
Spectral Methods for Correlated Topic Models

Forough Arabshahi
University of California Irvine

Animashree Anandkumar
University of California Irvine

Abstract

In this paper we propose guaranteed spectral methods for learning a broad range of topic models, which generalize the popular Latent Dirichlet Allocation (LDA). We overcome the limitation of LDA to incorporate arbitrary topic correlations, by assuming that the hidden topic proportions are drawn from a flexible class of Normalized Infinitely Divisible (NID) distributions. NID distributions are generated by normalizing a family of independent Infinitely Divisible (ID) random variables. The Dirichlet distribution is a special case obtained by normalizing a set of Gamma random variables. We prove that this flexible topic model class can be learnt via spectral methods using only moments up to the third order, with (low order) polynomial sample and computational complexity. The proof is based on a key new technique derived here that allows us to diagonalize the moments of the NID distribution through an efficient procedure that requires evaluating only univariate integrals, despite the fact that we are handling high dimensional multivariate moments. In order to assess the performance of our proposed Latent NID topic model, we use two real datasets of articles collected from New York Times and Pubmed. Our experiments yield improved perplexity on both datasets compared with the baseline.

Keywords: Latent variable models, spectral methods, tensor decomposition, moment matching, infinitely divisible, Lévy processes.

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

1 Introduction

Topic models are a popular class of exchangeable latent variable models for document categorization. Their goal is to uncover hidden topics based on the distribution of word occurrences in a document corpus. Topic models are *admixture* models, which go beyond the usual mixture models that allow for only one hidden topic to be present in each document. In contrast, topic models incorporate multiple topics in each document. It is assumed that each document has a latent proportion of different topics, and the observed words are drawn in a conditionally independent manner, given the set of topics.

Latent Dirichlet Allocation (LDA) is the most popular topic model [8], in which the topic proportions are drawn from the Dirichlet distribution. While LDA has widespread applications, it is limited by the choice of the Dirichlet distribution. Notably, Dirichlet distribution can only model negative correlations [6], and thus, is unable to incorporate arbitrary correlations among the topics that may be present in different document corpora. Another drawback is that the elements with similar means need to have similar variances. While there have been previous attempts to go beyond the Dirichlet distribution, e.g. [7, 21], their correlation structures are still limited, learning these models is usually difficult and no guaranteed algorithms exist. Furthermore, As discussed in [20], the correlation structure considered in [7], gives rise to spurious correlations resulting in a better perplexity on the held-out set even when the recovered topics are less interpretable. The work of [5] provides a provably correct algorithm for learning topic models that also allow for certain correlations among the topics, however, it requires “anchor word” separability assumptions for the proof of correctness.

In this work, we consider a flexible class of topic models, and propose guaranteed and efficient algorithms for learning them. We employ the class of Normalized Infinitely Divisible (NID) distributions to model the topic proportions [10, 18]. These are a class of distributions on the simplex, formed by normalizing a set of

independent draws from a family of positive Infinitely Divisible (ID) distributions. The draws from an ID distribution can be represented as a sum of an arbitrary number of i.i.d. random variables. The concept of infinite divisibility was introduced in 1929 by Bruno de Finetti, and the most fundamental results were developed by Kolmogorov, Lévy and Khintchine in the 1930s. The idea of using normalized random probability measures with independent increments have also been used in the context of non-parametric models to go beyond the Dirichlet Process [17].

The Gamma distribution is an example of an ID distribution, and the Dirichlet distribution is obtained by normalizing a set of independent draws from Gamma distributions. We show that the class of NID topic models significantly generalize the LDA model: they can incorporate both positive and negative correlations among the topics and they involve additional parameters to vary the variance and higher order moments, while fixing the mean.

There are mainly three categories of algorithms for learning topic models, viz., variational inference [7, 8], Gibbs sampling [9, 11, 19], and spectral methods [2, 22]. Among them, spectral methods have gained increasing prominence over the last few years, due to their efficiency and guaranteed learnability. In this paper, we develop novel spectral methods for learning latent NID topic models.

Spectral methods have previously been proposed for learning LDA [2], and other latent variable models such as Independent Component Analysis (ICA), Hidden Markov Models (HMM) and mixtures of ranking distributions [3]. The idea is to learn the parameters based on spectral decomposition of low order moment tensors (third or fourth order). Efficient algorithms for tensor decomposition have been proposed before [3], and implies consistent learning with (low order) polynomial computational and sample complexity.

The main difficulty in extending spectral methods to the more general class of NID topic models is the presence of arbitrary correlations among the hidden topics which need to be “untangled”. For instance, for a single topic model (i.e. each document has only one topic), the third order moment, which is the co-occurrence tensor of word triplets, has a CANDECOMP/PARAFAC (CP) decomposition, computing which yields an estimate of the topic-word matrix. In contrast, for the LDA model, such a tensor decomposition is obtained by a linear combination of moments up to third order. In other words, the moments of the LDA model need to be appropriately “centered” in order to have the tensor decomposition form.

Finding such a moment combination has so far been an

“art form”, since it is based on explicit manipulation of the moments of the hidden topic distribution. So far, there is no principled mechanism to automatically find the moment combination with the CP decomposition form. For arbitrary topic models, finding such a combination may not even be possible. In general, one requires all the higher order moments for learning.

