Practical Collapsed Variational Bayes Inference for Hierarchical Dirichlet Process

Issei Sato
The University of Tokyo, Japan
sato@r.dl.itc.u-tokyo.ac.jp

Kenichi Kurihara
Google
kenichi.kurihara@gmail.com

Hiroshi Nakagawa
The University of Tokyo, Japan
n3@dl.itc.u-tokyo.ac.jp

ABSTRACT

We propose a novel collapsed variational Bayes (CVB) inference for the hierarchical Dirichlet process (HDP). While the existing CVB inference for the HDP variant of latent Dirichlet allocation (LDA) is more complicated and harder to implement than that for LDA, the proposed algorithm is simple to implement, does not require variance counts to be maintained, does not need to set hyper-parameters, and has good predictive performance.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Nonparametric statistics

General Terms

Algorithms

Keywords

Latent Dirichlet Allocation, Nonparametric Bayes, Hierarchical Dirichlet Process, Collapsed Variational Bayes Inference

1. INTRODUCTION

Probabilistic models with latent variables have attracted attention in knowledge discovery and data mining because of their power and flexibility in modeling real world phenomena. The goals of such probabilistic modeling are to capture the underlying generation mechanism and the statistical relationships of data and to make predictions yet to be observed.

Latent Dirichlet allocation (LDA) [1] has been one of the most studied probabilistic latent variable models in the last decade. LDA was originally used to model the co-occurrence of words by using latent variables called topics where a document is represented as a "bag of words," meaning that the order of words is ignored. Now we have a wide variety of topic models in many fields: modeling authors and topics [2, 3], entities and topics [4], document and citations [5], hypertext and topics [6], annotated biological figures and topics [7], the dynamics of documents and topics [8, 9, 10, 11, 12], power-law and topics[13], and a partially labeled data [14]. A non-probabilistic formulation also have been proposed called a conditional topical model [15]. Topic modeling has also been applied to information retrieval [16, 17], measuring redundancy by LDA-based submodular function[18], augmenting social networks [19], measuring scholarly impact [20], predicting legislative roll calls [21], and recommendationing scientific article [22].

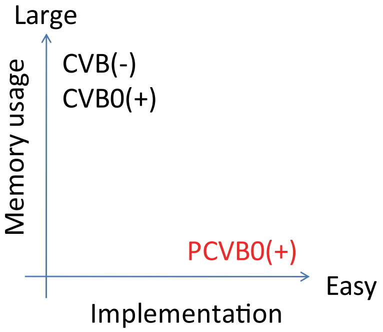

Figure 1: Overview of proposed and related inferences for HDP-LDA. CVB and CVB0 indicates second order and zero order approximation of the CVB inference. PCVB0 indicates the proposed (practical CVB0) inference. $(+)$ and $(-)$ mean that the predictive performance of $(+)$ is better than that of $(-)$ in terms of perplexity.

An important LDA extension is a nonparametric one using the hierarchical Dirichlet process (HDP) [23]. In actual applications, we do not know the number of topics a priori and HDP-LDA can determine this automatically. The purpose of this paper is to explore an efficient deterministic inference for HDP-LDA.

There are two common inferences for HDP-LDA: collapsed Gibbs sampler [23] and collapsed variational Bayes (CVB) inference [24]. The CVB inference is a deterministic algorithm for learning HDP-LDA that was originally developed for LDA [25]. The CVB inference is a variational approximation improved by collapsing parameters, which indicates integrating out the parameters as in a collapsed Gibbs sampler. The CVB inference was proposed for solving some problems of a sampling method; for example, sampling often requires many iterations and its averaging of topic-dependent quantities on the basis of samples is inefficient for estimating test data.

Teh et al. proposed the CVB inference for LDA [25] and HDP-LDA[24] by using a second-order Taylor expansion. Asuncion et al. [26, 27] proposed another approximation by using only the zero-order information, called the CVB0 inference. The CVB0 inference for LDA [26] is computationally faster and requires a smaller

KDD'12, August 12–16, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1462-6/12/08... $15.00.

memory than the CVB inferences for LDA [25] since it does not require calculating and holding variance counts, and converges more quickly than the collapsed Gibbs sampler since it is deterministic.. Furthermore, their empirical results suggest that the CVB0 inference learns models that are as good or better (predictively) than those learned by the collapsed Gibbs sampler.

Problems: It is easy to apply the CVB0 inference to HDP-LDA. However, the naive CVB0 inference for HDP-LDA does not provide the same efficiency as LDA because variance counts are used for estimating other statistics, e.g. $\mathbb{G}[\alpha_0\pi_k]$ and $\mathbb{G}[\beta_0\tau_{w_{d,i}}]$ in Eq.(19) (see the work of Teh et al.[24] for details). Calculating these statistics is extremely complicated and so requires large computational cost. Moreover, we needed to set hyper parameters for each dataset; however, we do not wish to tune them for each dataset in practice.

Contributions: We developed a novel CVB0 inference for HDP-LDA. We intended that the proposed algorithm be as simple and easy to implement as the previous algorithm for LDA [25, 26], and outperform the previous CVB inference for HDP-LDA [24]. The proposed algorithm has the following properties:

1. The proposed CVB0 inference for HDP-LDA does not require the variance counts.
2. Our algorithm does not require the setting of hyper-parameters.

These properties lead to a simple algorithm for learning HDP-LDA that will be useful to researchers in many scientific fields when they apply HDP-LDA to their problems. An overview of the proposed and related algorithms of HDP-LDA is shown in Fig.1. We define the notation symbols used in this paper in Table 1.

The remainder of this paper is organized as follows. Sections 2 and 3 overview LDA and HDP-LDA, respectively. Section 4 explains the CVB / CVB0 inference for HDP-LDA. Section 5 proposes the proposed algorithm. Section 6 evaluates algorithms in three kinds of experiments: document modeling, nearest neighbor search, and social network analysis.

2. OVERVIEW OF LDA

The following generative process is assumed with LDA.

First, document-topic distribution $\boldsymbol{\theta}_d$ and topic-word distribution $\boldsymbol{\phi}_k$ are generated by

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})\ (d = 1, \cdots, N), \quad \boldsymbol{\phi}_k \sim \text{Dir}(\boldsymbol{\beta})\ (k = 1, \cdots, K), \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$ is a K-dimensional vector and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_V)$ is a V-dimensional vector.

For each document d, generate the i-th topic $z_{d,i}$ and word $w_{d,i}$:

$$z_{d,i} \sim \text{Multi}(\boldsymbol{\theta}_d), \quad w_{d,i} \sim \text{Multi}(\boldsymbol{\phi}_{z_{d,i}}). \tag{2}$$

Wallach et al. [28] explored the effects of choosing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in LDA. The symmetric Dirichlet priors indicate $\alpha_k = \alpha_0/K$ for all k and $\beta_v = \beta_0/V$ for all v. In the asymmetric Dirichlet priors, they used a nonuniform base measure $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ and $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_V)$ instead of $1/K$ and $1/V$, i.e., $\alpha_k = \alpha_0\pi_k$ and $\beta_v = \beta_0\tau_v$. They found in Markov chain Monte Carlo (MCMC) simulations that using asymmetric $\boldsymbol{\alpha}$ and symmetric $\boldsymbol{\beta}$ results in better predictive performance of held-out documents. We investigate the effects of the choice of prior over the topic-word distributions in the CVB/CVB0 inference.

