# Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modelling

Zhe Chen
Department of Statistics
University of Florida

Hani Doss
Department of Statistics
University of Florida

**Abstract**

In Latent Dirichlet Allocation, the number of topics, $T$, is a hyperparameter of the model that must be specified before one can fit the model. The need to specify $T$ in advance is restrictive. One way of dealing with this problem is to put a prior on $T$, but unfortunately the distribution on the latent variables of the model is then a mixture of distributions on spaces of different dimensions, and estimating this mixture distribution by Markov chain Monte Carlo is very difficult. We present a variant of the Metropolis-Hastings algorithm that can be used to estimate this mixture distribution, and in particular the posterior distribution of the number of topics. We evaluate our methodology on synthetic data, and compare it with procedures that are currently used in the machine learning literature. We also give an illustration on two collections of articles from Wikipedia. Supplemental materials for the paper are available online.

*Key words and phrases:* Latent Dirichlet allocation, Markov chain Monte Carlo, mixture model, topic modelling, uniform ergodicity.

Research supported by NSF Grant DIIS-17-24174

# 1 Introduction

Latent Dirichlet Allocation (LDA, Blei et al. 2003) is a heavily used model that is used to describe high-dimensional sparse count data represented by feature counts. Although the model can be applied to many different kinds of data, for example collections of annotated images and social networks, for the sake of concreteness, here we focus on data consisting of a collection of documents. Suppose we have a corpus of documents, and these span several different topics, such as sports, medicine, politics, etc. We imagine that for each word in each document, there is a latent (i.e. unobserved) variable indicating a topic from which that word is drawn. There are several goals, but two principal ones are to recover an interpretable set of topics for the corpus, and to infer the "topic proportions" for each document.

To describe the LDA model, we first set up some terminology and notation. There is a vocabulary $\mathcal{V}$ of $V$ words; typically, this is taken to be the union of all the words in all the documents of the corpus, after removing stop (i.e. uninformative) words. There are $D$ documents in the corpus, and for $d = 1, \ldots, D$, document $d$ has $n_d$ words, $w_{d1}, \ldots, w_{dn_d}$. In total, the corpus has $N = \sum_{d=1}^{D} n_d$ words. The order of the words is considered uninformative, and so is neglected. Each word is represented as an index $1 \times V$ vector with a $1$ at the $v^{\text{th}}$ element, where $v$ denotes the term selected from the vocabulary. Thus, document $d$ is represented by the vector $\boldsymbol{w}_d = (w_{d1}, \ldots, w_{dn_d})$ and the corpus is represented by the vector $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D)$. A topic is, by definition, a distribution over $\mathcal{V}$, i.e. a point in $\mathbb{S}_{V-1}$, the $(V-1)$-dimensional simplex. The number of topics, $T$, is finite and known. For $d = 1, \ldots, D$, for each word $w_{di}$, $z_{di}$ is an index $1 \times T$ vector which represents the latent variable that denotes the topic from which $w_{di}$ is drawn. The distribution of $z_{d1}, \ldots, z_{dn_d}$ will depend on a document-specific variable $\theta_d$ which indicates a distribution on the topics for document $d$.

We will let $\mathrm{Dir}_L(a_1, \ldots, a_L)$ denote the finite-dimensional Dirichlet distribution on the simplex $\mathbb{S}_{L-1}$. Also, we will use $\mathrm{Mult}_L(b_1, \ldots, b_L)$ to denote the multinomial distribution with number of trials equal to $1$ and probability vector $(b_1, \ldots, b_L)$. Given $T$, we will form a $T \times V$ matrix $\boldsymbol{\beta}$, whose $t^{\text{th}}$ row is the $t^{\text{th}}$ topic (how $\boldsymbol{\beta}$ is formed will be described shortly). Thus, $\boldsymbol{\beta}$ will consist of vectors $\beta_1, \ldots, \beta_T$, all lying in $\mathbb{S}_{V-1}$. Formally, the LDA model indexed by $T$ is described by the following hierarchical model, in which $\eta \in (0, \infty)$ and $\alpha \in (0, \infty)$ are hyperparameters:

1. $\beta_t \overset{\text{iid}}{\sim} \mathrm{Dir}_V(\eta, \ldots, \eta)$, $t = 1, \ldots, T$.

2. $\theta_d \overset{\text{iid}}{\sim} \mathrm{Dir}_T(\alpha, \ldots, \alpha)$, $d = 1, \ldots, D$, and the $\theta_d$'s are independent of the $\beta_t$'s.

3. Given $\boldsymbol{\beta}$ and the $\theta_d$'s, $z_{di} \overset{\text{iid}}{\sim} \text{Mult}_T(\theta_d)$, $i = 1, \dots, n_d$, $d = 1, \dots, D$, and the $D$ vectors $(z_{11}, \dots, z_{1n_1}), \dots, (z_{D1}, \dots, z_{Dn_D})$ are independent.
4. Given $\boldsymbol{\beta}$ and the $z_{di}$'s, $w_{di}$ are independently drawn from the row of $\boldsymbol{\beta}$ indicated by $z_{di}$, $i = 1, \dots, n_d$, $d = 1, \dots, D$.

From the description of the model, we see that there is a latent topic variable for each word in the document. Thus it is possible that a document has several topics. However, because there is a single $\theta_d$ for document $d$, the model encourages different words in document $d$ to have the same topic. Also note that the hierarchical nature of the model encourages different documents to share the same topics. This is because $\boldsymbol{\beta}$ is chosen once, at the top of the hierarchy, and is shared among the $D$ documents.

For $d = 1, \dots, D$, let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_D)$, and $\boldsymbol{z}_d = (z_{d1}, \dots, z_{dn_d})$. Let $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_D)$; this is a vector of length $N$ determining the latent topic variables for all the words in the corpus. Lines 1–3 induce a prior distribution on the parameter vector $\boldsymbol{\psi} := (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$. We will denote this prior by $\nu_{\boldsymbol{\psi}}^{(T)}$. Line 4 gives the likelihood of $\boldsymbol{\psi}$, which we will denote by $\ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\psi})$. The words $\boldsymbol{w}$ are observed, and we are interested in the posterior distribution of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{z}$ given $\boldsymbol{w}$.

A serious limitation of the model is that one has to specify the number of topics, $T$, in advance, and there is no simple way of doing so (Blei, 2012). It is now well recognized that selecting the number of topics is one of the most problematic choices in topic modelling. If $T$ is taken to be smaller than the true number of topics, then the posterior distribution of the parameters will be inconsistent, i.e. it will not converge to a point mass at the true value of the parameters, even with an infinite amount of data. On the other hand, specifying $T$ to be too large results in a deterioration of the rate at which the posterior distribution contracts around the true value (and also results in increased computational costs when fitting the model). See Tang et al. (2014) and Nguyen (2015) for precise statements of the last two facts regarding asymptotics. Generally speaking, robustness of the LDA model to the choice of $T$ is not well understood (Wallach et al., 2009a).

There are two natural approaches for dealing with the uncertainty in $T$. One is frequentist and is described as follows. Let $m_{\boldsymbol{w}}(T)$ be the marginal likelihood of $T$, i.e. $m_{\boldsymbol{w}}(T) = \int \ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\psi}) \nu_{\boldsymbol{\psi}}^{(T)}(\boldsymbol{\psi}) \, d\boldsymbol{\psi}$. (Since $\boldsymbol{\psi}$ has continuous as well as discrete components, if $f$ is a function of $\boldsymbol{\psi}$, the notation $\int f(\boldsymbol{\psi}) \, d\boldsymbol{\psi}$ should be taken to mean a combination of integration and summation in the obvious way.) The marginal likelihood $m_{\boldsymbol{w}}(T)$ is the normalizing constant in the statement "the posterior is proportional to the likelihood times the prior." The parameter $T$ may be estimated by $\widehat{T} = \arg\max_T m_{\boldsymbol{w}}(T)$ and, in fact, using the LDA model indexed by $\widehat{T}$ amounts to empirical Bayes inference. Unfortunately, $m_{\boldsymbol{w}}(T)$ is a very high dimensional integral of very large sums, and cannot be calculated except in trivial cases.

Newton and Raftery (1994) presented the *harmonic mean estimator* (HME) of the marginal likelihood which, in the present context, is described as follows. Let $\nu^{(T)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$ denote the posterior distribution of $\boldsymbol{\psi}$ given $\boldsymbol{w}$. Suppose that $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ is an ergodic Markov chain with invariant distribution $\nu^{(T)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$. The HME of $m_{\boldsymbol{w}}(T)$ is $\widehat{m}_{\boldsymbol{w}}(T) = \big[(1/n)\sum_{i=1}^n (1/\ell^{(T)}_{\boldsymbol{w}}(\boldsymbol{\psi}_i))\big]^{-1}$. It is very easy to show that the HME is consistent: we have

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{\ell^{(T)}_{\boldsymbol{w}}(\boldsymbol{\psi}_i)} \xrightarrow{\text{a.s.}} \int \frac{1}{\ell^{(T)}_{\boldsymbol{w}}(\boldsymbol{\psi})} \nu^{(T)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})\, d\boldsymbol{\psi} = \int \frac{1}{\ell^{(T)}_{\boldsymbol{w}}(\boldsymbol{\psi})} \frac{\ell^{(T)}_{\boldsymbol{w}}(\boldsymbol{\psi})\nu^{(T)}_{\boldsymbol{\psi}}(\boldsymbol{\psi})}{m_{\boldsymbol{w}}(T)}\, d\boldsymbol{\psi} = \frac{1}{m_{\boldsymbol{w}}(T)},$$

and therefore $\widehat{m}_{\boldsymbol{w}}(T) \xrightarrow{\text{a.s.}} m_{\boldsymbol{w}}(T)$. To estimate $T$ we proceed as follows. For each $T$ in a finite range, we run a Markov chain to form the HME of $m_{\boldsymbol{w}}(T)$, and then take $\widehat{T} = \arg\max_T m_{\boldsymbol{w}}(T)$. This method has two significant defects. First, convergence of HME's is extremely slow; in fact, the rate is typically much slower than $n^{1/2}$ (Wolpert and Schmidler, 2012). Second, one has to run a separate Markov chain for each value of $T$. These problems are well known, at least in some circles, but nevertheless the method is often used (see, e.g., Griffiths and Steyvers (2004), among many others). Chib (1995) provided a general-purpose method for estimating marginal likelihoods. While this method is often used in topic modelling, to the best of our knowledge it has been neither used nor investigated for the purpose of selecting $T$ in the LDA model. In Section 5 we show that the method does not perform well in our problem, and explain why this is to be expected in view of the high dimension of the model.

The other natural approach is Bayesian: we change the four-level hierarchy that defines the model to a five-level hierarchy, in which the first level stipulates that $T$ is drawn from some prior distribution $\nu_T$. A major problem with this approach is that the parameter is now $\vartheta := (T, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$, in which the dimensions of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{z}$ all depend on $T$. Estimating the posterior distribution of $\vartheta$ by Markov chain Monte Carlo in this kind of situation is notoriously difficult, and involves so-called transdimensional chains, of which reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995) gives the main class of algorithms in current use. There are several problems with RJMCMC: (i) RJMCMC is not an automatic algorithm. There are many parameters to tune and these have a major effect on the efficiency of the algorithm. Unfortunately, tuning these parameters is typically very difficult and can be done only by trial and error. (ii) RJMCMC involves an accept-reject step since it is an implementation of the Metropolis-Hastings algorithm, and often even after extensive tuning, the acceptance rate is very low. Acceptance rates are low particularly when the dimension is high, and can be unacceptably low even when the dimension is much smaller than the dimension of the LDA model. (iii) Theoretical results on rates of convergence for RJMCMC are very hard to come by.

In this paper we use the Bayesian approach described above, in which we put a prior distribution $\nu_T$ on $T$, and we do this in the first step of the hierarchy. The unknown parameter is then $\vartheta = (T, \boldsymbol{\psi})$, but instead of using RJMCMC to estimate the posterior distribution of $\vartheta$, we focus entirely on the marginal posterior distribution of $T$, thus avoiding RJMCMC.

Before proceeding, we remark on our conventions regarding notation for distributions. If $\varphi$ is a subcomponent of $\vartheta$, $\nu_\varphi$ and $\nu_{\varphi \mid \boldsymbol{w}}$ will denote the marginal prior and posterior distributions on $\varphi$, respectively, induced by the five-level hierarchy. For example, $\nu_T$ and $\nu_{T \mid \boldsymbol{w}}$ denote the marginal prior and posterior distributions of $T$; also $\nu_{\boldsymbol{z}}$ and $\nu_{\boldsymbol{z} \mid \boldsymbol{w}}$ denote the marginal prior and posterior distributions of $\boldsymbol{z}$. For the LDA model in which $T$ is fixed, we will use the superscript "$(T)$"; thus, $\nu_{\boldsymbol{z}}^{(T)}$ and $\nu_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}$ denote the prior and posterior distributions of $\boldsymbol{z}$, respectively, stipulated in the original 4-line model.

Let $\mathcal{T}$ be the support of $\nu_T$, and suppose that $q_T(\cdot, \cdot)$ is a Markov transition function on $\mathcal{T}$. In principle, to generate a Markov chain with invariant distribution $\nu_{T \mid \boldsymbol{w}}$, a Metropolis-Hastings algorithm based on $q_T(\cdot, \cdot)$ would be as follows.

1. Let $T$ be the current state. Generate a proposal $T' \sim q_T(T, \cdot)$.

2. Compute the acceptance ratio

$$r(T, T') = \frac{\nu_{T \mid \boldsymbol{w}}(T') q_T(T', T)}{\nu_{T \mid \boldsymbol{w}}(T) q_T(T, T')}. \tag{1.1}$$

3. Accept $T'$ with probability $\min\{r(T, T'), 1\}$; otherwise stay at $T$.

Unfortunately, Step 2 is infeasible because $\nu_{T \mid \boldsymbol{w}}(T)$ is analytically intractable: it is obtained by integrating/summing out $\boldsymbol{\psi}$ from $\nu_{\vartheta \mid \boldsymbol{w}}$, which is not possible because of the high dimensions involved in the LDA model. To deal with this problem, we proceed as follows. First, we note that for every fixed $T \in \mathcal{T}$, there exists a Markov transition function $g_T(\cdot, \cdot)$ on $\boldsymbol{z}$ for generating a Markov chain with invariant distribution $\nu_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}$ (we are referring to the "collapsed Gibbs sampler" (CGS) of Griffiths and Steyvers (2004), which we discuss in Section 2). Let $\boldsymbol{z}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_m$ be a chain generated according to $g_T(\cdot, \cdot)$, and for any $t \in \mathcal{T}$, define $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t)$ by

$$\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t) = \frac{1}{m} \sum_{l=1}^{m} \frac{\nu_{T, \boldsymbol{z} \mid \boldsymbol{w}}(t, \boldsymbol{z}_l)}{g_T(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)}.$$

Now in the model in which $T$ is fixed, the posterior distribution of $\boldsymbol{z}$ is known up to a normalizing constant—this fact is the basis for the CGS of Griffiths and Steyvers (2004)—and as a consequence, in our model $\nu_{T, \boldsymbol{z} \mid \boldsymbol{w}}(t, \boldsymbol{z}_l)$ is known up to a normalizing constant $c$ which does not depend on $t$ or $\boldsymbol{z}_l$. We can show that as $m \to \infty$, $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t) \xrightarrow{\text{a.s.}} \nu_{T \mid \boldsymbol{w}}(t)$. It is tempting to do the following: within each iteration, we calculate $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(T')$ and $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(T)$, and use them instead of $\nu_{T \mid \boldsymbol{w}}(T')$ and $\nu_{T \mid \boldsymbol{w}}(T)$, respectively, in Step 2 of the infeasible algorithm described above. (The fact that the constant $c$ is unknown does not cause a problem, since the constant cancels in (1.1).) Unfortunately, this algorithm has no theoretical validity, and the chain it simulates may not even have an invariant distribution. We discuss this in more detail in Remark 1 in Section 4. Instead, we proceed differently, as follows. Within each iteration, we calculate *only* $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(T')$ and use it instead of $\nu_{T \mid \boldsymbol{w}}(T')$, while recycling $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(T)$ from the previous iteration and using it instead of $\nu_{T \mid \boldsymbol{w}}(T)$. We show that the resulting algorithm is theoretically valid. Specifically, let $\mu_T^{m,n}$ denote the distribution of $T$ after $n$ cycles of this algorithm with $T_0$ being the initial sample (note that $\mu_T^{m,n}$ depends on $T_0$ but we suppress this dependence here). It turns out that for any fixed $m$, as $n \to \infty$, $\mu_T^{m,n}$ does not converge to $\nu_{T \mid \boldsymbol{w}}$. But we show that for any starting point $T_0$, as $m, n \to \infty$, $\mu_T^{m,n}$ converges to $\nu_{T \mid \boldsymbol{w}}$, and thus the algorithm described above provides a viable way to estimate the posterior distribution $\nu_{T \mid \boldsymbol{w}}$.

The idea above is a variation of the idea behind the so-called pseudo-marginal Metropolis-Hastings algorithm (PMMH), originally proposed by Beaumont (2003) and developed theoretically by Andrieu and Roberts (2009). In Section 4 we discuss the connection between our approach and the previous proposals and also the differences in the theoretical developments. The fact that the asymptotic regime requires both $m$ and $n$ to go to infinity imposes a computational burden that is greater than that for standard MCMC algorithms. However, this drawback has to be considered in the context of two facts. First, there does not seem to be a usable alternative: the dimension of the problem appears to preclude a workable RJMCMC algorithm. Second, the Markov chain in the "inner loop," i.e. the chain on $\boldsymbol{z}$, is a collapsed Gibbs sampler which executes very fast and has very good mixing properties, as we establish theoretically in Section 2. We note that Beaumont (2003) developed his PMMH algorithm in order to deal with a class of genetics problems for which there are no practical MCMC algorithms—this is the same situation we face.

The paper is organized as follows. In Section 2 we give a brief review of Bayesian inference for the standard LDA model and an existing Gibbs sampler which is used to estimate posterior distributions in the model. We also investigate the uniform ergodicity of this Gibbs sampler. In Section 3 we extend the standard LDA model to the case where the number of topics $T$ is a parameter of the model under the Bayesian framework. In Section 4 we design a variant of the Metropolis-Hastings algorithm for specifying $T$ in the LDA model. We also provide theoretical justification of our algorithm by showing that it can be used to consistently estimate the true posterior distribution of $T$ given the words and, more generally, the posterior distribution of all the latent variables in the model. In Section 5 we evaluate the performance of our methodology by considering a synthetic data set generated according to an LDA model, and also two real data sets consisting of two collections of articles from Wikipedia. We show that our algorithm has excellent performance on those data sets. In Section 6 we point out that there exist models that are alternatives to LDA, in which the number of topics is possibly unbounded and is to be inferred from the data, and we discuss our work in the context of these models. With the exception of Theorem 2 and Corollary 1, the proofs of the theoretical results are in the Appendix.

