (9) that the optimal variational distribution of \(\log q(z_{si})\) is proportional to

\[
E_q(G_0) E_p(G_j|G_0) p(z_{si}|\hat{z}_{s,i-}) \\
+ E_q(\beta) \log p(x_t|z_{si}, \beta)
\tag{11}
\]

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 \( \hat{z}_{s,i} \), \( \hat{z}_{s,T_s} \), from the stable Markov chain to update the parameters of \( q(G_0) \) 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 \( \hat{z}_{s,1}, \ldots, \hat{z}_{s,T_s} \) correspondingly.

4.3 Adaptive Truncation

Finally, we seek to obtain the finite partition \( \Omega \) that could reach the limit inferior in NPELBO. Rather than having \( \Omega \) fixed on a universal truncation level, we enable the dimension of \( \Omega \) 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 \( \Omega \) 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, \( q(z_{si}) \in \Omega \) is positive given (11), and hence \( \hat{z}_{si} \) can be sampled within \( \phi_0 \), that is, \( \hat{z}_{si} \) is a new sample, distinct from \( \phi_1, \ldots, \phi_K \). If this happens, we draw a new \( \phi_{K+1} \) and refine the partition as \( \{\phi_0, \phi_1, \ldots, \phi_{K}, \phi_{K+1}\} \), where \( \phi_0 \) 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 \( \phi_0 \) after each update, to minimize the KL divergence, the posterior should be proportional to the prior on \( \hat{z}_{si} \), given \( q(G_0(\phi_0)) \propto p(G_0(\phi_0)) \). Moreover, if we further partition \( \phi_0 \) into \( \phi_0 = \sum_{k=1}^{\infty} p_k \phi_k \), the KL divergence stays the same. Thus, \( E(\text{ELBO}^p) = \text{ELBO}^p = \text{NPELBO} \). See Appendix C.3 for a justification.

We summarize the CATVI algorithm in Algorithm 1.

**Algorithm 1: CATVI Algorithm**

1. Initialize the partition \( \Omega \), the parameters for \( q(G_0), q(\beta) \) and set up the step-size \( \{\rho_t\}_{t \geq 1} \).
2. Randomly select \( x_1, \ldots, x_S \) from the dataset.
3. For \( s \in \{1, \ldots, S\} \)
   - Sample \( \hat{z}_{si} \mid q(G_0), q(\beta), \hat{z}_{s,i-1} \).
   - If sampling a new \( \hat{z}_{si} \) then
     - Refine the partition \( \Omega \).
   end if
4. End for
5. Until convergence
6. Resample \( \hat{z}_s = \{\hat{z}_{s,t}\}_{t=1}^{T_s} \).}

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 \( q(G_s|G_0, z_s) = \text{DP}(\sum_{k=1}^{\infty} n_{sk} \delta_{\phi_k} + G_0) \), 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, \( q(\beta_k) = \text{Dirichlet}(\lambda_k) \), where \( \lambda_k = (\lambda_{k1}, \ldots, \lambda_{kW})^T \) is the parameter of vocabulary distribution for topic \( k \). The variational distribution for topics without any observation remains the same as the prior, hence \( q(\beta_0) = \text{Dirichlet}(\nu) \). Third, we specify the variational family for \( G_0 \) using spike and slab distributions [Andersen et al., 2017] as \( q(G_0) = \sum_{k=1}^{K} m_k \delta_{\phi_k} + m_0 \text{DP}(\alpha H) \), such that \( \sum_{k=0}^{K} m_k = 1 \). Finally, following (11) we use Monte Carlo sampling to obtain samples \( \{\hat{z}_s\}_{s=1}^S \) avoiding the need to parametrize their variational distributions.

