The Discrete Infinite Logistic Normal Distribution for Mixed-Membership Modeling

John Paisley, Chong Wang and David Blei
Department of Computer Science, Princeton University
{jpaisley,chongw,blei}@princeton.edu

Abstract

We present the *discrete infinite logistic normal* distribution (DILN, "Dylan"), a Bayesian nonparametric prior for mixed membership models. DILN is a generalization of the hierarchical Dirichlet process (HDP) that models correlation structure between the weights of the atoms at the group level. We derive a representation of DILN as a normalized collection of gamma-distributed random variables, and study its statistical properties. We consider applications to topic modeling and derive a variational Bayes algorithm for approximate posterior inference. We study the empirical performance of the DILN topic model on four corpora, comparing performance with the HDP and the correlated topic model.

1 Introduction

The hierarchical Dirichlet process (HDP) has emerged as a powerful Bayesian nonparametric prior for grouped data (Teh *et al.*, 2007). Applied to topic modeling, it extends latent Dirichlet allocation (LDA) (Blei *et al.*, 2003) to allow each group of data (e.g., document) to be modeled as a mixture over an infinite collection of components (e.g., topics). The hallmark of the HDP is that components may be shared across groups—different documents will use the same population of topics. In posterior inference, the model selects the number of topics based on the data.

A drawback of the HDP, however, is that it cannot model correlations between components in the group level distributions (beyond the implicit correlations imposed by the infinite simplex). For example, the HDP cannot capture that the presence of a "sports" topic in a document positively correlates more with a "health" topic than with a "military" topic. To address this issue in the finite LDA model, Blei & Lafferty (2007) introduced the correlated topic model (CTM). The CTM replaces the finite Dirichlet prior on topic proportions with a finite logistic normal prior (Aitchison, 1982).

Our goal in this paper is to develop a hierarchical Bayesian nonparametric prior that can model correlations between the occurrences of the infinite collection of components, i.e., to develop an "infinite CTM." Unfortunately, the natural nonparametric extension of the logistic normal does not serve this purpose (Lenk, 1988). In the HDP, the sharing of components across groups arises because the group level distributions are sparse; the model of Lenk (1988) does not produce sparse distributions on the infinite simplex.

To address this problem, we develop the *discrete infinite logistic normal* (DILN, pronounced "Dylan"), a new Bayesian nonparametric prior for mixed-membership modeling. The DILN prior produces discrete probability distributions over an infinite collection of components and models an explicit correlation structure between the presence of those components in the group-level distributions. It retains the essential property that group-level distributions place mass on a shared collection of components.

When used in a topic model, variational inference for DILN results in an algorithm where the data determines the number of topics and where the presence of topics in a document exhibits an explicit correlation structure. On four corpora, we will show that this provides a better predictive model (Figure 2) and an effective new method for summarizing large collections of documents (Figure 3).

In more detail, the intuition behind DILN is that each component is located in a latent space, and the correlation structure between them is determined by the distances between their locations. We will first define DILN as a scaled HDP, where the scaling is determined by an exponentiated Gaussian process (GP) (Rasmussen & Williams, 2006) whose kernel is a function of the latent distance matrix between component locations. We then recast DILN with a gamma representation. With this representation, we can precisely characterize the a priori correlation structure assumed by

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

the model, and derive a variational inference algorithm to approximate the posterior distribution of components and their latent kernel matrix. As a by-product, this variational algorithm can be modified into a new posterior inference algorithm for traditional HDPs.

Related Work. There is a large literature on Bayesian nonparametric methods for learning dependent probability distributions, where dependence is defined on predictors observed for each data point, (e.g. MacEachern (1999); De Iorio *et al.* (2004); Gelfand *et al.* (2005); Dunson & Park (2008); Rao & Teh (2009)). These methods use the spatial or temporal information of the data to construct observation-specific distributions. In contrast, DILN does not assume the observations have locations. The latent locations of each component are instead used to induce correlation structure between components, which occurs before any data appears in the generative process.

We present DILN in Section 2 as a scaled HDP, and derive a representation based on the gamma process. We derive a variational posterior inference algorithm in Section 3 and in the appendix. We study the performance on four real-world corpora in Section 4.