# 2 The Collapsed Gibbs Sampler for the LDA Model: Description and Uniform Ergodicity

In this section we review the CGS of Griffiths and Steyvers (2004) and state a theorem that asserts that this Markov chain is uniformly ergodic. We begin by obtaining an expression for the marginal posterior distribution of $\boldsymbol{z}$ given $\boldsymbol{w}$. We need this development for two purposes: (i) to establish uniform ergodicity, and (ii) to obtain the joint marginal posterior distribution of $(T, \boldsymbol{z})$ given $\boldsymbol{w}$, which we need in order to construct the Markov chain on $T$, in Section 4. This section deals only with the LDA model in which $T$ is fixed; therefore, throughout the entire section we will not use the superscript "$(T)$" for distributions, as it is unnecessary and cumbersome.

We first express the posterior distributions of $\boldsymbol{\psi}$ and $\boldsymbol{z}$ in convenient forms. The posterior distribution $\nu_{\boldsymbol{\psi} \mid \boldsymbol{w}}$ of $\boldsymbol{\psi}$ satisfies

$$\nu_{\boldsymbol{\psi} \mid \boldsymbol{w}}(\boldsymbol{\psi}) \propto \ell_{\boldsymbol{w}}(\boldsymbol{\psi}) \nu_{\boldsymbol{\psi}}(\boldsymbol{\psi}). \tag{2.1}$$

From the hierarchical nature of the LDA model we have

$$\nu_{\boldsymbol{\psi}}(\boldsymbol{\psi}) = \nu_{\boldsymbol{\psi}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}) = \nu_{\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\beta}}(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\beta}) \, \nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \, \nu_{\boldsymbol{\beta}}(\boldsymbol{\beta})$$

in self-explanatory notation, where $\nu_{\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\beta}), \nu_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, and $\nu_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ are given by Lines 3, 2, and 1, respectively, of the LDA model.

Before proceeding, we set up some notation that we will need. Recall that $n_d$ is the number of words in document $d$, $w_{di}$ represents word $i$ in document $d$, and $z_{di}$ is the latent topic variable for word $i$ in document $d$. Also, $z_{dit}$ is component $t$ of the vector $z_{di}$, $w_{div}$ is component $v$ of the vector $w_{di}$, and $\beta_{tv}$ is component $v$ of $\beta_t$. Additionally, we define the following:

$n_{dt} = \sum_{i=1}^{n_d} z_{dit}$ is the number of words in document $d$ assigned to topic $t$;

$m_{dtv} = \sum_{i=1}^{n_d} z_{dit}w_{div}$ is the number of words in document $d$ for which the latent topic is $t$ and the index of the word in the vocabulary is $v$;

$m_{\cdot tv} = \sum_{d=1}^{D} m_{dtv}$ is the number of words in the corpus for which the latent topic is $t$ and the index of the word in the vocabulary is $v$;

$m_{\cdot t\cdot} = \sum_{v=1}^{V} m_{\cdot tv}$ is the number of words in the corpus for which the latent topic is $t$.

With this notation, using the Dirichlet and multinomial distributions specified in Lines 1–3 of the model, we have

$$\begin{aligned}\nu_{\boldsymbol{\psi}}(\boldsymbol{\psi}) &= \Biggl[\prod_{d=1}^{D}\prod_{t=1}^{T}\theta_{dt}^{n_{dt}}\Biggr]\Biggl[\prod_{d=1}^{D}\biggl(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}\theta_{dt}^{\alpha-1}\biggr)\Biggr]\Biggl[\prod_{t=1}^{T}\biggl(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\prod_{v=1}^{V}\beta_{tv}^{\eta-1}\biggr)\Biggr] \\ &= \Biggl[\prod_{d=1}^{D}\biggl(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}\theta_{dt}^{n_{dt}+\alpha-1}\biggr)\Biggr]\Biggl[\prod_{t=1}^{T}\biggl(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\prod_{v=1}^{V}\beta_{tv}^{\eta-1}\biggr)\Biggr].\end{aligned} \tag{2.2}$$

From Line 4 of the LDA model statement we have

$$\begin{aligned}\ell_{\boldsymbol{w}}(\boldsymbol{\psi}) = \prod_{d=1}^{D}\prod_{t=1}^{T}\prod_{v=1}^{V}\beta_{tv}^{\sum_{i=1}^{n_d} z_{dit}w_{div}} &= \prod_{d=1}^{D}\prod_{t=1}^{T}\prod_{v=1}^{V}\beta_{tv}^{m_{dtv}} = \prod_{t=1}^{T}\prod_{v=1}^{V}\prod_{d=1}^{D}\beta_{tv}^{m_{dtv}} \\ &= \prod_{t=1}^{T}\prod_{v=1}^{V}\beta_{tv}^{\sum_{d=1}^{D} m_{dtv}} = \prod_{t=1}^{T}\prod_{v=1}^{V}\beta_{tv}^{m_{\cdot tv}}.\end{aligned} \tag{2.3}$$

Plugging the likelihood (2.3) and the prior (2.2) into (2.1) and combining terms, we get

$$\nu_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi}) \propto \Biggl[\prod_{d=1}^{D}\biggl(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\prod_{t=1}^{T}\theta_{dt}^{n_{dt}+\alpha-1}\biggr)\Biggr]\Biggl[\prod_{t=1}^{T}\biggl(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\prod_{v=1}^{V}\beta_{tv}^{m_{\cdot tv}+\eta-1}\biggr)\Biggr]. \tag{2.4}$$

By inspection of the expression for $\nu_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})$ above, it can be seen that given $\boldsymbol{z}$ (and $\boldsymbol{w}$),

$$\begin{aligned}&\theta_1,\dots,\theta_D \text{ and } \beta_1,\dots,\beta_T \text{ are all independent, with} \\ &\theta_d \sim \mathrm{Dir}_T(n_{d1}+\alpha,\dots,n_{dT}+\alpha), \text{ for } d=1,\dots,D, \\ &\beta_t \sim \mathrm{Dir}_V(m_{\cdot t1}+\eta,\dots,m_{\cdot tV}+\eta), \text{ for } t=1,\dots,T.\end{aligned} \tag{2.5}$$

George and Doss (2018) show that from (2.4) the conditional $\nu_{\boldsymbol{z}\,|\,\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{w}}(\boldsymbol{z})$ may be obtained by inspection. Let $p_{dit} = \prod_{v=1}^{V} \bigl(\beta_{tv}\theta_{dt}\bigr)^{w_{div}}$. They show that given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ (and $\boldsymbol{w}$),

$$\begin{gathered} z_{11}, \dots, z_{1n_1}, z_{21}, \dots, z_{2n_2}, \dots, z_{D1}, \dots, z_{Dn_D} \text{ are all independent, with} \\ z_{di} \sim \text{Mult}_T(p_{di1}, \dots, p_{diT}). \end{gathered} \tag{2.6}$$

The two conditionals (2.5) and (2.6) enable the construction of a two-cycle Gibbs sampler that runs on the pair $(\boldsymbol{z}, (\boldsymbol{\beta}, \boldsymbol{\theta}))$. Although it is not the Markov chain we use in this paper, this Gibbs sampler has very interesting properties, which we discuss in Section 5.4. Integrating out $\theta_1, \dots, \theta_D$ and $\beta_1, \dots, \beta_T$ in (2.4), we obtain the marginal posterior distribution $\nu_{\boldsymbol{z}\,|\,\boldsymbol{w}}$ of $\boldsymbol{z}$ up to a normalizing constant:

$$\nu_{\boldsymbol{z}\,|\,\boldsymbol{w}}(\boldsymbol{z}) \propto \left[\prod_{d=1}^{D}\left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\frac{\prod_{t=1}^{T}\Gamma(n_{dt}+\alpha)}{\Gamma(n_d+T\alpha)}\right)\right]\left[\prod_{t=1}^{T}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\right)\right]. \tag{2.7}$$

This formula was previously obtained by Griffiths and Steyvers (2004). To summarize: (2.4) and (2.7) give the posterior distributions (up to normalizing constants) of $\boldsymbol{\psi}$ and $\boldsymbol{z}$, respectively, in the LDA model indexed by $T$.

The Gibbs sampler developed in Griffiths and Steyvers (2004) runs over the $N$-dimensional vector $(z_{11}, \dots, z_{1n_1}, \dots, z_{D1}, \dots, z_{Dn_D})$, updating one variable at a time, with $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ integrated out. To describe the needed conditionals, we first set up some notation. For $d = 1, \dots, D$ and $i = 1, \dots, n_d$, we use $\boldsymbol{w}_{(-di)}$ to denote the collection of all words except for $w_{di}$ and we use $\boldsymbol{z}_{(-di)}$ to denote the collection of latent topic variables for $\boldsymbol{w}_{(-di)}$. The counts $n_d$, $n_{dt}$, $m_{dtv}$, $m_{\cdot tv}$, and $m_{\cdot t\cdot}$ were defined just above display (2.2). For any $d, d' \in \{1, \dots, D\}$, the counts $n_{d'(-di)}$, $n_{d't(-di)}$, $m_{d'tv(-di)}$, $m_{\cdot tv(-di)}$, and $m_{\cdot t\cdot(-di)}$ are defined in the same way, except that they are based on $\boldsymbol{w}_{(-di)}$ and $\boldsymbol{z}_{(-di)}$; so they are based on the corpus in which we have removed a single word, namely word $i$ in document $d$. It is clear that for each $t$, $n_{dt(-di)} = n_{dt} - z_{dit}$, $n_{d(-di)} = \sum_{t=1}^{T} n_{dt(-di)} = n_d - 1$, and that for $d' \neq d$, $n_{d't(-di)} = n_{d't}$ and $n_{d'(-di)} = \sum_{t=1}^{T} n_{d't(-di)} = n_{d'}$. Let $p_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$ denote the joint distribution of $\boldsymbol{\psi}$ and $\boldsymbol{w}$ under the LDA model. If $\varphi$ is a subcomponent of $\boldsymbol{\psi}$, then $p_{\varphi\,|\,\boldsymbol{\psi}_{-\varphi},\boldsymbol{w}}$ denotes the conditional distribution of $\varphi$ given all the other components of $\boldsymbol{\psi}$, and $\boldsymbol{w}$. This notation follows our conventions; however, we will simply write $p$ in order to avoid cumbersome expressions, whenever the meaning is clear from context.

As mentioned earlier, in order to proceed with our development, we need the conditional distribution of $z_{di}$ given $\boldsymbol{z}_{(-di)}$ and $\boldsymbol{w}$, i.e. obtain $p(z_{dit} = 1 \,|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w})$ for every $t$. In Chen and Doss

(2018) we obtain the formula

$$p(z_{dit} = 1 \mid \boldsymbol{z}_{(-di)}, \boldsymbol{w}) \propto \left(\frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t \cdot (-di)} + V\eta}\right)\left(\frac{n_{dt(-di)} + \alpha}{n_d - 1 + T\alpha}\right) \qquad \text{for } t = 1, \dots, T, \tag{2.8}$$

where $v$ denotes the term which $w_{di}$ is observed to take, i.e. $v$ is such that $w_{div} = 1$. It is natural to ask why we derive this expression, in view of the fact that the expression is given in Griffiths and Steyvers (2004). The reason we do this is that Griffiths and Steyvers (2004) do not provide any derivation or justification for the expression at all, and as one can see in Chen and Doss (2018), the needed calculation is far from trivial.

Let $g_T$ denote the Markov transition function for the CGS, i.e. $g_T(\boldsymbol{z}_0, \cdot)$ is the distribution of $\boldsymbol{z}_1$ given $\boldsymbol{z}_0$, and let $g_T^m(\boldsymbol{z}_0, \cdot)$ denote the $m$-step Markov transition function. Also, let $\mathcal{Z}_T$ denote the set of all possible values of $\boldsymbol{z}$. Theorem 1 establishes *uniform ergodicity*, which is the very strong condition that there exist constants $M > 0$ and $c > 0$ such that $\|g_T^m(\boldsymbol{z}_0, \cdot) - \nu_{\boldsymbol{z} \mid \boldsymbol{w}}(\cdot)\|_{\text{TV}} \leq M(1-c)^m$ for all initial $\boldsymbol{z}_0 \in \mathcal{Z}_T$, where the total variation distance $\|\cdot\|_{\text{TV}}$ denotes the supremum over all subsets of $\mathcal{Z}_T$ (the geometric rate of convergence does not depend on the initial starting point $\boldsymbol{z}_0$). Uniform ergodicity is equivalent to the so-called Doeblin condition, which is that there exist a probability measure $\rho$ on $\mathcal{Z}_T$, an integer $k$, and a constant $c > 0$ such that $g_T^k(\boldsymbol{z}, \boldsymbol{z}') \geq c\rho(\boldsymbol{z}')$ for all $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}_T$. See Theorem 3 of Athreya et al. (1996).

**Theorem 1** *For each $T$, $g_T(\boldsymbol{z}, \boldsymbol{z}')$ satisfies the Doeblin condition with $k = 1$:*

$$g_T(\boldsymbol{z}, \boldsymbol{z}') \geq c_T \upsilon(\boldsymbol{z}') \qquad \text{for any } \boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}_T,$$

*where $\upsilon$ is the uniform distribution on $\mathcal{Z}_T$,*

$$c_T = \left(\frac{\eta}{N - 1 + V\eta}\right)^N \left[\prod_{d=1}^{D} \left(\frac{\sqrt{T}\alpha}{n_d - 1 + T\alpha}\right)^{n_d}\right], \tag{2.9}$$

*and recall that $N = \sum_{d=1}^{D} n_d$ denotes the total number of words in the corpus. The CGS is uniformly ergodic, with $\|g_T^m(\boldsymbol{z}_0, \cdot) - \nu_{\boldsymbol{z} \mid \boldsymbol{w}}(\cdot)\|_{TV} \leq (1 - c_T)^m$.*

The value of $c_T$ given by (2.9) is astronomically small for any reasonable values of the document sizes $n_1, \dots, n_D$, and the resulting bound on the total variation distance is useless from a practical point of view. Undoubtedly better constants can be found, as in our relatively short proof we have not tried to obtain the sharpest possible bound. Thus, in its present form Theorem 1 is primarily of theoretical interest. Empirical studies strongly suggest that the actual rate of convergence is very fast—for example, the popular MALLET package (McCallum, 2002) uses a default value of $m = 40$ for situations in which the CGS is used within a loop, with apparently good results.

# 3 A Bayesian Approach for Specifying the Number of Topics

In a Bayesian approach to making inference about $T$, we put a prior distribution on $T$ at the top of the four-line hierarchy that defines the original LDA model. The possible values for $T$ will be the positive integers, and we will put a proper prior on this set. Specification of a prior for $T$ at the top of the hierarchy induces a prior distribution, $\nu_\vartheta$, on the augmented parameter $\vartheta = (T, \boldsymbol{\psi})$: we have $\nu_\vartheta(\vartheta) = \nu_\vartheta(T, \boldsymbol{\psi}) = \nu_{\boldsymbol{\psi} \mid T}(\boldsymbol{\psi} \mid T)\, \nu_T(T) = \nu_{\boldsymbol{\psi}}^{(T)}(\boldsymbol{\psi}) \nu_T(T)$ for all $T \in \mathcal{T}$. We let $\ell_{\boldsymbol{w}}(\vartheta)$ denote the likelihood function of $\vartheta = (T, \boldsymbol{\psi})$. Clearly $\ell_{\boldsymbol{w}}(\vartheta) = \ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\psi})$ where, recall that $\ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\psi})$ is the likelihood function (2.3) under the LDA model indexed by $T$. The posterior distribution $\vartheta = (T, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$ given $\boldsymbol{w}$ is $\nu_{\vartheta \mid \boldsymbol{w}}(\vartheta) \propto \nu_\vartheta(\vartheta) \ell_{\boldsymbol{w}}(\vartheta) \propto \nu_{\boldsymbol{\psi}}^{(T)}(\boldsymbol{\psi}) \nu_T(T) \ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\psi})$; more specifically,

$$\nu_{\vartheta \mid \boldsymbol{w}}(\vartheta) \propto \left[ \prod_{d=1}^{D} \left( \frac{\Gamma\big(T\alpha\big)}{\Gamma(\alpha)^T} \prod_{t=1}^{T} \theta_{dt}^{n_{dt}+\alpha-1} \right) \right] \left[ \prod_{t=1}^{T} \left( \frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \prod_{v=1}^{V} \beta_{tv}^{m_{\cdot tv}+\eta-1} \right) \right] \nu_T(T).$$

Integrating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ out from the equation above, as we did to obtain (2.7), we get a closed-form expression for $\nu_{T, \boldsymbol{z} \mid \boldsymbol{w}}$ up to a normalizing constant; that is for each $T \in \mathcal{T}$,

$$\nu_{T, \boldsymbol{z} \mid \boldsymbol{w}}(T, \boldsymbol{z}) \propto \left[ \prod_{d=1}^{D} \left( \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \frac{\prod_{t=1}^{T} \Gamma(n_{dt}+\alpha)}{\Gamma\big(n_d + T\alpha\big)} \right) \right] \left[ \prod_{t=1}^{T} \left( \frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \frac{\prod_{v=1}^{V} \Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t \cdot} + V\eta)} \right) \right] \nu_T(T). \tag{3.1}$$

Unfortunately, even if the normalizing constant in (3.1) was available, we would not be able to obtain a closed-form expression for $\nu_{T \mid \boldsymbol{w}}(T)$, because summing out the latent variables $z_{11}, \ldots, z_{1n_1}, \ldots, z_{D1}, \ldots, z_{Dn_D}$ is computationally infeasible. Our objective is to design an effective PMMH algorithm for estimating the posterior distribution of the number of topics, $T$, in the LDA model, and to analyze its ergodicity properties.

# 4 A Pseudo-Marginal Metropolis-Hastings Algorithm for Estimating the Number of Topics

This section consists of three parts. In Section 4.1 we describe, in general terms, the PMMH algorithm. In Section 4.2 we design a variant of the algorithm particular to the problem of estimating the marginal posterior distribution of the number of topics in the LDA model. In Section 4.3 we show theoretically that the samples produced by our algorithm have a distribution that is close to the target distribution $\nu_{T \mid \boldsymbol{w}}$.