As different samples in \( \{\hat{z}_s\}_{s=1}^S \) 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 \( \Omega \), and cluster 0 is the topic without any observation. Given samples \( \{\hat{z}_s\}_{s=1}^S \), we define the number of topics with observations by \( \hat{K} = \sum_{k=0}^{\infty} I(\sum_{s=1}^{S} T_s \hat{n}_{sk,t} > 0) \), where \( \hat{n}_{sk,t} = \sum_{s=1}^{S} I(\hat{z}_{st} = \phi_k) \). In Appendix C.4, we rely on (10) to derive the empirical evidence lower bound with respect to \( q(G_0) \),

\[
\alpha \log m_0 - \sum_{k=0}^{K} \log m_k \\
+ \sum_{s=1}^{S} \sum_{k=1}^{K} \sum_{t=1}^{T_s} \frac{J}{ST_s} \log \frac{\Gamma(\gamma m_k + \hat{n}_{sk,t})}{\Gamma(\gamma m_k)} 
\]

(12)
up to a constant. According to Algorithm 1, we repeatedly select documents of a batch size $S$, sample $\{\hat{z}_s\}_{s=1}^S$, and update parameters for $G_0$ and $\beta$ 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 $\Omega$ to be $(K+1)$-dimensional, with $K$ single points $\{\phi_k\}_{k=1}^K$ and one complement set $\phi_0 = \Omega \setminus \{\phi_k\}_{k=1}^K$. During the training, this procedure is repeated until $\Omega$ 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 $\mu$ to approximate $G_0(\Omega)$. We choose the variational family for the global prior $G_0$ as $q(G_0) = \mu(\sum_{k=1}^{K} m_k \delta_{\phi_k} + m_0 \tilde{N}(\alpha H))$, where $\tilde{N}$ 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 $\Gamma$DP model, in Appendix B.3.

### 6 RELATIONSHIP TO RELATED WORKS

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), the split-merge variational inference (SMVI) (Bryant and Sudderth, 2012) and Gibbs sampling (GS) (Teh et al., 2004).

**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 44M 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

$$\text{perplexity} = \exp \left\{ -\sum_{j \in D_{\text{test}}} \log p(x_j^{\text{test}} | x_j^{\text{train}}, D_{\text{train}}) \right\},$$

where $D_{\text{train}}$ and $D_{\text{test}}$ represent the training and test data, respectively, $x_j^{\text{train}}$ and $x_j^{\text{test}}$ are the training and test words in test document $j$, respectively, and $|x_j^{\text{test}}|$ is the number of words in $x_j^{\text{test}}$ (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 cannot be computed exactly, the standard routine uses $D_{\text{train}}$ to compute the variational distribution for $\beta$ and $G_0$, then obtains the variational distribution for $G_j$ based on $G_0$ and $x_j^{\text{test}}$, and then approximates the likelihood by

$$p(x_j^{\text{test}} | x_j^{\text{train}}) = \prod_{w \in x_j^{\text{test}}} \sum_{k=0}^{K} \tilde{\gamma}_{jk} \tilde{\beta}_{kw},$$

where $\tilde{\gamma}_{jk}$ and $\tilde{\beta}_{kw}$ are the variational expectations of $G_{jk}$ and $\beta_{kw}$, 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 $\alpha$ and $\gamma$ are the concentration parameters for $G_0$ and $G_j$, respectively, and $\eta$ 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 $\gamma$. In the experiments, we remove clusters with fewer than 1 document during the training.

### 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, SMVI and GS provide much higher perplexities, we do not plot their results in Figure 2. Table 1 reports numerical summaries for all comparison methods. In particular, as GS cannot scale to large datasets, we use a subset with 500 documents to run the experiments. Several conclusions can be drawn here. First, on all three datasets, CATVI uniformly outperforms competing methods. The improvement is highly consequential, 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 hence 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.

<table>
<thead>
<tr>
<th></th>
<th>arXiv</th>
<th>Wiki</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tell</td>
<td>null</td>
</tr>
<tr>
<td>2</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>3</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>4</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>5</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>6</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>7</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>8</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>9</td>
<td>try</td>
<td>null</td>
</tr>
<tr>
<td>10</td>
<td>try</td>
<td>null</td>
</tr>
</tbody>
</table>

### 8 DISCUSSION

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.
Acknowledgments

We thank the anonymous reviewers for useful comments during the review process.

