The Generalized Dirichlet Distribution in Enhanced Topic Detection

Karla Caballero*
UC, Santa Cruz
Santa Cruz CA, USA
karla@soe.ucsc.edu

Joel Barajas
UC, Santa Cruz
Santa Cruz CA, USA
jbarajas@soe.ucsc.edu

Ram Akella
UC, Santa Cruz
Santa Cruz CA, USA
akella@soe.ucsc.edu

ABSTRACT

We present a new, robust and computationally efficient Hierarchical Bayesian model for effective topic correlation modeling. We model the prior distribution of topics by a Generalized Dirichlet distribution (GD) rather than a Dirichlet distribution as in Latent Dirichlet Allocation (LDA). We define this model as GD-LDA. This framework captures correlations between topics, as in the Correlated Topic Model (CTM) and Pachinko Allocation Model (PAM), and is faster to infer than CTM and PAM. GD-LDA is effective to avoid over-fitting as the number of topics is increased. As a tree model, it accommodates the most important set of topics in the upper part of the tree based on their probability mass. Thus, GD-LDA provides the ability to choose significant topics effectively. To discover topic relationships, we perform hyper-parameter estimation based on Monte Carlo EM Estimation. We provide results using Empirical Likelihood (EL) in 4 public datasets from TREC and NIPS. Then, we present the performance of GD-LDA in ad hoc information retrieval (IR) based on MAP, P@10, and Discounted Gain. We discuss an empirical comparison of the fitting time. We demonstrate significant improvement over CTM, LDA, and PAM for EL estimation. For all the IR measures, GD-LDA shows higher performance than LDA, the dominant topic model in IR. All these improvements with a small increase in fitting time than LDA, as opposed to CTM and PAM.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Document Filtering*; G.3 [**Probability and Statistics**]: Statistical Computing

General Terms

Algorithms, Experimentation

Keywords

Statistical Topic Modeling, Document Representation

*Main contact.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

1. INTRODUCTION

Topic modeling has been widely studied in Machine Learning and Text Mining as an effective approach to extract latent topics from unstructured text documents. The key idea underlying topic modeling is to use term co-occurrences in documents to discover associations between those terms. The development of Latent Dirichlet Allocation (LDA) [3, 8] enabled the rigorous prediction of new documents first time. Consequently, variants and extensions of LDA have been an active area of research in topic modeling. This research has 2 main streams in document representation: 1) The exploration of super and subtopics as in Pachinko Model Allocation (PAM) [9, 10]; and 2) the correlation of topics [2, 9, 10]. However, despite the improvement of these approaches over LDA, they are rarely used in applications that handle large datasets, such as Information Retrieval. Major gaps in these models include: modeling correlated topics in a computationally effective manner; and a robust approach to ensure good performance for a wide range of document numbers, vocabulary size and average word length per document.

In this paper we develop a new model to meet these needs. We use the Generalized Dirichlet (GD) distribution as a prior distribution of the document topic mixtures, leading to GD-LDA. We show that the Dirichlet distribution is a special case of GD. As a result, GD-LDA is deemed to be a generalized case of LDA. Our goal is to provide a more flexible model for topics while retaining the conjugacy properties, which are desirable in inference. The features of the GD-LDA model include: 1) An effective method to represent sparse topic correlations in natural language documents. 2) A model which handles global topic correlations with time complexity $O(KW)$, adding minimal computational cost respect to LDA. This results in a fast and robust approach compared to CTM and PAM. 3) A hierarchical tree structure that accommodates the most significant topics, based on probability mass, at the upper levels allowing us to reduce the number of topics efficiently. Note that GD is a special case of Dirichlet Trees [13, 5]. This distribution has been used previously in topic modeling to add domain knowledge to the probability of words given a topic [1]. In contrast, we use the GD distribution to model topic correlations. Thus, this approach is complementary to GD-LDA.

To validate our model, we use Empirical Likelihood (EL) in four data sets with different characteristics of document length, vocabulary size, and total number of documents. GD-LDA outperforms CTM, PAM and LDA consistently in all the datasets. In addition, we test the performance of GD-LDA in adhoc Information Retrieval, obtaining superior

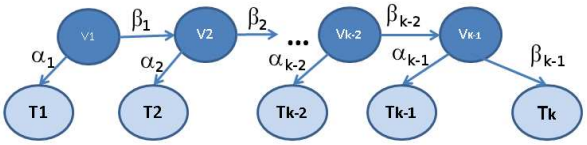

Figure 1: Tree structure of the GD distribution.

results to those in the literature. We show a significant difference in running times between GD-LDA, CTM and PAM which makes GD-LDA as viable as LDA for large data sets.

This paper is organized as follows: In section 2, we present the definition and features of the GD distribution. We will use this definition to develop the methodology of the GD-LDA model. Section 3 depicts our proposed approach with the proper derivations. Validation criteria, experimental settings, and results are presented in section 4. Finally the discussion and conclusion are presented in section 5.

2. THE GENERALIZED DIRICHLET DISTRIBUTION

2.1 Properties and Intuition

The Generalized Dirichlet (GD) distribution was introduced by Connor and Mosimann in [4]. The GD distribution is motivated by the limitations of the Dirichlet distribution in modeling covariances. In the case of the Dirichlet distribution, all the entries of the random vector must share a common variance, and they must sum to one. When we use the Dirichlet distribution as a prior for the multinomial distribution we have only one degree of freedom, the total prior sample size, to incorporate our confidence in the prior knowledge. As a consequence, we can not add individual variance information for each entry of the random vector. In addition, all entries are always negatively correlated. In other words, if the probability of one entry increases each of the other probabilities must either decrease or remain the same to sum to one. Despite these limitations, the Dirichlet distribution is widely used given that this is a conjugate prior of the multinomial distribution.

The GD distribution allows us to sample each entry of the random vector of proportions from independent Beta distributions. This independence is the key property that provides more flexibility than the Dirichlet distribution. Formally this distribution is defined as:

$$p(\theta|\alpha,\beta) = \prod_{j=1}^{K-1} \frac{\Gamma(\alpha_j+\beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \theta_j^{\alpha_j-1}(1-\theta_1-\cdots-\theta_j)^{\eta_j} \quad (1)$$

where $\theta_1+\theta_2+\ldots+\theta_{K-1}+\theta_K=1$, $\eta_j=\beta_j-\alpha_{j+1}-\beta_{j+1}$ for $1\leq j\leq K-2$ and $\eta_{K-1}=\beta_{K-1}-1$.

To illustrate the properties of the GD distribution we define $Z_1=\theta_1$ and $Z_k=\theta_k/V_k$ for $k=2,3,\ldots,K-1$ where $V_k=1-\theta_1-\cdots-\theta_{k-1}$. Let T_k be the discrete random variable with multinomial distribution with parameter $\theta_1\ldots\theta_K$ for K different categories. We start at node V_1, and at this node we sample T_1 with probability Z_1 and V_2 with probability $1-Z_1$. Conditional on V_2, we sample T_2 with probability Z_2 and V_3 with probability $1-Z_2$. In the general case, conditional on V_k, we sample T_k with probability Z_k and V_{k+1} with probability $1-Z_k$ for $k=1\ldots K-1$. If we now add a prior Beta distribution with parameters α_k,β_k for each conditional Binomial distribution of the nodes V_k, we have a GD distribution where this set of Beta distributions is conjugate to the set of Binomial distributions.

$$Z_k\sim Beta(\alpha_k,\beta_k) \quad T_k\sim Bin(Z_k,N_k) \text{ for } N_k=N-\textstyle\sum_{i<k}T_i \quad (2)$$

where N is the total number of observations and N_k is the number of observations remaining from previous categories in the tree[20].

The GD distribution is a special case of the Dirichlet Tree distribution [13, 5] where a cascade hierarchy is employed in the generative process of the distribution. To interpret the parameters of the GD distribution we refer to its tree representation shown in Fig 1. Conditional on V_k, α_k is the sample size assigned to the discrete output T_k, and β_k is the sample size assigned to the next level in the tree. If β_k is too small compared to α_k, we could discard the rest of the tree. This is a desirable property as it facilitates dimensionality reduction in the number of topics. By setting $\beta_k=\alpha_{k+1}+\beta_{k+1}$, $\beta_{K-1}=\alpha_K$, we obtain a Dirichlet distribution. This is why GD distribution "generalizes" the Dirichlet distribution [4]. Although a general Dirichlet Tree structure that is inferred by data is highly appealing, in practice this structure determines the number and properties of the parameters to fit. Thus, the tree must be fixed before inferring the model [13, 14]. A key property of the GD cascade structure is that it facilitates topic reduction. This is because the tree can be pruned based on the conditional probability of going to the following level.