## 4.1 Pseudo-Marginal Metropolis-Hastings Algorithms

Let $X$ be a random variable defined on a (fixed-dimensional) measurable space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, let $\Pi_X$ denote the distribution of $X$, and let $\pi_X$ be the density of $\Pi_X$ with respect to a $\sigma$-finite measure $\mu_X$. We are interested in simulating from $\Pi_X$. In the ideal situation in which $\pi_X$ is known except for a normalizing constant, the Metropolis-Hastings algorithm can be used to simulate samples whose distribution is approximately $\Pi_X$. Let $Q_X(\cdot, \cdot)$ be a Markov transition function defined on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, and suppose that the distributions $Q_X(x, \cdot)$ have densities $q_X(x, \cdot)$ with respect to $\mu_X$. The Metropolis-Hastings algorithm is described as follows.

1. Let $x$ be the current state. Generate a proposal $x' \sim Q_X(x, \cdot)$.
2. Compute the acceptance ratio
$$r(x, x') = \frac{\pi_X(x')q_X(x', x)}{\pi_X(x)q_X(x, x')}. \tag{4.1}$$
3. Accept $x'$ with probability $\alpha(x, x') = \min\{r(x, x'), 1\}$; otherwise stay at $x$.

Unfortunately, when $\pi_X$ is analytically intractable we cannot evaluate the acceptance ratio (4.1), which makes Step 2 in the algorithm above infeasible. So we call this algorithm the ideal Metropolis-Hastings algorithm, and we use $P_{\text{MH}}$ to denote its one-step Markov transition function.

We can deal with this problem using an idea related to Data Augmentation. Assume that there exists a measurable space $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$, and a distribution $\Pi_{X,Z}$ on $\mathcal{X} \times \mathcal{Z}$ which has density $\pi_{X,Z}$ with respect to the product measure $\mu_X \times \mu_Z$, where $\mu_Z$ is a $\sigma$-finite measure on $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$. Assume further that (i) the $X$-marginal of $\pi_{X,Z}$ is $\pi_X$, i.e.

$$\int_{\mathcal{Z}} \pi_{X,Z}(x, z)\, d\mu_Z(z) = \pi_X(x), \tag{4.2}$$

and (ii) it is feasible to simulate from the conditional densities $\pi_{Z \mid X=x}$, $x \in \mathcal{X}$. If $(X, Z) \sim \Pi_{X,Z}$, then the random variable $Z$ is called an auxiliary variable. The development of a PMMH algorithm is based on such a variable, and a rough and brief preliminary description of the method is as follows. Within each iteration, let $x'$ be as in Step 2 of the ideal Metropolis-Hastings algorithm, and let $m$ be some large integer (the choice of $m$ is discussed later). We generate $(Z'_1, \ldots, Z'_m)$ from some distribution $Q_{\boldsymbol{Z}}^{(m,x')}$ on $\mathcal{Z}^m$, which depends on $m$ and $x'$; we let $\tilde{\pi}_X^{(m)}(x')$ be a certain linear combination of $\pi_{X,Z}(x', Z'_1), \ldots, \pi_{X,Z}(x', Z'_m)$ with random coefficients (this is similar to what is done in classical importance sampling); we estimate $\pi_X(x')$ by $\tilde{\pi}_X^{(m)}(x')$; and we estimate

$\pi_X(x)$ by an estimate $\tilde{\pi}_X^{(m)}(x)$, constructed in an analogous manner, *and recycled from the previous iteration*—this is a very important point, on which we comment further in Remark 1 below. The PMMH algorithm is then the same as the infeasible Metropolis-Hastings algorithm described earlier, except that we replace $\pi_X(x')$ and $\pi_X(x)$ by $\tilde{\pi}_X^{(m)}(x')$ and $\tilde{\pi}_X^{(m)}(x)$, respectively, and the algorithm runs on $\mathcal{X} \times \mathcal{Z}^m$ instead of $\mathcal{X}$.

To provide the details and explain why this works, we need to set up some notation. For any $m \in \mathbb{N}$, let $\{Q_{\boldsymbol{Z}}^{(m,x)}, x \in \mathcal{X}\}$ be a family of distributions on $\mathcal{Z}^m$ which are dominated by $\mu_Z^m$, the $m$-fold product of $\mu_Z$, and let $q_{\boldsymbol{Z}}^{(m,x)}$ be the density of $Q_{\boldsymbol{Z}}^{(m,x)}$ with respect to $\mu_Z^m$. If $\boldsymbol{z} = (z_1, \dots, z_m) \in \mathcal{Z}^m$, we let $\boldsymbol{z}_{(-l)} = (z_0, \dots, z_{l-1}) \in \mathcal{Z}^l$ for $l = 1, \dots, m$. Also, if $\boldsymbol{Z} \sim Q_{\boldsymbol{Z}}^{(m,x)}$, we denote the conditional distribution of $Z_l$ given $\boldsymbol{Z}_{(-l)} = (z_0, Z_1, \dots, Z_{l-1})$ by $Q_{Z_l \mid \boldsymbol{Z}_{(-l)}}^{(m,x)}$, and let $q_{Z_l \mid \boldsymbol{Z}_{(-l)}}^{(m,x)}$ denote its density with respect to $\mu_Z$.

We will refer to Conditions A1, A2, and A3 below. The first is needed to properly define $\tilde{\pi}_X^{(m)}(x)$. The second and third are standard requirements for proving ergodicity of Markov chains, and we will need them later when we establish the ergodicity properties of the PMMH algorithm.

- A1 For all $m \in \mathbb{N}$, $x \in \mathcal{X}$, $l = 1, \dots, m$, the conditional distribution of $Z$ given $X = x$ is absolutely continuous with respect to the conditional distribution of $Z_l$ given $\boldsymbol{Z}_{(-l)} = \boldsymbol{z}_{(-l)}$, for any $\boldsymbol{z}_{(-l)} \in \mathcal{Z}^l$, i.e. $\Pi_{Z \mid X}(\cdot \mid x) \ll Q_{Z_l \mid \boldsymbol{Z}_{(-l)}}^{(m,x)}\big(\cdot \bigm| \boldsymbol{z}_{(-l)}\big)$.
- A2 The ideal algorithm is $\Pi_X$-irreducible; that is, for any $x \in \mathcal{X}$ and $A \in \mathcal{B}_{\mathcal{X}}$ such that $\Pi_X(A) > 0$, there exists an integer $n = n(x, A)$ such that $P_{\text{MH}}^n(x, A) > 0$.
- A3 The ideal algorithm is aperiodic; that is, there exists a probability measure $\rho$ on $\mathcal{X}$ such that

$$\text{g.c.d.}\{n : \text{there is an } \epsilon_n > 0 \text{ such that } P_{\text{MH}}^n(x, \cdot) \geq \epsilon_n \rho(\cdot) \text{ for each } x \in \mathcal{X}\} = 1,$$

  where g.c.d. denotes greatest common divisor.

Under A1, for any $m \in \mathbb{N}$, and $\boldsymbol{Z} = (Z_1, \dots, Z_m) \in \mathcal{Z}^m$, we can define a "pseudo-marginal density of $X$" by

$$\tilde{\pi}_X^{(m)}(x) = \frac{1}{m} \sum_{l=1}^m \frac{\pi_{X,Z}(x, Z_l)}{q_{Z_l \mid \boldsymbol{Z}_{(-l)}}^{(m,x)}\big(Z_l \bigm| \boldsymbol{Z}_{(-l)}\big)} \qquad \text{for all } x \in \mathcal{X},$$

which depends on $\boldsymbol{Z}$, although this dependence is suppressed in the notation. A key point regarding $\tilde{\pi}_X^{(m)}(x)$ is that if $\boldsymbol{Z} \sim Q_{\boldsymbol{Z}}^{(m,x)}$, then $E\big(\tilde{\pi}_X^{(m)}(x)\big) = \pi_X(x)$ for all $x \in \mathcal{X}$. To see this, we imagine that $\boldsymbol{Z}$ is generated component-wise: given an initial point $z_0$, $Z_1$ is generated according to $Q^{(m,x)}_{Z_1 \mid \boldsymbol{Z}_{(-1)}}\big(\cdot \mid \boldsymbol{Z}_{(-1)}\big)$, etc. Then, for $l = 1, \ldots, m$, we have

$$E\Bigg(\frac{\pi_{X,Z}(x, Z_l)}{q^{(m,x)}_{Z_l \mid \boldsymbol{Z}_{(-l)}}\big(Z_l \mid \boldsymbol{Z}_{(-l)}\big)}\Bigg) = \int \frac{\pi_{X,Z}(x, z_l)}{q^{(m,x)}_{Z_l \mid \boldsymbol{Z}_{(-l)}}\big(z_l \mid \boldsymbol{z}_{(-l)}\big)} q^{(m,x)}_{Z_l \mid \boldsymbol{Z}_{(-l)}}\big(z_l \mid \boldsymbol{z}_{(-l)}\big)\, d\mu_Z(z_l) = \pi_X(x), \quad (4.3)$$

where the last equality is from (4.2). However, without conditions on $Q^{(m,x)}_{\boldsymbol{Z}}$, there is no reason to expect that $\tilde{\pi}^{(m)}_X(x) \xrightarrow{\text{a.s.}} \pi_X(x)$ as $m \to \infty$. Our development in Sections 4.2 and 4.3, in the context of the LDA model, shows that for a certain choice of the distributions $Q^{(m,x)}_{\boldsymbol{Z}}$, we do have $\tilde{\pi}^{(m)}_X(x) \xrightarrow{\text{a.s.}} \pi_X(x)$ as $m \to \infty$ for every $x \in \mathcal{X}$.

With the notation above, the PMMH algorithm can now be described as follows:

1. Given the current $x$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ (and hence the current pseudo-marginal density $\tilde{\pi}^{(m)}_X(x)$), generate a proposal $x' \sim Q_X(x, \cdot)$.
2. Using the proposal $x'$, generate $\boldsymbol{Z}' = (Z'_1, \ldots, Z'_m) \sim Q^{(m,x')}_{\boldsymbol{Z}}$.
3. Compute the pseudo-marginal density at $x'$, which is given by

$$\tilde{\pi}^{(m)}_X(x') = \frac{1}{m} \sum_{l=1}^{m} \frac{\pi_{X,Z}(x', Z'_l)}{q^{(m,x')}_{Z_l \mid \boldsymbol{Z}_{(-l)}}\big(Z'_l \mid \boldsymbol{Z}'_{(-l)}\big)}. \quad (4.4)$$

4. Compute the acceptance ratio

$$\tilde{r}^{(m)}(x, x') = \frac{\tilde{\pi}^{(m)}_X(x') q_X(x', x)}{\tilde{\pi}^{(m)}_X(x) q_X(x, x')}. \quad (4.5)$$

5. Accept $x'$ and $\boldsymbol{Z}'$ with probability $\min\{\tilde{r}^{(m)}(x, x'), 1\}$, and with the remaining probability to stay at $x$ and $\boldsymbol{Z}$.

**Remark 1** The purpose of the PMMH algorithm is to deal with the problem that the quantities $\pi_X(x)$ and $\pi_X(x')$, needed in the acceptance ratio of the ideal Metropolis-Hastings algorithm, are not available analytically; the PMMH algorithm produces estimates of these quantities. The PMMH algorithm is complicated, and it is perhaps natural to ask why not proceed via what is called the Monte Carlo within Metropolis-Hastings algorithm (MCWMH), which is very simple and is described as follows. Having proposed $x' \sim Q_X(x, \cdot)$ in the ideal Metropolis-Hastings algorithm, we generate $\boldsymbol{Z} \sim Q^{(m,x)}_{\boldsymbol{Z}}$ and $\boldsymbol{Z}' \sim Q^{(m,x')}_{\boldsymbol{Z}}$, we form $\tilde{\pi}_X(x)$ and $\tilde{\pi}_X(x')$, calculate the acceptance ratio $\tilde{r}^{(m)}(x, x')$ given by (4.5), and we accept or reject $x'$ based on $\tilde{r}^{(m)}(x, x')$. We discard $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ and the algorithm runs on $\mathcal{X}$. The reason we do not use the MCWMH algorithm is that this algorithm has no theoretical validity. It is not a Metropolis-Hastings algorithm in any sense (there are no detailed balance conditions that are satisfied—see Section 6 of Andrieu and Roberts (2009)), and indeed it is not clear that it has an invariant distribution. In contrast, as we shall see in the next paragraph, the PMMH algorithm always has an invariant distribution, and as we shall see in Section 4.3, its ergodicity properties can be rigorously established.

In order to further analyze the properties of the PMMH algorithm, we first need to set up some notation. For each $m, n \in \mathbb{N}$, we let $P^{m,n}$ denote the $n$-step Markov transition function for the PMMH algorithm in which the number of inner loops is $m$. We also define a joint density $\tilde{\pi}^{(m)}_{X,\boldsymbol{Z}}$ on $\mathcal{X} \times \mathcal{Z}^m$ by

$$\tilde{\pi}^{(m)}_{X,\boldsymbol{Z}}(x, \boldsymbol{z}) = \tilde{\pi}^{(m)}_{X}(x) q^{(m,x)}_{\boldsymbol{Z}}(\boldsymbol{z}), \tag{4.6}$$

and a Markov transition function $q^{(m)}_{X,\boldsymbol{Z}}(\cdot;\cdot)$ by

$$q^{(m)}_{X,\boldsymbol{Z}}(x, \boldsymbol{z}; x', \boldsymbol{z}') = q_X(x, x') q^{(m,x')}_{\boldsymbol{Z}}(\boldsymbol{z}') \qquad \text{for all } (x, \boldsymbol{z}), (x', \boldsymbol{z}') \in \mathcal{X} \times \mathcal{Z}^m. \tag{4.7}$$

Then for the Markov chain generated according to $P^{m,1}$: (i) the invariant density is $\tilde{\pi}^{(m)}_{X,\boldsymbol{Z}}$ and (ii) the Markov transition function for generating a proposal from $\mathcal{X} \times \mathcal{Z}^m$ is $q^{(m)}_{X,\boldsymbol{Z}}$. To see points (i) and (ii), we first recall the following well-known fact regarding the Metropolis-Hastings algorithm. Suppose that $\pi$ is a density on some space $\mathcal{Y}$, $p(\cdot,\cdot)$ is a Markov transition function on $\mathcal{Y}$, and the function $R$ is defined by

$$R(y, y') = \frac{\pi(y')p(y', y)}{\pi(y)p(y, y')}.$$

If $p(\cdot,\cdot)$ is used to generate proposals from $y$ to $y'$, and these are accepted with probability $\min\{R(y, y'), 1\}$, then the resulting chain has $\pi$ as invariant density. With this in mind, we rewrite the acceptance ratio (4.5) in the PMMH algorithm as

$$\begin{aligned} \tilde{r}^{(m)}(x, x') = \frac{\tilde{\pi}^{(m)}_X(x') q_X(x', x)}{\tilde{\pi}^{(m)}_X(x) q_X(x, x')} &= \frac{\left[\tilde{\pi}^{(m)}_X(x') q^{(m,x')}_{\boldsymbol{Z}}(\boldsymbol{z}')\right] \left[q_X(x', x) q^{(m,x)}_{\boldsymbol{Z}}(\boldsymbol{z})\right]}{\left[\tilde{\pi}^{(m)}_X(x) q^{(m,x)}_{\boldsymbol{Z}}(\boldsymbol{z})\right] \left[q_X(x, x') q^{(m,x')}_{\boldsymbol{Z}}(\boldsymbol{z}')\right]} \\ &= \frac{\tilde{\pi}^{(m)}_{X,\boldsymbol{Z}}(x', \boldsymbol{z}') q^{(m)}_{X,\boldsymbol{Z}}(x', \boldsymbol{z}'; x, \boldsymbol{z})}{\tilde{\pi}^{(m)}_{X,\boldsymbol{Z}}(x, \boldsymbol{z}) q^{(m)}_{X,\boldsymbol{Z}}(x, \boldsymbol{z}; x', \boldsymbol{z}')}, \end{aligned}$$

where the last equality is from (4.6) and (4.7). Points (i) and (ii) now follow.

## 4.2 A Pseudo-Marginal Metropolis-Hastings Algorithm for Simulating $T$

Consider now the LDA model. We will first describe an implementation of the PMMH algorithm particular to the setup of this model, where $T$ plays the role of $X$, and the vector of latent topic indicators $\boldsymbol{z} = (z_{11}, \dots, z_{1n_1}, \dots, z_{D1}, \dots, z_{Dn_D})$ plays the role of $Z$. Then we will discuss important differences between our method and the original method introduced in Beaumont (2003), and explain why our method is very efficient.

Henceforth, in order to avoid confusion, we will reserve the symbol "$T$" for the random variable denoting the number of topics, and we will use "$t$" to denote a fixed point in $\mathcal{T}$. Let $t \in \mathcal{T}$, and recall that $g_t(\cdot, \cdot)$ denotes the Markov transition function of the CGS of Griffiths and Steyvers (2004), which runs over $\mathcal{Z}_t$. Let $\boldsymbol{z}_0$ be the starting point. Let $\boldsymbol{\zeta} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_m)$ be the initial segment of the CGS. We will take $q_{\boldsymbol{\zeta}}^{(m,t)}$ to be the distribution of $\boldsymbol{\zeta}$. Thus,

$$q_{\boldsymbol{\zeta}}^{(m,t)}(\boldsymbol{\zeta}) = g_t(\boldsymbol{z}_0, \boldsymbol{z}_1) \times g_t(\boldsymbol{z}_1, \boldsymbol{z}_2) \times \cdots \times g_t(\boldsymbol{z}_{m-1}, \boldsymbol{z}_m), \tag{4.8}$$

and for any $l > 0$ the conditional distribution of $\boldsymbol{z}_l$ given $\boldsymbol{\zeta}_{(-l)}$ is $q_{\boldsymbol{z}_l \mid \boldsymbol{\zeta}_{(-l)}}^{(m,t)}\big(\cdot \bigm| \boldsymbol{\zeta}_{(-l)}\big) = g_t(\boldsymbol{z}_{l-1}, \cdot)$. It is obvious that the mechanism above for generating the auxiliary variable $\boldsymbol{\zeta}$ guarantees Condition A1 so, hereafter, we will not mention this condition except when we need to be explicit.