Table 1: Notation Table

Symbol	Definition
N	Total number of documents
V	Vocabulary size
K	Total number of topics
T	Truncation level of stick-breaking process
d	Document index, i.e., $d = 1, \cdots, N$
v	Vocabulary index, i.e., $v = 1, \cdots, V$
n_d	Number of words in documents d
$n_{d,k}$	Number of times topic k appears in document d
$n_{k,v}$	Number of times word v appears in topic k
$w_{d,i}$	i-th word in document d
$\boldsymbol{w}$	Set of all words, i.e., $\{\boldsymbol{w}_d\}_{d=1}^N$
$z_{d,i}$	Assigned topic at i-th word in document d
$\boldsymbol{z}$	Set of all latent topic variables, i.e., $\{\boldsymbol{z}_d\}_{d=1}^N$
$\theta_{d,k}$	Probability of topic k appearing in document d
$\boldsymbol{\theta}_d$	K-dimensional probability vector
$\phi_{k,v}$	Probability of word v appearing in topic k
$\boldsymbol{\phi}_k$	V-dimensional probability vector
$\boldsymbol{\alpha}$	K-dimensional vector
α_0	$= \sum_k \alpha_k$, or concentration parameter of DP
γ_0	concentration parameter of DP or SBP
$\boldsymbol{\beta}$	V-dimensional vector
β_0	$= \sum_v \beta_v$
$\mathbb{E}[x]$	Expectation of x
$\mathbb{G}[x]$	Geometric expectation $\exp(\mathbb{E}[\log x])$
$\mathbb{V}[x]$	Variance $\mathbb{E}[x^2] - \mathbb{E}[x]^2$
$\Gamma(x)$	Gamma function
$\Psi(x)$	Digamma function.
Beta$(\cdot)$	Beta distribution
Multi$(\cdot)$	Multinomial distribution
Dir$(\cdot)$	Dirichlet distribution
DP$(\cdot)$	Dirichlet process
SBP$(\cdot)$	Stick-breaking process formulated as Eq.(9)
TSBP$(\cdot)$	Truncated SBP formulated as Eq.(12)
KL$[\cdot\|\|\cdot]$	Kullback-Leibler divergence

3. OVERVIEW OF HDP-LDA

We have a nonparametric Bayes model of LDA by using the HDP, called HDP-LDA [23]. The generation process of HDP-LDA is

$$\begin{aligned} G_0 &\sim DP(\gamma_0, Dir(\boldsymbol{\beta})), && (3)\\ G_d &\sim DP(\alpha_0, G_0), && (4)\\ \phi_{z_{d,i}} &\sim G_d, && (5)\\ w_{d,i} &\sim Mult(\phi_{z_{d,i}}). && (6) \end{aligned}$$

Using a stick-breaking process (SBP), G_0 and G_d are represented as sums of point masses given by

$$G_0 = \sum_{k=1}^{\infty} \boldsymbol{\pi}_k \delta_{\boldsymbol{\phi}_k},\ G_d = \sum_{k=1}^{\infty} \boldsymbol{\theta}_{d,k} \delta_{\boldsymbol{\phi}_k}, \tag{7}$$

$$\boldsymbol{\theta}_d \sim \text{DP}(\alpha_0, \boldsymbol{\pi}),\ \phi_k \sim \text{Dir}(\boldsymbol{\beta}), \tag{8}$$

$$\pi_k = \tilde{\pi}_k \prod_{l=1}^{k-1} (1 - \tilde{\pi}_l),\ \tilde{\pi}_k \sim \text{Beta}(1, \gamma_0). \tag{9}$$

The construction of $\boldsymbol{\pi}$ in Eq.(9) is called the stick-breaking construction and is denoted by $\boldsymbol{\pi} \sim \text{SBP}(\gamma_0)$.

We describe a truncated stick-breaking process (TSBP) construction of HDP-LDA because we truncate an infinite number of components to apply the CVB inference to HDP. The TSBP has the advantage of being able to estimate running time per iteration, which is a useful property in actual applications. Using the TSB representation, G_0 and G_j are represented as sums of point masses given by

$$G_0 = \sum_{k=1}^{T} \boldsymbol{\pi}_k \delta_{\boldsymbol{\phi}_k},\ G_j = \sum_{k=1}^{T} \boldsymbol{\theta}_{j,k} \delta_{\boldsymbol{\phi}_k}, \tag{10}$$

$$\boldsymbol{\theta}_d \sim Dir(\alpha_0 \boldsymbol{\pi}),\ \phi_k \sim \mathrm{Dir}(\boldsymbol{\beta}), \tag{11}$$

$$\pi_k = \tilde{\pi}_k \prod_{l=1}^{k-1} (1 - \tilde{\pi}_l),\ \tilde{\pi}_k \sim Beta(1, \gamma_0),\ \tilde{\pi}_T = 1. \tag{12}$$

Note that since T is not the number of topics but the truncation level in HDP-LDA, we can estimate the effective number of topics K if we set $T > K$. The mixing proportions $\boldsymbol{\pi}$ distributed according to the TSBP, denoted by $\boldsymbol{\pi} \sim$ TSBP(γ_0), decreases exponentially fast at the tail of the distribution. This motivates the use of the TSBP, which is easier to handle than the infinite case. Ishwaran and James [29] give a bound for the error introduced by truncating the SBP; they showed that truncating the number of mixture components to a moderate level is sufficient to successfully approximate the DP.

In the collapsed Gibbs sampler for HDP-LDA, the topic assignment of the i-th word in document d using α_0 and $\boldsymbol{\pi}$ is given by

$$p(z_{d,i} = k | w_{d,i} = v, \boldsymbol{z}^{-d,i}, \boldsymbol{w}^{-d,i}) = \frac{n_{d,k}^{-d,i} + \alpha_0 \pi_k}{n_d^{-d,i} + \alpha_0} \frac{n_{k,v}^{-d,i} + \beta_v}{n_{k,\cdot}^{-d,i} + \beta_0}, \tag{13}$$

where the superscription "$-d,i$" denotes the corresponding variables or counts with $w_{d,i}$ and $z_{d,i}$ excluded, e.g., $\boldsymbol{w}^{-d,i} = \boldsymbol{w} \backslash \{w_{d,i}\}$, $\boldsymbol{z}^{-d,i} = \boldsymbol{z} \backslash \{z_{d,i}\}$, and $n_{k,v}^{-d,i}$ is the number of observations of word v assigned to topic k leaving out $z_{d,i}$.

We need to use the Chinese restaurant process (CRP) procedure for sampling α_0 and $\boldsymbol{\pi}$, where a document, word, and topic represent respectively a restaurant, a customer, and a dish served at a table. The procedure in which the i-th word in document d is assigned to topic k indicates that the i-th customer sits at a table serving dish k in restaurant d. However, sampling the seating arrangements in this CRP procedure is time consuming. Fortunately, we only need the table numbers serving topic k at restaurant d, denoted by $m_{d,k}$, not the seating arrangements of customers. Therefore, we have alternates consisting of three sampling stages: (1) sampling the topic assignments $z_{d,i}$, (2) sampling the number of tables $m_{d,k}$, and (3) sampling α_0 and $\boldsymbol{\pi}$.