Opinions expressed in this paper are those of the authors, and do not necessarily reflect the view of J.P. Morgan. Opinions and estimates constitute our judgement as of the date of this Material, are for informational purposes only and are subject to change without notice. This Material is not the product of J.P. Morgan’s Research Department and therefore, has not been prepared in accordance with legal requirements to promote the independence of research, including but not limited to, the prohibition on the dealing ahead of the dissemination of investment research. This Material is not intended as research, a recommendation, advice, offer or solicitation for the purchase or sale of any financial product or service, or to be used in any way for evaluating the merits of participating in any transaction. It is not a research report and is not intended as such. Past performance is not indicative of future results. Please consult your own advisors regarding legal, tax, accounting or any other aspects including suitability implications for your particular circumstances. J.P. Morgan disclaims any responsibility or liability whatsoever for the quality, accuracy or completeness of the information herein, and for any reliance on, or use of this material in any way. Important disclosures at: www.jpmorgan.com/disclosures.

References


Artificial Intelligence and Statistics, volume 2, pages 564–571.


A Short Review of Completely Random Measures

Suppose that \((\Omega, \mathcal{F})\) is a Polish sample space, \(\Theta\) is the set of all bounded measures on \((\Omega, \mathcal{F})\) and \(\mathcal{M}\) is a \(\sigma\)-algebra on \(\Theta\). 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

\[
(P(A_1), \ldots, P(A_n)) \sim \text{Dirichlet}(P_0(A_1), \ldots, P_0(A_n))
\]

for any partition \(\Omega = (A_1, \ldots, A_n)\) of \(\Omega\), 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 \(\Omega\). 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,

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

where \(A\) is any measurable subset of \(\Omega\) 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 \(\Omega\) 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 \(q(\beta)\) and \(q(z_{si})\)

To maximize the NPELBO, we iterate the following three steps: (i) randomly select a small batch from the entire data, (ii) sample \(\{\hat{z}_{si}\}_{s=1}^S\) by Monte Carlo method, and (iii) update \(q(G^i_0)\) and \(q(\beta)\) in the stochastic variational inference framework.

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

\[
\liminf_{\alpha} \left\{ \sum_{j=1}^J E_{q(\beta)}E_{q(z_j)} \log p(x_j | z_j, \beta) - KL(q(\beta)\|p(\beta)) \right\},
\]  

(A.1)
and the empirical evidence lower bound with respect to \( q(\beta), \) ELBO, is,

\[
\sum_{s=1}^{S} \sum_{t=1}^{T_s} \frac{J}{ST_s} E_{q(\beta)} \log p(x_s | z_{s,t}, \beta) - KL(q(\beta) | p(\beta)) \tag{A.2}
\]

to a constant and then we can update its parameter with the corresponding random gradient in a similar way. Moreover, the NPELBO with respect to \( q(z_{si}) \) is

\[
\lim_{s} \inf \left\{ E_{q(G_0^i)} E_{q(z_j)} \log E_{p(G_j^i | G_0^i)} p(z_j | G_0^i) + E_{q(z_j)} E_{q(\beta)} \log p(x_j | z_j, \beta) - E_{q(z_j)} \log q(z_j) \right\}, \tag{A.3}
\]

Factorizing \( E_{q(z_j)} \) as \( E_{q(z_j)} E_{q(z_j_t)} \) leads to (11). 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 \( \tau \)-th iteration, we first define

\[
m^{*}_{k} \propto J S^{-1} \gamma \sum_{s=1}^{S} \{ T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m^{(\tau)}_k + \hat{n}_{s,t}) - \Phi(\gamma m^{(\tau)}_k) \} m^{(\tau)}_k - 1 \quad k = 1, \ldots, K,
\]

\[
m^{*}_{k} \propto J S^{-1} \gamma \sum_{s=1}^{S} \{ T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m^{(\tau)}_k + \hat{n}_{s,t}) - \Phi(\gamma m^{(\tau)}_k) \} m^{(\tau)}_k - 1 \quad k = 0,
\tag{B.1}
\]

where \( \Phi(\cdot) \) denotes the log-gamma function, such that \( \sum_{k=0}^{K} m^{*}_k = 1 \), and then we update the parameters by \( m^{(\tau+1)} = (1 - \rho) m^{(\tau)} + \rho \cdot m^{*}_k \), where \( \rho \) 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 \( \beta \).** By (A.2), we update the parameters for \( q(\beta) \) using samples \( \{z_{s}\}_{s=1}^{S} \). We define \( \lambda^{*}_{kw} \) for topic \( k \) and word \( w \) as,