In this work, we show that surprisingly, for the flexible class of NID topic models, moments up to third order suffice for learning, and we provide an efficient algorithm for computing the coefficients to combine the moments. The algorithm is based on computation of a univariate integral, that involves the Lévy measure of the underlying ID distribution. The integral can be computed efficiently through numerical integration since it is only univariate, and has no dependence on the topic or word dimensions. Intriguingly, this can be accomplished, even when there exists no closed form probability density functions for the NID variables.

2 Latent Normalized Infinitely Divisible Topic Models

Topic models incorporate relationships between words $\mathbf{x}_1, \mathbf{x}_2 \dots \in \mathbb{R}^d$ and a set of k hidden topics. We represent the words \mathbf{x}_i using one-hot encoding, i.e. $\mathbf{x}_i = \mathbf{e}_j$ if the j^{th} word in the vocabulary occurs, and \mathbf{e}_j is the standard basis vector. The proportions of topics in a document is represented by vector $\mathbf{h} \in \mathbb{R}^k$. We assume that \mathbf{h} is drawn from an NID distribution. The detailed generative process of a latent NID topic model for each document is as follows

1. Draw k independent variables, z_1, z_2, \dots, z_k from a family of ID distributions.
2. Set \mathbf{h} to $(\frac{z_1}{Z}, \dots, \frac{z_k}{Z})$ where $Z = \sum_{i \in [k]} z_i$.
3. For each word \mathbf{x}_i ,
 - (a) Choose a topic $\zeta_i \sim \text{Multi}(\mathbf{h})$ and represent it with one-hot encoding.
 - (b) Choose a word vector, \mathbf{x}_i , to be the standard basis vector with probability,

$$\mathbb{E}(\mathbf{x}_i | \zeta_i) = \mathbf{A}\zeta_i, \quad (1)$$

conditioned on the drawn topic ζ_i , where $\mathbf{A} \in \mathbb{R}^{d \times k}$ is the topic-word matrix.

From (1), we also have

$$\mathbb{E}(\mathbf{x}_i | \mathbf{h}) = \mathbb{E}[\mathbb{E}(\mathbf{x}_i | \mathbf{h}, \zeta_i)] = \mathbb{E}(\mathbf{x}_i | \zeta_i) \mathbb{E}(\zeta_i | \mathbf{h}) = \mathbf{A}\mathbf{h} \quad (2)$$

When the z_i is drawn from the Gamma($\alpha_i, 1$) distribution, we obtain the Dir(α) distribution for the hidden

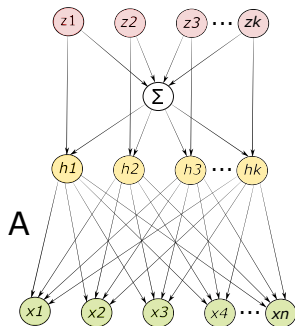


Figure 1: Graphical Model Representation of the Latent NID Topic Model. z_1, z_2, \dots, z_k are a collection of independent Infinitely Divisible positive variables that are characterized by the collection of their corresponding Lévy measures $\alpha_1\nu, \alpha_2\nu, \dots, \alpha_k\nu$ And h_1, h_2, \dots, h_k are the resulting NID variables representing topic proportions in a document of length N with words x_1, \dots, x_N

vector $\mathbf{h} = (h_1, \dots, h_k)$, and the LDA model through the above generative process. Our goal is to recover the topic-word matrix \mathbf{A} given the document collection. In the following section we introduce the class of NID distributions and discuss its properties.

3 Properties of NID distributions

NID distributions are a flexible class of distributions on the simplex and have been applied in a range of domains. This includes hierarchical mixture modeling with Normalized Inverse-Gaussian distribution [16], and modeling overdispersion with the normalized tempered stable distribution [14], both of which are examples of NID distributions. For more applications, see [10]. Let us first define the concept of infinite divisibility and present the properties of an ID distribution, and then consider the NID distributions.

3.1 Infinitely Divisible Distributions

If random variable z has an Infinitely Divisible (ID) distribution, then for any $n \in \mathbb{N}$ there exists a collection of i.i.d random variables y_1, \dots, y_n such that $z \stackrel{d}{=} y_1 + \dots + y_n$. In other words, an Infinitely Divisible distribution can be expressed as the sum of an arbitrary number of independent identically distributed random variables.

The Poisson distribution, compound Poisson, the negative binomial distribution, Gamma distribution, and the trivially degenerate distribution are examples of Infinitely Divisible distributions; as are the normal distribution, Cauchy distribution, and all other members of the stable distribution family. The Student's

t -distribution is also another example of Infinitely Divisible distributions. The uniform distribution and the binomial distribution are not infinitely divisible, as are all distributions with bounded (finite) support.

The special decomposition form of ID distributions makes them natural choices for certain models or applications. E.g. a compound Poisson distribution is a Poisson sum of i.i.d random variables. The discrete compound Poisson distribution, also known as the stuttering Poisson distribution, can model batch arrivals (such as in a bulk queue [1]) and can incorporate Poisson mixtures. In the sequel, we limit the discussion to ID distributions on \mathbb{R}^+ in order to ensure that the Normalized ID variables are on the simplex. Let us now present how ID distributions can be characterized.

Lévy measure: A σ -finite Borel measure ν on \mathbb{R}^+ is called a Lévy measure if $\int_0^\infty \min(1, x)\nu(dx) < \infty$. According to the Lévy-Khintchine representation given below, the Lévy measure uniquely characterizes an ID distribution along with a constant scale τ . This implies that every Infinitely Divisible distribution corresponds to a Lévy process, which is a stochastic process with independent increments.