The number of tables serving topic k in document d, i.e., $m_{d,k}$, is sampled using the Stirling numbers of the first kind denoted by $str(\cdot)$ [23]:

$$p(m_{d,k} = m | n_{d,k}, \alpha_0, \pi_k) = str(n_{d,k}, m)(\alpha_0 \pi_k)^m \frac{\Gamma(\alpha_0 \pi_k)}{\alpha_0 \pi_k + n_{d,k}}. \tag{14}$$

The conditional posterior for the mixing proportions $\boldsymbol{\pi}$ is also given as a stick-breaking construction as follows.

$$\tilde{\pi}_k \sim Beta(1 + m_{\cdot,k}, \gamma_0 + \sum_{l=k+1}^{T} m_{\cdot,l}),\ \tilde{\pi}_T = 1, \tag{15}$$

$$\pi_1 = \tilde{\pi}_1,\ \pi_k = \tilde{\pi}_k \prod_{l=1}^{k-1} (1 - \tilde{\pi}_l). \tag{16}$$

Algorithm 1 PCVB0 inference for HDP-LDA

```
for each iteration do
    for each document $d$ do
        for each word $w_{d,i}$ do
            Update $q(z_{d,i})$ by using Eq.(35).
        end for
    end for
    Update $q(\tilde{\pi}_k)$ by using Eq.(24)
    Update $\alpha_0$ by using Eq.(29).
    Update $\gamma_0$ by using Eq.(33).
    Update $\beta 0$ and $\tau_v$ by using Eqs.(30) and Eqs.(32).
end for
```

The concentration parameter, α_0, can be estimated by auxiliary variable sampling [23, 30].

4. CVB/CVB0 INFERENCE FOR HDP-LDA

Teh et al. [24] proposed the CVB inference for HDP-LDA. They marginalize over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ in the CVB inference as in a collapsed Gibbs sampler. They assumed an asymmetric Dirichlet prior over the topic-word distributions, i.e., $\boldsymbol{\beta} = \beta_0 \boldsymbol{\tau}$. By integrating out $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$, a join distribution over $\boldsymbol{w}$ and $\boldsymbol{z}$ is given by

$$p(\boldsymbol{z}, \boldsymbol{w} | \alpha_0, \beta_0, \boldsymbol{\pi}, \boldsymbol{\tau}) = \left[\prod_{d=1}^{N} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_d)} \prod_{k=1}^{T} \frac{\Gamma(\alpha_0 \pi_k + n_{d,k})}{\Gamma(\alpha_0 \pi_k)} \right] \left[\prod_{k=1}^{T} \frac{\Gamma(\beta_0)}{\Gamma(\beta_0 + n_{k,\cdot})} \prod_{v=1}^{V} \frac{\Gamma(\beta_0 \tau_v + n_{k,v})}{\Gamma(\beta_0 \tau_v)} \right]. \tag{17}$$

The CVB inference for $q(z_{d,i} = k)$ is given by

$$q(z_{d,i} = k) \propto \frac{\mathbb{E}[n_{k,w_{d,i}}^{-d,i}] + \mathbb{G}[\beta_0 \tau_{w_{d,i}}]}{\mathbb{E}[n_{k,\cdot}^{-d,i}] + \mathbb{G}[\beta_0]} \frac{\mathbb{E}[n_{d,k}^{-d,i}] + \mathbb{G}[\alpha_0 \pi_k]}{n_d^{-d,i} + \mathbb{G}[\alpha_0]} \exp\left(- \frac{\mathbb{V}[n_{k,w_{d,i}}^{-d,i}]}{2(\mathbb{E}[n_{k,w_{d,i}}^{-d,i}] + \mathbb{G}[\beta_0 \tau_{w_{d,i}}])^2} + \frac{\mathbb{V}[n_{k,\cdot}^{-d,i}]}{2(\mathbb{E}[n_{k,\cdot}^{-d,i}] + \mathbb{G}[\beta_0])^2} \right) \exp\left(- \frac{\mathbb{V}[n_{d,k}^{-d,i}]}{2(\mathbb{E}[n_{d,k}^{-d,i}] + \mathbb{G}[\alpha_0 \pi_k])^2} \right), \tag{18}$$

where "-d,i" denotes subtracting $q(z_{d,i} = k)$ and $q(z_{d,i} = k)(1 - q(z_{d,i} = k))$. The calculations of $\mathbb{G}[\alpha_0 \pi_k]$ and $\mathbb{G}[\beta_0 \tau_{w_{d,i}}]$ are extremely complicated and so require large computational cost (see the work of Teh et al. [24] for details).

Asuncion et al. [26, 27] showed the usefulness of an approximation using only zero-order information, called the CVB0 inference, in LDA. It is easy to apply the CVB0 inference to HDP-LDA. The update using only zero-order information for HDP-LDA is given by

$$q(z_{d,i} = k) \propto \frac{\mathbb{E}[n_{k,w_{d,i}}^{-d,i}] + \mathbb{G}[\beta_0 \tau_{w_{d,i}}]}{\mathbb{E}[n_{k,\cdot}^{-d,i}] \mathbb{G}[\beta_0]} \frac{\mathbb{E}[n_{d,k}^{-d,i}] + \mathbb{G}[\alpha_0 \pi_k]}{n_d^{-d,i} + \mathbb{G}[\alpha_0]}. \tag{19}$$

However, the problem is that this CVB0 inference also requires the calculations of $\mathbb{G}[\alpha_0 \pi_k]$ and $\mathbb{G}[\beta_0 \tau_{w_{d,i}}]$, i.e., it requires complicated calculations and must keep variance counts for them. Consequently, in using it, we cannot avoid the computational drawbacks of the CVB inference. We solve this problem in the next section.

5. PROPOSED INFERENCE

This section explains our inference which is an approximation of the existing CVB inference[24]. First, we describe our motivation for developing this approximation and then explain the derivation.

The proposed inference can be summarized as follows. We marginalize over $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ and estimate a factorized variational posterior over $\boldsymbol{z}$ and $\boldsymbol{\pi}$, i.e., $q(\boldsymbol{z}, \boldsymbol{\pi}) = \prod_{j,i} q(z_{j,i}) \prod_k q(\tilde{\pi}_k)$. We use the point estimation for $\alpha_0, \beta_0, \gamma_0$ and $\boldsymbol{\tau}$ because we do not wish to set any hyper parameters. The proposed update algorithms are given in Algorithm 1.

5.1 Motivation

We were strongly motivated by David Sontag and Daniel Roy, who in [31] theoretically showed that small parameters of Dirichlet distribution over document-topic distribution, i.e. $\boldsymbol{\alpha}$, encourage sparsity. The meaning of "sparsity" here is that the topic distribution will be large on as few topics as necessary to explain every word of a document and otherwise will be close to zero. They reported that when they applied LDA to a NIPS corpus with 200 topics, the parameters found range from 0.0009 to 0.135 with the median being 0.01. We think HDP-LDA also has this property because HDP-LDA, actually a stick-breaking process, induces sparsity of the topic distributions. We surveyed the parameters $\{\mathbb{G}[\alpha_0 \pi_k]\}_{k=1}^T$ of HDP-LDA with the existing CVB inference [24] on several datasets and found that the parameters took small values (shown in Sec.6). We utilize this property for deriving an approximation of the CVB inference.