\[
\lambda^{*}_{kw} = \eta + \sum_{s=1}^{S} \sum_{t=1}^{T_s} \frac{J}{ST_s} I(z_{s,t} = \phi_k, x_{s,t} = w), \tag{B.2}
\]

and update the parameter \( \lambda_k \) by \( \lambda^{(\tau+1)}_k = (1 - \rho_k) \lambda^{(\tau)}_k + \rho \cdot \lambda^{*}_k \) for each \( k \), where \( \lambda^{*}_k = (\lambda^{*}_{k1}, \ldots, \lambda^{*}_{kw}) \).

**Sampling for \( z \).** According to (11) we sample \( z_{si} \) conditional on \( q(G_0) \) and \( z_{si-} \) by

\[
q(z_{si} = \phi_k) \propto \left( \frac{\gamma m_k + \hat{n}_{s,i-}^{w}}{\gamma m_0 + \sum_{w=1}^{W} \lambda^{*}_kw} \right)^k - 1 \quad k = 1, \ldots, K, \tag{B.3}
\]

\[
q(z_{si} = \phi_k) \propto \left( \frac{\gamma m_k + \hat{n}_{s,i-}^{w}}{\gamma m_0 + \sum_{w=1}^{W} \lambda^{*}_kw} \right)^k - 1 \quad k = 0,
\]

\[
to construct the Markov chain, where \( \hat{n}_{s,i-}^{w} = \sum_{1 \leq t \leq T_s, t \neq i} I(z_{s,t} = \phi_k) \). Whenever the sampled \( z_{si} \) is in \( \phi_0 \), meaning \( z_{si} \) 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 \( \phi_{K+1} \). Otherwise the partition dimension remains the same. Iterating the sampling scheme till convergence, we obtain the samples \( \{z_{s,t}\}_{1 \leq s \leq S, 1 \leq t \leq T_s} \) and corresponding \( \{\hat{n}_{s,t}\}_{1 \leq s \leq S, 1 \leq k \leq K, 1 \leq t \leq T_s} \) for the selected chunk.

### B.3 CATVI for the HDP Model

HDP releases the constraint of fixed concentration parameter \( \gamma \) in HDP. Therefore, the CATVI algorithm for HDP inherits the steps in (B.1) and (B.3), except that a parameter \( \mu \) replaces the concentration parameter \( \gamma \) in both formulas.
We derive the empirical evidence lower bound in Appendix C.7 with respect to $q(G_0)$ as,

$$
\sum_{k=1}^{K} \log v(\mu m_k) + \log u(\mu o_0) + \sum_{s=1}^{S} \sum_{i=1}^{T_s} \log \Gamma(\mu) + N_s \quad \sum_{s=1}^{S} \sum_{k=1}^{K} \sum_{i=1}^{T_s} \log \frac{\Gamma(\mu m_k + \hat{n}_{sk,t})}{\Gamma(\mu m_k)} + K \log \mu \quad (B.4)
$$

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

$$
m_k^* \propto J S^{-1} \mu \sum_{s=1}^{S} \{ T_s^{-1} \sum_{i=1}^{T_s} \Phi(\mu m_k^{(r)} + \hat{n}_{sk,t}) - \Phi(\mu m_{k}^{(r)}) \} m_k^{(r)} - 1 \quad k = 1, \ldots, K, \quad k = 0, \quad (B.5)
$$

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

$$
q(z_{si} = \phi_k) \propto \left\{ \begin{array}{ll}
(\mu m_k + \hat{n}_{si}) \exp \left( \Phi(\lambda_{k,x_{si}}) - \Phi(\sum_{w=1}^{W} \lambda_{kw}) \right) & k = 1, \ldots, K, \\
\mu o_0 \exp \left( \Phi(\eta) - \Phi(W \eta) \right) & k = 0.
\end{array} \right. \quad (B.6)
$$

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

$$
g'(\mu) = \frac{\alpha - 1}{\mu} - 1 + \frac{J S}{S} \sum_{s=1}^{S} \left\{ \Phi(\mu) - \Phi(\mu + N_s) + \sum_{k=1}^{K} \frac{1}{T_s} \sum_{i=1}^{T_s} m_k \left( \Phi(\mu m_k + \hat{n}_{sk,t}) - \Phi(\mu m_k) \right) \right\},
$$

and then update $\mu$ by $\mu^{(r+1)} = \mu^{(r)} + \rho_t g'(\mu^{(r)})$.