Lévy-Khintchine representation [Theorem 16.14 [13]] Let $\mathcal{M}_1(\Lambda)$ and $\mathcal{M}_\sigma(\Lambda)$ indicate the set of probability measures and the set of σ -finite measures on a non-empty set Λ , respectively. Let $\mu \in \mathcal{M}_1((0, \infty))$ and let $\Psi(u) = -\log \int_0^\infty e^{-uz} d(\mu)$ be the log-Laplace transform of μ . Then μ is Infinitely Divisible, if and only if there exists a $\tau \geq 0$ and a σ -finite measure $\nu \in \mathcal{M}_\sigma((0, \infty))$ with

$$\int_0^\infty \min(1, z)\nu(dz) < \infty, \quad (3)$$

such that

$$\Psi(u) = \tau u + \int_0^\infty (1 - e^{-uz})\nu(dz) \quad \text{for } u \geq 0, \quad (4)$$

In this case the pair (τ, ν) is unique, ν is called the Lévy measure of μ and τ is called the deterministic part. It can be shown that $\tau = \sup\{z \geq 0 : \mu([0, z]) = 0\}$.

In particular, let $\Phi_{z_i}(u) = \mathbb{E}[e^{\iota uz_i}] = \int_0^\infty e^{\iota uz_i} f(z_i) dz_i$ indicate the characteristic function of an Infinitely Divisible random variable z_i with pdf $f(z_i)$ and corresponding pair (τ_i, ν_i) , where ι is the imaginary unit. Based on the Lévy-Khintchine representation it holds that $\Phi_{z_i}(\iota u) = \mathbb{E}[e^{-uz_i}] = e^{-\Psi_i(u)}$ where $\Psi_i(u) = \tau_i u + \int_0^\infty (1 - e^{-uz})\nu_i(dz)$ is typically referred to as the

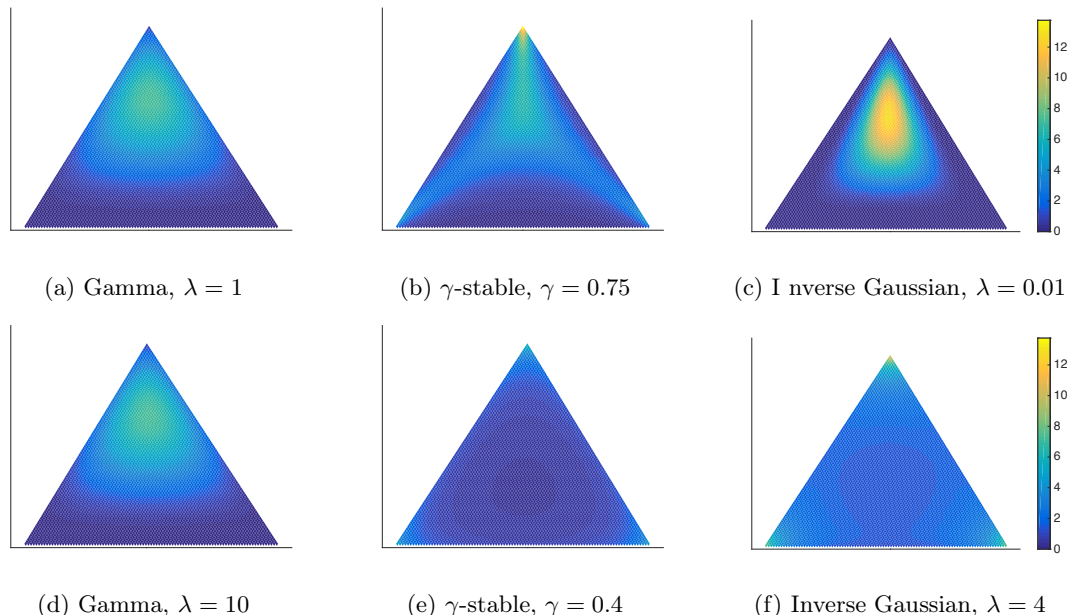


Figure 2: Heat map of the pdf of three examples of the NID class that have closed form pdf. All the figures have $\alpha = (2, 2, 4)$. For the Inverse Gaussian the distribution moves from the center to the vertices of the simplex as λ goes from 0 to ∞ with fixed α and for the γ -stable we have the same behavior when γ changes from 1 to 0 with fixed α .

Laplace exponent of z_i . This implies that the Laplace exponent of an ID variable is also completely characterized by pair (τ_i, ν_i) . It holds for ID variables that if ν_i is a well-defined Lévy measure, so is $\alpha_i \nu_i$ for any $\alpha_i > 0$, which indicates that $\alpha_i \Psi_i(u)$ is also a well-defined Laplace exponent of an ID variable.

3.2 Normalized Infinitely Divisible Distributions

As defined in [10], a Normalized Infinitely Divisible (NID) random variable is a random variable that is formed by normalizing independent draws of strictly positive (not necessarily coinciding) ID distributions. More specifically, let z_1, \dots, z_k be a set of independent strictly positive Infinitely Divisible random variables and $Z = z_1 + \dots + z_k$. An NID distribution is defined as the distribution of the random vector $\mathbf{h} = (h_1, \dots, h_k) := (\frac{z_1}{Z}, \dots, \frac{z_k}{Z})$ on the $(k-1)$ -dimensional simplex, denoted as Δ^{k-1} . The strict positivity assumption implies that \mathbf{h} is on the simplex [10, 18].