5.2 Derivation

Note that $\boldsymbol{\pi}$ in the Dirichlet distribution Dir$(\alpha_0 \boldsymbol{\pi})$ is generated from the TSBP, i.e., $\boldsymbol{\pi} \sim$ TSBP(γ_0). The difficulty in the derivation of variational inference, in fact the estimation for $q(\boldsymbol{\pi})$, is that the Dirichlet distribution Dir$(\alpha_0 \boldsymbol{\pi})$ does not conjugate to the stick-breaking process TSBP(γ_0). Therefore, we use the following approximation for calculating the Gamma function to estimate $q(\boldsymbol{\pi})$.

Suppose $\alpha < 1$ and $n \geq 1$. We can use the approximations

$$\Gamma(\alpha + n) \approx \Gamma(n), \; \Gamma(\alpha) \approx \frac{1}{\alpha}. \tag{20}$$

These approximations are obtained as follows. By using $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$, we have $\Gamma(\alpha) = \frac{\Gamma(\alpha+1)}{\alpha}$. Thus, when $\alpha < 1$, in particular, $\Gamma(\alpha + 1) \approx \Gamma(1) = 1$, we have $\Gamma(\alpha) \approx \frac{1}{\alpha}$. Furthermore, $\Gamma(\alpha + n) = (\alpha + n - 1) \cdots (\alpha + 1)\alpha\Gamma(\alpha) \approx (n-1) \cdots 1\alpha\frac{1}{\alpha} = \Gamma(n)$. Note that in fact $\Gamma(\alpha) \leq \frac{1}{\alpha}$ $(\alpha \leq 1)$, $\Gamma(n+\alpha) \geq \Gamma(n)$ $(\alpha \leq 1, n \geq 2)$, $\Gamma(\alpha + 1) \leq \Gamma(1) = 1$ $(\alpha \leq 1)$. These approximations are also described in [32, 33]

These approximations enable us to obtain a closed form update for $q(\tilde{\boldsymbol{\pi}})$ in Eq.(24). In the experimental section, we look into $\mathbb{E}[\alpha_0 \pi_k]$ of the CVB inference in HDP-LDA to verify these approximations. Although we use the same approximation for β_v as α_k, we can also use Minka's fixed point iterations for estimating β_v used in [26].

Using Eq.(20), Eq.(17) is approximated by

$$\begin{aligned}
&\tilde{p}(\boldsymbol{z}, \boldsymbol{w} | \alpha_0, \beta_0, \boldsymbol{\pi}, \boldsymbol{\tau}) = \\
&\left[\prod_{d=1}^{N} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + n_d)} \prod_{k=1}^{T} [\Gamma(n_{d,k})\alpha_0 \pi_k]^{I(n_{d,k}>0)} \right] \\
&\left[\prod_{k=1}^{T} \frac{\Gamma(\beta_0)}{\Gamma(\beta_0 + n_{k,\cdot})} \prod_{v=1}^{V} [\Gamma(n_{k,v})\beta_0 \tau_v]^{I(n_{k,v}>0)} \right].
\end{aligned} \tag{21}$$

The variational lower bound on log likelihood is approximately given by

$$\begin{aligned}
&\log p(\boldsymbol{w} | \alpha_0, \beta_0, \gamma_0, \boldsymbol{\tau}) \\
&\geq \mathbb{E}[\log p(\boldsymbol{z}, \boldsymbol{w} | \alpha_0, \beta_0, \boldsymbol{\pi}, \boldsymbol{\tau}) - \log q(\boldsymbol{z})] - \mathrm{KL}[q(\boldsymbol{\pi}) || p(\boldsymbol{\pi} | \gamma_0)], && (22) \\
&\approx \mathbb{E}[\log \tilde{p}(\boldsymbol{z}, \boldsymbol{w} | \alpha_0, \beta_0, \boldsymbol{\pi}, \boldsymbol{\tau}) - \log q(\boldsymbol{z})] - \mathrm{KL}[q(\boldsymbol{\pi}) || p(\boldsymbol{\pi} | \gamma_0)]. && (23)
\end{aligned}$$

Taking the functional derivative of the lower bound Eq.(23) with respect to $q(\tilde{\boldsymbol{\pi}}_k)$ and equating them to zero, we have

$$q(\tilde{\boldsymbol{\pi}}_k) = Beta(a_k, b_k) \; (k = 1, \cdots, T-1), \; q(\tilde{\pi}_T = 1) = 1, \tag{24}$$

$$a_k = 1 + \sum_d \mathbb{E}[I(n_{d,k} \geq 1)], \tag{25}$$

$$b_k = \gamma_0 + \sum_{l=k+1}^{T} \sum_d \mathbb{E}[I(n_{d,l} \geq 1)], \tag{26}$$

$$\begin{aligned}
&\mathbb{E}[I(n_{d,k} \geq 1] = q(n_{d,k} \geq 1) = 1 - q(n_{d,k} = 0) \\
&= 1 - \prod_i q(z_{d,i} \neq k) = 1 - \exp\left(\sum_i \log(1 - q(z_{d,i} = k)) \right).
\end{aligned} \tag{27}$$

Note that

$$\mathbb{E}[\pi_1] = \mathbb{E}[\tilde{\pi}_1], \; \mathbb{E}[\pi_k] = \mathbb{E}[\tilde{\pi}_k \prod_{l=1}^{k-1} (1 - \tilde{\pi}_l)] = \frac{a_k}{a_k + b_k} \prod_{l=1}^{k-1} \frac{b_l}{a_l + b_l}. \tag{28}$$

Moreover, taking derivatives of $\mathbb{E}[\log \tilde{p}(\boldsymbol{z}, \boldsymbol{w} | \alpha_0, \beta_0, \boldsymbol{\pi}, \boldsymbol{\tau})]$ in Eq.(23) with respect to α_0, β_0 and τ_v and equating them to zero, we obtain the following fixed point iteration equations

$$\alpha_0^{\text{new}} = \frac{\sum_{d,k} \mathbb{E}[I(n_{d,k} \geq 1)]}{\sum_d [\Psi(n_d + \alpha_0^{\text{old}}) - \Psi(\alpha_0^{\text{old}})]}, \tag{29}$$

$$\beta_0^{\text{new}} = \frac{\sum_{k,v} \mathbb{E}[I(n_{k,v} \geq 1)]}{\sum_k [\Psi(\mathbb{E}[n_{k,\cdot}] + \beta_0^{\text{old}}) - \Psi(\beta_0^{\text{old}})]}, \tag{30}$$

$$\tau_v \propto \sum_k \mathbb{E}[I(n_{k,v} \geq 1)], \tag{31}$$

$$\mathbb{E}[I(n_{k,v} \geq 1)] = 1 - \exp\left(\sum_{d,i} I(w_{d,i} = v) \log(1 - q(z_{d,i} = k)) \right). \tag{32}$$

We update γ_0 by taking derivatives of KL$[q(\tilde{\boldsymbol{\pi}}_k) || p(\tilde{\boldsymbol{\pi}}_k | \gamma_0)]$ in Eq.(23) with respect to γ_0 and equating them to zero.