2 The Discrete Infinite Logistic Normal

In this section, we develop the discrete infinite logistic normal (DILN) and derive some of its properties. We first review the HDP and its construction as a normalized gamma process. We then present DILN as an alternative prior for mixed membership models. We define DILN as a scaled HDP, with scaling determined by an exponentiated Gaussian process (Rasmussen & Williams, 2006), and then show how it naturally fits in the family of normalized gamma representations of discrete probability distributions.

2.1 The Gamma Process Construction of the HDP

The Dirichlet process (Ferguson, 1973) is a prior on discrete probability distributions $G = \sum_{k=1}^{K} \pi_k \delta_{\eta_k}$, where atoms $\{\eta_k\}_{k=1}^{K}$ are drawn *iid* from a base distribution G_0, weights $\{\pi_k\}_{k=1}^{K}$ depend on a scaling parameter $\alpha > 0$, and K is typically infinity.

The hierarchical Dirichlet process (Teh *et al.*, 2007) is a Dirichlet process that has a base probability measure which is also a Dirichlet process. The hierarchical representation of this process is

$$G \sim \mathrm{DP}(\alpha G_0), \quad G'_m \overset{iid}{\sim} \mathrm{DP}(\beta G), \tag{1}$$

where α and β are scaling parameters and m indexes multiple draws. In the context of mixed-membership modeling, let $f(\cdot)$ be a distribution and let $X_n^{(m)}$ denote the nth observation in the mth group. Then

$$\theta_n^{(m)} \sim G'_m, \quad X_n^{(m)} \sim f(\theta_n^{(m)}). \tag{2}$$

When used to model document collections, the HDP provides a topic model. The observation $X_n^{(m)}$ is the nth word in the mth document and words are drawn from $\theta_n^{(m)}$ which is a distribution over the vocabulary, and is equal to one of the atoms η_k indexed by G'_m. The base distribution G_0 is usually a symmetric Dirichlet over the vocabulary simplex; repeated draws from G_0 give the infinite set of topics that are used and reused in generating the documents. Given a document collection, posterior inference yields a set of shared topics and per-document proportions for each topic. Unlike its finite counterpart, latent Dirichlet allocation (Blei *et al.*, 2003), the HDP topic model determines the number of topics from the data.

The hierarchical structure of the HDP ensures that each G'_m has probability mass distributed across a shared set of atoms.[1] By stipulating that the base distribution of these multiple group-level DPs is also a DP, the base is a discrete measure. Therefore, the same subset of atoms will be used frequently by all groups, but with different probabilities for each group. This hierarchical process can be defined to an arbitrary depth, but we focus on two-level HDPs here. In the context of topic modeling, a set of topics (atoms) describes the collection, and each document exhibits a subset of these topics with different proportion.

The HDP has several representations. We present the representation that will be useful in the remainder of this paper. We represent the top-level Dirichlet process using the stick-breaking construction of the DP (Sethuraman, 1994),

$$\begin{aligned} G &= \sum_{k=1}^{\infty} V_k \prod_{j=1}^{k-1} (1 - V_j) \delta_{\eta_k}, \\ V_k &\overset{iid}{\sim} \mathrm{Beta}(1, \alpha), \quad \eta_k \overset{iid}{\sim} G_0. \end{aligned} \tag{3}$$

Let $p_k := V_k \prod_{j=1}^{k-1}(1 - V_j)$. Each second level DP is constructed with a normalized gamma process,

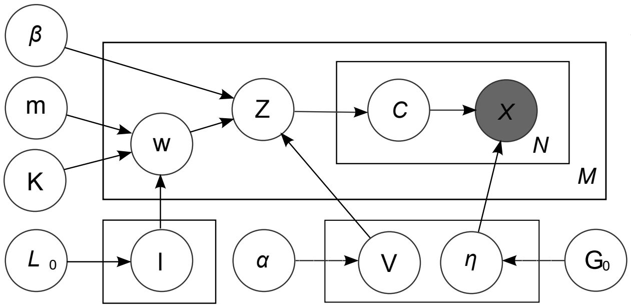
$$\begin{aligned} G'_m &= \sum_{k=1}^{\infty} \frac{Z_k^{(m)}}{\sum_{j=1}^{\infty} Z_j^{(m)}} \delta_{\eta_k}, \\ Z_k^{(m)} &\overset{ind}{\sim} \mathrm{Gamma}(\beta p_k, 1); \end{aligned} \tag{4}$$

Recall that the density of a gamma random variable is

$$\mathrm{Gamma}(x|a, b) = b^a x^{a-1} \exp\{-bx\}/\Gamma(a). \tag{5}$$

The gamma process representation of the DP is discussed in detail in Ferguson (1973), Kingman (1993) and Ishwaran & Zarepour (2002). Note, however, that this representation has not been specifically applied to the HDP in the literature. We will return to this representation in Section 2.3.