The GD distribution is a conjugate prior distribution to the Multinomial distribution, then we can integrate out the parameter of the Multinomial distribution leading to:

$$p(T|\alpha,\beta)=\int p(T|\theta)p(\theta)d\theta=\prod_{k=1}^{K-1}\frac{\Gamma(\alpha_k+\beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)}\prod_{k=1}^{K-1}\frac{\Gamma(\alpha'_k)+\Gamma(\beta'_k)}{\Gamma(\alpha'_k+\beta'_k)} \quad (3)$$

where $\alpha'_k=\alpha_k+T_k$, $\beta'_k=\beta_k+T_{k+1}+\ldots+T_K$. The ability to estimate this integral in closed form is crucial for GD-LDA model because this enables us to perform Gibbs sampling. This is the product of Beta-Binomial distributions for Z_k independent random variables. As a result, this expression is factorized into independent ratios of Gamma functions , which allows us to find the Maximum Likelihood Estimation (MLE) of α,β in a simple manner (independently) which is not the case of the Dirichlet distribution.

2.2 Covariances: Properties and Constraints

In natural language documents, most of the times only a few topics co-occur leading to sparse topic correlations [19]. This implies that a very few random topics may suffice, rather than the full joint distribution including all the topics.

When considering useful distributions for topic modeling, a desirable feature is to model only a few topics in a document. For instance, the topic *domestic politics* could frequently co-occur with *middle east politics* and *health care reform*. This implies that if we observe *domestic politics*, the conditional probabilities of observing *middle east politics* and *health care reform* will increase. However, if a document contains *domestic politics* and *middle east politics*, it could very rarely contain *health care reform*.

GD has the computational advantage of only having $2K-1$ parameters. This also implies that the distribution covariances are constrained. The question is whether it is possible that these constraints imply that only a few topics co-occur

Algorithm 1 GD-LDA Generative Model

```
for topic k ← 1 to K do
    draw φ_k ~ Dir(γ)
end for
for document j ← 1 to D do
    draw θ_j ~ GenDir(α, β)
    for word w ← 1 to N_j do
        draw z_wj ~ Mult(θ_j)
        draw w|z_wj = k ~ Mult(φ_k)
    end for
end for
```

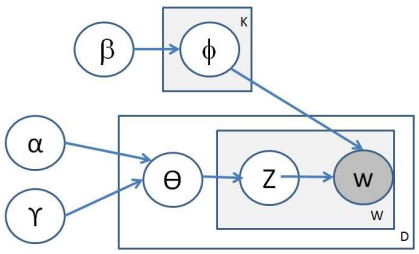

Figure 2: Graphical Model for GD-LDA. α and β are vectors of size $K-1$ where K is the number of topics. γ is a vector of size V (vocabulary size).

in a document. In this respect, the covariance properties described in Appendix A.1, particularly the third one, indicate that the covariances are dependent on the expected value of two topics, then several covariances might be close to zero if the right hand side ratio is small. This is an important feature of the GD distribution which we propose to exploit in topic modeling. A detailed discussion of the covariance constraints of GD distribution is provided in Appendix A.

3. GD-LDA: METHODOLOGY

In this section, we depict the parameter estimation process to fit the GD as a prior distribution for topics. We show that GD-LDA can be computed with a computational cost similar to that of LDA. Fig 2 shows the graphical model for GD-LDA and its generative model is described in Algorithm 1. We follow a Monte Carlo Expectation Maximization (MCEM) approach to fit our model. Conditional on the hyper-parameters, we develop a Gibbs sampling approach to infer the topic assignments to each word in the corpus. Then, assuming that the topic assignment expectations as given, we optimize the hyper-parameters of the model.

3.1 Notation

Let K be the number of topics, D the number of documents, and V the vocabulary size. We use the indices: i to denote a word or term index in the vocabulary, k to denote a specific topic, j to refer to a document, and w to identify a specific observed word. $N_{i,k}$ defines the number of observed words which correspond to term i and have been assigned to topic k. $N_{j,k}$ is the frequency of topic k in document j. We refer to $p(w, z|\cdot)$ as the joint probability of all the words w in the corpus and their topic assignments z. γ represents a vector of size V. α and β are vectors of size $K-1$.

3.2 Model and Gibbs Sampling

We define the joint probability associated with the graphical model defined in Fig 2 with joint distribution:

$$p(w, z|\alpha, \beta, \gamma) = p(w|z, \gamma)p(z|\alpha, \beta) \tag{4}$$

This expression allows us to decompose the problem into two hierarchical models that can be treated and optimized separately based on these conditional probabilities. The probability of words given topics is:

$$p(w|z, \gamma) = \int_{\phi} p(w|\phi, z)\, p(\phi|\gamma, z)\, d\phi = \prod_{k=1}^{K} \frac{\Gamma(\sum_{i=1}^{V}\gamma_i)}{\prod_{i=1}^{V}\Gamma(\gamma_i)} \frac{\prod_{i=1}^{V}\Gamma(\gamma_i + N_{k,i})}{\Gamma(\sum_{i=1}^{V}(\gamma_i + N_{k,i}))} \tag{5}$$

For the probability of topics, we have a GD prior distribution for the topic mixtures in each document which is assumed to be Multinomial. Thus we have:

$$p(z|\alpha, \beta) = \int_{\theta} p(z|\theta)\, p(\theta|\alpha, \beta)\, d\theta = \prod_{j=1}^{D}\prod_{k=1}^{K-1} \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \prod_{k=1}^{K-1} \frac{\Gamma(\alpha_k^j) + \Gamma(\beta_k^j)}{\Gamma(\alpha_k^j + \beta_k^j)} \tag{6}$$

where $\alpha_k^j = \alpha_k + N_{j,k}$, $\beta_k^j = \beta_k + N_{j,k+1} + ... + N_{j,K}$.

The topic assignments z are not observed. Then, we define a Gibbs sampling method to infer z as follows:

$$p(z_{wj} = k|z^{\neg wj}, \alpha, \beta, \gamma) = \frac{p(w|z, \gamma)p(z|\alpha, \beta)}{p(w|z^{\neg wj}, \gamma)p(z^{\neg wj}|\alpha, \beta)} \tag{7}$$

Here, $z^{\neg wj}$ represents the topic assignments for all the words except the word w from document j. This analysis leads us to the following distributions for Gibbs sampling:

$$\frac{p(w|z, \gamma)}{p(w|z^{\neg wj}, \gamma)} \propto \frac{N_{i,k}^{\neg wj} + \gamma_i}{\sum_{i=1}^{V}\left(N_{i,k}^{\neg wj} + \gamma_i\right)} \tag{8}$$

$$\frac{p(z|\alpha, \beta)}{p(z^{\neg wj}|\alpha, \beta)} \propto \begin{cases} \dfrac{\alpha_k + N_{j,k}^{\neg wj}}{\alpha_k + \beta_k + \sum_{l=1}^{K} N_{j,l}^{\neg wj}}, & k = 1 \\ \dfrac{\alpha_k + N_{j,k}^{\neg wj}}{\alpha_k + \beta_k + \sum_{l=k}^{K} N_{j,l}^{\neg wj}} \prod_{m=1}^{k-1} \dfrac{\beta_m + \sum_{l=m+1}^{K} N_{j,l}^{\neg wj}}{\alpha_m + \beta_m + \sum_{l=m}^{K} N_{j,l}^{\neg wj}} & 1 < k < K \\ \prod_{m=1}^{K-1} \dfrac{\beta_m + \sum_{l=m+1}^{K} N_{j,l}^{\neg wj}}{\alpha_m + \beta_m + \sum_{l=m}^{K} N_{j,l}^{\neg wj}} & k = K \end{cases} \tag{9}$$

In LDA, the topic distribution depends on α_k alone for the current topic k. Intuitively, if a new sample is assigned to a topic k in LDA, there is no effect on the sampling distribution of other topics. However, if a new sample is assigned to a topic k in GD-LDA, it will affect other topics through the evaluation of the product in Eq 9. The impact of this assignment will depend on the parameters α, β.