$$\gamma_0 = \frac{T-1}{-\sum_{k=1}^{T-1} \mathbb{E}[\log(1 - \tilde{\pi}_k)]} = \frac{T-1}{\sum_{k=1}^{T-1} \Psi(a_k + b_k) - \Psi(b_k)}. \tag{33}$$

By using a second-order Taylor expansion as an approximation, we have the CVB inference for $q(z_{d,i})$ given by

$$q(z_{d,i}=k) \propto \frac{\mathbb{E}[n_{k,w_{d,i}}^{-d,i}]+\beta_0\tau_{w_{d,i}}}{\mathbb{E}[n_{k,\cdot}^{-d,i}]+\beta_0} \frac{\mathbb{E}[n_{d,k}^{-d,i}]+\alpha_0\mathbb{E}[\pi_k]}{n_d^{-d,i}+\alpha_0}$$
$$\exp\left(-\frac{\mathbb{V}[n_{k,w_{d,i}}^{-d,i}]}{2(\mathbb{E}[n_{k,v_{d,i}}^{-d,i}]+\beta_0\tau_{w_{d,i}})^2}+\frac{\mathbb{V}[n_{k,\cdot}^{-d,i}]}{2(\mathbb{E}[n_{k,\cdot}^{-d,i}]+\beta_0)^2}\right)$$
$$\exp\left(-\frac{\mathbb{V}[n_{d,k}^{-d,i}]}{2(\mathbb{E}[n_{d,k}^{-d,i}]+\alpha_0\mathbb{E}[\pi_k])^2}\right), \tag{34}$$

The CVB0 inference for $q(z_{d,i})$ can be made by using only the zero-order information as follows.

$$q(z_{d,i}=k) \propto \frac{\mathbb{E}[n_{k,w_{d,i}}^{-d,i}]+\beta_0\tau_{w_{d,i}}}{\mathbb{E}[n_{k,\cdot}^{-d,i}]+\beta_0}(\mathbb{E}[n_{d,k}^{-d,i}]+\alpha_0\mathbb{E}[\boldsymbol{\pi}_k]). \tag{35}$$

Note that this CVB0 inference for HDP-LDA does not require the maintenance of variance counts, the same as the CVB0 inference for LDA [26]. As previously mentioned, although the CVB0 inference described in Sec.4 does not use variance counts for estimating $q(\boldsymbol{z})$, it needs to estimate other parameters ($\mathbb{G}[\alpha_0\pi_k]$ and $\mathbb{G}[\beta_0\tau_v]$ using variance counts), which negatively affects its good properties of the CVB0 inference. We clarify the relationship between our CVB0 inference and the collapsed Gibbs sampler Eq.(13) in the next section.

5.3 Interpretation

The differences between Eq.(13) and Eq.(35) can be interpreted as the difference between the calculations for the number of tables in the CRP representation, although Eq.(13) is stochastic while Eq.(35) is deterministic. From the analogy to $m_{\cdot,k}$ of Eq.(15), $\sum_{d=1}^{N}\mathbb{E}[I(n_{d,k}\geq 1)]$ in Eq.(24), (25), and (26) indicates the number of tables where $\mathbb{E}[I(n_{d,k}\geq 1)]$ means the number of tables at which dish k is served is limited to one in a restaurant(document), i.e., $\sum_{d=1}^{N}\mathbb{E}[I(n_{d,k}\geq 1)] \leq m_{\cdot,k}$. In other words, it can be said that we approximate the number of tables $m_{\cdot,k}$ by using lower bound $\sum_{d=1}^{N}\mathbb{E}[I(n_{d,k}\geq 1)]$ in our inference. This limitation seems to be a good approximation throughout a whole corpus.

Equation (35) also helps us understand the sparsity of topics. Intuitively , a topic with low document frequency is eliminated because $\sum_{d=1}^{N}\mathbb{E}[I(n_{d,k}\geq 1)]$ also indicates the document frequency of topic k, i.e., the number of unique documents in which topic k occurs. The constraint $\sum_k q(z_{j,i}=k)=1$, $\forall k$ will lead to some topics with small assignment probabilities, which makes $\mathbb{E}[n_{d,k}\geq 1]$ and $\mathbb{E}[n_{d,k}]$ small for some topics $\{k\}$. Moreover, $\mathbb{E}[\pi_k]$ will take very small value for some topics due to the constraint $\sum_k\mathbb{E}[\pi_k]=1$. As a result, $\mathbb{E}[n_{d,k}]+\alpha_0\mathbb{E}[\pi_k]$ will also become smaller for some topics, which makes $q(z_{j,i}=k)$ more sharply distributed and gives topics almost no chance to become bigger $q(z_{j,i}=k)$ in the future. In this way, we obtain only a smaller number of effective topics than a truncation level.

6. EXPERIMENTS

We evaluated the proposed inferences on three tasks: (1) document modeling in terms of perplexity, (2) a nearest neighbor search on the topic simplex, and (3) a link prediction in a social network. All results are averaged values from five experimental runs with random initialization. We initialized $\alpha_0\boldsymbol{\pi}_k=0.1/T$, $\beta_0\tau_v=0.1/T$ and $q(z_{d,i}=k)\propto 0.1+u$ where u is generated from the uniform distribution over $[0,1]$. We set the number of iterations to 100 for each inference. First, we give preliminaries for reading experimental results, next describe datasets, and then discuss the experimental results. In this section, we empirically show that the results of our inference algorithms are similar to those of the existing algorithms, which means our approximation works well.

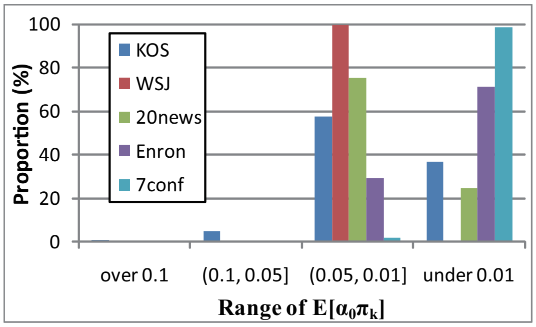

Figure 2: The proportion of $\mathbb{E}[\alpha_0\pi_k]$ $(k=1,\cdots,T=300)$ of CVB-AA estimated from five datasets.
For example, almost 99% of $\mathbb{E}[\alpha_0\pi_k]$ $(k=1,\cdots,T=300)$ are in the range $(0.05, 0.01]$ in WSJ. For the lack of space, we eliminated the results of $\mathbb{E}[\beta_v]$. Almost $\mathbb{E}[\beta_v]$ $(v=1,\cdots,V)$ of CVB-AA also takes small value as in $\mathbb{E}[\alpha_k]$. Maximum values of $\mathbb{E}[\alpha_k]$ is 0.1532 in KOS, 0.0610 in WSJ, 0.0309 in 20news, 0.0176 in Enron, and 0.0122 in 7conf. Note that $\Gamma(1)=1$, $\Gamma(1+0.05)=0.9735$, $\Gamma(0.05)=19.47$, and $1/0.05=20$.