### C Technical Proofs and Derivations

#### C.1 Proof for (4)

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

$$
\int_{\Theta} \log \frac{dq'}{dp'} \cdot dq' = \int_{\Theta} \log \frac{dq}{dp} \cdot dq.
$$

It follows from $\lim \sup \frac{dq'}{dp'} = \frac{dQ}{dP}$ and the monotone convergence theorem that

$$
\lim \sup \int_{\Theta} \log \frac{dq'}{dp'} \cdot dq' = \int_{\Theta} \log \frac{dQ}{dP} \cdot dQ.
$$

Combining the above equations yields (4). Furthermore, suppose there exists a sequence of partition $\{\Omega_i\}_{i \geq 1}$ such that $\lim \sup \Omega_i = \Omega$, we have

$$
\lim \sup \int_{\Theta} \log \frac{dq^{(n)}}{dp^{(n)}} \cdot dq^{(n)} = \lim \sup \int_{\Theta} \log \frac{dq^{(n)}}{dp^{(n)}} \cdot dQ = \int_{\Theta} \log \frac{dq'}{dp} \cdot dq = \int_{\Theta} \log \frac{dQ}{dP} \cdot dq.
$$

Hence $\lim \sup_{\Theta} \mathcal{KL}(q^{(n)} || p^{(n)}) = \mathcal{KL}(q^{(n)} || p^{(n)})$, which will be used in Appendix C.4

#### C.2 Proof for (6)

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

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

Taking the limit inferior on both sides, we have

$$
\lim \inf \int \log \frac{p(X, Z)}{q(Z)} q(dZ) = \log p(X) - \lim \sup_{\Theta} \left\{ - \int \log \frac{p(Z | X)}{q(Z)} q(dZ) \right\}.
$$

Combining the above equation with the definition of NP$\text{ELBO}$ in (5) and the KL divergence in (4) yields (6).
C.3 Derivation for (9)

By \( p(G_0', \{z_j\}_{j=1}^d) = \int \cdots \int p(G_0', \{G_j\}_{j=1}^d, \{z_j\}_{j=1}^d) dG_1 dG_2 \cdots dG_J \) and the hierarchical generative structure, the evidence lower bound under partition \( \Omega \) with respect to \( q(G_0') \) equals,

\[
\text{ELBO}' = E_{q(G_0')} E_{q(\{z_j\}_{j=1}^d)} \log p(G_0', \{z_j\}_{j=1}^d) - E_{q(G_0')} \log q(G_0') + \text{constant}
\]

\[
= E_{q(G_0')} \prod_{j=1}^d \int p(G_j | G_0') p(z_j | G_j) dG_j - E_{q(G_0')} \log q(G_0') + \text{constant}
\]

\[
= \sum_{j=1}^d E_{q(G_0')} \log E_{p(G_j | G_0') | G_j} p(z_j | G_j) + E_{q(G_0')} \log p(G_0') - E_{q(G_0')} \log q(G_0') + \text{constant}.
\]

Furthermore, based on the equation above, (8) can be expressed as \( \text{NPELBO} = \lim \inf_n \text{ELBO}' \).

C.4 Derivation for (12)

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

\[
\text{ELBO} = \sum_{k=1}^K \Gamma(\gamma) \prod_{k=1}^K \frac{\Gamma(\gamma G_{0k} + \hat{n}_{sk,t})}{\Gamma(\gamma G_{0k})}.
\]