Sampling from the distribution in Eq 9 results in a high computational cost due to the cumulative product defined there. Since this distribution is not standard, we need to estimate its normalization constant. To perform this task with a computational cost comparable to collapsed Gibbs sampling in LDA [8], we compute the cumulative product for each $k > 1$ iterations and pass it to the next iteration of the evaluation. This is illustrated in Algorithm 2. From this pseudo code, $O((2K+1)W)$ operations are required for each Gibbs sampling draw for the whole corpus, where W is the total number of words. Thus, the time complexity for each iteration is $O(KW)$ as in the case of LDA [15]. This complexity is possible due to the cascade structure of the GD

Algorithm 2 GD-LDA Gibbs Sampling

$cumFactor \leftarrow 1$, $k \leftarrow 1$

$CDF[k] \leftarrow \frac{\alpha_k + N_{j,k}^{\neg wj}}{\alpha_k+\beta_k+\sum_{l=1}^{K} N_{j,l}^{\neg wj}} \times \frac{N_{i,k}^{\neg wj}+\gamma_i}{\sum_{i=1}^{V}\left(N_{i,k}^{\neg wj}+\gamma_i\right)}$

for $k \leftarrow 2$ to $K-1$ **do**

$cumFactor \leftarrow cumFactor \times \frac{\beta_{k-1}+\sum_{l=k}^{K} N_{j,l}^{\neg wj}}{\alpha_{k-1}+\beta_{k-1}+\sum_{l=k-1}^{K} N_{j,l}^{\neg wj}}$

$CDF[k] \leftarrow CDF[k-1] + cumFactor \times \frac{\alpha_k + N_{j,k}^{\neg wj}}{\alpha_k+\beta_k+\sum_{l=1}^{K} N_{j,l}^{\neg wj}} \times \frac{N_{i,k}^{\neg wj}+\gamma_i}{\sum_{i=1}^{V}\left(N_{i,k}^{\neg wj}+\gamma_i\right)}$

end for

$CDF[K] \leftarrow CDF[K-1] + cumFactor \times \frac{N_{i,K}^{\neg wj}+\gamma_i}{\sum_{i=1}^{V}\left(N_{i,K}^{\neg wj}+\gamma_i\right)}$

distribution, which facilitates the calculation of the cumulative factor in Algorithm 2. In contrast, the time complexity of a Gibbs draw for PAM depends on the number of super-topics, S, and sub-topics K as $O(SKW)$ [11]. The general recommendation is to set $S = K/2$[1] leading to $O(K^2W)$.

3.3 Parameter Estimation

Previous research has assumed uniform priors for the topic mixtures θ and the vocabulary distribution for topics ϕ without estimating them (constant values for α and γ) [8]. Wallach *et al.* in [16] concludes that parameter estimation with asymmetric Dirichlet prior probability of topics provides an improvement in the fitting. In GD-LDA, we estimate the parameters α, β of the GD distribution to discover topic correlations. We estimate the parameters for the prior distribution of words given topics γ. Ideally, we should maximize the likelihood $p(w|\alpha, \beta, \gamma)$ for observations w and hyper-parameters α, β, γ directly. Unfortunately, this distribution is intractable for this model. To solve this issue, we augment the likelihood to $p(w, z|\alpha, \beta, \gamma)$ and use Monte Carlo Expectation Maximization (MCEM) [18]. Conditional on hyper-parameters α, β, γ, we use Gibbs sampling to estimate the posterior topic assignment distribution for each word (E-step). Then, given the expected topic assignments and words, we optimize $p(w, z|\alpha, \beta, \gamma)$ (M-step). Algorithm 3 describes these iterations.

To fit γ, we maximize the joint distribution described in Eq 5 conditional on the expected topic assignments, $\overline{z} = E(z|\alpha, \beta, \gamma)$, estimated from Gibbs sampling. Then we have the following optimal function:

$$\gamma^{new} = \arg\max_{\gamma} \prod_{k=1}^{K} \frac{\Gamma(\sum_{i=1}^{V}\gamma_i)}{\prod_{i=1}^{V}\Gamma(\gamma_i)} \frac{\prod_{i=1}^{V}\Gamma(\gamma_i + \overline{N}_{k,i})}{\Gamma(\sum_{i=1}^{V}(\gamma_i + \overline{N}_{k,i}))} \tag{10}$$

where: $\overline{N}_{k,i} = f(\overline{z})$

We follow the Newton-based approach proposed by Minka in Eqs 56-60 of [12]. Here, we have a DCM distribution to fit from K observed vectors of dimension V. To initialize the search, we use the method of moments based on the observed proportions $p_{k,i} = \overline{N}_{k,i} / \sum_{i=1}^{V} \overline{N}_{k,i}$.

Similarly, we estimate the parameters of the GD distribution by maximizing the joint distribution:

$$\alpha^{new}, \beta^{new} = \arg\max_{\alpha,\beta} \prod_{j=1}^{D} \prod_{k=1}^{K-1} \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \prod_{k=1}^{K-1} \frac{\Gamma(\alpha_k^j) + \Gamma(\beta_k^j)}{\Gamma(\alpha_k^j + \beta_k^j)} \tag{11}$$

[1] $K/2$ super-topics is the recommendation by Mallet toolbox, `http://mallet.cs.umass.edu/`

Algorithm 3 Monte Carlo EM

Start with an initial guess for α, β, γ and z

repeat

Run Gibbs Sampling using Eqs. 8 and 9

Find Expected value for topic assignments $E(z|\alpha, \beta, \gamma) = \overline{z}$

Choose γ to maximize complete Likelihood Eqs 56-60 in [12]

Choose α, β that maximize complete Likelihood using Eq 27

until convergence of α, β, γ

Choose topic assignments z^* with highest probability

Set $\alpha^* = \alpha, \beta^* = \beta, \gamma^* = \gamma$

return $\alpha^*, \beta^*, \gamma^*, z^*$

where $\alpha_k^j = \alpha_k + \overline{N}_{j,k}$, $\beta_k^j = \beta_k + \overline{N}_{j,k+1} + ... + \overline{N}_{j,K}$.

We develop a Newton-based method in Appendix B. To initialize the search, we use the method of moments based on the conditional beta distributions from the tree representation of the GD distribution [21]. The proportions $p_{k,i} = \overline{N}_{k,i} / \sum_{i=1}^{V} \overline{N}_{k,i}$ are employed for this initialization.

One key component of this optimization is that each pair α_k, β_k is optimized separately. Thus, the time complexity for this optimization is linear in time with K. Parameter fitting in PAM is performed using the method of moments due the high model complexity and computational cost of the optimization [10, 11]. In addition, the number of parameters to fit for CTM and PAM is K^2 [2] [19], assuming $K/2$ super-topics for PAM. As a result, these methods are highly prone to over-fitting, as many parameters are being fit.

3.4 Predictive Distributions

Given the optimal parameters (α^*, β^*) obtained from Algorithm 3, and the word-topic observations (w, z) the predictive distribution for document j, $\hat{\theta}_j$, is estimated as:

$$\hat{\theta}_{jk} = \begin{cases} \frac{\alpha_k^* + \overline{N}_{j,k}}{\alpha_k^* + \beta_k^* + \sum_{l=1}^{K} \overline{N}_{j,l}}, & \text{if } k = 1 \\ \frac{\alpha_k^* + \overline{N}_{j,k}}{\alpha_k^* + \beta_k^* + \sum_{l=k}^{K} \overline{N}_{j,l}} \times \prod_{m=1}^{k-1} \frac{\beta_m^* + \sum_{l=m+1}^{K} \overline{N}_{j,l}}{\alpha_m^* + \beta_m^* + \sum_{l=m}^{K} \overline{N}_{j,l}} & \text{if } 1 < k < K \\ \prod_{m=1}^{K-1} \frac{\beta_m^* + \sum_{l=m+1}^{K} \overline{N}_{j,l}}{\alpha_m^* + \beta_m^* + \sum_{l=m}^{K} \overline{N}_{j,l}} & \text{if } k = K \end{cases} \tag{12}$$

for the topics $k = 1 \ldots K$ and the documents $j = 1 \ldots D$.