6.1 Preliminaries

The purpose of the experiments is to investigate the performance of our approximation and the effect of the choice of prior over the topic-word distributions in HDP-LDA with the (practical) CVB / CVB0 inferences. In this section's figures, "CVB" and "CVB0" indicates second order and zero order approximation of the CVB inference for HDP, while "PCVB" and "PCVB0" indicate the proposed (practical CVB and CVB0) inferences where we use the second order Taylor approximation in PCVB as in CVB.

Wallach et al.[28] explored the effects of the choice of prior (symmetric versus asymmetric Dirichlets) over the document-topic distributions, denoted by $\boldsymbol{\theta}$, and topic-word distributions, denoted by $\boldsymbol{\phi}$, in LDA. They introduced notations, SS, AS and AA, as the choice of prior over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. SS uses symmetric priors over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. AS uses asymmetric priors over $\boldsymbol{\theta}$, and symmetric prior over $\boldsymbol{\phi}$. AA uses asymmetric priors over $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. It is assumed with HDP-LDA that a prior over the document-topic distributions is asymmetric, i.e., we consider AS and AA models. Using these notations, CVB-AA was proposed in Teh et al. [24].

Wallach et al.[28] found in MCMC simulations that the AS model results in better predictive performance of held-out documents. Asuncion et al. showed in [26] that using the appropriate the Dirichlet parameters plays a large role in learning accurate topic models in the SS model. Asuncion also showed in [27] that the predictive performance of the CVB0 inference clearly outperformed that of the CVB inference when they used the AA model. We investigated the AS model in our experiments because this choice performed well in MCMC simulations reported by Wallach et al.[28].

6.2 Datasets

We used five sets of text data with different properties. The first was 'KOS blog corpus (KOS)" where the number of documents was $N=3,430$ and the vocabulary size was $V=6,906$. The second was “The Wall Street Journal (WSJ)” where we randomly chose $N = 3,000$ ($V = 30,576$) documents. The third was “Enron email corpus (Enron)” made by Frank and Asuncion [34] where we randomly chose $N = 10,000$ ($V = 15,258$) documents. The fourth was “20 news group corpus (20news)[1]” where we randomly chose $N = 10,000$ ($V = 13,178$). The fifth was “7 Conference paper corpus (7conf)” where we used seven conference abstracts (KDD ($N = 1252$), ICDM (N=867), SIGMOD ($N = 2155$), SIGIR ($N = 2074$, WWW ($N = 1064$), ICML ($N = 1525$), and CVPR ($N = 1411$)) collected by Deng et al. [35]. In total, we obtained $N = 10,311$ ($V = 4,950$) documents. Stop words were eliminated.

As previously mentioned, we were strongly motivated by David Sontag and Daniel Roy, who in [31] showed that small parameters of Dirichlet distribution over document-topic distribution $\boldsymbol{\theta}_d$ encourage sparsity. We looked into $\mathbb{E}[\alpha_0\pi_k]$ of the existing CVB inference proposed by Teh et al. [24] with $T = 300$, noting that a large truncation level makes the TSBP approach the SBP. Figure 2 shows that $\mathbb{E}[\alpha_0\pi_k]$ takes small values in datasets. This is because in HDP-LDA even if a whole corpus contains many topics, a document usually does not, which means the document-topic distribution has sparsity. What we found from Fig. 2 inspired us to assume $\alpha_k(= \alpha_0\pi_k) < 1$ to derive our practical inference.

6.3 Document modeling

The comparison metric we used for document modeling was the perplexity that indicates the prediction performance for held-out words that was used by Teh et al. [25, 24]. We randomly split the words in a document into training words $\boldsymbol{w}_d^{train}$(80%) and test words $\boldsymbol{w}_d^{test}$ (20%). The perplexity of $\{\boldsymbol{w}_d^{test}\}$ is given by

$$\exp\left[-\frac{1}{\sum_{d=1}^{N} n_d^{test}} \sum_{d=1}^{N} \sum_{i=1}^{n_d^{test}} \log p(w_{d,i}^{test}|\boldsymbol{w}_d^{train})\right]. \quad (36)$$

The predictive distributions of our inferences are given by

$$p(w^*|\boldsymbol{w}_d) = \sum_{k=1}^{K} \frac{\beta_0\tau_v + \mathbb{E}[n_{k,w^*}]}{\beta_0 + \mathbb{E}[n_{k,\cdot}]} \frac{\alpha_0\mathbb{E}[\pi_k] + \mathbb{E}[n_{d,k}]}{\alpha_0 + n_d}. \quad (37)$$

Figure 3 shows the experimental results we obtained for perplexity. The left line indicates the results for test set perplexity in terms of the truncation level ($T = 100, 200, 300$) in each corpus where the number of iterations is 100. The right line shows the relationships between test set perplexity and the number of iterations where the truncation level was $T = 300$. We show CVB-AA,CVB0-AA, and PCVB0-AA in figure because the AS and AA models and the CVB and PCVB inferences were almost the same and it became difficult to distinguish their respective lines. We show 50 iterations to clarify the differences among algorithms, although we ran 100 iterations.

CVB0 and PCVB0 basically outperformed CVB and PCVB in terms of perplexity. In the experiments, the convergence rate of CVB0 and PCVB0 were respectively slower than those of CVB and PCVB. The performance of PCVB and PCVB0 were respectively similar to those of CVB and CVB0. The difference between AS and AA was small in our experiments.

6.4 Nearest Neighbor Search on Topic Simplex

The purpose of this experiment was to evaluate the estimation performance of topic distributions for test documents. We estimated a topic distribution $\boldsymbol{\theta}^{\text{test}}$ of a test document and searched for the nearest neighbor of a test document in the training documents by using the distance between $\boldsymbol{\theta}^{\text{test}}$ and $\boldsymbol{\theta}_d$ for each training document d.

LDA can be regarded as a tool for reducing the dimensions of a document from a word space to a topic space. One application of LDA as a dimension reduction tool is to perform the nearest neighbor search in the topic simplex. Generally, the nearest neighbor search is fast as the dimension of a data point is low.

We used the Hellinger distance as the distance metric of topic distributions. The Hellinger distance is a symmetric measure and that is often used in statistics to quantify the similarity between two probability distributions given by

$$H^2(\boldsymbol{\theta}_d, \boldsymbol{\theta}^{\text{test}}) = \sum_{k=1}^{T}(\sqrt{\theta_{d,k}} - \sqrt{\theta_k^{\text{test}}})^2, \quad (38)$$

in which we actually used the expectation of $\theta_{d,k}$, e.g., $E[\theta_{d,k}] = \frac{\mathbb{E}[\alpha_0\pi_k]+\mathbb{E}[n_{d,k}]}{\sum_k \mathbb{E}[\alpha_0\pi_k]+\mathbb{E}[n_{d,k}]}$.

Using the 7conf and 20news corpora, we randomly split both data sets into training documents (90%) and test documents (10%). We evaluated the estimation performance of topic distributions for test documents with the nearest neighbor classification task. We categorized test documents into the same category to which the nearest neighbor document is labeled. We had seven labels in 7conf and 20 labels in 20news. Figure 4 shows the accuracy of the models, from which it can be seen that (P)CVB0 outperformed (P)CVB and the accuracy of PCVB and PCVB0 were respectively similar to those of CVB and CVB0. It is important to note that the performance of our approximation algorithms was similar to that of the existing ones, which means that the estimated topic distributions were similar and that our approximation does not negatively affects the performance of the existing algorithms.