Given $t$ and $\boldsymbol{\zeta}$, we define the pseudo-marginal distribution of $T$ evaluated at $t$ by

$$\tilde{\nu}_{T \mid \boldsymbol{w}}^{(m)}(t) = \frac{1}{m} \sum_{l=1}^{m} \frac{\nu_{T, \boldsymbol{z} \mid \boldsymbol{w}}(t, \boldsymbol{z}_l)}{g_t(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)}. \tag{4.9}$$

This quantity depends implicitly on the vector of auxiliary variables $\boldsymbol{\zeta}$. The numerator on the right side of (4.9) is known up to a normalizing constant; see the explicit expression (3.1). It is important to note that this normalizing constant does not depend on either $t$ or $\boldsymbol{\zeta}$. Having specified a Markov transition function $q_T(\cdot, \cdot)$ on $\mathcal{T}$, our pseudo-marginal Metropolis-Hastings algorithm proceeds as follows:

1. Given the current $t$ and $\boldsymbol{\zeta} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_m)$ (and hence the current $\tilde{\nu}_{T \mid \boldsymbol{w}}^{(m)}(t)$), propose $t' \sim q_T(t, \cdot)$.
2. Given $t'$, generate $\boldsymbol{\zeta}' = (\boldsymbol{z}'_1, \dots, \boldsymbol{z}'_m)$ according to $q_{\boldsymbol{\zeta}}^{(m,t')}$.
3. Compute $\tilde{\nu}_{T \mid \boldsymbol{w}}^{(m)}(t')$, which is given by

$$\tilde{\nu}_{T \mid \boldsymbol{w}}^{(m)}(t') = \frac{1}{m} \sum_{l=1}^{m} \frac{\nu_{T, \boldsymbol{z} \mid \boldsymbol{w}}(t', \boldsymbol{z}'_l)}{g_{t'}(\boldsymbol{z}'_{l-1}, \boldsymbol{z}'_l)}.$$

4. Compute the acceptance ratio

$$\tilde{r}^{(m)}(t, t') = \frac{\tilde{\nu}_{T \mid \boldsymbol{w}}^{(m)}(t') q_T(t', t)}{\tilde{\nu}_{T \mid \boldsymbol{w}}^{(m)}(t) q_T(t, t')}.$$

5. Accept $t'$ and $\boldsymbol{\zeta}'$ with probability $\min\{\tilde{r}^{(m)}(t,t'),1\}$, and with the remaining probability stay at $t$ and $\boldsymbol{\zeta}$.

While Step 3 cannot really be carried out because, as noted earlier, we don't know the normalizing constant for $\nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}$, Step 4 is feasible because the constant cancels in the calculation of $\tilde{r}^{(m)}(t,t')$.

We now return to the description of a generic PMMH algorithm given in Section 4.1, and consider the choice of the function $Q^{(m,x)}_{\boldsymbol{Z}}$ that is used in the algorithm. Ideally, we would want $Q^{(m,x)}_{\boldsymbol{Z}}$ to be such that $q^{(m,x)}_{Z_l\,|\,\boldsymbol{Z}_{(-l)}}$ is equal to $\pi_{Z|X}(\cdot\,|\,x)$, for then in (4.4), on the right side each summand would be equal to $\pi_X(x)$. In the LDA model this choice is infeasible, of course, because $\pi_{Z|X}$ is then the posterior distribution of $\boldsymbol{z}$ in the model indexed by $T$, and it is not possible to simulate from the posterior distribution of $\boldsymbol{z}$ given the words.

Returning to the LDA setup, we note that our algorithm has the following desirable feature. By Theorem 1, the Markov chain generated by the CGS is uniformly ergodic. Hence as $m\to\infty$, $g^m_T(\boldsymbol{z}_0,\cdot)$ converges uniformly to $\nu_{\boldsymbol{z}\,|\,T,\boldsymbol{w}}$, so $\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}$ converges rapidly to $\nu_{T\,|\,\boldsymbol{w}}$. In sharp contrast, in the algorithm introduced in Beaumont (2003), the distribution $Q^{(m,x)}_{\boldsymbol{Z}}$ is taken to be an importance sampling distribution, for example a product measure on $\mathcal{Z}^m$. For the LDA model, in which the dimension of $\mathcal{Z}_T$ is very high, the estimate $\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}$ resulting from such a choice would have very high variance, rendering the PMMH algorithm impractical. We return to this point in Section 5.2. In the Appendix we discuss the differences between our theoretical results and those of Andrieu and Roberts (2009).

We also note that our algorithm is fairly automatic: the only choices involved are the prior on $T$, the Markov transition function $q_T(\cdot,\cdot)$ on $\mathcal{T}$ and the choice of the CGS as the Markov chain on $\boldsymbol{z}$. In contrast, RJMCMC involves a number of parameters which are difficult to tune properly when dealing with a problem of the scale of the LDA model.

Our PMMH algorithm can be used not only to estimate $\nu_{T\,|\,\boldsymbol{w}}$, but also $\nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}$, the joint posterior distribution of $T$ and $\boldsymbol{z}$ and, in fact, it can be used to estimate the posterior distribution of the entire set of latent variables in the LDA model. We now explain this, and we begin with $\nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}$. Suppose that $m$ and $n$ are both large. Let $(T_n,\boldsymbol{\zeta}_n)$ be the output of the $n^{\text{th}}$ cycle in the outer loop, where $\boldsymbol{\zeta}_n=(\zeta_{n,1},\dots,\zeta_{n,m})$. A consequence of Theorem 3 below is that the distribution of $T_n$ is nearly equal to $\nu_{T\,|\,\boldsymbol{w}}$ and, because the CGS mixes rapidly, the distribution of $(T_n,\zeta_{n,m})$ is nearly equal to $\nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}$. Suppose now that in addition, $(\boldsymbol{\beta}^{(n)},\boldsymbol{\theta}^{(n)})$ are generated according to (2.5) (note that the $n_{ds}$'s and $m_{\cdot sv}$'s in (2.5) depend on the latent topic indicator variables, although the notation suppresses this dependence). Then the distribution of $(T_n,\zeta_{n,m},\boldsymbol{\beta}^{(n)},\boldsymbol{\theta}^{(n)})$ is nearly $\nu_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}\,|\,\boldsymbol{w}}$.

The estimate of the posterior distribution $\nu_{T \mid \boldsymbol{w}}$ produced by our PMMH algorithm is subject to two kinds of error. One is from the inner loop, where the pseudo-marginal distribution $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}$ replaces $\nu_{T \mid \boldsymbol{w}}$ in the calculation of the acceptance ratio, and the other is from the outer loop. The first error is controlled by the value of $m$, and we note that this error is small because the CGS mixes very fast, a fact which can be seen both theoretically (Theorem 1) and empirically. The second error is pervasive in all Monte Carlo algorithms. In Section 4.3 we show that as $m, n \to \infty$, the estimate of $\nu_{T \mid \boldsymbol{w}}$ produced by our algorithm converges to $\nu_{T \mid \boldsymbol{w}}$.

## 4.3 Ergodicity of Our Pseudo-Marginal Metropolis-Hastings Algorithm

This section deals with the convergence properties of the PMMH algorithm developed in Section 4.2. The algorithm proceeds as follows. Having fixed $m$, we generate $T_0$ from the prior $\nu_T$ and then generate $\boldsymbol{\zeta}_0 \sim q_{\boldsymbol{\zeta}}^{(m,T_0)}$, to initialize the algorithm. We then generate $(T_1, \boldsymbol{\zeta}_1) \sim P^{m,1}(T_0, \boldsymbol{\zeta}_0; \cdot, \cdot)$, then $(T_2, \boldsymbol{\zeta}_2) \sim P^{m,1}(T_1, \boldsymbol{\zeta}_1; \cdot, \cdot)$, and continue in this manner. This produces a Markov chain $\{T_n, \boldsymbol{\zeta}_n\}_{n=0}^{\infty}$. Whether or not this chain is useful depends on the answers to the following two questions:

I Does the pseudo-marginal distribution $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}$ converge to the true marginal distribution $\nu_{T \mid \boldsymbol{w}}$ as $m \to \infty$?

II Does the marginal distribution of $T_n$ converge to the marginal posterior distribution $\nu_{T \mid \boldsymbol{w}}$ for all starting values $T_0$? More generally, can we say that for any starting values $T_0$ and $\boldsymbol{\zeta}_0$, the distribution of $(T_n, \zeta_{n,m})$ converges to $\nu_{T,\boldsymbol{z} \mid \boldsymbol{w}}$ as $m, n \to \infty$?

We first consider Question I. Theorem 2 which follows establishes a Strong Law of Large Numbers (SLLN) for $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t)$ for each $t \in \mathcal{T}$, and hence provides justification for substituting $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}$ in place of $\nu_{T \mid \boldsymbol{w}}$ in the acceptance ratio (4.1) of the ideal Metropolis-Hastings algorithm.

**Theorem 2** *Let $t \in \mathcal{T}$, and let $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots$ be a Markov chain generated according to the CGS for the LDA model indexed by $t$ (see (4.8)). Then*

$$\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t) = \frac{1}{m} \sum_{l=1}^{m} \frac{\nu_{T,\boldsymbol{z} \mid \boldsymbol{w}}(t, \boldsymbol{z}_l)}{g_t(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)} \xrightarrow{\text{a.s.}} \nu_{T \mid \boldsymbol{w}}(t) \qquad \text{as } m \to \infty. \tag{4.10}$$

Theorem 2 states that $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t) \xrightarrow{\text{a.s.}} \nu_{T \mid \boldsymbol{w}}(t)$ for each $t \in \mathcal{T}$, and it is therefore natural to ask why then do we need a PMMH algorithm to estimate $\nu_{T \mid \boldsymbol{w}}$. The answer is that the numerator of each summand in (4.10) is known only up to a multiplicative constant, say $c_{\boldsymbol{w}}$. This is a problem when trying to use $\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)$ to estimate $\nu_{T\,|\,\boldsymbol{w}}(t)$, but is not a problem when using $\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)$ in the acceptance ratio (4.5).

A referee pointed out that when $\mathcal{T}$ is finite we have

$$\frac{c_{\boldsymbol{w}}\tilde{\nu}^{(m)}_{T\,|\,w}(t)}{\sum_{t'\in\mathcal{T}} c_{\boldsymbol{w}}\tilde{\nu}^{(m)}_{T\,|\,w}(t')} = \frac{\tilde{\nu}^{(m)}_{T\,|\,w}(t)}{\sum_{t'\in\mathcal{T}} \tilde{\nu}^{(m)}_{T\,|\,w}(t')} \xrightarrow{\text{a.s.}} \frac{\nu_{T\,|\,w}(t)}{\sum_{t'\in\mathcal{T}} \tilde{\nu}_{T\,|\,w}(t')} = \nu_{T\,|\,w}(t), \tag{4.11}$$

where the left side is computable, and the convergence is by Theorem 2. Therefore the left side of (4.11) may be used to estimate $\nu_{T\,|\,\boldsymbol{w}}$, and so bypass the PMMH algorithm entirely. Our empirically experience with this estimator is that it has large variance and does not provide a useful estimate of $\nu_{T\,|\,\boldsymbol{w}}$. At the same time, curiously, it is accurate enough to enable a useful PMMH algorithm.

**Proof of Theorem 2** An outline of the proof is as follows. Let $\boldsymbol{z}_0 \in \mathcal{Z}_t$ be the initial sample. The sequence $\boldsymbol{z}_0, \boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$ induces the sequence $(\boldsymbol{z}_0, \boldsymbol{z}_1), (\boldsymbol{z}_1, \boldsymbol{z}_2), (\boldsymbol{z}_2, \boldsymbol{z}_3), \ldots$, and it is easy to see that this is a Markov chain on the state space $\mathcal{Z}_t \times \mathcal{Z}_t$. Let $h\colon \mathcal{Z}_t \times \mathcal{Z}_t \to \mathbb{R}$ be defined by $h(\boldsymbol{x}, \boldsymbol{y}) = \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t, \boldsymbol{y})/g_t(\boldsymbol{x}, \boldsymbol{y})$. Then the average in (4.10) is equal to $(1/m)\sum_{l=1}^m h(\boldsymbol{u}_l)$, where $\boldsymbol{u}_l = (\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)$. To prove Theorem 2, we will apply an ergodic theorem for Markov chains, i.e. apply a result that gives a SLLN for Markov chains. The literature has several that we could use. We will use Theorem 2 of Athreya et al. (1996) because that theorem is particularly amenable to our setup. The theorem states the following, with notation adapted to our context. Suppose that $\boldsymbol{u}_0, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ is a Markov chain on $\mathcal{Z}_t \times \mathcal{Z}_t$, and let $K(\boldsymbol{u}, \boldsymbol{u}')$ be the Markov transition function for the chain. If

$$\pi \text{ is an invariant distribution for } K, \tag{4.12}$$

and there exist a probability mass function $\rho$, a constant $c > 0$, and a positive integer $r$ such that

$$K^r(\boldsymbol{u}, \boldsymbol{u}') \geq c\rho(\boldsymbol{u}') \qquad \text{for all } \boldsymbol{u}, \boldsymbol{u}' \in \mathcal{Z}_t \times \mathcal{Z}_t, \tag{4.13}$$

then

$$\frac{1}{m}\sum_{l=1}^m h(\boldsymbol{u}_l) \xrightarrow{\text{a.s.}} \int h(\boldsymbol{u})\, d\pi(\boldsymbol{u}) \qquad \text{for } [\pi]\text{-almost every starting point } \boldsymbol{u}_0.$$

(We have taken the set $A$ in the statement of Theorem 2 of Athreya et al. (1996) to be the entire state space $\mathcal{Z}_t \times \mathcal{Z}_t$.) Our plan is to identify the Markov transition function $K$ and the invariant distribution $\pi$ for our chain $\boldsymbol{u}_0, \boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$, and to find a $(\rho, c, r)$ triple that works. It will turn out that $\pi$ gives positive mass to each point in the state space, so that the "almost every" proviso does not impose any restriction. We now proceed with the details.

Consider now the Markov chain $(\boldsymbol{z}_0, \boldsymbol{z}_1), (\boldsymbol{z}_1, \boldsymbol{z}_2), (\boldsymbol{z}_2, \boldsymbol{z}_3), \ldots$, let $K$ be the Markov transition function, and let us ask what is the formula for $K\big((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')\big)$. Since $\boldsymbol{x}'$ must equal $\boldsymbol{y}$, and $\boldsymbol{y}'$ is generated according to $g_t(\boldsymbol{x}', \cdot)$, it is clear that

$$K\big((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')\big) = \delta_{\boldsymbol{y}}(\boldsymbol{x}') g_t(\boldsymbol{x}', \boldsymbol{y}') \qquad \text{for all } (\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{Z}_t \times \mathcal{Z}_t. \tag{4.14}$$

Here, $\delta_{\boldsymbol{y}}$ is the Dirac delta at $\boldsymbol{y}$: $\delta_{\boldsymbol{y}}(\boldsymbol{x}') = 1$ if $\boldsymbol{x}' = \boldsymbol{y}$ and is zero otherwise. Define $\pi\colon \mathcal{Z}_t \times \mathcal{Z}_t \to \mathbb{R}$ by

$$\pi(\boldsymbol{x}, \boldsymbol{y}) = \nu^{(t)}_{\boldsymbol{z} \mid \boldsymbol{w}}(\boldsymbol{x}) g_t(\boldsymbol{x}, \boldsymbol{y}) \qquad \text{for all } (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}_t \times \mathcal{Z}_t. \tag{4.15}$$

It is easy to check that $\sum_{\boldsymbol{x} \in \mathcal{Z}_t} \sum_{\boldsymbol{y} \in \mathcal{Z}_t} \nu^{(t)}_{\boldsymbol{z} \mid \boldsymbol{w}}(\boldsymbol{x}) g_t(\boldsymbol{x}, \boldsymbol{y}) = 1$, so that $\pi$ is a valid probability mass function. We will show that $\pi$ is the invariant distribution for the Markov transition function $K$. To this end, let $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}_t \times \mathcal{Z}_t$, and consider $\sum_{\boldsymbol{u} \in \mathcal{Z}_t} \sum_{\boldsymbol{v} \in \mathcal{Z}_t} K\big((\boldsymbol{u}, \boldsymbol{v}), (\boldsymbol{x}, \boldsymbol{y})\big)$. We have

$$\begin{aligned}
\sum_{\boldsymbol{u} \in \mathcal{Z}_t} \sum_{\boldsymbol{v} \in \mathcal{Z}_t} K\big((\boldsymbol{u}, \boldsymbol{v}), (\boldsymbol{x}, \boldsymbol{y})\big) \pi(\boldsymbol{u}, \boldsymbol{v}) &= \sum_{\boldsymbol{u} \in \mathcal{Z}_t} \sum_{\boldsymbol{v} \in \mathcal{Z}_t} \delta_{\boldsymbol{v}}(\boldsymbol{x}) g_t(\boldsymbol{x}, \boldsymbol{y}) \nu^{(t)}_{\boldsymbol{z} \mid \boldsymbol{w}}(\boldsymbol{u}) g_t(\boldsymbol{u}, \boldsymbol{v}) \\
&= \sum_{\boldsymbol{u} \in \mathcal{Z}_t} g_t(\boldsymbol{x}, \boldsymbol{y}) \nu^{(t)}_{\boldsymbol{z} \mid \boldsymbol{w}}(\boldsymbol{u}) g_t(\boldsymbol{u}, \boldsymbol{x}) \\
&= g_t(\boldsymbol{x}, \boldsymbol{y}) \sum_{\boldsymbol{u} \in \mathcal{Z}_t} \nu^{(t)}_{\boldsymbol{z} \mid \boldsymbol{w}}(\boldsymbol{u}) g_t(\boldsymbol{u}, \boldsymbol{x}) \\
&= g_t(\boldsymbol{x}, \boldsymbol{y}) \nu^{(t)}_{\boldsymbol{z} \mid \boldsymbol{w}}(\boldsymbol{x}) \\
&= \pi(\boldsymbol{x}, \boldsymbol{y}),
\end{aligned}$$

where the second-to-last equality follows from the fact that $\nu^{(t)}_{\boldsymbol{z} \mid \boldsymbol{w}}$ is the invariant distribution for the CGS for the LDA model indexed by $t$. This shows that equation (4.12) is satisfied, with $K$ and $\pi$ defined by (4.14) and (4.15), respectively.