Let $[k]$ denote natural numbers $1, \dots, k$. As stated by the Lévy-Khintchine theorem, a collection of ID positive variables z_i for $i \in [k]$ is completely characterized by the collection of the corresponding Lévy measures ν_1, \dots, ν_k . It was shown in [18] that this also holds for the normalized variables h_i for $i \in [k]$.

In this paper, we assume that the ID variables z_1, \dots, z_k are drawn independently from ID distributions that are characterized with the corresponding

collection of Lévy measures $\alpha_i \nu, \dots, \alpha_k \nu$, respectively. Which in turn translates respectively to variables with Laplace exponents $\alpha_1 \Psi(u), \dots, \alpha_k \Psi(u)$. Variables α_i will allow the distribution to vary in the interior of the simplex, providing the asymmetry needed to model latent models. The homogeneity assumption on the Lévy measure or the Laplace exponent provides the structure needed for guaranteed learning (Theorem 1). The overall graphical model representation is shown in Figure 1 If the original ID variables z_i have probability densities f_i for all $i \in [k]$, then the distribution of vector \mathbf{h} , where $h_k = 1 - \sum_{i \in [k-1]} h_i$ is,

$$f(\mathbf{h}) = \int_0^\infty \prod_{i \in [k]} f_i(h_i Z) Z^{k-1} dZ.$$

There are only three members of the NID class that have closed form densities namely, the Gamma distribution, $\text{Gamma}(\alpha_i, \lambda)$, the Inverse Gaussian distribution, $\text{IG}(\alpha_i, \lambda)$, and the 1/2-stable distribution $\text{St}(\gamma, \beta, \alpha_i, \mu)$ with $\gamma = 1/2$. $\mu = 1$ and $\beta = 1$ to ensure positive support for the Stable distribution. As noted earlier, $\text{Gamma}(\alpha_i, 1)$ reduces to the Dirichlet distribution. An interested reader is referred to [10, 18] for the closed form of each distribution.

Figure 2 depicts the heatmap of the density of these distributions on the probability simplex for different value of their parameters. Note that all the distributions have the same α parameter and hence, the same mean values. However, their concentration properties are widely varying, showing that the NID class can in-

corporate variations in higher order moments through additional parameters.

Gamma ID distribution: When the ID distribution is Gamma with parameters $(\alpha_i, 1)$, we have the Dirichlet distribution as the resulting NID distribution. The Laplace exponent for this distribution will, therefore, be $\Psi_i(u) = \alpha_i \ln(1 + u)$.

γ -stable ID distribution: The variables are drawn from the positive stable distribution $St(\gamma, \beta, \alpha_i, \mu)$ with $\mu = 0$, $\beta = 1$ and $\gamma < 1$ which ensures that the distribution is on \mathbb{R}^+ . The Laplace exponent of this distribution is $\Psi_i(u) = \alpha_i \frac{\Gamma(1-\gamma)}{\sqrt{2\pi\gamma}} u^\gamma$. Note that the γ -stable distribution can be represented in closed form for $\gamma = \frac{1}{2}$.

Inverse Gaussian ID distribution: The random variables are drawn from the Inverse-Gaussian (IG) distribution $IG(\alpha_i, \lambda)$. The Laplace exponent of this distribution is $\Psi_i(u) = \alpha_i (\sqrt{2u + \lambda^2} - \lambda)$. *Note:* The Dirichlet distribution, the 1/2-Stable distribution and the Inverse Gaussian distribution are all special cases of the generalized Inverse Gaussian distribution [10].

As mentioned earlier, the class of NID distributions is capable of modeling positive and negative correlations among the topics. This property is depicted in Figure 3. These figures show the proportion of positively correlated pairs for the three presented distributions for different parameters of each distribution. As we can see, the Inverse Gaussian NID distribution can capture both positive and negative correlations.

4 Learning NID Topic Models through Spectral Methods

In order to be able to guarantee efficient learning using higher order moments, the moments need to have a very specific structure. Namely, the moment of the underlying distribution of \mathbf{h} needs to form a diagonal tensor. If the components of \mathbf{h} were indeed independent, this is obtained through the cumulant tensor. On the other hand, for LDA, it has been shown by Anandkumar et. al. [2] that a linear combination of moments of up to third order of \mathbf{h} forms a diagonal tensor for the Dirichlet distribution. Below, we extend the result to the more general class of NID distributions.

4.1 Consistency of Learning through Moment Matching

Assumption 1 *ID random variables z_i for $i \in [k]$ are said to be partially homogeneous if they share the same Lévy measure. This implies that the corresponding Laplace exponent of variable z_i is given $\alpha_i \Psi(u)$ for*

some $\alpha_i \in \mathbb{R}^+$, and $\Psi(u)$ is the Laplace exponent of the common Lévy measure.

Under the above assumption, we prove guaranteed learning of NID models through spectral methods. This is based on the following moment forms for NID models, which admit a CP tensor decomposition. The components of the decomposition will be the columns of the topic-word matrix: $\mathbf{A} := [\mathbf{a}_1 | \mathbf{a}_2 | \dots | \mathbf{a}_k]$. Define

$$\Omega(m, n, p) = \int_0^\infty u^m \frac{d^n}{du^n} \Psi(u) \left(\frac{d}{du} \Psi(u) \right)^p e^{-\alpha_0 \Psi(u)} du, \quad (5)$$

where $\Psi(u)$ is the Laplace exponent of the NID distribution and $\alpha_0 = \sum_{i \in [k]} \alpha_i$.