The predictive distribution for the probability of words given topics, $\hat{\phi}_k$ is estimated as follows [8]:

$$\hat{\phi}_{ki} = \frac{\overline{N}_{i,k} + \gamma_i^*}{\sum_{i=1}^{V}\left(\overline{N}_{i,k} + \gamma_i^*\right)} \tag{13}$$

Notice that the probability of topics $\hat{\theta}_j$ is document dependent. On the other hand, the probability of words given their topic $\hat{\phi}_k$ is topic dependent. This implies that for an unseen document, we need to estimate its predictive distribution of topics while the probability of words given their topics remains the same.

4. EXPERIMENTAL RESULTS

4.1 Validation

The main challenge to validate statistical topic models is the lack of reliable observed topic labels for each word of a corpus. The standard approach is to estimate the likelihood of held-out data [17]. In addition, model performance in a

Algorithm 4 EL Estimation for model M

for $s \leftarrow 1$ to N_s **do**
  Sample document d_s with mixture θ_s given model M
  Find $p(w = i|\theta_s, \hat{\phi}, M) = \sum_{k=1}^{K} \theta_{sk}\hat{\phi}_{ki}$ for $i = 1, \ldots, V$
end for
Set logEL= 0
for $t \leftarrow 1$ to D_t **do**
  Find $p(d_t|d_s, M) = \prod_{w_t=1}^{N_t} p(w_t|\theta_s, \hat{\phi}, M)$
  Find $p(d_t|M) = \frac{1}{N_s}\sum_{s=1}^{N_s} p(d_t|d_s, M)$
  Update logEL=logEL+$Log(p(d_t|M))$
end for
return logEL

supervised task has also been used [22]. Here, we validate the GD-LDA performance addressing these two forms. We estimate the likelihood of completely unseen documents using Empirical Likelihood, and we compare the performance of topic models in the ad hoc Information Retrieval as developed in [19].

4.1.1 Empirical Likelihood (EL)

There has been a debate about the evaluation of topic models with methods ranging from the harmonic mean of complete likelihood for topic assignments, to perplexity and empirical likelihood. As discussed in [17], perplexity falls into the category of document completion where a portion each document must be observed to estimate the likelihood of the remaining content. Similarly, a "left to right" evaluation has been proposed to estimate the probability of words in a test document incrementally [17]. For validation, we use Empirical Likelihood (EL) criterion. Our intent is to predict the likelihood of fully unseen documents. We do not use perplexity or "left to right" evaluation because they are based on the word order inside the document, in contrast to "bag of words". In addition, EL has been shown to be a more pessimistic approximation for the probability of held-out documents than "left to right" method [17].

We estimate EL as described in [11]. Here, we generate a set of pseudo documents, θ_s using the estimated prior topic distribution of the model being tested using training data. Then, a word distribution is estimated for each θ_s based on $\hat{\phi}$. This is used to estimate the probability of seeing the test set. We define this process by means of Algorithm 4. Here, D_t represents the number of documents for a test set, d_t is the document $t = 1, \ldots, D_t$, w_t are the N_t words of d_t, and N_s is the number of pseudo documents given model M. We use the generative approach based on the conditional Beta distributions of the tree representation for GD as in Fig 1. The main limitation of EL estimation is that the number of pseudo documents should be sufficiently large to cover the parameter space of θ given the trained model. Then, N_s determines the accuracy of the approximation.

4.1.2 Ad hoc Retrieval

Information Retrieval represents a hard problem where the gains of new models are often small or nil. This application represents a good test of the power of our approach. We compare the performance of different topic models in ad hoc Information Retrieval (IR). We use the approach proposed in [19] and incorporate our model by replacing the LDA model described in the method. Based on the predictive distribution estimators for topics $\hat{\phi}$, and for each document $\hat{\theta}_j$, we provide a topic-based language model for each document, $P_{TM}(w|\theta_j)$, as follows:

$$P_{TM}(w|\hat{\theta}_j, \hat{\phi}) = \sum_{k=1}^{K} P(w|z = k, \hat{\phi_k})P(z = k|\hat{\theta_j}) \quad (14)$$

This is augmented with the maximum likelihood estimate for the language model based on document terms D_j, $P_{ML}(w|D_j)$, and for the language model based on the corpus C, $P_{ML}(w|C)$, leading to:

$$P_{IR}(w|D_j, \hat{\theta}_j, \hat{\phi}) = \lambda\left(\frac{N_j}{N_j+\mu}P_{LM}(w|D_j) + \frac{\mu}{N_j+\mu}P_{ML}(w|C)\right) + (1-\lambda)P_{TM}(w|\hat{\theta}_j, \hat{\phi}) \quad (15)$$

where μ is a smoothing parameter, N_j is the number of words in document j, and λ is a parameter for the linear combination. Therefore, the ranking function for query Q with terms q is given by:

$$P(Q|D_j, \hat{\theta}_j, \hat{\phi}) = \prod_{q \in Q} P_{IR}(w = q|D_j, \hat{\theta}_j, \hat{\phi}) \quad (16)$$

We use the parameter values: μ=1000, λ=0.7. These are the values recommended in [19].

Table 1: Features of the datasets analyzed. Mean document includes the 95% interval

Dataset	NIPS	NYT	APW	OHSUMED
# documents	1840	5553	14657	196404
# unique terms	13649	11229	18471	38900
Mean doc. length	1322 ± 274	274 ± 132	169 ± 76	186 ± 36

Table 2: Number of Gibbs samples per EM iteration

MCEM iteration	1-4	5-6	7-8	9	10
Burn-in Samples	10	15	20	40	50
Gibbs Samples	50	100	200	500	700

4.2 Experimental Settings

We use 4 different datasets to test our model. NIPS conference papers dataset[2] which contains long documents (8-10 pages). Two news datasets, NYT and APW, obtained from TREC-3. These collections contain shorter documents (1 page) with different vocabulary size. A fourth dataset, OHSUMED from TREC-9, consists of abstracts from medical papers. In contrast to the other datasets, the number of documents is much larger (more than ten times). Table 1 shows the features of these datasets. We remove standard stop words and perform stemming in all datasets.

We test 4 variants of the GD-LDA algorithm. In two of these cases we fix the value of γ and optimize the parameters α and β as in Algorithm 3 based on two variants: 1) using the empirical expected topic assignments $\bar{z}$ [18], *GD-LDA Fixed Gamma Mean*; and 2) using the empirical mode z^* [7] in the M-step, *GD-LDA Fixed Gamma Max*. In the other two cases, we optimize all the three parameters (α,β and γ) based on the expected topic assignments, *GD-LDA Mean*, and the topic assignment mode *GD-LDA Max*.

We compare GD-LDA, LDA with parameter optimization, CTM and PAM. For LDA, we use our own implementation. We follow the collapsed Gibbs sampling approach [8], with asymmetric Dirichlet prior distributions for both, the probability of words given a topic, and for the probability of topics. These parameters are optimized as discussed by authors in [6]. For optimization, we follow the Newton-based approach proposed by Minka in Eqs 56-60 [12]. For both

[2]Available at: `http://cs.nyu.edu/~roweis/data.html`

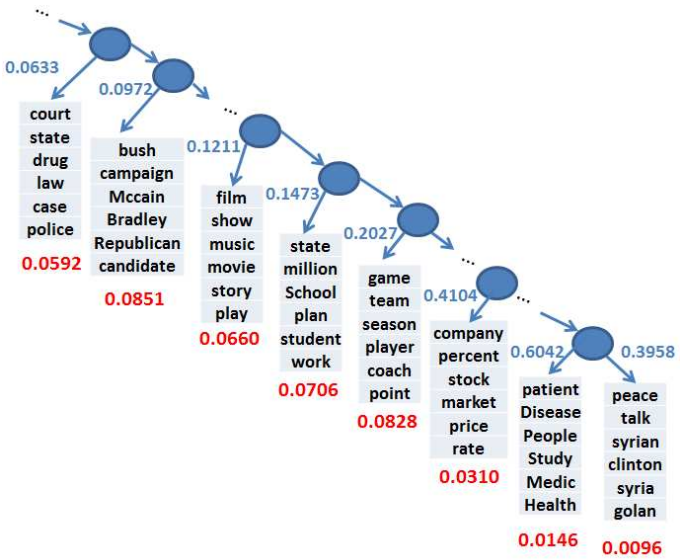

Figure 3: Qualitative example of GD-LDA for the NYT dataset with 20 topics. Probability of topics: marginal probability as defined by Eq 18 (red), conditional probability given the parent node as defined by Eq 22 (blue).