To show (4.13), let $\rho(\boldsymbol{x}, \boldsymbol{y}) = t^{-N} g_t(\boldsymbol{x}, \boldsymbol{y})$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}_t \times \mathcal{Z}_t$. It is not difficult to see that $\sum_{\boldsymbol{x} \in \mathcal{Z}_t} \sum_{\boldsymbol{y} \in \mathcal{Z}_t} g_t(\boldsymbol{x}, \boldsymbol{y}) = t^N$, and so $\rho$ is a probability mass function on $\mathcal{Z}_t \times \mathcal{Z}_t$. Consider now the two-step Markov transition function $K^2$. We have

$$K^2\big((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')\big) = g_t(\boldsymbol{y}, \boldsymbol{x}') g_t(\boldsymbol{x}', \boldsymbol{y}') \qquad \text{for all } (\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{Z}_t \times \mathcal{Z}_t. \tag{4.16}$$

Informally, this is because two steps of the Markov chain $(\boldsymbol{z}_0, \boldsymbol{z}_1), (\boldsymbol{z}_1, \boldsymbol{z}_2), (\boldsymbol{z}_2, \boldsymbol{z}_3), \ldots$ take us from $(\boldsymbol{z}_0, \boldsymbol{z}_1)$ to $(\boldsymbol{z}_2, \boldsymbol{z}_3)$, so we need a CGS-transition from $\boldsymbol{z}_1$ to $\boldsymbol{z}_2$ followed by a CGS-transition from $\boldsymbol{z}_2$ to $\boldsymbol{z}_3$, and this is given by $g_t(\boldsymbol{z}_1, \boldsymbol{z}_2) g_t(\boldsymbol{z}_2, \boldsymbol{z}_3)$, which is equivalent to (4.16). Alternatively, we may write

$$K^2\big((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')\big) = \sum_{\boldsymbol{u} \in \mathcal{Z}_t} \sum_{\boldsymbol{v} \in \mathcal{Z}_t} K\big((\boldsymbol{u}, \boldsymbol{v}), (\boldsymbol{x}', \boldsymbol{y}')\big) K\big((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{u}, \boldsymbol{v})\big),$$

and apply the definition of $K$ (equation (4.14)) to obtain (4.16). By Theorem 1, $g_t(\boldsymbol{y}, \boldsymbol{x}') \geq c_t t^{-N}$ with $c_t$ given by (2.9); therefore from (4.16) we get

$$K^2\big((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')\big) \geq c_t t^{-N} g_t(\boldsymbol{x}', \boldsymbol{y}') = c_t \rho(\boldsymbol{x}', \boldsymbol{y}').$$

Hence (4.13) is satisfied, with $r = 2$.

We now apply Theorem 2 of Athreya et al. (1996) to conclude that

$$\begin{aligned}
\frac{1}{m} \sum_{l=1}^{m} \frac{\nu_{T, \boldsymbol{z} \,|\, \boldsymbol{w}}(t, \boldsymbol{z}_l)}{g_t(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)} \xrightarrow{\text{a.s.}} & \sum_{\boldsymbol{x} \in \mathcal{Z}_t} \sum_{\boldsymbol{y} \in \mathcal{Z}_t} \frac{\nu_{T, \boldsymbol{z} \,|\, \boldsymbol{w}}(t, \boldsymbol{y})}{g_t(\boldsymbol{x}, \boldsymbol{y})} \nu^{(t)}_{\boldsymbol{z} \,|\, \boldsymbol{w}}(\boldsymbol{x}) g_t(\boldsymbol{x}, \boldsymbol{y}) \\
= & \sum_{\boldsymbol{x} \in \mathcal{Z}_t} \sum_{\boldsymbol{y} \in \mathcal{Z}_t} \nu_{T, \boldsymbol{z} \,|\, \boldsymbol{w}}(t, \boldsymbol{y}) \nu^{(t)}_{\boldsymbol{z} \,|\, \boldsymbol{w}}(\boldsymbol{x}) \\
= & \left( \sum_{\boldsymbol{y} \in \mathcal{Z}_t} \nu_{T, \boldsymbol{z} \,|\, \boldsymbol{w}}(t, \boldsymbol{y}) \right) \left( \sum_{\boldsymbol{x} \in \mathcal{Z}_t} \nu^{(t)}_{\boldsymbol{z} \,|\, \boldsymbol{w}}(\boldsymbol{x}) \right) \\
= & \ \nu_{T \,|\, \boldsymbol{w}}(t). \qquad \square
\end{aligned}$$

Corollary 1 below states that in addition to almost sure convergence of $\tilde{\nu}^{(m)}_{T \,|\, \boldsymbol{w}}(t)$ to $\nu_{T \,|\, \boldsymbol{w}}(t)$, we also have convergence in $L_1$. Besides being of interest in its own right, this result is used in the proof of Theorem 3. Corollary 1 and Lemma 1 are proved under the assumption that $\mathcal{T}$ is finite. These two results are needed in our proof of Theorem 3. Actually, Theorem 3 can be stated without the finiteness assumption, but the proof is then much more complicated and is given in the supplementary document Chen and Doss (2018). (The results under the finiteness assumption are of interest in their own right—when dealing with the number of components in mixture modelling, the finiteness assumption is sometimes made; see, e.g., Fernández and Green (2002) and Richardson and Green (1997).)

**Corollary 1** *Suppose that $\mathcal{T}$ is finite. Let $t \in \mathcal{T}$, and let $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$ be a Markov chain generated according to the CGS for the LDA model indexed by $t$. Then*

$$\sum_{\boldsymbol{\zeta} \in \mathcal{Z}_t^m} \big| \tilde{\nu}^{(m)}_{T \,|\, \boldsymbol{w}}(t) - \nu_{T \,|\, \boldsymbol{w}}(t) \big| q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \to 0 \qquad \text{as } m \to \infty \tag{4.17}$$

*and*

$$\sum_{t \in \mathcal{T}} \sum_{\boldsymbol{\zeta} \in \mathcal{Z}_t^m} \big| \tilde{\nu}^{(m)}_{T \,|\, \boldsymbol{w}}(t) - \nu_{T \,|\, \boldsymbol{w}}(t) \big| q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \to 0 \qquad \text{as } m \to \infty. \tag{4.18}$$

**Proof** We will show that the sequence $\{\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)\}_{m=1}^{\infty}$ is uniformly integrable. Since almost sure convergence in the presence of uniform integrability implies $L_1$ convergence, the result will follow. By Theorem 1, $g_t(\boldsymbol{z}, \boldsymbol{z}') \geq c_t(1/t)^N$; therefore we have

$$\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t) = \frac{1}{m}\sum_{l=1}^{m} \frac{\nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t, \boldsymbol{z}_l)}{g_t(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)} \leq \frac{1}{m}\sum_{l=1}^{m} \frac{\nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t, \boldsymbol{z}_l)}{c_t/t^N} \leq \frac{1}{m}\sum_{l=1}^{m} \frac{1}{c_t/t^N} = \frac{t^N}{c_t} < \infty.$$

Since $\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)$ is bounded by a constant which does not depend on $m$, the sequence $\{\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)\}_{m=1}^{\infty}$ is uniformly integrable, so (4.17) follows, and because $\mathcal{T}$ is finite, (4.18) follows from (4.17). $\square$

Lemma 1 and Theorem 3 deal with Question II. Lemma 1 is a straightforward consequence of Theorems 1 and 8 in Andrieu and Roberts (2009). The results in Theorem 3 are similar to those in Theorem 6 and Corollary 7 in Andrieu and Roberts (2009), but with important differences, which we discuss later. To state our results, we need some additional notation. For each $m \in \mathbb{N}$, define

$$\begin{aligned} \tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t, \boldsymbol{\zeta}) &= \tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t) q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \qquad \text{for all } t \in \mathcal{T},\ \boldsymbol{\zeta} \in \mathcal{Z}_t^m, \\ \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t, \boldsymbol{\zeta}) &= \nu_{T\,|\,\boldsymbol{w}}(t) q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \qquad \text{for all } t \in \mathcal{T},\ \boldsymbol{\zeta} \in \mathcal{Z}_t^m. \end{aligned} \tag{4.19}$$

Note that $\tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t, \boldsymbol{\zeta})$ was defined earlier in a general setting, and from the discussion just before (4.3), we know that $E(\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)) = \nu_{T\,|\,\boldsymbol{w}}(t)$, where the expectation is taken with respect to $q^{(m,t)}_{\boldsymbol{\zeta}}$. From the discussion following (4.7), we know that $\tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}$ is the invariant distribution of the PMMH algorithm with $m$ inner loops. Further, define the sub-stochastic kernel $K_T$ on $\mathcal{T}$ by

$$K_T(t, t') = \min\{1, r(t, t')\} q_T(t, t'). \tag{4.20}$$

In brief, Lemma 1 states that the ergodicity properties of the ideal Metropolis-Hastings algorithm get passed on to the PMMH algorithm (whose invariant distribution is $\tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}$—see the discussion following (4.7)): (1) if the ideal Metropolis-Hastings algorithm is ergodic, then for each $m \in \mathbb{N}$ so is the PMMH algorithm; and (2) if $K_T$ given by (4.20) satisfies a Doeblin-like condition, then the PMMH algorithm is uniformly ergodic. Lemma 1 is a "fixed-$m$" result. Theorem 3 gives results for the case where both $m, n \to \infty$. One of these is as follows. Let $\mu_T^{m,n}$ be the distribution of $T$ after $n$ cycles of the PMMH algorithm with $m$ inner loops. Then, as $m, n \to \infty$, $\mu_T^{m,n}$ converges to $\nu_{T\,|\,\boldsymbol{w}}$, i.e. the posterior distribution of $T$, in absolute deviation norm. Thus, the PMMH algorithm can be used to estimate $\nu_{T\,|\,\boldsymbol{w}}$. The theorem also states that the PMMH algorithm can be used to estimate $\nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}$, and even $\nu_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}\,|\,\boldsymbol{w}}$.

**Lemma 1** *Suppose that $\mathcal{T}$ is finite. Assume Conditions A2, and A3 in the context of the LDA model. Then:*

1. *The PMMH algorithm is ergodic, i.e.*

$$\lim_{n\to\infty} \left\| P^{m,n}(t,\boldsymbol{\zeta};\cdot,\cdot) - \tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot,\cdot) \right\| = 0 \qquad \text{for all } t \in \mathcal{T},\ \boldsymbol{\zeta} \in \mathcal{Z}_t^m. \tag{4.21}$$

2. *Suppose in addition that there exist a probability measure $\varrho_T$ on $\mathcal{T}$, a positive integer $n_0$, and a constant $\delta > 0$ such that*

$$K_T^{n_0}(t,\cdot) \geq \delta \varrho_T(\cdot) \qquad \text{for all } t \in \mathcal{T}.$$

*Then the PMMH algorithm is uniformly ergodic: there exists a constant $\kappa \in (0,1)$ such that*

$$\left\| P^{m,n}(t,\boldsymbol{\zeta};\cdot,\cdot) - \tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot,\cdot) \right\| \leq (1-\kappa)^{[n/n_0]} \qquad \text{for all } t \in \mathcal{T},\ \boldsymbol{\zeta} \in \mathcal{Z}_t^m.$$

*Moreover, the constant $\kappa$ does not depend on $m$.*

Consider now the PMMH algorithm with $m$ inner loops. Recall that $\mu_T^{m,n}$ denotes the distribution of $T_n$. Let $\mu_{T,\boldsymbol{z}}^{m,n}$ denote the distribution of $(T_n, \zeta_{n,m})$. Given $(T_n, \zeta_{n,m})$, we generate $(\boldsymbol{\beta}^{(n)}, \boldsymbol{\theta}^{(n)})$ according to (2.5). Let $\mu_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}}^{m,n}$ denote the distribution of $(T_n, \zeta_{n,m}, \boldsymbol{\beta}^{(n)}, \boldsymbol{\theta}^{(n)})$. These distributions all depend on the starting point $(t_0, \boldsymbol{\zeta}_0) \in \mathcal{S}$, but this dependence is suppressed in the notation.

**Theorem 3** *Suppose that $\mathcal{T}$ is finite. Assume Conditions A2 and A3 in the context of the LDA model, the conditions of Part 2 of Lemma 1, and that the mechanism for generating $\boldsymbol{\zeta}$ is the CGS (see (4.8)). Then:*

1. *For any $\epsilon > 0$ there exist positive integers $M(\epsilon)$ and $N(\epsilon)$ such that for any $m \geq M(\epsilon)$ and any $n \geq N(\epsilon)$, for any initial points $t_0 \in \mathcal{T}$ and $\boldsymbol{\zeta}_0 \in \mathcal{Z}_t^m$ we have*

$$\left\| P^{m,n}(t_0,\boldsymbol{\zeta}_0;\cdot,\cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot,\cdot) \right\| \leq \epsilon.$$

2. $$\left\| \mu_T^{m,n}(\cdot) - \nu_{T\,|\,\boldsymbol{w}}(\cdot) \right\| \to 0 \qquad \text{as } m,n \to \infty.$$

3. $$\left\| \mu_{T,\boldsymbol{z}}^{m,n}(\cdot,\cdot) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(\cdot,\cdot) \right\| \to 0 \qquad \text{as } m,n \to \infty.$$

4. *For any $t \in \mathcal{T}$, $\boldsymbol{z} \in \mathcal{Z}_t$,*

$$\left\| \mu_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}}^{m,n}(t,\boldsymbol{z},\cdot,\cdot) - \nu_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}\,|\,\boldsymbol{w}}(t,\boldsymbol{z},\cdot,\cdot) \right\| \to 0 \qquad \text{as } m,n \to \infty.$$

As is the case for Theorem 1, we do not provide rates for the convergence.

# 5 Performance of the PMMH Algorithm for Selecting the Number of Topics

This section provides an evaluation of our methodology, and consists of three parts. In Section 5.1 we briefly review existing methods for estimating $T$. In Section 5.2 we consider a synthetic corpus generated from the LDA model. The reason for considering synthetic corpora is that for such corpora the parameters are known, thus enabling an evaluation of performance and comparison with other methods. In Section 5.3 we consider two real corpora, each consisting of a set of articles from Wikipedia. For each corpus, the topics spanned by the documents in the corpus are fairly close to each other. For these corpora, the true number of topics is not known with certainty, so assessment of any methodology is difficult to perform. To carry out the evaluation, we use a criterion called "Posterior Predictive Checking" (PPC), which we discuss in Section 5.1.

The following facts are obvious, but it is perhaps worthwhile to state them explicitly. The PMMH algorithm is not, properly speaking, an estimator of $T$; rather, it is a Monte Carlo method for approximating the posterior distribution of $T$, and hence the mode (or mean) of that posterior. Likewise, the harmonic mean estimator and Chib's method are not estimators of $T$; rather, they are procedures for approximating the intractable likelihood $m_{\boldsymbol{w}}(T)$, and hence the standard frequentist estimator of $T$, which is $\arg\max_T m_{\boldsymbol{w}}(T)$. To express this more forcefully, if we were to compare the PMMH algorithm to Chib's method and the HME using extremely large simulation sizes, ultimately we would be comparing the mode of the posterior and $\arg\max_T m_{\boldsymbol{w}}(T)$ as estimators of $T$, and our comparison would have nothing to do with the PMMH algorithm, Chib's method, and the HME. Therefore, the PMMH algorithm, Chib's method, and the HME should be judged on their efficiency (computational and statistical) in carrying out the approximations for which they were designed. We should also keep in mind the well-known fact that, under a uniform prior, the mode of the posterior and the maximum likelihood estimate are nearly equal, at least for large data sets.

In Sections 5.1 and 5.2 we show that the PMMH algorithm performs very well on both synthetic and real data, and compares very favorably with all the other methods that we review, where evaluation is in the sense described in the paragraph above. All the data sets we consider are large, and consequently Bayes and frequentist estimators of $T$ are both nearly perfect, and the issue is computability of these estimators. Generally speaking, we find that the methods we discuss for approximating the frequentist estimator do not fare well on the LDA model, because of the high dimensions involved.

## 5.1 Other Methods for Selecting the Number of Topics

Below, we discuss harmonic mean estimators and Chib's (1995) method. We also discuss PPC, which is a general-purpose Bayesian method for model evaluation and checking.

Harmonic mean estimation was discussed in Section 1, and here we mention that there are two ways to implement the method. In the first implementation, we let $\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(n)}$ be generated according to the CGS indexed by $T$. The invariant distribution for this chain is $\nu_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}$. Let $L_{\boldsymbol{w}}^{(T)}(\boldsymbol{z})$ denote the likelihood of $\boldsymbol{z}$. For this implementation, which we denote by HME-$\boldsymbol{z}$, we form the estimator $\widehat{m}_{\boldsymbol{w}}(T) = \bigl[(1/n) \sum_{i=1}^n \bigl(1/L_{\boldsymbol{w}}^{(T)}(\boldsymbol{z}^{(i)})\bigr)\bigr]^{-1}$. In the second implementation, we consider the state space of the entire latent variable $\boldsymbol{\psi}$. We run a Markov chain with invariant distribution $\nu_{\boldsymbol{\psi} \mid \boldsymbol{w}}^{(T)}$. Let $\ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\psi})$ denote the likelihood of $\boldsymbol{\psi}$. This implementation, which we denote by HME-$\boldsymbol{\psi}$, is then exactly as originally described in Section 1. Chen (2015) provides details, including explicit expressions for the two HME's, and also discusses the Markov chain that is used for HME-$\boldsymbol{\psi}$.

*Chib's Method* This is a generic method for estimating the marginal likelihood from the output of a Gibbs sampler. Here, we briefly explain the method for the problem of estimating $m_{\boldsymbol{w}}(T)$. Suppose that $\boldsymbol{z}^{(1)}, \dots, \boldsymbol{z}^{(n)}$ are generated according to the CGS indexed by $T$. Note that for any fixed but arbitrary point $(\boldsymbol{\beta}, \boldsymbol{z})$, we have the identity

$$m_{\boldsymbol{w}}(T) = \frac{\ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\beta}, \boldsymbol{z})\, \nu_{\boldsymbol{\beta},\boldsymbol{z}}^{(T)}(\boldsymbol{\beta}, \boldsymbol{z})}{\nu_{\boldsymbol{\beta},\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}(\boldsymbol{\beta}, \boldsymbol{z})} = \frac{\ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\beta}, \boldsymbol{z})\, \nu_{\boldsymbol{\beta},\boldsymbol{z}}^{(T)}(\boldsymbol{\beta}, \boldsymbol{z})}{\nu_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}(\boldsymbol{z})\, \nu_{\boldsymbol{\beta} \mid \boldsymbol{z},\boldsymbol{w}}^{(T)}(\boldsymbol{\beta} \mid \boldsymbol{z})}. \tag{5.1}$$

Of the four terms on the right side of (5.1), all are known except for $\nu_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}(\boldsymbol{z})$, which we estimate by $\hat{\nu}_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}(\boldsymbol{z}) = \bigl[(1/n) \sum_{i=1}^n g_T(\boldsymbol{z}^{(i)}, \boldsymbol{z})\bigr]$. Details, including explicit expressions for $\ell_{\boldsymbol{w}}^{(T)}(\boldsymbol{\beta}, \boldsymbol{z})$, $\nu_{\boldsymbol{\beta},\boldsymbol{z}}^{(T)}(\boldsymbol{\beta}, \boldsymbol{z})$, and $\nu_{\boldsymbol{\beta} \mid \boldsymbol{z},\boldsymbol{w}}^{(T)}(\boldsymbol{\beta} \mid \boldsymbol{z})$, and an explanation of why $\hat{\nu}_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}(\boldsymbol{z}) \xrightarrow{\text{a.s.}} \nu_{\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}(\boldsymbol{z})$ are given in Chen (2015). For the method to work successfully, Chib (1995) recommended that the distinguished point $(\boldsymbol{\beta}, \boldsymbol{z})$ be taken to be a "high density point" under the posterior distribution, and to fully specify the method, we need to state how we choose this point. We choose it by running a small pilot experiment in which we generate a Markov chain $(\boldsymbol{\beta}^{(1)}, \boldsymbol{z}^{(1)}), \dots, (\boldsymbol{\beta}^{(k)}, \boldsymbol{z}^{(k)})$ with invariant distribution $\nu_{\boldsymbol{\beta},\boldsymbol{z} \mid \boldsymbol{w}}^{(T)}$. We take the average of $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(k)}$, and use this as the fixed point $\boldsymbol{\beta}$. We specify the fixed point $\boldsymbol{z}$ as follows. For each word $w_{di}$, we let $t = \arg\max_s \sum_{j=1}^k z_{dis}^{(j)}$, and then let $z_{dit} = 1$ and $z_{dit'} = 0$ for $t' \neq t$. Chib's method is not automatic: its efficiency depends heavily on the choice of the distinguished point $(\boldsymbol{\beta}, \boldsymbol{z})$. Unfortunately, because of the high dimension of the LDA model, there may not exist a point of sufficiently high density, and even if such a point exists, it may be hard to identify it.