6.5 Link Prediction in Social Networks

We used two social network datasets. The first one was a collaboration network of the Arxiv Astro Physics category (AstroPh) used in [36, 37], where nodes represent scientists, edges represent collaborations (co-authoring a paper), and there were 18,772 nodes and 396,160 links. The second one is the Enron email network (EnronNet) used in [36], where an link indicated that email was exchanged, and there are 36,692 nodes and 367,662 links. In network modeling by LDA, we assumed that a node indicates a document and that a link to other node indicates a word, i.e., we represented a node as a “bag of links.” We evaluated the algorithms with a recommendation task that ranks link-nodes with high link probability for each query-node using its link history. We ranked link-nodes for each query-node with a predictive distribution used in evaluating the perplexity in the previous section, i.e.,

$$p(\text{query-node } j \text{ links link-node } l|\text{link history of query-node } j). \quad (39)$$

We performed an information retrieval based evaluation to compare the algorithms. We used the mean average precision (MAP), which is often used in information retrieval. The MAP reflects the overall retrieval accuracy. To calculate the MAP, we need query sets, candidate item sets for ranking, and relevance judgments lists. We constructed the test collection as follows. We used nodes with over 20 links as query sets. We split the link data of each query into

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html\#news20

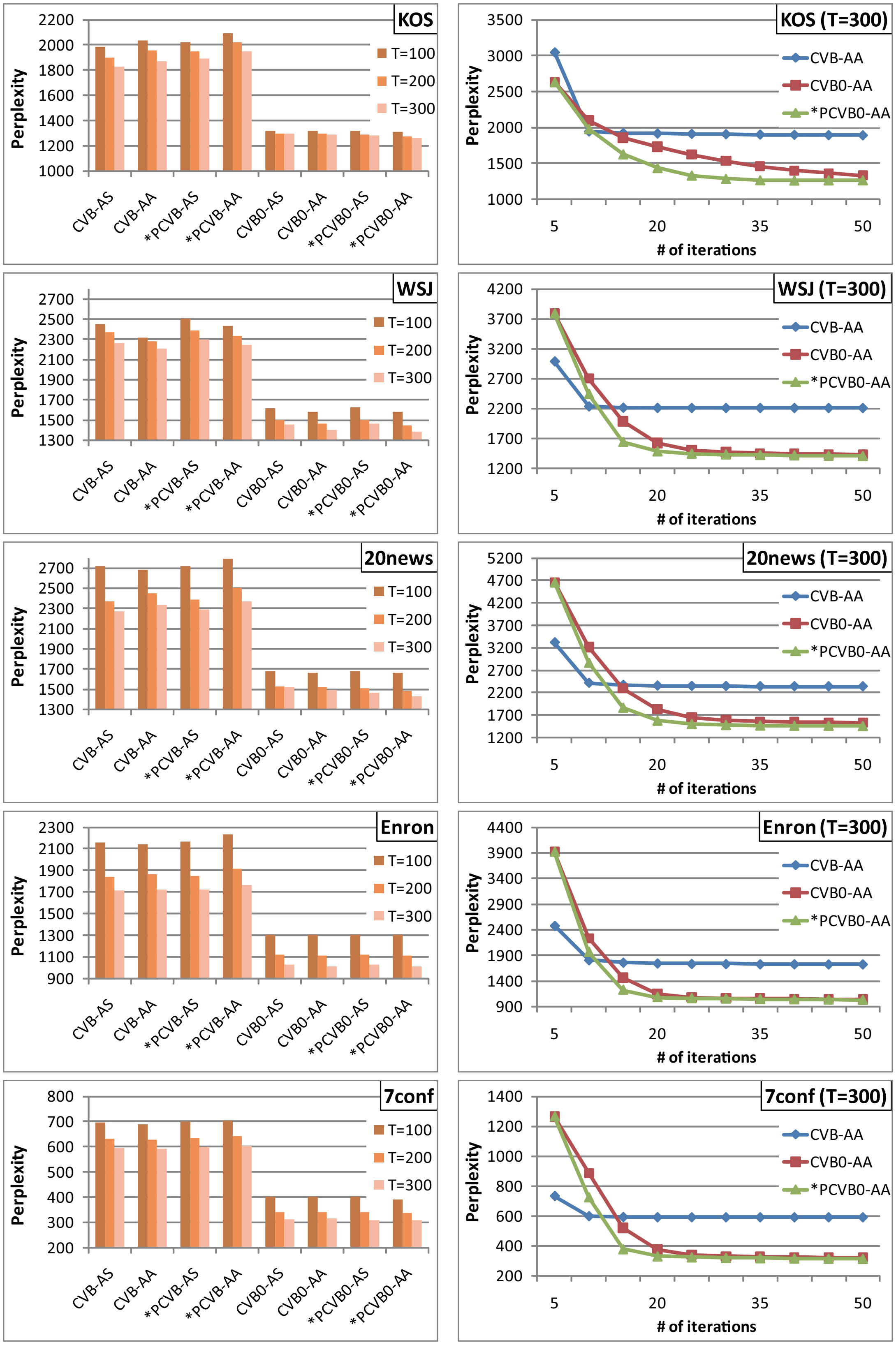

Figure 3: Experiment results for document modeling in five datasets where * indicates our inferences. T denotes the truncation level of TSBP. Lower perplexity indicates better performance.

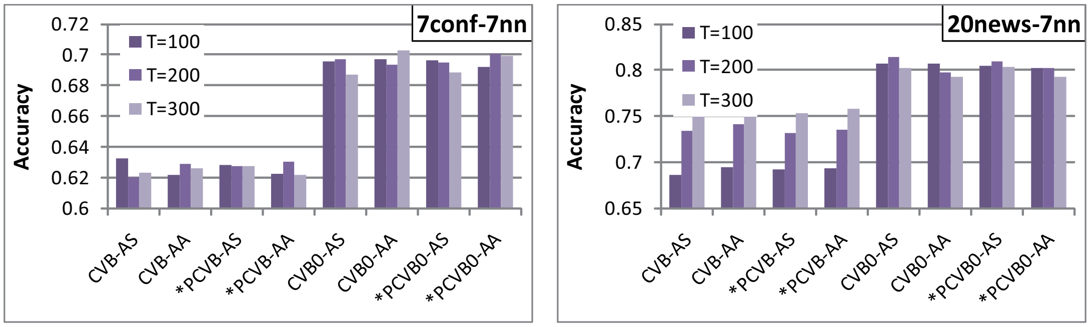

Figure 4: Evaluations for Nearest Neighbor Search.
Results for accuracy among inference algorithms by changing the number of topics in 7-nearest neighbors in the 7conf and 20news corpus. Higher accuracy indicates better classification performance.