GD-LDA and LDA, we set a maximum of 1,000 Newton iterations for parameter optimization. The schedule of the Gibbs sampling is detailed in Table 2. We initialize the values $\alpha_k = 2/K$ for $k = 1, \ldots, K-1$, and $\beta_{K-1} = 2/K$ in GD-LDA to obtain the special case of Dirichlet distribution; we start with LDA as prior model, and allow the data to adapt the model in the MCEM iterations to the more general GD. For CTM[3], we use the default settings with parameter estimation: a maximum of 1,000 EM iterations with convergence of 10^{-5} and maximum of 20 variational iterations with a convergence rate of 10^{-6}. For PAM[4] we use 1000 Gibbs samples and $K/2$ super topics. This implementation supports multi-threading; to enable a fair comparison we use only one thread. We modify the implementation to obtain the fitted parameters since they are not provided by default. For EL estimation, we perform 10-fold cross-validation and we sample $N_s = 10,000$ pseudo documents.

4.3 Results

4.3.1 Qualitative Results

Fig 3 shows a qualitative example of GD-LDA in the NYT dataset. The figure represents the empirically estimated probability of topics for a subset of the topics. Each box represents a topic and the terms displayed are the ones with highest posterior predictive probability $\hat{\phi}_k$. Given the fitted GD-LDA model, we calculate the point estimate of the total probability for each topic (in blue) and its conditional probability given a parent node (red). Here, conditional on the current node, we move deeper into the tree or observe a word from the topic at that level. As we move down into the tree, the conditional probability of picking the left hand topic increases since the probability mass of the remaining topics is less. This is a desirable property which facilitates

[3]We use the CTM implementation provided by Blei at `http://www.cs.princeton.edu/~blei/ctm-c/`.

[4]We use the implementation provided by Mallet `http://mallet.cs.umass.edu/`.

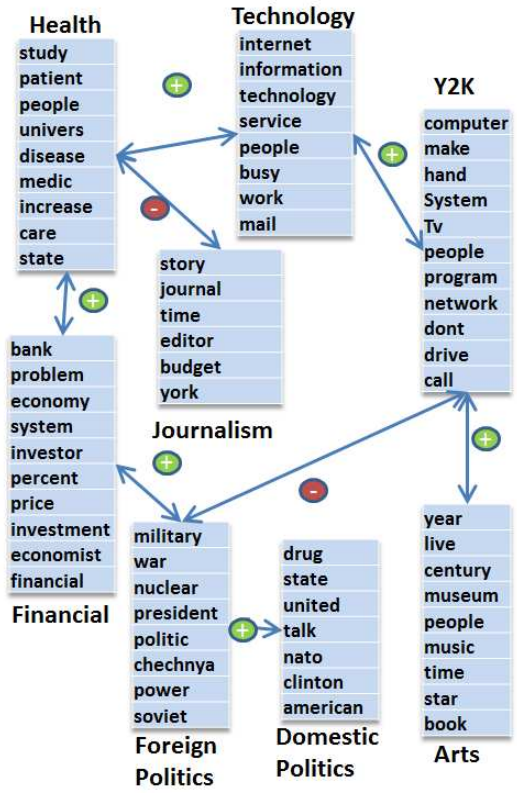

Figure 4: Correlation graph of a subset of topics inferred by GD-LDA for the NYT dataset with 30 topics.

dimensionality reduction. If this probability is *large* compared with the probability of exploring further the tree, we can discard the remaining tree.

Fig 4 shows some of the positive and negative correlations between different topics inferred by GD-LDA. We observe that *Health* is positively correlated with *Technology* and *Financial*. *Y2K* is positively correlated with *Technology* and *Arts* but negatively correlated with *Foreign Politics*. Note that by default LDA assumes negative correlations, thus the main value of topic correlated models is to discover positive correlations. Fig 5(a) shows the decomposition of a *sports* document from the NYT dataset into topics using PAM with 10 super-topics and 20 sub-topics, CTM and GD-LDA with 20 topics. This shows that CTM and GD-LDA assign a high probability to a single topic (sports). On the other hand, PAM provides 3 sub-topics and most of the super-topics with a significant probability mass. This is not desirable since an important objective with topic modeling is to cluster documents based on the topic mixture. Fig 5(b) shows a comparison between CTM and GD-LDA for a document about the *Y2K* problem and *airport functionality*. We observe that GD-LDA provides better segmentation of the document based on the topics proportions and content.

In addition, we calculate the distribution of the number of significant topics in a document. We rank the topics based on their topic probability mass in each document. Then, we estimate the number of topics which accounts for 95% of this probability mass. Fig 6 shows the distribution of this number of topics inferred by LDA, CTM, PAM, and GD-LDA. For PAM, we consider the number of sub-topics. Here, we observe that CTM favors high number of topics for

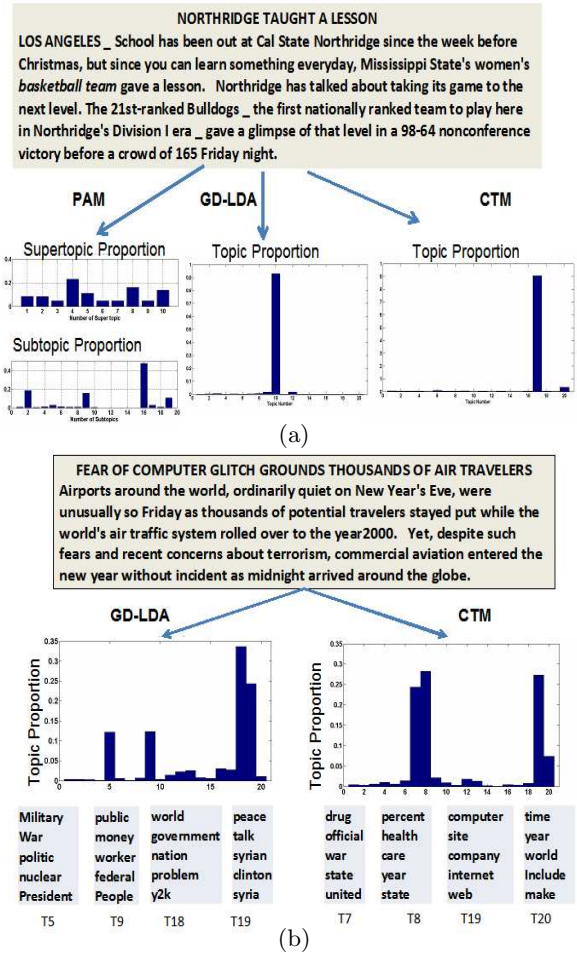

Figure 5: Qualitative comparison of the topic distribution for two documents from the NYT dataset. (a) From left to right: PAM, CTM, and GD-LDA (b) From left to right: GD-LDA and CTM

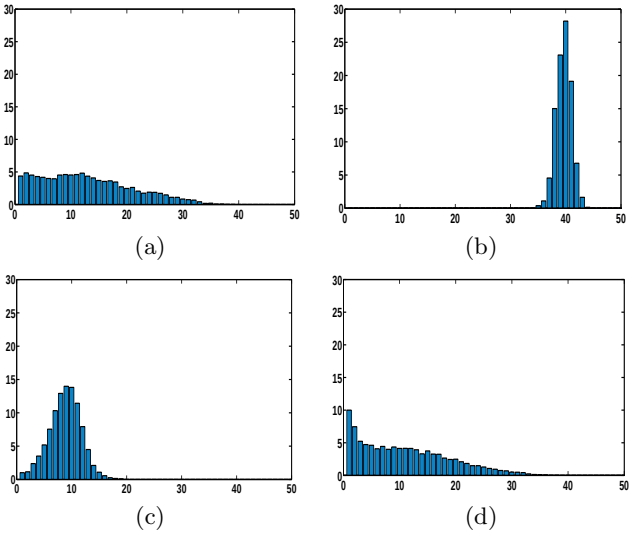

Figure 6: Distribution of the number of significant topics per document for the APW corpus inferred by: (a)LDA, (b)CTM, (c)PAM, and (d)GDLDA using $K = 50$ topics. X-axis is number of topics per documents and Y-axis is the percentage of documents in the corpus.

each documents which introduces noise when the goal is to characterize documents as in Information Retrieval. In PAM the number of topics per document is highly dependent of the number of super topics used. In order to handle topic correlation sparsity, PAM prunes the relationships between a super topic and subtopics [11]. This reduces the amount of correlations that can be modeled. In contrast, GD-LDA favors smaller number of topics per documents and without any constraint in the inference. The key property for this behavior is the cascade structure of the distribution. As discussed in section 2.2, this validates empirically the intuition of few significant topics for natural language documents, and sparse correlations.