*Posterior Predictive Checking* PPC is a Bayesian model checking method which, when applied to the LDA context, is described as follows. For $d = 1, \dots, D$, let $\boldsymbol{w}_{(-d)}$ denote the corpus consisting of all the documents except for document $d$. To evaluate a given model (in our case the LDA model indexed by a given $T$) through posterior predictive checking, in essence we see how well the model based on $\boldsymbol{w}_{(-d)}$ predicts document $d$, the held-out document. We do this for $d = 1, \dots, D$, and take the geometric mean. We formalize this as follows. For the LDA model indexed by $T$, the predictive likelihood of the held-out document is

$$L^{(T)}(d) = \int \ell^{(T)}_{\boldsymbol{w}_d}(\boldsymbol{\psi}) \nu_{\boldsymbol{\psi} \,|\, \boldsymbol{w}_{(-d)}}(\boldsymbol{\psi}) \, d\boldsymbol{\psi}. \tag{5.2}$$

We form the score $S(T) = \left[\prod_{d=1}^{D} L^{(T)}(d)\right]^{1/D}$. Two different values of $T$ are compared via their scores. Unfortunately, calculation of $S(T)$ is computationally extremely demanding. In the machine learning literature, $L^{(T)}(d)$ is often estimated by $\ell^{(T)}_{\boldsymbol{w}_d}(\widehat{\boldsymbol{\psi}})$, where $\widehat{\boldsymbol{\psi}}$ is a single point estimate that "summarizes the distribution $\nu^{(T)}_{\boldsymbol{\psi} \,|\, \boldsymbol{w}_{(-d)}}$" in some sense. Approximations of this sort can be woefully inadequate. Conceptually, it is easy to estimate $L^{(T)}(d)$ by direct Monte Carlo: let $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots$ be an ergodic Markov chain with invariant distribution $\nu^{(T)}_{\boldsymbol{\psi} \,|\, \boldsymbol{w}(-d)}$. We then approximate the integral by $(1/n) \sum_{i=1}^{n} \ell^{(T)}_{\boldsymbol{w}_d}(\boldsymbol{\psi}_i)$. Care needs to be exercised, however, because in (5.2), the variable $\boldsymbol{\psi}$ in the term $\ell^{(T)}_{\boldsymbol{w}_d}(\boldsymbol{\psi})$ has a dimension that is different than that of the variable $\boldsymbol{\psi}$ in the rest of the integral. Chen (2015) gives a careful description of a Monte Carlo scheme for estimating the integral in (5.2).

## 5.2 A Synthetic Data Set

Here we evaluate the performance of our PMMH algorithm on a corpus generated synthetically, using the following specifications of the LDA model: the number of topics is $T = 20$, the hyper-parameter is $h = (\alpha, \eta) = (0.1, 0.5)$, the vocabulary size is $V = 100$, the number of documents is $D = 2000$, and the document lengths are $n_d = 300$, $d = 1, \dots, D$. We generated the latent variables and the documents according to the LDA model with these specifications. For this corpus, we took the prior on $T$ to be the uniform distribution on $\{t_{\text{min}}, \dots, t_{\text{max}}\}$, where $t_{\text{min}} = 2$ and $t_{\text{max}} = 100$, and for our PMMH algorithm, we took the Markov transition function $q_T$ to be defined as follows. For $t$ not equal to one of the boundary points $t_{\text{min}}$ or $t_{\text{max}}$, $q_T(t, \cdot)$ gives mass $1/2$ to

$t-1$ and $t+1$; and $q_T(t_{\text{min}}, \cdot)$ and $q_T(t_{\text{max}}, \cdot)$ give mass 1 to $t_{\text{min}}+1$ and $t_{\text{max}}-1$, respectively. We ran the algorithm with $m = 50$, 75, and 100, each time for $n = 10{,}000$ iterations, and taking the starting value for the number of topics to be $T_0 = 30$. The reason for considering multiple values of $m$ is to evaluate the effect of $m$ on the performance of the algorithm, and in particular to determine whether a value of $m$ as small as 50 gives good results.

The top two panels in Figure 1 give plots of running means for $T$ produced by the PMMH algorithm using the three values of $m$. The plots differ only in the scaling of the $x$-axis. As mentioned earlier, because of the size of the corpus, we presume that the posterior distribution is essentially a point mass at the true value of $T$ (which is 20) and also that the marginal likelihood $m_{\boldsymbol{w}}(\cdot)$ is maximized at the true value. The plots show that even when $m = 50$, it takes less than 4000 iterations for the mean to reach 20; and once it reaches 20 the mean essentially remains there. The plots also show that convergence is faster when $m$ is larger, but that the gain in efficiency is relatively minor, and a value of $m$ as small as 50 produces good results. To investigate the effect of the vocabulary size, we also produced the plot at the bottom of Figure 1. For this plot, the corpus has the same configuration as for the top two plots, except that we took $V = 3000$ (and we took $m = 50$). The plot shows that with this new configuration we still have convergence; although it is a bit slower, it is still well within the acceptable range. In our experiments we used the true value of $h$. The reason we use this oracle value is that here our focus is on estimation of $T$, and considering various methods for estimating $h$ would obscure our results (in the next section we discuss the issue of estimating $h$). In any case, there is no qualitative change in our results if we use the empirical Bayes estimate of $h$ (discussed in the next subsection). In this example, where $D = 2000$, the posterior distribution of $T$ is fairly concentrated. In Chen and Doss (2018) we investigate the behavior of the algorithm in situations where the posterior is more diffuse. A brief summary of our findings is as follows. Running means of $T$ produced by our algorithm still converge, although convergence slows as the posterior becomes more diffuse. Nevertheless, the algorithm still produces good results.

Here we remark on the choice of using the CGS in the denominator of (4.9). In principle, any importance sampling distribution can be used in that denominator, but unless the distribution is chosen carefully, one should not expect good results. Although it is not advisable to do this, in Bayesian statistics the prior is sometimes used as an importance sampling distribution. When we used the prior $\nu_{\boldsymbol{z}}$ in the denominator of (4.9), the performance of the algorithm was abysmal: starting at $T_0 = 30$, a traceplot of $T$ reached 2 in less than 100 iterations and stayed there.

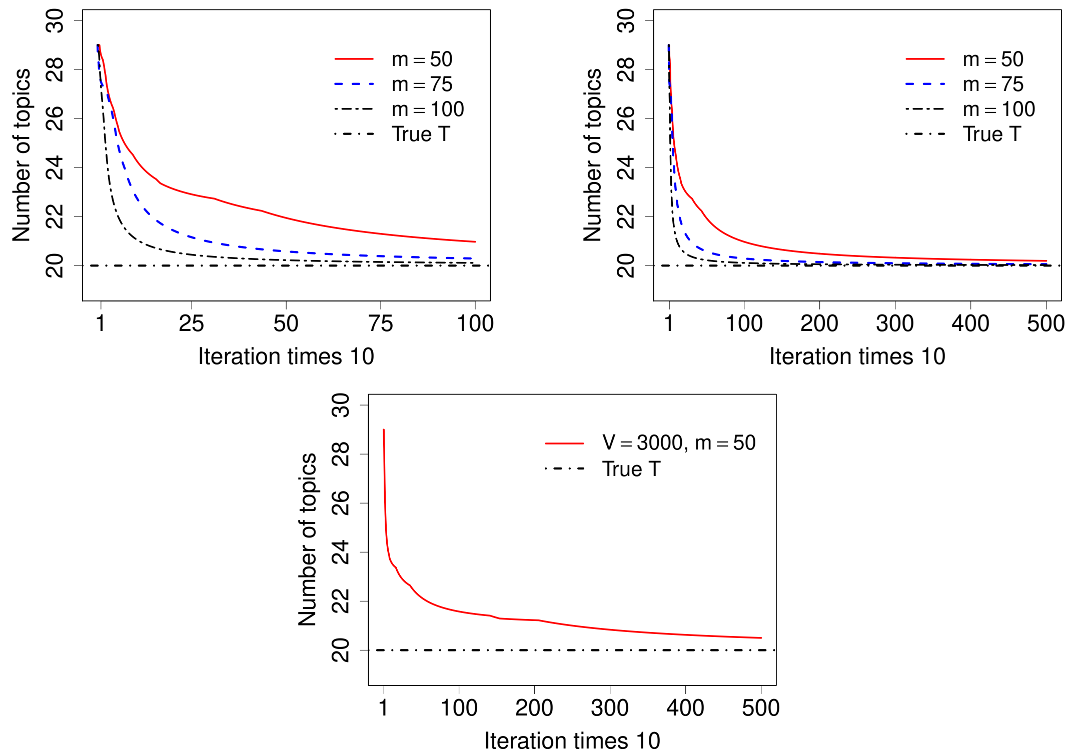
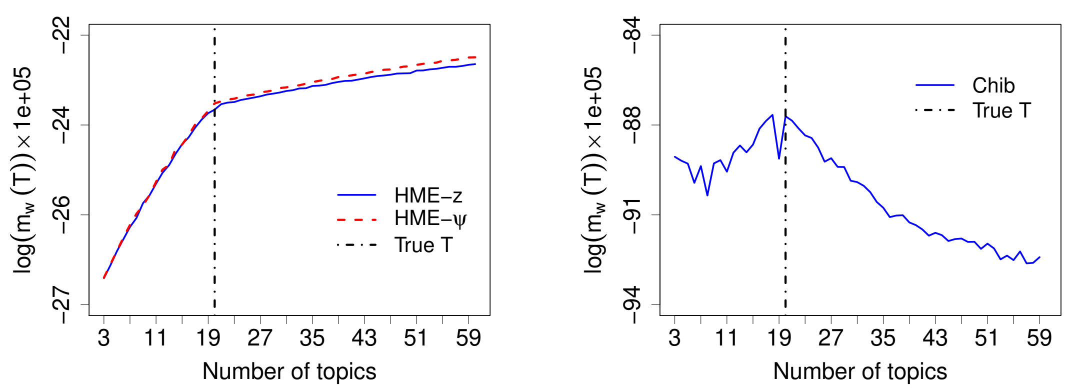



Figure 1: Running means for $T$ produced by the PMMH algorithm. The top two plots, which differ only in the scale of the $x$-axis, show the running means using $m = 50$, $75$, and $100$ for the corpus described in the first paragraph of this subsection. The bottom plot gives the running mean for the case where $V$ is increased to $3000$.

We now compare the performances of the PMMH algorithm, the two harmonic mean estimators, and the Chib estimator, keeping in mind the caveat stated at the beginning of Section 5. To this end we calculated the HME-$\boldsymbol{z}$, HME-$\boldsymbol{\psi}$, and Chib estimates for $T = 3, \dots, 60$, using Markov chains of lengths $3500$, $3500$, and $1000$, respectively. For the Chib method, we used a Markov chain of length $300$ for the pilot study needed to select the high-density point that is used in the main simulation. With these Markov chain lengths, the running times of all these methods are approximately equal, and this common time is about five times the running time of the PMMH algorithm using $m = 50$. Information on timing for each algorithm, data set, and parameter configuration used in this paper is given in Section 5.4.

Figure 2 shows the results. The HME-$\boldsymbol{z}$ and HME-$\boldsymbol{\psi}$ estimates of $\arg\max_T m_{\boldsymbol{w}}(T)$ are almost identical, and both increase with $T$ over the entire range $\{3, \dots, 60\}$. So the HME method gives hopelessly bad estimates. The poor performance of the HME has been noted in the literature before (Wallach et al., 2009b), although not for the problem of choosing $T$, and we included harmonic mean estimation in our study because the method does get used in the machine learning literature (Griffiths and Steyvers (2004), Griffiths et al. (2004), Wallach (2006), among others). Figure 2(b) gives a plot of Chib's estimate of the marginal likelihood on the log scale. The maximum is reached at $T = 18$ and the estimate at $T = 20$ is a local minimum. It should be noted that the ratio of estimate of the marginal likelihood at $18$ to the estimate at $20$ is $\exp(16732)$ (the ratio is understated by the appearance of the plot, which is on the log scale); thus, Chib's method effectively rules out the true value of $T$. The poor performance of Chib's method for topic modelling has been noted in the literature before (Wallach et al., 2009b) and is not surprising in view of the high dimension of the LDA model, as was discussed in Section 5.1.

The PPC method is evaluated empirically in the supplementary document Chen and Doss (2018), where we give an illustration on three synthetic corpora which are similar to the corpus described here, except that the number of topics is only $6$ instead of $20$, and the number of documents is only $300$ instead of $2000$. In that illustration all the methods are considered and compared, and the results are, roughly speaking, as follows. The HME and Chib methods perform very poorly; however the PPC method gives reasonable results, although the estimates it produces are not as accurate as those produced by the PMMH algorithm. Unfortunately, the PPC method is computationally infeasible except on small corpora, and in particular, we were not able to implement it on the corpus considered in this section.




(a) HME's of $\log(m_{\boldsymbol{w}}(T))$ (b) Chib estimate of $\log(m_{\boldsymbol{w}}(T))$

Figure 2: HME-$\boldsymbol{\psi}$, HME-$\boldsymbol{z}$, and Chib estimates of $\log(m_{\boldsymbol{w}}(T))$ for the models indexed by different $T$'s. The vertical line at $20$ in each plot represents the true value of $T$.

There is one more point that should be mentioned. The HME's, Chib's estimator, and the PPC criterion all have to be computed for a range of values of $T$'s, and this range needs to include the true value. In our experiments we used the range $\{3, \ldots, 60\}$, and this decision was informed by our knowledge that the true value is $20$. When dealing with a real data set we will not have an oracle which presents us with a good range to use, and we may have to use a large range of values of $T$, making these methods slow and unwieldy. By contrast, our algorithm does not suffer from this difficulty. To illustrate this, we pretended that we believed that the true value of $T$ might be around $200$, so we ran the PMMH algorithm using $m = 50$ with starting value for the number of topics being $T = 200$. A traceplot of $T$ reached $T = 30$ in $470$ iterations and from then on looked very similar to the traceplot of $T$ for which the starting value is $T = 30$.

## 5.3 Wikipedia Data Sets

In this section, we illustrate the use of our PMMH algorithm on two collections of real documents from the English Wikipedia. When a Wikipedia web article is created, it is typically tagged to one or more categories, one of which is the "primary category." The two corpora we used were created by George (2015). (They are included in compact form in our software; the original files, with descriptions, are available at `https://github.com/clintpgeorge/ldamcmc/tree/master/data-raw`.) The first corpus, which we call R-1, is a subset of the articles under the Wikipedia category *Whales*. Each of these articles is tagged to one of the following five subcategories: *Baleen Whale*, *Dolphins*, *Oceanic Dolphins*, *Whaling*, and *Whale Products*. The other corpus, which we call R-2, is a subset of the articles under the Wikipedia category *Birds of Prey*. Each of these articles is tagged to one of the following seven subcategories: *Eagles*, *Falco (genus)*, *Falconry*, *Harriers*, *Hawks*, *Kites*, and *Owls*. We will act as if the number of topics is unknown, and the objective is to infer it. We remark that each of these corpora is hard to analyze because in each corpus the subcategories are similar to each other—a *Baleen Whale* article and a *Dolphin* article are far more similar than are two New York Times articles, one from *Sports* and the other from *Politics*. We selected these corpora precisely because they are difficult to cluster, and we are interested in seeing how our methodology works for such corpora.

The following point is obvious, but nevertheless is well worth emphasizing. For these real corpora, there is no such thing as a "true number of topics": the variable $T$ is a hyperparameter of the LDA model, and there is no reason to think that either of the two corpora follows this model. For example, it may be reasonable to believe, a priori, that two of the $7$ subcategories under *Birds of Prey* should be lumped together, or that one of the subcategories should be split in two. There are actually two distinct goals: (1) estimate the true value of $T$, and (2) identify the value of $T$, say $T_0$, for which the LDA model based on $T_0$ outperforms LDA models based on any other value of $T$. The first goal is not meaningful if the LDA model does not hold, while the second is meaningful even in that case. (This is analogous to a variable selection situation in linear regression. One goal is to identify those regression coefficients which are exactly zero, and a distinct goal is to select a set of variables for which the corresponding model has the best predictive ability—the second goal is meaningful even if the linear regression model does not hold. See Yang (2005) for a discussion of these points.) Our interest, therefore, is in whether the mode of the posterior distribution of $T$, as estimated by the PMMH algorithm, gives a model with good predictive ability. In this regard, we will consider the PPC score, as this criterion provides a useful method for evaluation of predictive ability. In essence, we want to determine whether accomplishing the first goal also accomplishes the second goal.