Theorem 1 (Moment Forms for NID models)

Let \mathbf{M}_2 and \mathbf{M}_3 be respectively the following matrix and tensor constructed from the moments of the data,

$$\mathbf{M}_2 = \mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2] + v \cdot \mathbb{E}[\mathbf{x}_1] \otimes \mathbb{E}[\mathbf{x}_2], \quad (6)$$

$$\begin{aligned} \mathbf{M}_3 = & \mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3] + v_2 \cdot \mathbb{E}[\mathbf{x}_1] \otimes \mathbb{E}[\mathbf{x}_2] \otimes \mathbb{E}[\mathbf{x}_3] \\ & + v_1 [\mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2] \otimes \mathbb{E}[\mathbf{x}_3] + \mathbb{E}[\mathbf{x}_1] \otimes \mathbb{E}[\mathbf{x}_2 \otimes \mathbf{x}_3] \\ & + \mathbb{E}[\mathbf{x}_1 \otimes \mathbb{E}[\mathbf{x}_2] \otimes \mathbf{x}_3]] \end{aligned} \quad (7)$$

where,

$$v = \frac{\Omega(1, 1, 1)}{(\Omega(0, 1, 0))^2}, \quad v_1 = -\frac{\Omega(2, 2, 1)}{2\Omega(1, 2, 0)\Omega(0, 1, 0)}, \quad (8)$$

$$v_2 = \frac{-0.5\Omega(2, 1, 2) + 3v_1\Omega(1, 1, 1)\Omega(0, 1, 0)}{(\Omega(0, 1, 0))^3}, \quad (9)$$

Then given Assumption 1,

$$\mathbf{M}_2 = \sum_{j \in [k]} \kappa_j (\mathbf{a}_j \otimes \mathbf{a}_j), \quad \mathbf{M}_3 = \sum_{j \in [k]} \lambda_j (\mathbf{a}_j \otimes \mathbf{a}_j \otimes \mathbf{a}_j). \quad (10)$$

for a set of κ_j 's and λ_j 's which are a function of the parameters of the distribution.

Remark 1: efficient computation of v, v_1 and v_2

What makes Theorem 1 specially intriguing is the fact that weights v, v_1 and v_2 can be computed through univariate integration, which can be computed efficiently, regardless of the dimensionality of the problem.

Remark 2: investigation of special cases

When the ID distribution is Gamma with parameters $(\alpha_i, 1)$, we have the Dirichlet distribution as the resulting NID distribution. Weights v_1 and v_2 reduce to the results of Anandkumar et. al. [2] for the Gamma($\alpha_i, 1$) distribution, which are $v_1 = -\frac{\alpha_0}{\alpha_0+2}$ and $v_2 = \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)}$.

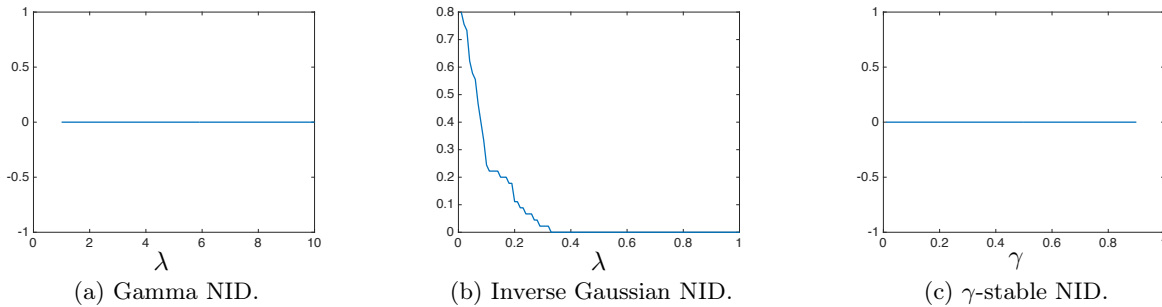


Figure 3: Proportion of positively correlated pairs of 10-dimensional NID distributed variables with respect to the parameters of the Laplace exponent for a fixed randomly drawn vector α .

When the variables are drawn from the positive stable distribution $St(1/2, \beta, \alpha_i, \mu)$ weights v_1 and v_2 in Theorem 1 can be represented in closed form as $v_1 = -\frac{1}{4}$ and $v_2 = -\frac{5}{8}$.

It is hard to find the closed form representation of the weights for other stable distributions and the Inverse Gaussian distribution. Therefore, we give the form of the weights with respect to the parameters of each distribution in Figure 4. As it can be seen in Figures 2e and 2b, as γ increases, the distribution gets more centralized on the simplex. Therefore, as depicted in Figure 4a the weight becomes more negative to compensate for it. The same holds in Figure 4b.

The above result immediately implies guaranteed learning for non-degenerate topic-word matrix \mathbf{A} .