4.3.2 *Empirical Likelihood Results*

Fig 7 shows EL estimations for the 4 variants of GD-LDA, CTM, PAM and LDA with asymmetric prior and parameter optimization in the four datasets. For the NIPS dataset, LDA is not shown since its predictive likelihood is extremely low. Similarly, PAM performance for the OHSUMED dataset is not shown for the same reason. The CTM model would not run for the OHSUMED dataset (mid-size dataset) based on the number of documents and vocabulary size.

We observe that optimizing γ has a low impact as more data becomes available. Conceptually, a prior distribution represents the prior knowledge with a given sample size. Consequently adding more data decreases the impact of the prior information. This is consistent with the conclusions discussed in [16]. We then compare the use of the topic assignments, $\bar{z}$ (posterior mean), and the topic assignment mode, z^* (maximum a posteriori MAP). We find that *GD-LDA mean* performs better, when the number of documents and the vocabulary size are relatively small (NIPS dataset). In contrast, *GD-LDA Max* shows the highest performance for the other datasets, Fig 7(b)-(d). In particular, this method performance is clearly superior for all the topics when tested for the largest dataset (OHSUMED). *GD-LDA Mean* requires fewer topics to achieve a superior performance compared to LDA, CTM and PAM. Moreover, EL decreases more smoothly than these methods. It also remains fairly constant when the number of topics attains a high value.

An important difference between the datasets analyzed is the average document length and number of unique terms. The worst performance of CTM, as well as PAM, is found in the case of APW dataset. This dataset has a larger vocabulary size, and significantly shorter documents than NIPS dataset, where both CTM and PAM show a similar performance to that of *GD-LDA Mean*. This suggests over-fitting by CTM where a full topic covariance matrix is estimated, and by PAM where a matrix of K subtopics by $K/2$ supertopics is estimated. In addition, PAM uses the method of moments, instead of optimization, for parameter estimation. This is a limitation of PAM when EL based evaluation is performed since the fitted parameters do not optimize the likelihood. As discussed in section 4.3.1, CTM favors larger number of topics for each document than the other methods. This behavior introduces noise in EL because correlations are sparse, and only a few topics are present in natural language documents.

4.3.3 *Ad hoc Information Retrieval Results*

We show the application of GD-LDA in ad hoc IR using the OHSUMED dataset. As discussed above, we use the ap-

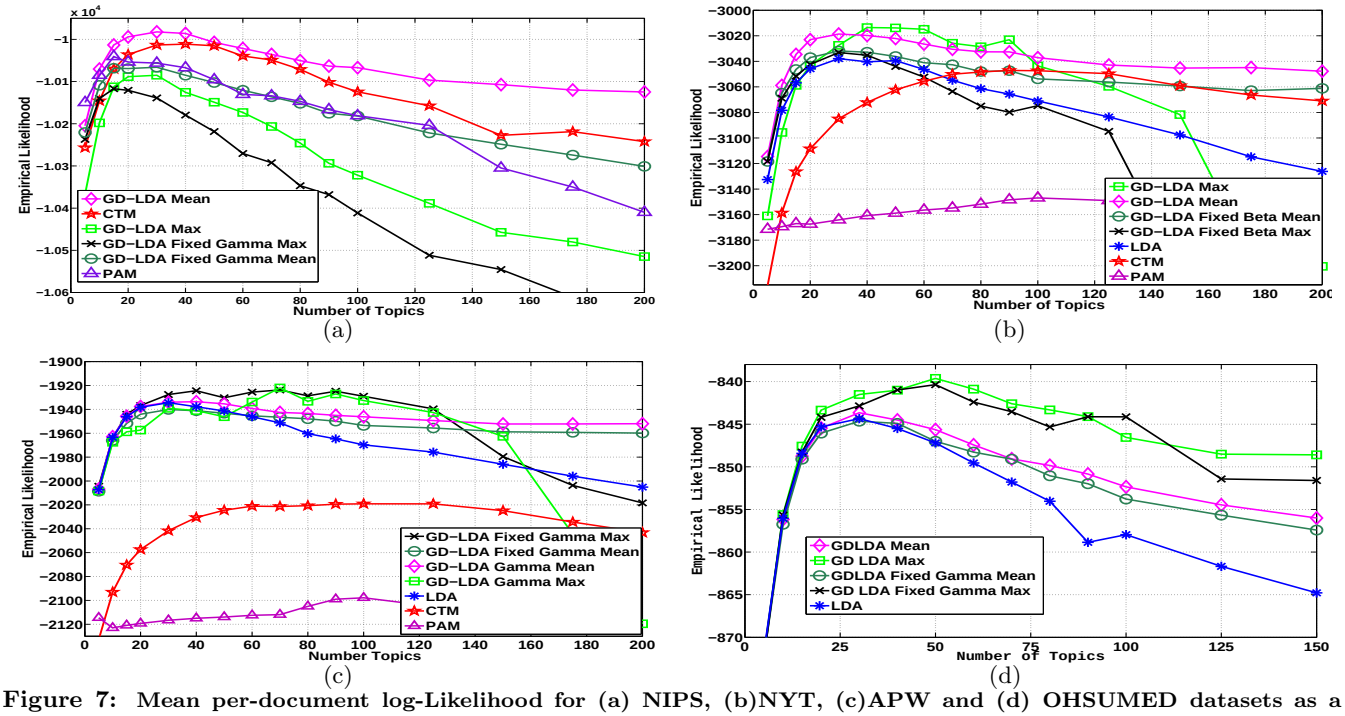

Figure 7: Mean per-document log-Likelihood for (a) NIPS, (b)NYT, (c)APW and (d) OHSUMED datasets as a function of the number of topics.

Table 3: Results of ad hoc IR using GDLDA, LDA and PAM models using K topics for the OHSUMED dataset

Model	K	P@10	MAP	DG
GD-LDA	50	**0.502**±0.06	**0.496**±0.06	**0.790**±0.07
LDA	50	0.470±0.06	0.470±0.06	0.780±0.08
PAM	50	0.434±0.06	0.439±0.05	0.671±0.04
GD-LDA	75	0.470±0.06	**0.491**±0.06	0.742±0.04
LDA	75	0.470±0.06	0.461±0.06	0.741±0.04
PAM	75	0.431±0.06	0.438±0.05	0.651±0.03

proach from [19] as a benchmark for comparison. We train the *GD-LDA Max* using $K = 50$ and $K = 75$ topics, and compare its performance with LDA and PAM. We use standard IR measures: Precision at 10 (P@10), Mean Average Precision (MAP) and Discounted Gain (DG), to compare these methods. The OHSUMED dataset is considered to be a medium size dataset in the IR literature. This contains 63 topical queries with 35,000 relevant labels.

We have not modified the retrieval model of [19] to exploit the power of the new GD-LDA model. Despite this, the performance improvement for all the measures is significant as we observe in Table 3. Here, the best performance is for $K = 50$ topics, where EL estimation peaks in Fig 7(d). For this case, there is agreement between the adhoc IR and EL performance. Respect to LDA, GD-LDA shows improvement of: **6.3%** for P@10, **5.5%** for MAP, and **1.1%** for DG. Note that PAM shows lower performance than LDA in IR for all the measures. This is consistent with the IR performance of PAM reported in [22].

4.3.4 Computational Cost Comparison

One advantage of GD-LDA over CTM and PAM is that its computational complexity is linear in the number of topics. This is not the case for CTM and PAM which scale quadratically with the number of topics. As discussed in section 3, GD-LDA should add minimal computational cost to LDA. Fig 8 shows the computational time in minutes to fit the model for LDA, CTM, PAM and GD-LDA in the datasets we consider. A non-linear increase in time is observed after 50 topics for CTM in NYT and APW datasets, and after 80 topics in NIPS dataset. We observe that the computational cost of PAM grows quadratically after 20 topics. In general, the computational cost of GD-LDA is comparable to that of LDA and is less than CTM and PAM. This is a significant advantage since GD-LDA provides a more flexible model structure when compared with LDA. Moreover, variances and a number of covariances are modeled more effectively, and with minimal increase in computational cost. This property makes GD-LDA more suitable than CTM or PAM for larger scale applications, such as IR.