For each of these corpora, we ran the PMMH algorithm with $m = 50$ for $n = 10{,}000$ iterations, using the same specifications for the prior $\nu_T$ and the Markov transition function $q_T$ as in the experiments done on the synthetic corpus described in Section 5.2. An issue of implementation is the choice of the hyperparameter $h = (\alpha, \eta)$. The literature gives several default choices (Griffiths and Steyvers, 2004; Asuncion et al., 2009; Řehůřek and Sojka, 2010); these are all ad-hoc, i.e. not based on any statistical principle. The gold standard is the empirical Bayes choice: let $m(h)$ be the marginal likelihood of the data under the model specified by the hyperparameter $h$ (this is the likelihood with the parameter $\boldsymbol{\psi}$ integrated out). The empirical Bayes choice is by definition $\arg\max_h m(h)$, but unfortunately, it is analytically intractable. Wallach (2006) proposed a "Gibbs-EM" algorithm, in which the E-step is approximated by an MCMC estimate, and Blei et al. (2003) developed a "variational-inference EM" algorithm, in which the E-step is approximated via variational methods. For neither of these schemes has theoretical validity been established. An MCMC algorithm for estimating $\arg\max_h m(h)$ is developed in George and Doss (2018), and this is the algorithm we use. Table 1 presents some information on these Wikipedia corpora, including the value of the empirical Bayes choice of $h$.

Table 2 gives the distributions of the samples of $T$ produced by the algorithm for corpora R-1 and R-2. As can be seen from the table, for R-1 the posterior distribution of $T$ is concentrated at $5$, with $96.35\%$ of its mass at that point; and for R-2, it is concentrated at 7, with $97.2\%$ of its mass at that point.

Figure 3 gives plots of the running means of $T$ produced by the PMMH algorithm for corpora

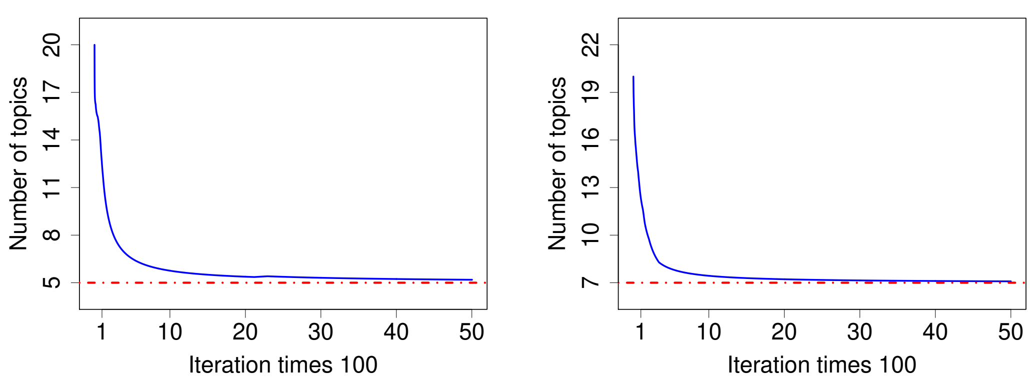
| *Corpus* | *Wikipedia Category* | $N$ | $h = (\alpha, \eta)$ | *Wikipedia Subcategories* |
|---|---|---|---|---|
| R-1 | Whales (153) | 52,107 | $(.11, .41)$ | Baleen Whale (40), Dolphins (10), Oceanic Dolphins (61), Whaling (32), Whale Products (10) |
| R-2 | Birds of Prey (304) | 116,135 | $(.11, .25)$ | Eagles (62), Falco (genus) (55), Falconry (52), Harriers (21), Hawks (16), Kites (22), Owls (76) |

Table 1: The two Wikipedia corpora. The numbers shown in parentheses after the category and subcategory names are the numbers of documents associated with the corresponding categories and subcategories, and $N$ is the total number of words in the corpus.

| *Corpus* | $T = 5$ | $T \geq 6$ | *Corpus* | $T \leq 6$ | $T = 7$ | $T \geq 8$ |
|---|---|---|---|---|---|---|
| R-1 | $96.35\%$ | $3.64\%$ | R-2 | $1.5\%$ | $97.22\%$ | $1.28\%$ |

Table 2: Posterior distributions of $T$ given by the PMMH algorithm with $m = 50$ for corpora R-1 and R-2.

R-1 and R-2. The left panel shows that for R-1 it takes about $5000$ iterations for the mean to reach the posterior mode of $5$, and that once it reaches $5$, it essentially stays there. We see the same effect for R-2 (right panel), except that the number of iterations is $3000$.



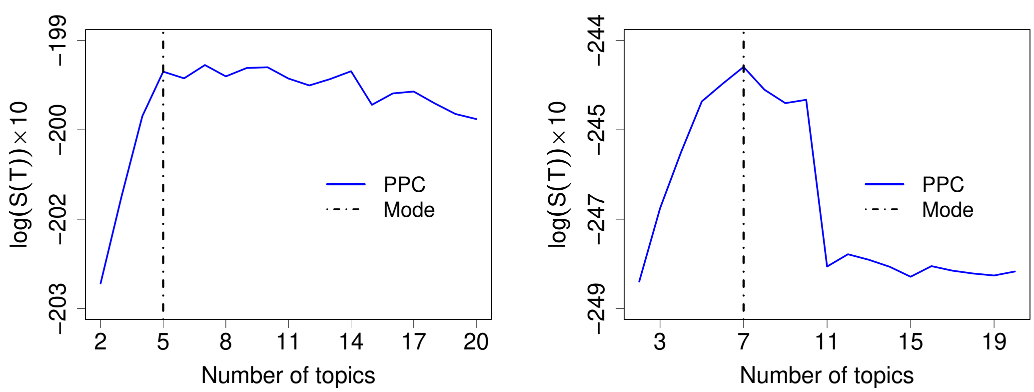

(a) Corpus R-1

(b) Corpus R-2

Figure 3: Running means for $T$ produced by the PMMH algorithm with $m = 50$ for the Wikipedia corpora.

Table 2 shows that for each corpus the mode of the posterior distribution of $T$, as estimated by the PMMH algorithm, is the number of subcategories; but as noted earlier, one cannot conclude that the estimate produced by the algorithm is correct, because the true number of topics is not a well-defined entity. It is therefore of interest to evaluate the estimate produced by the algorithm via the PPC score. Figure 4 gives a plot of the PPC score $S(T)$ for $T = 2, \dots, 20$ on the log scale, for each corpus. For each $T$, the score $S(T)$ was estimated using Markov chains of length $100$. (The reason the chains are this short is that for each $T$, we need to run a chain for each document in the corpus, as indicated in Section 5.1.) The plots show that for corpus R-1, according to the PPC criterion the LDA model with $T = 5$ fits the data well (and is nearly optimal), and for corpus R-2, the model with $T = 7$ gives the best fit. We conclude that our algorithm produces very reasonable results on these two real corpora.

When looking at Figure 4, one notes that as $T$ moves away from the mode, the drop is sharper for R-2 than for R-1. Chen (2015) gives an interesting explanation for this. He shows that the $L_1$ distances between the topic vectors are smaller on average for R-1 than they are for R-2. Thus, estimation of $T$ is harder for R-1, and this explains both the relative flatness of the plot of the PPC score near its maximizer and the slower convergence of the running mean for R-1.




(a) Corpus R-1

(b) Corpus R-2

Figure 4: PPC score (on the log scale) for $T = 2, 3, \dots, 20$ for the Wikipedia corpora. The vertical lines indicate the mode of the posterior distribution of $T$, as estimated by the PMMH algorithm.

## 5.4 Speed of Execution and a Grouped Gibbs Sampler for Distributed Computing

Our empirical studies were conducted through the R programming language (R Core Team, 2017), using Rcpp (Eddelbuettel and François, 2011) and RcppArmadillo (Eddelbuettel and Sanderson, 2014), on a 3.40GHz quad-core Intel i7-3770 CPU running Linux. Table 3 gives the amount of time, in hours, for each experiment.

| *Data Set* | *Method* | *Time* |
|---|---|---|
| Synthetic data | PMMH ($n = 5000, m = 50$) | 22.7 |
| | PMMH ($n = 5000, m = 75$) | 33.6 |
| | PMMH ($n = 5000, m = 100$) | 44.5 |
| | Chib estimator | 22.7 |
| | HME-$\boldsymbol{z}$ | 22.7 |
| | HME-$\boldsymbol{\psi}$ | 22.7 |
| Whales corpus | PMMH ($n = 10{,}000, m = 50$) | 1.1 |
| | PPC | 3.3 |
| Birds of prey corpus | PMMH ($n = 10{,}000, m = 50$) | 1.2 |
| | PPC | 10.6 |

Table 3: Length of time, in hours, it takes to carry out the various methods on the three corpora studied in Sections 5.2 and 5.3. For the HME, Chib and PPC methods, the parameters used are as described in those sections.

As is clear from the table, the amount of time it takes to execute the PMMH algorithm is quite significant, and we now address this problem. The bottleneck is the CGS, which is a Gibbs sampler that runs on the vector $\boldsymbol{z} = (z_{11}, \ldots, z_{1n_1}, \ldots, z_{D1}, \ldots, z_{Dn_D})$, updating each variable sequentially, with $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ integrated out. So there is a node for each word of each document. As mentioned in Section 2, George and Doss (2018) considered a Markov chain for estimating the posterior distribution of $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$ that runs on the *pair* $(\boldsymbol{z}, (\boldsymbol{\beta}, \boldsymbol{\theta}))$: the chain is a two-cycle Gibbs sampler based on the conditionals (2.5) and (2.6). This scheme, which we will call "Grouped Gibbs Sampler" (GGS), has the very attractive feature that it can be parallelized: given $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\boldsymbol{w}$, all the components of $\boldsymbol{z}$ are independent, so can be updated simultaneously by different processors; and given $\boldsymbol{z}$ and $\boldsymbol{w}$, all the $\theta_d$'s and $\beta_t$'s can be updated simultaneously by different processors.

The scheme was mentioned briefly by Newman et al. (2009), who dismissed it on the grounds that the CGS has superior mixing properties because, according to a theorem in Liu et al. (1994), collapsing improves the mixing rate. Liu et al. (1994) consider a Gibbs sampling situation involving three variables $X$, $Y$, and $Z$. They show that a Gibbs sampler on the pair $(X, Y)$ (with $Z$ integrated out), which they call a collapsed Gibbs sampler, is superior (in terms of mixing) to a Gibbs sampler on the triple $(X, Y, Z)$. They also show that a two-cycle Gibbs sampler on the pair $(X, (Y, Z))$, which they call a grouped Gibbs sampler, is superior to a Gibbs sampler on the triple $(X, Y, Z)$. Using the terminology of Liu et al. (1994), if we take our base to be the Gibbs sampler that runs through $(\beta_1, \ldots, \beta_T, \theta_1, \ldots, \theta_D, z_{11}, \ldots, z_{1n_1}, \ldots, z_{D1}, \ldots, z_{Dn_D})$, then the CGS is a collapsed version, while the GGS is a grouped version. The theorem in Liu et al. (1994) says nothing about a comparison between the CGS and the GGS, because the CGS is not a collapsed version of the GGS in any sense at all. George (2015) compared the mixing rates for various parameters empirically, and found that the mixing rate for the CGS is faster, but not much faster.

To conclude, if we have access to distributed computing, the execution time of the PMMH algorithm can be reduced by several orders of magnitude if we use the GGS instead of the CGS. How much of an improvement we get depends on the extent of distributed computing that is available.

# 6 Discussion

In order to use LDA, one has to specify the number of topics, and the development of a principled method for doing so is not easy. Therefore, it is perhaps natural to ask why not abandon LDA altogether and use a Bayesian model in which the number of topics does not need to be specified. Indeed, topic models based on hierarchical Dirichlet processes (HDP's) (Teh et al., 2006) have precisely this feature.

In the literature, one often sees statements to the effect that HDP's are a nonparametric extension of LDA. Actually, HDP-based models are not extensions of LDA models in any sense: for no hyperparameter setting does an HDP-based model reduce to an LDA model. The models are simply different, and whether HDP-based models should be preferred is an open question: there are no formal studies that show that their performance is superior to that of LDA. While HDP-based models have proven to be quite useful, the various MCMC implementations have all been computationally very intensive. This problem precludes fitting these models on very large corpora, at least by MCMC (however note that in this regard, Stochastic Variational Inference (Hoffman et al., 2013) gives useful approximations even on massive corpora).

It has sometimes been proposed that HDP's be used to estimate the number of topics. This suggestion deals with a subtle point. Generally speaking, for a given corpus the number of topics is not a well-defined quantity, for the same reason that for a given finite set of points in Euclidean space, the number of clusters is not well defined. However, for an LDA model in which $T$ is fixed but unknown, $T$ is a well-defined parameter. On the other hand, HDP-based models inherently involve infinite mixtures, so using them to estimate a parameter $T$ known to be a fixed finite number is not sensible in the first place. Thus, these models should not be used to infer the number of topics for an LDA model, and in fact the literature has results that say that, generally speaking, Dirichlet process mixture models should not be used to estimate the number of components in a mixture (Miller and Harrison, 2013, 2014). These results should not be too surprising.

# Appendix

**Proof of Theorem 1**

We first establish the lower bound on $g_T(\boldsymbol{z}, \boldsymbol{z}')$. By the nature of the CGS, $g_T(\boldsymbol{z}, \boldsymbol{z}')$ can be expressed as

$$\begin{aligned} g_T(\boldsymbol{z}, \boldsymbol{z}') = p\big(z'_{11} \,\big|\, \boldsymbol{z}_1/\{z_{11}\}, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_D, \boldsymbol{w}\big) \times p\big(z'_{12} \,\big|\, z'_{11}, \boldsymbol{z}_1/\{z_{11}, z_{12}\}, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_D, \boldsymbol{w}\big) \times \cdots \\ \times p\big(z'_{1n_1} \,\big|\, z'_{11}, z'_{12}, \ldots, z'_{1,n_1-1}, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_D, \boldsymbol{w}\big) \times \\ \cdots\cdots \\ \times p\big(z'_{D1} \,\big|\, \boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_{D-1}, \boldsymbol{z}_D/\{z_{D1}\}, \boldsymbol{w}\big) \\ \times p\big(z'_{D2} \,\big|\, \boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_{D-1}, z'_{D1}, \boldsymbol{z}_D/\{z_{D1}, z_{D2}\}, \boldsymbol{w}\big) \times \cdots \\ \times p\big(z'_{Dn_D} \,\big|\, \boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_{D-1}, z'_{D1}, z'_{D2}, \ldots, z'_{D,n_D-1}, \boldsymbol{w}\big). \end{aligned}$$

From the expressions for the full conditional distributions in Section 2 (see (2.8)), we know that for each $d = 1, \ldots, D$, $i = 1, \ldots, n_d$, and $t = 1, \ldots, T$,

$$\begin{aligned} & p\big(z_{dit} = 1 \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}\big) \\ & \quad = \left(\frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t \cdot (-di)} + V\eta}\right)\left(\frac{n_{dt(-di)} + \alpha}{n_d - 1 + T\alpha}\right)\left[\sum_{t=1}^{T}\left(\frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t \cdot (-di)} + V\eta}\right)\left(\frac{n_{dt(-di)} + \alpha}{n_d - 1 + T\alpha}\right)\right]^{-1}. \end{aligned} \tag{A.1}$$

Applying the Cauchy-Schwartz inequality to the square of the sum term between the brackets in (A.1), we get

$$
\begin{aligned}
\left[\sum_{t=1}^{T} \frac{m_{\cdot tv(-di)}+\eta}{m_{\cdot t\cdot(-di)}+V\eta} \cdot \frac{n_{dt(-di)}+\alpha}{n_d-1+T\alpha}\right]^2 &\leq \left[\sum_{t=1}^{T}\left(\frac{m_{\cdot tv(-di)}+\eta}{m_{\cdot t\cdot(-di)}+V\eta}\right)^2\right]\left[\sum_{t=1}^{T}\left(\frac{n_{dt(-di)}+\alpha}{n_d-1+T\alpha}\right)^2\right] \\
&\leq \left[\sum_{t=1}^{T} 1^2\right]\left[\sum_{t=1}^{T}\left(\frac{n_{dt(-di)}+\alpha}{n_d-1+T\alpha}\right)^2\right] \\
&= T\left[\frac{\sum_{t=1}^{T}(n_{dt(-di)}+\alpha)^2}{(n_d-1+T\alpha)^2}\right] \\
&\leq T\left[\frac{\left(\sum_{t=1}^{T}(n_{dt(-di)}+\alpha)\right)^2}{(n_d-1+T\alpha)^2}\right] \\
&= T\left[\frac{(n_d-1+T\alpha)^2}{(n_d-1+T\alpha)^2}\right] = T.
\end{aligned}
$$

Hence from (A.1) we obtain

$$
\begin{aligned}
p\big(z_{dit}=1 \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}\big) &\geq \left(\frac{m_{\cdot tv(-di)}+\eta}{m_{\cdot t\cdot(-di)}+V\eta}\right)\left(\frac{n_{dt(-di)}+\alpha}{n_d-1+T\alpha}\right)\frac{1}{\sqrt{T}} \\
&\geq \left(\frac{\eta}{N-1+V\eta}\right)\left(\frac{\alpha}{n_d-1+T\alpha}\right)\frac{1}{\sqrt{T}} \\
&= \left(\frac{\eta}{N-1+V\eta}\right)\left(\frac{\sqrt{T}\alpha}{n_d-1+T\alpha}\right)\frac{1}{T}, \qquad \text{(A.2)}
\end{aligned}
$$

where (A.2) does not depend on $\boldsymbol{z}$ or $\boldsymbol{z}'$. Therefore,

$$
\begin{aligned}
g_T(\boldsymbol{z}, \boldsymbol{z}') &\geq \prod_{d=1}^{D}\prod_{i=1}^{n_d}\left(\frac{\eta}{N-1+V\eta}\right)\left(\frac{\sqrt{T}\alpha}{n_d-1+T\alpha}\right)\frac{1}{T} \\
&= \left(\frac{\eta}{N-1+V\eta}\right)^N\left[\prod_{d=1}^{D}\left(\frac{\sqrt{T}\alpha}{n_d-1+T\alpha}\right)^{n_d}\right]\left(\frac{1}{T}\right)^N = c_T\upsilon(\boldsymbol{z}'),
\end{aligned}
$$

where $c_T$ is given by

$$
c_T = \left(\frac{\eta}{N-1+V\eta}\right)^N\left[\prod_{d=1}^{D}\left(\frac{\sqrt{T}\alpha}{n_d-1+T\alpha}\right)^{n_d}\right].
$$

The lower bound on $g_T(\boldsymbol{z}, \boldsymbol{z}')$ now gives $\|g_T^m(\boldsymbol{z}_0, \cdot) - \nu_{\boldsymbol{z}\,|\,\boldsymbol{w}}(\cdot)\|_{\text{TV}} \leq (1-c_T)^m$ (see Theorem 3 of Athreya et al. (1996)).

**Proof of Lemma 1**

*Proof of Part 1* Under Conditions A2, and A3, all the assumptions of Theorem 1 of Andrieu and Roberts (2009) are satisfied, and we may apply that theorem to conclude that (4.21) is true.