Assumption 2 *Topic-word matrix $\mathbf{A} \in \mathbb{R}^{d \times k}$ has linearly independent columns and the parameters $\alpha_i > 0$.*

Corollary 1 (Guaranteed Learning of NID Topic Models using Spectral Methods) *Given empirical versions of moments \mathbf{M}_2 and \mathbf{M}_3 in (6) and (7), using tensor decomposition algorithm from [3], under Assumption 2, we can consistently estimate topic-word matrix \mathbf{A} and parameters α with polynomial computational and sample complexity. (Algorithm 1)*

Algorithm 1 Parameter Learning

Input: Chosen NID distribution and hidden dimension k

Output: Parameters of NID distribution α and topic-word matrix \mathbf{A}

- 1: Estimate empirical moments $\hat{\mathbb{E}}(\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3)$, $\hat{\mathbb{E}}(\mathbf{x}_1 \otimes \mathbf{x}_2)$ and $\hat{\mathbb{E}}(\mathbf{x}_1)$.
 - 2: Compute weights v , v_1 and v_2 in (8) and (9) for the given NID distribution by numerical integration.
 - 3: Estimate tensors \mathbf{M}_2 and \mathbf{M}_3 in (6) and (7).
 - 4: Decompose tensor \mathbf{M}_3 into its rank-1 components using the algorithm in [3] that requires \mathbf{M}_2 .
 - 5: Return columns of \mathbf{A} as the components of the decomposition.
-

Remark 3: third order moments suffice For the flexible class of latent NID topic models, only moments up to the third order suffice for efficient learning.

Remark 4: Sample Complexity Following [2], Algorithm 1 can recover matrix \mathbf{A} under Assumption 2 with polynomial sample complexity.

Remark 5: Implementation Efficiency In order to make the implementation efficient we use the discussion in [3]. Specifically, as mentioned in [3], we can find a whitening transformation from matrix \mathbf{M}_2 that lowers the data dimension from the vocabulary space to the topic space. We then use the same whitening transformation to go back to the original space and recover the parameters of the model.

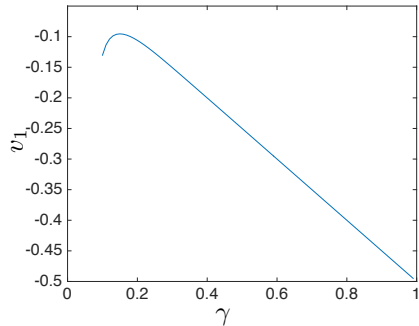
Overview of the proof of Theorem 1 We begin the proof by forming the following second order and third order tensors using the moments of the NID distribution given in Lemma 1.

$$\mathbf{M}_2^{(\mathbf{h})} = \mathbb{E}(\mathbf{h} \otimes \mathbf{h}) + v\mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h}), \quad (11)$$

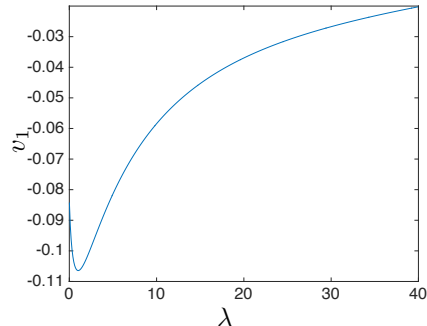
$$\begin{aligned} \mathbf{M}_3^{(\mathbf{h})} &= \mathbb{E}(\mathbf{h} \otimes \mathbf{h} \otimes \mathbf{h}) + v_2\mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h}) \\ &+ v_1 [\mathbb{E}(\mathbf{h} \otimes \mathbf{h}) \otimes \mathbb{E}(\mathbf{h}) + \mathbb{E}(\mathbf{h} \otimes \mathbb{E}(\mathbf{h}) \otimes \mathbf{h}) \\ &+ \mathbb{E}(\mathbf{h}) \otimes \mathbb{E}(\mathbf{h} \otimes \mathbf{h})] \end{aligned} \quad (12)$$

Weights v , v_1 and v_2 are as in Equations (8) and (9). They are computed by setting the off-diagonal entries of matrix $\mathbf{M}_2^{(\mathbf{h})}$ in Equation 11 and $\mathbf{M}_3^{(\mathbf{h})}$ in Equation 12 to 0. Due to the homogeneity assumption, all the off-diagonal entries can be simultaneously made to vanish with these choices of coefficients for v , v_1 and v_2 . We obtain $\mathbf{M}_2^{(\mathbf{h})} = \sum_{i \in [k]} \kappa'_i \mathbf{e}_i^{\otimes 2}$ and $\mathbf{M}_3^{(\mathbf{h})} = \sum_{i \in [k]} \lambda'_i \mathbf{e}_i^{\otimes 3}$ where \mathbf{e}_i 's are the standard basis vectors, and this implies they are diagonal tensors. Due to this fact and the exchangeability of the words given topics according to (2), Equations 10 follow.

The exact forms of v , v_1 and v_2 are obtained by the following moment forms for NID distributions.



(a) Weight v_1 of theorem 1 for a Stable ID distribution $St(\gamma, \beta, \alpha_i, \mu)$ with $\mu = 0$, $\beta = 1$, $\gamma < 1$ and $\alpha_i > 0$ vs. γ for $\alpha_0 = 1$



(b) Weight v_1 of theorem 1 for an Inverse Gaussian distribution $IG(\alpha_i, \lambda)$ vs. $\lambda > 0$ and $\alpha_i \geq 0$ for $\alpha_0 = 1$

Figure 4: Weight v_1 for two different examples of the NID distribution. Weights v and v_2 in the theorem have similar behavior w.r.p the parameters.