4.3.5 Choosing the Number of Topics

An open problem is how to select the optimal number of topics to train a model. In general, an exhaustive experimentation needs to be performed to select the optimal number of topics. Due to the tree structure of the GD distribution, GD-LDA tends to accommodate the most relevant topics, based on probability mass, at the upper levels of the tree.

Fig 9 shows the cumulative probability vs the number of topics of the expected topic mixture of documents given the fitted parameters for GD-LDA and CTM. We observe that GD-LDA favors a smaller number of topics. Notice that, after a certain number of topics, the expected contribution of the remaining ones is not significant. This prevents GD-LDA from over-fitting. When comparing GD-LDA with CTM, we observe that CTM favors uniform topic mixtures in each document. Thus, if we fit CTM for larger number of topics, we would observe the same linear behavior as seen in graphs from Fig 9. This prevents CTM from discarding any topic easily or suggesting an optimal topic range as opposed to GD-LDA. In addition, this behavior makes CTM less effective in handling over-fitting.

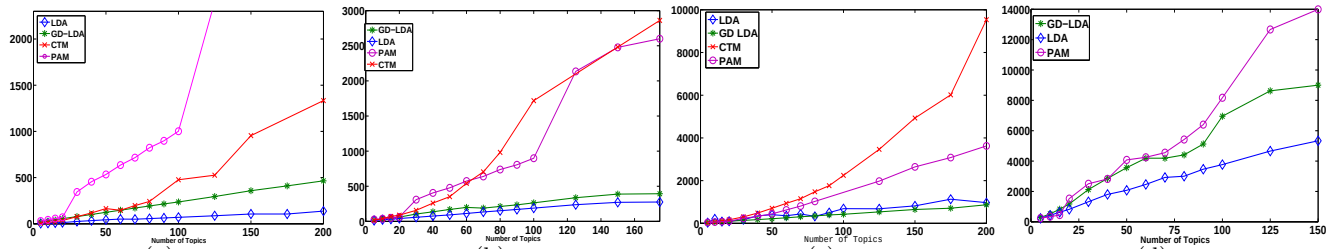

Figure 8: Average running times in minutes of CTM, GD-LDA and LDA for: (a) NIPS, (b) APW, (c) NYT, (d) OHSUMED. x-axis represents the number of topics. All the experiments were performed using a Quad Intel machine with 2.5GHz with 8GB of memory.

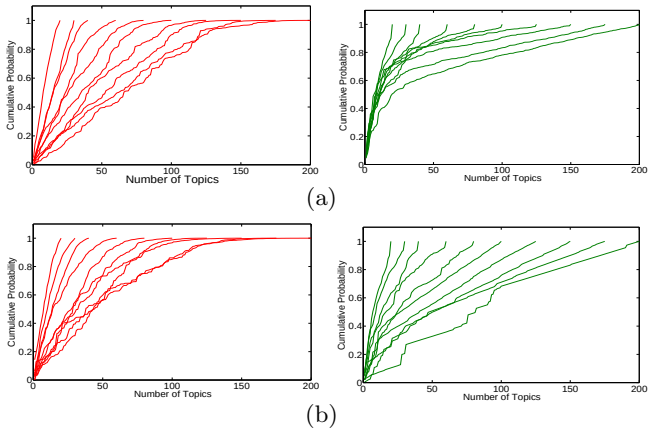

Figure 9: Cumulative distribution of the topic mixtures based on the fitted prior distribution parameters. From left to right GD-LDA and CTM for: (a) NIPS and (b) APW. The distributions are shown for $\{20, 30, 40, 60, 80, 100, 125, 150, 175, 200\}$ topics.

5. DISCUSSION AND CONCLUSION

We have introduced the use of the GD distribution in probabilistic topic modeling. The advantages of the GD over the Dirichlet distribution, and the benefits when compared with the estimation of the full covariance matrix in CTM have been described. The apparent constraints on covariances in GD actually results in modeling better sparse topic correlations in natural language documents, as our empirical validation indicates. This results in better performance in empirical likelihood and IR measures. We have developed an efficient Gibbs sampling model which uses the conjugacy property of GD with the Multinomial distribution. We have demonstrated that the running time of GD-LDA is comparable to LDA and less than CTM and PAM. This provides a model computationally competitive and with better performance than these methods.

We have shown that the impact of optimizing the vocabulary parameter γ decreases when the vocabulary size and the number of documents in the corpus is large. Due to the tree structure of the GD distribution, GD-LDA proves to be powerful in handling over-fitting with a large number of topics as its performance remains fairly high even when the number of topics is increased. This is not the case for CTM, PAM, and LDA. As a consequence, we can reduce the number of topics, by using the conditional probability of the remaining topics when we are moving down into the tree. A natural extension of the model is to allow the tree structure to be fitted. A direction of improvement is to modify the Dirichlet tree to have a comparable notion of subtopics and super topics as in PAM with the same conjugacy and computational cost as GD-LDA.

We have shown that the use of GD-LDA in adhoc IR increases the performance significantly, in contrast to earlier incorporations of topic models. Future directions include the use of topic mixture instead the probability of words in the ranking function. In addition, we plan to explore the impact of GD-LDA in Interactive Information Retrieval.

6. ACKNOWLEDGEMENTS

This work is partially funded by CONACYT grant 207751 and CONACYT UC-MEXUS grant 194880.

7. REFERENCES

[1] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *ICML*, pages 25–32, 2009.

[2] D. Blei and Lafferty. Correlated topic models. *NIPS*, 18:147–154, 2006.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022, 2003.

[4] R. J. Connor and J. E. Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of American Statistics*, 64:194–206, 1969.

[5] S. Y. Dennis. A bayesian analysis of tree-structured statistical decision problems. *Journal of Statistical Planning and Inference*, 53(3):323–344, 1996.

[6] G. Doyle and C. Elkan. Accounting for burstiness in topic models. *Proceedings of the 26th ICML*, 2009.

[7] C. Elkan. Clustering documents with an exponential family approximation of the dirichlet compound multinomial distribution. *In Proceedings of 23rd ICML*, 2006.

[8] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the NAS*, 101(6):5228–5235, 2004.

[9] W. Li, D. Blei, and A. McCallum. Nonparametric Bayes pachinko allocation. In *23rd UAI Conference*, 2007.

[10] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of ICML 06*, pages 577–584, 2006.

[11] W. Li and A. McCallum. Pachinko allocation: Scalable mixture models of topic correlations. *Journal of Machine Learning*, 2008. Submitted.

[12] T. P. Minka. Estimating a dirichlet distribution. 2003.

[13] T. P. Minka. The dirichlet-tree distribution. 2004.

[14] B. Null. The nested dirichlet distribution: Properties and applications. *Submitted*, 2008.

[15] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. *Proceedings of the KDD 2008*, 2008.

[16] H. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In *NIPS 22*, pages 1973–1981. 2009.

[17] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of ICML '09*, pages 1105–1112, 2009.
[18] G. C. G. Wei and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.
[19] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. *In Proceedings SIGIR 2006*, 2006.
[20] T.-T. Wong. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, 97:165–181, 1998.
[21] T.-T. Wong. Parameter estimation for generalized dirichlet distributions from the sample estimates of the first and the second moments of random variables. *Comput. Stat. Data Anal.*, 54(7):1756–1765, 2010.
[22] X. Yi and J. Allan. Evaluating topic models for information retrieval. *In Proceedings CIKM*, 2008.