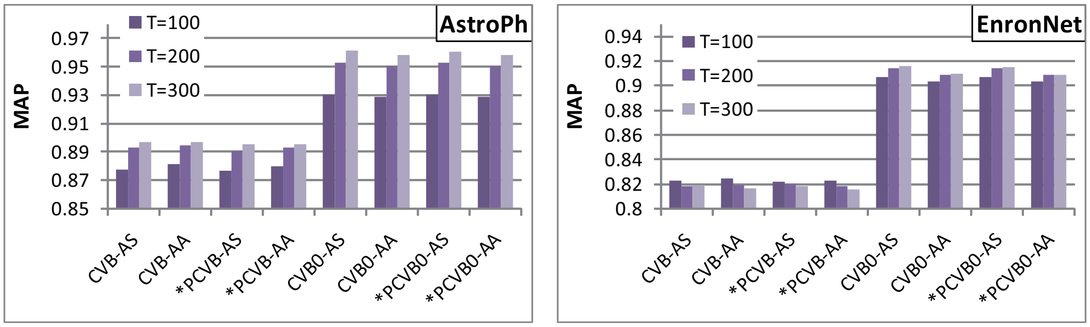

Figure 5: Evaluations for Link Prediction in Social Networks.
Results for MAP among inference algorithms by changing the number of topics in the co-author network in Astro Physics (AstroPh) and the Enron communication networks (EnronNet). Higher MAP indicates better recommendation (ranking) performance.

80% training links and 20% test links. We used the 20% test links for positive examples and labeled those link-nodes "relevance". We constructed negative examples by using randomly selected nodes that were not linked from a query node and labeled those nodes "non-relevance". The number of negative examples was equal to that of the 80% training sets. This is because the MAP does not depend on the number of links for each query-node. The MAP results for various methods are shown in Fig.5. The performance of the link prediction is similar to that of perplexity.

6.6 Discussion

Asunction [27] described the interest relationships between the CVB0 inference and other inferences, e.g., belief propagation, to explain the performance of the CVB0 inference. Here, we explain the (P)CVB0 performance from another aspect. From the results shown on the right side of Fig.3, we see that the convergence of CVB is faster than that of (P)CVB0. This rapid convergence seems to be the reason that the performance of CVB is worse than that of (P)CVB0, i.e., CVB seems to rapidly stuck in poor local optima. This is because for estimating $q(z_{d,i})$, CVB uses the exponential function $\exp(-\mathbb{V}/x^2)$ which makes $q(z_{d,i})$ a sharper distribution when variance $\mathbb{V}$ takes a larger value (note that variance is large in initial iterations). This is similar to the inverse temperature of simulated annealing, which induces the rapid convergence of stochastic inference. We think that the poor local optima can negatively affects the performance of the HDP modeling, which means the HDP model will not work well even if we use a high truncation level.

7. CONCLUSION

We proposed a practical deterministic inference for HDP-LDA. Moreover, we investigated the effects of the choice of prior (symmetric versus asymmetric Dirichlets) over the topic-word distributions in the CVB/CVB0 inference. Although the choice of a symmetric Dirichlet prior over the topic-word distribution model performed well in MCMC simulations reported by Wallach et al.[28], the choice of prior has a small difference in our variational settings. The purpose of this study was to explore a simple learning algorithm for HDP-LDA because the CVB inference of HDP-LDA [24] is more complicated and harder to implement than the variational inference for LDA. The PCVB0 inference is particularly simple to implement, does not require variance counts to be maintained, does not need to set hyper-parameters, and has good predictive performance. Consequently, we recommend the PCVB0 inference in practice and hope that the proposed inference will alleviate researchers' difficulties in using HDP-LDA in a wide range of fields. For future work, we extend the PCVB0 inference into an online algorithm such as [38, 39, 40].

8. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[3] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315. ACM Press, 2004.

[4] D. Newman, C. Chemudugunta, and P. Smyth. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686. ACM Press, 2006.

[5] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550. ACM, 2008.

[6] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Latent Topic Models for Hypertext. In *UAI*, pages 230–239, 2008.

[7] A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy. Structured correspondence topic models for mining captioned figures in biological literature. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 39–48. ACM, 2009.

[8] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[9] C. Wang, D. M. Blei, and D. Heckerman. Continuous Time Dynamic Topic Models. In *UAI*, pages 579–586, 2008.

[10] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 937–946. ACM, 2009.

[11] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda. Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 663–672. ACM, 2010.

[12] L. Hong, D. Yin, J. Guo, and B. D. Davison. Tracking trends: incorporating term volume into temporal topic models. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 484–492. ACM, 2011.

[13] I. Sato and H. Nakagawa. Topic Models with Power-Law Using Pitman-Yor Process. In *Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[14] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465. ACM, 2011.

[15] J. Zhu, N. Lao, N. Chen, and E. P. Xing. Conditional topical coding: an efficient topic model conditioned on rich features. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 475–483. ACM, 2011.

[16] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185. ACM, 2006.

[17] D. Andrzejewski and D. Buttler. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 600–608. ACM, 2011.

[18] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin. Turning down the noise in the blogosphere. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 289–298. ACM, 2009.

[19] J. Chang, J. L. Boyd-Graber, and D. M. Blei. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 169–178. ACM, 2009.

[20] S. Gerrish and D. M. Blei. A Language-based Approach to Measuring Scholarly Impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML2010), year = 2010, pages = 375-382*.

[21] S. Gerrish and D. M. Blei. Predicting Legislative Roll Calls from Text. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pages 489–496, 2011.

[22] C. Wang and D. M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 448–456, 2011.

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

[24] Y. W. Teh, K. Kurihara, and M. Welling. Collapsed Variational Inference for HDP. In *Advances in Neural Information Processing Systems 20*, 2008.

[25] Y. W. Teh, D. Newman, and M. Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 19*, 2007.

[26] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On Smoothing and Inference for Topic Models. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2009.

[27] A. Asuncion. Approximate Mean Field for Dirichlet-Based Models. In *Topic Models Workshop, ICML*.

[28] H. Wallach, D. Mimno, and A. McCallum. Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems 22*, pages 1973–1981. 2009.

[29] H. Ishwaran and L. F. James. Gibbs Sampling Methods for Stick Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[30] Escobar and West. Bayesian Density Estimation and Inference using Mixtures. *Journal of the American Statistical Association*, 90, 1995.

[31] D. Sontag and D. Roy. Complexity of Inference in Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 24*, pages 1008–1016. 2011.

[32] T. P. Minka. Estimating a Dirichlet distribution. Technical report, Microsoft, 2000.

[33] C. Elkan. Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the 23rd international conference on Machine learning*, pages 289–296, 2006.

[34] A. Frank and A. Asuncion. UCI Machine Learning Repository, 2010.

[35] H. Deng, J. Han, B. Zhao, Y. Yu, and C. X. Lin. Probabilistic topic models with biased propagation on heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1271–1279, 2011.

[36] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 177–187. ACM, 2005.

[37] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 1(1), 2007.

[38] M. D. Hoffman, D. M. Blei, and F. R. Bach. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010*, pages 856–864, 2010.

[39] I. Sato, K. Kurihara, and H. Nakagawa. Deterministic Single-Pass Algorithm for LDA. In J. Lafferty, C. K. I. Williams, R. Zemel, J. Shawe-Taylor, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2074–2082. 2010.

[40] C. Wang, J. W. Paisley, and D. M. Blei. Online Variational Inference for the Hierarchical Dirichlet Process. *Journal of Machine Learning Research - Proceedings Track*, 15:752–760, 2011.