Lemma 1 ([18]) *The moments of NID variables h_1, \dots, h_k satisfy*

$$\mathbb{E}(h_1^{r_1} h_2^{r_2} \dots h_k^{r_k}) = \frac{1}{\Gamma(r)} \int_0^\infty u^{r-1} e^{-\alpha_0 \Psi(u)} \prod_{j \in [k]} B_{r_j}^j du, \quad (13)$$

where $r = \sum_{i \in [k]} r_i$ and $B_{r_j}^j$ can be written in terms of the partial Bell polynomial as

$$B_r^i = B_r(-\alpha_i \Psi^{(1)}(u), \dots, -\alpha_i \Psi^{(r)}(u)), \quad (14)$$

where $\Psi^{(l)}(u)$ is the l^{th} derivative of $\Psi(u)$ w.r.t. u .

5 Experiments

In this section we apply our proposed latent NID topic modeling algorithm to New York Times and Pubmed articles [15]. The New York Times dataset contains about 300,000 documents and the pubmed data contains around 8 million documents. The vocabulary size for both the datasets are around 100,000.

Table 1: Perplexity comparison for different datasets

Dataset	NYtimes	Pubmed
NID	3.5702e + 03	4.0771e + 03
LDA	4.8464e + 03	4.3702e + 03
CTM	1.0669e + 04	N/A

Table 2: PMI comparison for different datasets

Dataset	NYtimes	Pubmed
NID	0.2439	0.3080
LDA	0.2362	0.4487
CTM	0.497	N/A

Hyper-parameter Tuning: In this sub-section we propose a hyper-parameter tuning approach (Algorithm 2) for learning the class of NID topic models. In

Table 3: 10 Shared words - New York times dataset

Shared Words
boston-globe, tonight, question, newspaper, spot, percent, file, diane, copy, fall

order to find the best matching NID distribution that describes the data, we can directly tune for weights v , v_1 and v_2 of Equations 8 and 9. This is possible by altering the decomposition step of the learning algorithm (Step 4, Algorithm 1) such that it optimizes not only over the low rank components of the moment tensor, but also over the weights v_1 and v_2 . To be more specific we optimize the following least squares problem over the weights v_1 and v_2 and the components of the tensor a_j for $j \in [k]$:

$$\min_{v_1, v_2, a_j} \|\mathbf{T} + v_1 \mathbf{T}_1 + v_2 \mathbf{T}_2 - \sum_{j \in [k]} \lambda_j (\mathbf{a}_j \otimes \mathbf{a}_j \otimes \mathbf{a}_j)\|_F, \quad (15)$$

where, $\|\cdot\|_F$ is the Frobenius norm of a tensor and

$$\mathbf{T} = \mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2 \otimes \mathbf{x}_3] \quad (16)$$

$$\mathbf{T}_1 = \mathbb{E}[\mathbf{x}_1 \otimes \mathbf{x}_2] \otimes \mathbb{E}[\mathbf{x}_3] + \mathbb{E}[\mathbf{x}_1] \otimes \mathbb{E}[\mathbf{x}_2 \otimes \mathbf{x}_3] + \mathbb{E}[\mathbf{x}_1 \otimes \mathbb{E}[\mathbf{x}_2] \otimes \mathbf{x}_3] \quad (17)$$

$$\mathbf{T}_2 = \mathbb{E}[\mathbf{x}_1] \otimes \mathbb{E}[\mathbf{x}_2] \otimes \mathbb{E}[\mathbf{x}_3] \quad (18)$$

This optimization problem will return the components of tensor \mathbf{M}_3 , a_j , which are the columns of matrix \mathbf{A} in Equation 1 and weights v_1 and v_2 . As can be seen in Equations 8 and 9, these weights are functions of the Laplace exponent of the underlying NID distribution. Therefore, the underlying Laplace exponent can be recovered from the weights in a non-parametric manner using a held out data set. We can use the non-parametric representation of the Laplace exponent to extract the properties of the NID distribution such as correlations. As will be shown in the results, this efficient heuristic for hyper-parameter tuning shows a

good empirical performance on multiple real data sets.

Algorithm 2 Hyper-parameter Optimization (HPO)

Input: train data set, held out data set and hidden dimension k

Output: topic-word matrix \mathbf{A} , Laplace exponent $\Psi(u)$

- 1: Estimate empirical moments and compute \mathbf{T} , \mathbf{T}_1 and \mathbf{T}_2 in Equations 16, 17 and 18 from the train data set.
 - 2: Solve the optimization problem of Equation 15 and output \mathbf{A} , v_1 and v_2
 - 3: Estimate the Laplace exponent $\Psi(u)$ from v_1 and v_2 using non-parametric density estimation on the held out data set.
-

Results: We refer to our proposed method as *NID* and compare it against spectral *LDA* [2], where the distribution of the hidden space is fixed to be Dirichlet, and *CTM* [7] where the hidden topic distribution is logistic normal. It has been shown in [12] that spectral *LDA* is more efficient and achieves better perplexity compared to the conventional *LDA* [8]. Table 4 in the appendix provides a sketch of the top words per topics recovered by our latent *NID* topic model on the New York times dataset and Table 5 in the appendix shows the top words recovered from the pubmed dataset. We have also provided the top words recovered by *LDA* for the New York times dataset for comparison purposes in Table 6 in the appendix. Besides the top words, we also present the shared words among the recovered topics for the New York Times dataset in Table 3. The presence of words such as “tonight”, “question” and “fall” among these words are not far from expectation, since they are general words that are not usually indicative of any specific topic.