APPENDIX

A. FIRST AND SECOND MOMENTS OF THE GD DISTRIBUTION

We derive the first and second moments of the GD distribution based on the tree representation of Eq 2. We write:

$$\theta_1 = Z_1 \quad \theta_k = Z_k(1-\theta_1-\cdots-\theta_{k-1}) = Z_k\prod_{m=1}^{k-1}(1-Z_m) \tag{17}$$

Let $S_k = E(Z_k)$, $R_k = E(Z_k^2)$. Since Z_k's are independent:

$$E(\theta_k) = S_k\prod_{m=1}^{k-1}(1-S_m), \quad E(\theta_k^2) = R_k\prod_{m=1}^{k-1}(1-2S_m+R_m) \tag{18}$$

Similarly for the crossproducts where $k<j$:

$$\begin{aligned}
&E(\theta_k\theta_j) = E\left\{Z_k\left[\prod_{m=1}^{k-1}(1-Z_m)\right]Z_j\left[\prod_{m=1}^{j-1}(1-Z_m)\right]\right\}\\
&= E\left\{\left[\prod_{m=1}^{k-1}(1-Z_m)^2\right]Z_k(1-Z_k)\left[\prod_{m=i+1}^{j-1}(1-Z_m)\right]Z_j\right\}\\
&= \left[\prod_{m=1}^{k-1}(1-2S_m+R_m)\right](S_k-R_k)\left[\prod_{m=i+1}^{j-1}(1-S_m)\right]S_j
\end{aligned} \tag{19}$$

Therefore, we have for the variance and covariance:

$$\begin{aligned}
Var(\theta_k) &= R_k\left[\prod_{m=1}^{k-1}(1-2S_m+R_m)\right] - S_k^2\prod_{m=1}^{k-1}(1-S_m)^2\\
&= R_kE(V_k^2) - S_k^2(E(V_k))^2
\end{aligned} \tag{20}$$

$$\begin{aligned}
&Cov(\theta_k,\theta_j) = S_j\left[\prod_{m=i+1}^{j-1}(1-S_m)\right]\\
&\left\{(S_k-R_k)\prod_{m=1}^{k-1}(1-2S_m+R_m) - (S_k-S_k^2)\prod_{m=1}^{k-1}(1-S_m)^2\right\}\\
&= \left[E(\theta_j)/E(V_{k+1}^2)\right]\left[S_kVar(V_k) - Var(\theta_k)\right]
\end{aligned} \tag{21}$$

for $k=1,\ldots,K$ in Eq 20, and for $k=1,\ldots,K-1$ and $j=k+1,\ldots,K$ in Eq 21. We estimate S_k, R_k based on the independent Beta distributions of Z_k. From the moment generating function for the Beta distribution we have:

$$S_k = \frac{\alpha_k}{\alpha_k+\beta_k} \quad R_k = \frac{\alpha_k(\alpha_k+1)}{(\alpha_k+\beta_k)(\alpha_k+\beta_k+1)} \quad k=1,\ldots,K \tag{22}$$

where $\alpha_K = 1, \beta_K = 0$ [4]. Note that the variances and the expected value are not constrained.

A.1 Covariance Properties of GD Distribution

Three key properties can be derived from Eqs 20-22:

1. For $j>k$, $Cov(\theta_k,\theta_j) > 0$ if and only if $Var(V_k) > Var(\theta_k)$. Thus, $Cov(\theta_k,\theta_j) > 0$ if and only if the summation of the first $k-1$ θ's varies more than θ_k
2. For $j>1$, $Cov(\theta_1,\theta_j) = -Var(\theta_1)E(\theta_j)/E(1-\theta_1) < 0$ since θ_1 is at the root level of the tree
3. For $j>k+1$, $Cov(\theta_k,\theta_j) = Cov(\theta_k,\theta_{k+1})E(\theta_j)/E(\theta_{k+1})$. Thus, $Cov(\theta_k,\theta_j)$ at deeper levels of the tree will have the same sign as $Cov(\theta_k,\theta_{k+1})$

The GD distribution models the $Cov(\theta_k,\theta_{k+1})$ for consecutive tree levels without constraints. The covariances for deeper levels in the tree $Cov(\theta_k,\theta_j), j>k+1$ are constrained. Since θ_1 is used as base category, it is always negatively correlated with the other categories. This is a typical constraint of the Logistic Normal distribution used in CTM[2]. The GD distribution constrains the sign of $Cov(\theta_k,\theta_j)$ to be the same as that of $Cov(\theta_k,\theta_{k+1})$ for $j>k+1$. These constraints imply that probability of co-occurrence of these topics is very small.

B. MAXIMIZATION OF α AND β FOR GD

To estimate the parameters of the GD distribution, the log-likelihood of Eq. 6 is optimized using the Newton method.

$$\begin{aligned}
L(\alpha,\beta) = &\sum_{j=1}^{D}\sum_{k=1}^{K-1}\log\Gamma(\alpha_k+\beta_k) - \sum_{k=1}^{K-1}(\log\Gamma(\alpha_k)+\log\Gamma(\beta_k))\\
&+\sum_{k=1}^{K-1}\left(\log\Gamma(\alpha_k^j)+\log\Gamma(\beta_k^j)\right) - \sum_{k=1}^{K-1}\log\Gamma(\alpha_k^j+\beta_k^j)
\end{aligned} \tag{23}$$

By taking the derivative with respect to α,β we have:

$$\begin{aligned}
\frac{\partial L(\alpha,\beta)}{\partial\alpha_k} &= D\Psi(\alpha_k+\beta_k) - D\Psi(\alpha_k) + \sum_{j=1}^{D}\Psi(\alpha_k^j) - \sum_{j=1}^{D}\Psi(\alpha_k^j+\beta_k^j)\\
\frac{\partial L(\alpha,\beta)}{\partial\beta_k} &= D\Psi(\alpha_k+\beta_k) - D\Psi(\beta_k) + \sum_{j=1}^{D}\Psi(\beta_k^j) - \sum_{j=1}^{D}\Psi(\alpha_k^j+\beta_k^j)
\end{aligned} \tag{24}$$

Recall that $\alpha_k^j = \alpha_k + \overline{N}_{j,k}$, $\beta_k^j = \beta_k + \overline{N}_{j,k+1} + \ldots + \overline{N}_{j,K}$. Notice that the derivative of $L(\alpha,\beta)$ with respect to α_k just depends on α_k and β_k as opposed to the Dirichlet distribution. For the second derivative we have:

$$\begin{aligned}
\frac{\partial^2 L(\alpha,\beta)}{(\partial\alpha_k)^2} &= D\Psi'(\alpha_k+\beta_k) - D\Psi'(\alpha_k) + \sum_{j=1}^{D}\Psi'(\alpha_k^j) - \sum_{j=1}^{D}\Psi'(\alpha_k^j+\beta_k^j)\\
\frac{\partial^2 L(\alpha,\beta)}{(\partial\beta_k)^2} &= D\Psi'(\alpha_k+\beta_k) - D\Psi'(\beta_k) + \sum_{j=1}^{D}\Psi'(\beta_k^j) - \sum_{j=1}^{D}\Psi'(\alpha_k^j+\beta_k^j)\\
\frac{\partial^2 L(\alpha,\beta)}{\partial\alpha_k\partial\beta_k} &= D\Psi'(\alpha_k+\beta_k) - \sum_{j=1}^{D}\Psi'(\alpha_k^j+\beta_k^j)\\
\frac{\partial^2 L(\alpha,\beta)}{\partial\alpha_k\partial\alpha_l} &= 0, \qquad \frac{\partial^2 L(\alpha,\beta)}{\partial\beta_k\partial\beta_l} = 0 \quad \text{for } l\neq k
\end{aligned} \tag{25}$$

Therefore, the Hessian matrix can be written as:

$$H(\alpha,\beta) = \text{block-diag}\left[H_1(\alpha_1,\beta_1),\ldots,H_{K-1}(\alpha_{K-1},\beta_{K-1})\right] \tag{26}$$

where H_k is the Hessian matrix for α_k,β_k. Following a similar logic as in the case of a Dirichlet distribution described in Eqs (56-60) of [12], we have the following Newton iteration:

$$\begin{aligned}
&\alpha_k^{new} = \alpha_k - \left(H_k^{-1}g_k\right)_1, \qquad \beta_k^{new} = \beta_k - \left(H_k^{-1}g_k\right)_2,\\
&q_{11}^k = \sum_{j=1}^{D}\Psi'(\alpha_k^j) - D\Psi'(\alpha_k), \qquad q_{22}^k = \sum_{j=1}^{D}\Psi'(\beta_k^j) - D\Psi'(\beta_k),\\
&a_k = \frac{g_1^k/q_{11}^k + g_2^k/q_{22}^k}{b_k^{-1} + (q_{11}^k)^{-1} + (q_{22}^k)^{-1}}, \quad b_k = D\Psi'(\alpha_k+\beta_k) - \sum_{j=1}^{D}\Psi'(\alpha_k^j+\beta_k^j)\\
&\left(H_k^{-1}g_k\right)_l = \frac{g_l^k - a_k}{q_{ll}^k} \qquad \text{for } l=1,2
\end{aligned} \tag{27}$$

where $g_1^k = dL(\alpha,\beta)/d\alpha_k$ and $g_2^k = dL(\alpha,\beta)/d\beta_k$ from Eq 24.