*Proof of Part 2* Define

$$\gamma^{(m)}(t, \boldsymbol{\zeta}) = \frac{1}{m} \sum_{l=1}^{m} \frac{\nu_{\boldsymbol{z} \mid T, \boldsymbol{w}}(\boldsymbol{z}_l \mid t)}{g_t(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)},$$

in self-explanatory notation. Using the result in Theorem 1 that $g_t(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l) \geq c_t/t^N$, we get

$$\gamma^{(m)}(t, \boldsymbol{\zeta}) = \frac{1}{m} \sum_{l=1}^{m} \frac{\nu_{\boldsymbol{z} \mid T, \boldsymbol{w}}(\boldsymbol{z}_l \mid t)}{g_t(\boldsymbol{z}_{l-1}, \boldsymbol{z}_l)} \leq \frac{1}{m} \sum_{l=1}^{m} \frac{1}{c_t/t^N} = \frac{t^N}{c_t} \leq \max_{t \in \mathcal{T}} \frac{t^N}{c_t} := \gamma_*,$$

where $\gamma_*$ on the right side does not involve $m$ and is finite. Define

$$\varrho_{T,\boldsymbol{\zeta}}^{(m)}(t, \boldsymbol{\zeta}) = \varrho_T(t) q_{\boldsymbol{\zeta}}^{(m,t)}(\boldsymbol{\zeta}) \gamma^{(m)}(t, \boldsymbol{\zeta}) \qquad \text{for all } t \in \mathcal{T} \text{ and } \boldsymbol{\zeta} \in \mathcal{Z}_t^m.$$

Then, $\varrho_{T,\boldsymbol{\zeta}}^{(m)}$ is a probability measure on $\mathcal{S} = \cup_{t \in \mathcal{T}} \big(\{t\} \times \mathcal{Z}_t^m\big)$. This is because

$$\begin{aligned} \sum_{t \in \mathcal{T}} \sum_{\boldsymbol{\zeta} \in \mathcal{Z}_t^m} \varrho_T(t) q_{\boldsymbol{\zeta}}^{(m,t)}(\boldsymbol{\zeta}) \gamma^{(m)}(t, \boldsymbol{\zeta}) &= \sum_{t \in \mathcal{T}} \varrho_T(t) \sum_{\boldsymbol{\zeta} \in \mathcal{Z}_t^m} q_{\boldsymbol{\zeta}}^{(m,t)}(\boldsymbol{\zeta}) \gamma^{(m)}(t, \boldsymbol{\zeta}) \\ &= \sum_{t \in \mathcal{T}} \varrho_T(t) \\ &= 1, \end{aligned} \tag{A.3}$$

where the second equality in (A.3) follows because $\sum_{\boldsymbol{\zeta} \in \mathcal{Z}_t^m} q_{\boldsymbol{\zeta}}^{(m,t)}(\boldsymbol{\zeta}) \gamma^{(m)}(t, \boldsymbol{\zeta}) = E(\gamma^{(m)}(t, \boldsymbol{\zeta}))$, and $E(\gamma^{(m)}(t, \boldsymbol{\zeta})) = 1$, which can be proved in the same way we proved that $E\big(\tilde{\pi}_X^{(m)}(x)\big) = \pi_X(x)$—see (4.3). Let $\gamma_*^{(m)} = \max_{t,\boldsymbol{\zeta}} \gamma^{(m)}(t, \boldsymbol{\zeta})$. Clearly $\gamma_*^{(m)} \leq \gamma_*$. Applying the inequality at the end of the proof of Theorem 8 of Andrieu and Roberts (2009) in our LDA context, we see that for all $(t', \boldsymbol{\zeta}')$ and $(t, \boldsymbol{\zeta})$ we have

$$\begin{aligned} P^{m,n_0}(t', \boldsymbol{\zeta}'; t, \boldsymbol{\zeta}) &\geq \delta \big(\gamma_*^{(m)}\big)^{-n_0} \varrho_T(t)\, q_{\boldsymbol{\zeta}}^{(m,t)}(\boldsymbol{\zeta})\, \gamma^{(m)}(t, \boldsymbol{\zeta}) \\ &\geq \delta \gamma_*^{-n_0}\, \varrho_T(t)\, q_{\boldsymbol{\zeta}}^{(m,t)}(\boldsymbol{\zeta})\, \gamma^{(m)}(t, \boldsymbol{\zeta}) \\ &= \kappa\, \varrho_{T,\boldsymbol{\zeta}}^{(m)}(t, \boldsymbol{\zeta}), \end{aligned} \tag{A.4}$$

where $\kappa := \delta \gamma_*^{-n_0}$, and does not depend on $m$. Details on why the first inequality in (A.4) follows from the inequality at the end of the proof of Theorem 8 of Andrieu and Roberts (2009) are given in Chen (2015). From (A.4) we conclude that the PMMH algorithm is uniformly ergodic:

$$\big\| P^{m,n}(t, \boldsymbol{\zeta}; \cdot, \cdot) - \tilde{\nu}_{T,\boldsymbol{\zeta} \mid \boldsymbol{w}}^{(m)}(\cdot, \cdot) \big\| \leq (1 - \kappa)^{[n/n_0]}. \qquad \square$$

**Proof of Theorem 3**

*Proof of Part 1* For any integers $m, n$, and any initial points $t_0 \in \mathcal{T}$ and $\boldsymbol{\zeta}_0 \in \mathcal{Z}_t^m$, we have

$$\left\|P^{m,n}(t_0, \boldsymbol{\zeta}_0; \cdot, \cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\| \le \left\|P^{m,n}(t_0, \boldsymbol{\zeta}_0; \cdot, \cdot) - \tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\| + \left\|\tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\|. \tag{A.5}$$

Part 2 of Lemma 1 states that there exists a constant $\kappa \in (0, 1)$, which does not depend on $m$, such that $\left\|P^{m,n}(t_0, \boldsymbol{\zeta}_0; \cdot, \cdot) - \tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\| \le (1 - \kappa)^{[n/n_0]}$ for all $m$ and $n$. Therefore if $N(\epsilon) \ge n_0 \log(\epsilon/2)/ \log(1 - \kappa)$, then

$$\left\|P^{m,n}(t_0, \boldsymbol{\zeta}_0; \cdot, \cdot) - \tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\| \le \epsilon/2 \qquad \text{for all } n \ge N(\epsilon), \text{ all } m. \tag{A.6}$$

From the definitions of $\tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}$ and $\nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}$ given in (4.19), we have

$$\begin{aligned}
\left\|\tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\| &= \sup_{C \in \mathcal{B}_{\mathcal{S}}} \left| \sum_{(t,\boldsymbol{\zeta}) \in C} \left( \tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)\, q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) - \nu_{T\,|\,\boldsymbol{w}}(t)\, q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \right) \right| \\
&\le \sup_{C \in \mathcal{B}_{\mathcal{S}}} \sum_{(t,\boldsymbol{\zeta}) \in C} \left| \tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)\, q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) - \nu_{T\,|\,\boldsymbol{w}}(t)\, q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \right| \\
&= \sum_{t \in \mathcal{T}} \sum_{\boldsymbol{\zeta} \in \mathcal{Z}_t^m} \left| \tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t)\, q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) - \nu_{T\,|\,\boldsymbol{w}}(t)\, q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \right| \\
&= \sum_{t \in \mathcal{T}} \sum_{\boldsymbol{\zeta} \in \mathcal{Z}_t^m} \left| \tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t) - \nu_{T\,|\,\boldsymbol{w}}(t) \right| q^{(m,t)}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta}) \\
&\to 0 \qquad \text{as } m \to \infty,
\end{aligned}$$

where the convergence statement follows from (4.18). Hence, there exists an integer $M(\epsilon)$ such that

$$\left\|\tilde{\nu}^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\| < \epsilon/2 \qquad \text{for all } m > M(\epsilon). \tag{A.7}$$

Combining (A.5), (A.6), and (A.7), we see that for $m \ge M(\epsilon)$ and $n \ge N(\epsilon)$,

$$\left\|P^{m,n}(t_0, \boldsymbol{\zeta}_0; \cdot, \cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot, \cdot)\right\| \le \epsilon/2 + \epsilon/2 = \epsilon.$$

*Proof of Part 2* For any $t \in \mathcal{T}$ we have

$$\left|\mu^{m,n}_T(t) - \nu_{T\,|\,\boldsymbol{w}}(t)\right| = \left|P^{m,n}(t_0, \boldsymbol{\zeta}_0; t, \mathcal{Z}_t^m) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t, \mathcal{Z}_t^m)\right| \to 0 \qquad \text{as } m, n \to \infty,$$

where the convergence statement follows from Part 1 of the theorem. Part 2 of the theorem follows from the fact that $\mathcal{T}$ is finite.

*Proof of Part 3* For any $t \in \mathcal{T}$ and $\boldsymbol{z} \in \mathcal{Z}_t$, take $B = \mathcal{Z}_t^{m-1} \times \{\boldsymbol{z}\}$. We then have $\mu_{T,\boldsymbol{z}}^{m,n}(t, \boldsymbol{z}) = P^{m,n}(t_0, \boldsymbol{\zeta}_0; t, B)$, so

$$\begin{aligned}\left|\mu_{T,\boldsymbol{z}}^{m,n}(t,\boldsymbol{z}) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t,\boldsymbol{z})\right| &= \left|P^{m,n}(t_0,\boldsymbol{\zeta}_0;t,B) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t,\boldsymbol{z})\right| \\ &\le \left|P^{m,n}(t_0,\boldsymbol{\zeta}_0;t,B) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t,B)\right| + \left|\nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t,B) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t,\boldsymbol{z})\right|. \qquad \text{(A.8)}\end{aligned}$$

Now

$$\left|P^{m,n}(t_0,\boldsymbol{\zeta}_0;t,B) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t,B)\right| \le \left\|P^{m,n}(t_0,\boldsymbol{\zeta}_0;\cdot,\cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(\cdot,\cdot)\right\| \to 0 \qquad \text{as } m,n \to \infty,$$

where the convergence statement follows from Part 1 of the theorem, and we are slightly abusing notation by using $\nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}$ to denote both a probability mass function and a probability measure.

Consider the second term on the right side of (A.8). From the definition (4.19), we have

$$\nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t,B) = \nu_{T\,|\,\boldsymbol{w}}(t) q_{\boldsymbol{\zeta}}^{(m,t)}\big(\boldsymbol{z}_1 \in \mathcal{Z}_t, \ldots, \boldsymbol{z}_{m-1} \in \mathcal{Z}_t,\ \boldsymbol{z}_m \in \{\boldsymbol{z}\} \,\big|\, \boldsymbol{z}_0\big) = \nu_{T\,|\,\boldsymbol{w}}(t) g_t^m(\boldsymbol{z}_0, \boldsymbol{z}).$$

Therefore,

$$\begin{aligned}\left|\nu^{(m)}_{T,\boldsymbol{\zeta}\,|\,\boldsymbol{w}}(t,B) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t,\boldsymbol{z})\right| &= \left|\nu_{T\,|\,\boldsymbol{w}}(t) g_t^m(\boldsymbol{z}_0,\boldsymbol{z}) - \nu_{T\,|\,\boldsymbol{w}}(t)\nu_{\boldsymbol{z}\,|\,T,\boldsymbol{w}}(\boldsymbol{z}\,|\,t)\right| \\ &= \nu_{T\,|\,\boldsymbol{w}}(t)\left|g_t^m(\boldsymbol{z}_0,\boldsymbol{z}) - \nu_{\boldsymbol{z}\,|\,T,\boldsymbol{w}}(\boldsymbol{z}\,|\,t)\right| \\ &\le \left|g_t^m(\boldsymbol{z}_0,\boldsymbol{z}) - \nu_{\boldsymbol{z}\,|\,T,\boldsymbol{w}}(\boldsymbol{z}\,|\,t)\right| \\ &\to 0 \qquad \text{as } m \to \infty,\end{aligned}$$

where the convergence statement follows from Theorem 1 and is uniform in $\boldsymbol{z}_0$. Thus, as $m, n \to \infty$, both terms on the right side of (A.8) converge to $0$, and this proves Part 3 of the theorem since $\mathcal{S}$ is finite.

*Proof of Part 4* Let $A \in \mathcal{B}_{\mathbb{S}^t_{V-1}}$ and $B \in \mathcal{B}_{\mathbb{S}^D_{t-1}}$ (here $\mathcal{B}_{\mathbb{S}^t_{V-1}}$ is the Borel $\sigma$-field on the $t$-fold product of the $(V-1)$-dimensional simplex $\mathbb{S}_{V-1}$, and $\mathcal{B}_{\mathbb{S}^D_{t-1}}$ is defined analogously). We have

$$\begin{aligned}&\left|\mu^{m,n}_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}}(t,\boldsymbol{z},A,B) - \nu_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}\,|\,\boldsymbol{w}}(t,\boldsymbol{z},A,B)\right| \\ &\qquad = \left|\mu^{m,n}_{T,\boldsymbol{z}}(t,\boldsymbol{z})\,\nu_{\boldsymbol{\beta},\boldsymbol{\theta}\,|\,T,\boldsymbol{z},\boldsymbol{w}}(A,B\,|\,t,\boldsymbol{z}) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t,\boldsymbol{z})\,\nu_{\boldsymbol{\beta},\boldsymbol{\theta}\,|\,T,\boldsymbol{z},\boldsymbol{w}}(A,B\,|\,t,\boldsymbol{z})\right| \\ &\qquad = \left|\mu^{m,n}_{T,\boldsymbol{z}}(t,\boldsymbol{z}) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t,\boldsymbol{z})\right|\nu_{\boldsymbol{\beta},\boldsymbol{\theta}\,|\,T,\boldsymbol{z},\boldsymbol{w}}(A,B\,|\,t,\boldsymbol{z}) \\ &\qquad \le \left|\mu^{m,n}_{T,\boldsymbol{z}}(t,\boldsymbol{z}) - \nu_{T,\boldsymbol{z}\,|\,\boldsymbol{w}}(t,\boldsymbol{z})\right| \\ &\qquad \to 0 \qquad \text{as } m,n \to \infty,\end{aligned}$$

where the first equality is from (2.5), and the convergence statement is from Part 3 of the theorem, and is uniform in $A$ and $B$. This proves Part 4 of the theorem. $\square$

Part 1 of Theorem 3 is similar to Theorem 6 and Corollary 7 of Andrieu and Roberts (2009); however, there are major differences between our proof and theirs, which we now discuss. First, the proofs in Andrieu and Roberts (2009) are at a general level, whereas ours is particular to the LDA model. As a consequence, whereas their proofs are very technical and involved, we are able to take advantage of certain features of the LDA model which enable a relatively simple proof. A second difference is as follows. Using our notation, the goal is to bound the error $\left\| P^{m,n}(t_0, \boldsymbol{\zeta}_0; \cdot, \cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta} \mid \boldsymbol{w}}(\cdot, \cdot) \right\|$. For Andrieu and Roberts (2009), the rates at which $m$ and $n$ go to infinity are related, whereas in our result, $m$ and $n$ can go to infinity in an arbitrary manner. This is important in our application, because the convergence rate of the CGS is so high that it is sensible to take $m$ to be much smaller than $n$, so we prefer to not have any restrictions on the relationship between $m$ and $n$ as these go to infinity. Lastly, in Andrieu and Roberts (2009) there are restrictions on the starting points: $t_0$ may be chosen freely, but the choice of $\boldsymbol{\zeta}_0$ depends on several quantities, including $t_0$. Moreover, the dependence is very complicated, and in high-dimensional situation such as what we have in the LDA model, it is very difficult to determine what are the acceptable choices of $\boldsymbol{\zeta}_0$. In our theorem, there are no restrictions at all on the starting points $t_0$ and $\boldsymbol{\zeta}_0$.

# Supplementary Materials

**R Code and Data** The supplemental files for this article include files containing R code and data for reproducing all the empirical studies in the paper. The Readme file contained in the zip file gives a description of all the other files in the archive. (lda-ntopics-code.zip, zip archive)

**Appendix** The supplemental files include an Appendix which gives the following: (i) a derivation of an expression for the conditional distributions needed to run the Collapsed Gibbs Sampler of Griffiths and Steyvers (2004), (ii) a version of Theorem 3 without the condition that $\mathcal{T}$ is finite, and (iii) additional simulations to compare the PMMH algorithm with other methods for choosing the number of topics. (lda-ntopics-supp.pdf)


## Acknowledgments

We are grateful to the two referees for their very helpful constructive criticism.

# References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* **37** 697–725.

Asuncion, A., Welling, M., Smyth, P. and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09, AUAI Press, Arlington, Virginia, United States.

Athreya, K. B., Doss, H. and Sethuraman, J. (1996). On the convergence of the Markov chain simulation method. *The Annals of Statistics* **24** 69–100.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* **55** 77–84.

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

Chen, Z. (2015). *Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modelling*. Ph.D. thesis, University of Florida.

Chen, Z. and Doss, H. (2018). Supplement to "Inference for the number of topics in the latent Dirichlet allocation model via Bayesian mixture modelling".

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90** 1313–1321.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* **40** 1–18. `http://www.jstatsoft.org/v40/i08/`.

Eddelbuettel, D. and Sanderson, C. (2014). Rcpparmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis* **71** 1054–1063. `http://dx.doi.org/10.1016/j.csda.2013.02.005`.

Fernández, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: A Bayesian approach. *Journal of the Royal Statistical Society,* Series B **64** 805–826.

George, C. P. (2015). *Latent Dirichlet Allocation: Hyperparameter Selection and Applications to Electronic Discovery*. Ph.D. thesis, University of Florida.

George, C. P. and Doss, H. (2018). Principled selection of hyperparameters in the latent Dirichlet allocation model. *Journal of Machine Learning Research* **18**, No. 162, 1–38.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* **101** 5228–5235.

Griffiths, T. L., Steyvers, M., Blei, D. M. and Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Advances in Neural Information Processing Systems*.

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

Liu, J. S., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems 26* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger, eds.). Curran Associates, Inc., 199–206.

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research* **15** 3333–3370.

Newman, D., Asuncion, A., Smyth, P. and Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research* **10** 1801–1828.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society,* Series B **56** 3–48.

Nguyen, X. (2015). Posterior contraction of the population polytope in finite admixture models. *Bernoulli* **21** 618–646.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society,* Series B **59** 731–792.

Tang, J., Meng, Z., Nguyen, X., Mei, Q. and Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*.

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

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06, ACM, New York, NY, USA.

Wallach, H. M., Mimno, D. and McCallum, A. (2009a). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems* **22** 1973–1981.

Wallach, H. M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009b). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM.

Wolpert, R. L. and Schmidler, S. C. (2012). $\alpha$-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statistica Sinica* **22** 1233–1251.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92** 937–950.