We use the well-known likelihood perplexity measure [8] to evaluate the generalization performance of our proposed topic modeling algorithm as well as the Pointwise Mutual Information (PMI) score [4] to assess the coherence of the recovered topics. Perplexity is defined as the inverse of the geometric mean per-word of the estimated likelihood. It should be noted that lower perplexity indicates better generalization performance and higher PMI indicates better topic coherence. Tables 1 and 2 show the perplexity and PMI scores, respectively, for the New York Times and Pubmed datasets for a fixed number of topics across each dataset. More comparisons for different number of topics is presented in Figure 5 in the appendix for the New York Times corpus. It should be noted that *CTM* does not scale to the Pubmed dataset. The results suggest that if we allow the corpus to choose the best underlying topic distribution, we can get better

generalization performance as well as topic coherence on the held-out set compared to fixing the underlying distribution to Dirichlet. The improved perplexity of our proposed method is indicative of correlations in the underlying documents that are not captured by the Dirichlet distribution. Thus, latent *NID* topic models are capable of successfully capturing correlations within topics while providing guarantees for exact recovery and efficient learning as proven in Section 4.

Last but not least, the naive Variational Inference implementation of [8]¹, does not scale to the current datasets used in this paper. The naive implementation of the spectral *LDA*, however, takes only about a minute to run on the NYtimes dataset and about 15 minutes to run on the Pubmed dataset. It is, therefore, of great importance to have a class of models that can be learned using spectral methods mainly because of their inherent scalability, ease of implementation and statistical guarantees. As we show in this paper, latent *NID* topic models are such a class of models. The correlated topic model framework of [7] also uses Variational Inference to perform learning and it is limited to the logit-normal distribution. latent *NID* topic models not only scale to large data sets, but also model arbitrary correlations without requiring a fixed prior distribution over the topic space.

Acknowledgments

F. Arabshahi is supported by ONR Award N00014-14-1-0665 and DARPA Award D17AP00002. A. Anandkumar is supported by Microsoft Faculty Fellowship, NSF CAREER Award CCF-1254106, ONR Award N00014-14-1-0665, DARPA Award D17AP00002, Air Force Award FA9550-15-1-0221 and ARO Award W911NF-16-1-0134.

6 Conclusion

In this paper we establish the new class of Latent Normalized Infinitely Divisible (*NID*) topic models that generalize previously proposed topic models such as *LDA*. We provide guaranteed efficient learning for this class of distributions using spectral methods through untangling the dependence of the hidden topics. We provide evidence that our proposed *NID* topic model overcomes the shortcomings of the Dirichlet distribution by allowing for both positive and negative correlations among the topics. In the end we use two real world datasets to validate our claims in practice. The improved likelihood perplexity score indicates that if we allow the model to pick the underlying distribution we will get better generalization results.

¹available at: <http://www.cs.princeton.edu/blei/lda-c/>

References

- [1] RM Adelson. Compound poisson distributions. *OR*, 17(1):73–75, 1966.
- [2] Anima Anandkumar, Yi-kai Liu, Daniel J Hsu, Dean P Foster, and Sham M Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.
- [3] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.
- [4] Animashree Anandkumar, Ragupathyraj Valluvan, et al. Learning loopy graphical models with latent variables: Efficient methods and guarantees. *The Annals of Statistics*, 41(2):401–435, 2013.
- [5] Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*, pages 280–288, 2013.
- [6] Ali Shojaee Bakhtiari and Nizar Bouguila. Online learning for two novel latent topic models. In *Information and Communication Technology*, pages 286–295. Springer, 2014.
- [7] David Blei and John Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.
- [8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
- [9] Jianfei Chen, Jun Zhu, Zi Wang, Xun Zheng, and Bo Zhang. Scalable inference for logistic-normal topic models. In *Advances in Neural Information Processing Systems*, pages 2445–2453, 2013.
- [10] Stefano Favaro, Georgia Hadjicharalambous, and Igor Prünster. On a class of distributions on the simplex. *Journal of Statistical Planning and Inference*, 141(9):2987–3004, 2011.
- [11] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.
- [12] Furong Huang. Discovery of latent factors in high-dimensional data using tensor methods. *arXiv preprint arXiv:1606.03212*, 2016.
- [13] Achim Klenke. Infinitely divisible distributions. In *Probability Theory*, pages 331–349. Springer, 2014.
- [14] Michalis Kolossiatos, Jim E Griffin, and Mark FJ Steel. Modeling overdispersion with the normalized tempered stable distribution. *Computational Statistics & Data Analysis*, 55(7):2288–2301, 2011.
- [15] M. Lichman. UCI machine learning repository, 2013.
- [16] Antonio Lijoi, Ramsés H Mena, and Igor Prünster. Hierarchical mixture modeling with normalized inverse-gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291, 2005.
- [17] Antonio Lijoi and Igor Prünster. Models beyond the dirichlet process. *Bayesian nonparametrics*, 28:80, 2010.
- [18] Francesca Mangili and Alessio Benavoli. New prior near-ignorance models on the simplex. *International Journal of Approximate Reasoning*, 56:278–306, 2015.
- [19] David Mimno, Hanna M Wallach, and Andrew McCallum. Gibbs sampling for logistic normal topic models with graph-based priors. 2008.
- [20] Alexandre Passos, Hanna M Wallach, and Andrew McCallum. Correlations and anticorrelations in lda inference. 2011.
- [21] Issei Sato and Hiroshi Nakagawa. Topic models with power-law using pitman-yor process. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–682. ACM, 2010.
- [22] Hsiao-Yu Tung and Alex J Smola. Spectral methods for indian buffet process inference. In *Advances in Neural Information Processing Systems*, pages 1484–1492, 2014.