Based on the points \( \{\phi_k\}_{k=1}^K \) defined in Section 5.1, we propose a sequence of partition \( \{\Omega_c : \Omega_c = \bigcup_{k=1}^K \Omega_{ck} \}_{c \geq 1} \) to approach \( \Omega \), where \( \Omega_{ck} = \{\phi_c - c^{-1}, \phi_c + c^{-1}\} \) for \( k = 1, \ldots, K \) and \( \Omega_c \) is the corresponding complement. Under \( \Omega_c \), \( q(G_0') = d_{K+1}(m_0^{-1}(G_0' - M_0')) \) and \( p(G_0') = d_{K+1}(G_0') \), where \( d_{K+1}() \) denotes the density function for \((K+1)\)-dimensional Dirichlet distribution, \( M = \sum_{k=1}^K m_k \delta_{\phi_k} \) and \( M_0' \) is the corresponding induced random variable. By (10), the empirical evidence lower bound under \( \Omega_c \) is

\[
E_{q(G_0')} \left\{ \sum_{k=1}^K (\alpha H_k^{m_0} - 1) \log \frac{m_0 G_{0k}}{G_{0k} - m_k} + (\alpha H_k^{m_0} - 1) \log m_0 \right. \\
+ \left. \sum_{s=1}^S \sum_{k=1}^K \sum_{t=1}^{T_s} \frac{1}{ST_s} \log \frac{\Gamma(\gamma G_{0k} + \hat{n}_{sk,t})}{\Gamma(\gamma G_{0k})} \right\} + \text{constant},
\]

where \( H_k^{m_0} = H(\Omega_{ck}) \). Since \( (G_{0k} - m_k)/m_0 \sim \text{Beta}(H_k^{m_0}) \) under \( q(G_0') \), the term \( E_{q(G_0')} (\alpha H_k^{m_0} - 1) \log m_0 (G_{0k} - m_k)^{-1} \) is constant with respect to parameters \( \{m_k\}_{k=0}^K \). Taking \( \lim \sup_n \) on both sides of the above equation with \( \lim \sup_n E_{q(G_0')} (\log G_{0k}) = \lim m_k, \lim \sup_n H_k^{m_0} = 0 \) for \( k > 0 \) and \( \lim \sup_n H_0^{m_0} = 1 \), 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 \( \phi_0 \), and hence we have

\[
q(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K)) \propto p(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K)) p(\mathbf{x} | G_0(\phi_1), \ldots, G_0(\phi_K)).
\]

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

\[
\KL(q(G_0') \| p(G_0' | \mathbf{x})) = \int \log \frac{q(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K))}{p(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K))} dq(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K))
\]

\[
= - \log \mathcal{N},
\]
because
\[ q(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K)) = p(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K))p(x \mid G_0(\phi_1), \ldots, G_0(\phi_K)) / N, \]
where \( N \) is the normalization constant,
\[
\mathcal{N} = \int \cdots \int p(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K))p(x \mid G_0(\phi_1), \ldots, G_0(\phi_K))dG_0(\phi_1)dG_0(\phi_1)\cdots dG_0(\phi_K)
\]
\[
= \int p(x \mid G_0(\phi_1), \ldots, G_0(\phi_K))dp(G_0(\phi_0), G_0(\phi_1), \ldots, G_0(\phi_K))
\]
\[
= \int p(x \mid G_0(\phi_1), \ldots, G_0(\phi_K))dp(G_0(\phi_1), \ldots, G_0(\phi_K)).
\]
It is obvious that \( \mathcal{N} \) is independent of \( p(G_0(\phi_0)) \). Therefore, if we partition \( \phi_0 \) into \( \phi_0^1 \cup \phi_0^2 \), the normalization constant \( \mathcal{N} \) will not change, that is, the KL divergence under \( \Omega \) and \( \Omega' = (\phi_0^1, \phi_0^2, \phi_1, \ldots, \phi_K) \) are the same. Consequently, the partition \( \Omega \) enables the limit superior of KL divergence to be reached. By [6], the limit inferior of NP-ELBO is also attained.

C.6 Derivation for (B.1)

Consider the Lagrange multiplier of constrained optimization,
\[
L' = -\sum_{k=1}^{K} \log m_k + (\alpha - 1) \log m_0 + \sum_{s=1}^{S} \sum_{k=1}^{K} \frac{J}{ST_s} \sum_{t=1}^{T_s} \log \frac{\Gamma(\gamma m_k + \hat{n}_{sk,t})}{\Gamma(\gamma m_k)} - \lambda \left( \sum_{k=0}^{K} m_k - 1 \right),
\]
its first order conditions satisfy,
\[
\begin{cases}
JS^{-1} \gamma \sum_{s=1}^{S} \left\{ T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m_k + \hat{n}_{sk,t}) - \Phi(\gamma m_k) \right\} m_k - 1 = m_k \lambda, & k = 1, \ldots, K, \\
\alpha - 1 = m_0 \lambda, & k = 0.
\end{cases}
\]
Dividing \( \lambda \) 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 \( \partial m_k / \partial \theta_k = m_k - m_k^* \), \( \partial m_l / \partial \theta_k = -m_k m_l \) for \( l \neq k \), and the chain rule, we have
\[
\frac{\partial L}{\partial \theta_k} = \begin{cases}
JS^{-1} \gamma \sum_{s=1}^{S} \left\{ T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m_k + \hat{n}_{sk,t}) - \Phi(\gamma m_k) \right\} m_k - 1 - \lambda m_k & k = 1, \ldots, K, \\
\alpha - 1 - \lambda m_k & k = 0,
\end{cases}
\]
where \( L \) denotes ELBO'' in [12] and
\[
\Lambda = \alpha - 1 + \sum_{k=1}^{K} \left[ JS^{-1} \gamma \sum_{s=1}^{S} \left\{ T_s^{-1} \sum_{t=1}^{T_s} \Phi(\gamma m_k + \hat{n}_{sk,t}) - \Phi(\gamma m_k) \right\} m_k - 1 \right].
\]
As \( \partial L / \partial \theta_k = \Lambda (m_{k}^* - m_k) \), \( (m_{k}^* - m_k) \) represents the gradient with respect to \( \theta_k \) 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
\[
\log \mathbb{E}_{p(G_0 \mid G_0)} p(z_{s,t} \mid G_0^*) = \log \frac{\Gamma(\sum_{k=0}^{K} G_{0k})}{\Gamma(\sum_{k=0}^{K} G_{0k} + N_s)} \prod_{k=1}^{K} \frac{\Gamma(G_{0k} + \hat{n}_{sk,t})}{\Gamma(G_{0k})},
\]
In analogy to Appendix C.4, the empirical evidence lower bound under $\Omega_c$ is

$$K \log \mu + E_{q(G_{\theta_0}^c)} \left\{ \sum_{k=1}^{K} \log p(G_{\theta_0}^c_k) + \log p(c_{\theta_0}^c) + \frac{J}{S} \sum_{s=1}^{S} \sum_{k=1}^{K} T_s^{-1} \sum_{t=1}^{T_s} \log \frac{\Gamma(\gamma G_{\theta_0}^c + \hat{n}_{sk,t})}{\Gamma(\gamma G_{\theta_0}^c_k)} \right\},$$

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, $\limsup_s p(G_{\theta_0}^c_k) = v(G_{\theta_0}^c_k)$ and $\limsup_s p(G_{\theta_0}^c_0) = u(G_{\theta_0}^c_0)$. Therefore, we can obtain (B.4) by $\limsup \Omega_c p(G_{\theta_0}^c_k) = v(G_{\theta_0}^c_k)$ and $\limsup \Omega_c p(G_{\theta_0}^c_0) = u(G_{\theta_0}^c_0)$. Therefore, we can obtain (B.4) by $\limsup \Omega_c p(G_{\theta_0}^c_k) = v(G_{\theta_0}^c_k)$ and $\limsup \Omega_c p(G_{\theta_0}^c_0) = u(G_{\theta_0}^c_0)$. Therefore, we can obtain (B.4) by $\limsup \Omega_c p(G_{\theta_0}^c_k) = v(G_{\theta_0}^c_k)$ and $\limsup \Omega_c p(G_{\theta_0}^c_0) = u(G_{\theta_0}^c_0)$.

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 $O(K + T_s K N_S)$, where $N_S$ is the average number of words in a document, and $T_s$ 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://github.com/yiruiliu110/ConditionalVI. 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 MOV and SMVI using the Python package ‘bnpy’ under 3-clause BSD license, which is available at https://github.com/bnpy/bnpy. Finally, the codes to run GS are available at https://github.com/linkstrife/HDP.

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.