[1] This is not the case for multiple DP draws when the prior measure is continuous. Continuous priors are common when a DP is defined on an infinite parameter space for a mixture model, e.g., in the topic model.



notation	description
m , K	mean and kernel functions for GP
w	a draw from GP(m,K)
α , β	concentration parameters
$V_{1:\infty}$	top-level stick-breaking proportions
Z	atom proportions from gamma process
η_i , l_i	topic and its location
$G_0 \times L_0$	base distribution for topics and locations
C	topic index for words
X	observed words

Figure 1: A graphical model of the normalized gamma representation of the DILN topic model.

2.2 The Discrete Infinite Logistic Normal

The gamma process used to construct each group-level distribution of the HDP is an example of a completely random measure (Kingman, 1993)—all random variables are independent, as are all summations of subsets. This means that the HDP inherently cannot model correlation structure between the components at the group level. We therefore introduce DILN as a modification of the HDP.

To draw from the DILN prior, a top-level Dirichlet process is first drawn with a product base measure,

$$G \sim \mathrm{DP}(\alpha G_0 \times L_0). \tag{6}$$

As discussed, G_0 is a base distribution over parameter values $\eta \in \Omega$, while the probability measure L_0 gives a distribution over locations $\ell \in \mathbb{R}^d$. We can think of G as a distribution over parameters $\{\eta_k\}$ that are given a set of corresponding locations $\{\ell_k\}$ in a latent space.

In the second level of the process, both the probability measure G and the locations of the atoms are used to construct group-level probability distributions. First, as in the HDP, draw from a Dirichlet process $G_m^{\mathrm{DP}} \sim \mathrm{DP}(\beta G)$. This provides a new distribution on the atoms of G. Second, draw from a Gaussian process,

$$w^{(m)}(\ell) \sim \mathrm{GP}(\mathbf{m}(\ell), \mathbf{K}(\ell, \ell')), \tag{7}$$

which is defined on the locations of the atoms of G. The result is a random function $w^{(m)}(\cdot)$ that can be evaluated using the location of each atom. The covariance between $w^{(m)}(\ell)$ and $w^{(m)}(\ell')$ is determined by a kernel function, $\mathbf{K}(\ell, \ell')$, on their respective locations.

Finally, form the group-level distribution by scaling the probabilities of the Dirichlet process by the exponentiated values of the Gaussian process,

$$G'_m(\{\eta, \ell\}) \propto G_m^{\mathrm{DP}}(\{\eta, \ell\}) \exp\{w^{(m)}(\ell)\}. \tag{8}$$

We satisfy two objectives with this representation: (i) the probability measure G'_m is discrete, owing to the discreteness of G_m^{DP}; and (ii) the probabilities in G'_m are *explicitly* correlated, due to the exponentiated Gaussian process. Since these correlations arise from latent locations, learning these correlations is part of inference.

2.3 A Normalized Gamma Representation of DILN

We now turn to a gamma representation of DILN. We show that the DILN prior uses the second parameter of the normalized gamma representation of the HDP to model the covariance structure between the components of G'_m. This representation facilitates approximate posterior inference with variational inference, and helps clarify the covariance properties of the group-level distributions over atoms. The two levels of the construction are given below.

The top-level distribution of DILN follows from equation (3), and is the constructive representation of (6),

$$\begin{aligned} G &= \sum_{k=1}^{\infty} V_k \prod_{j=1}^{k-1} (1 - V_j) \delta_{\{\eta_k, \ell_k\}} \\ V_k &\overset{iid}{\sim} \mathrm{Beta}(1, \alpha) \\ \eta_k &\overset{iid}{\sim} G_0, \quad \ell_k \overset{iid}{\sim} L_0. \end{aligned} \tag{9}$$

The group-level distribution is similar to the gamma-process representation of the DP, but uses the second parameter of the gamma distribution,

$$\begin{aligned} G'_m &= \sum_{k=1}^{\infty} \frac{Z_k^{(m)}}{\sum_{j=1}^{\infty} Z_j^{(m)}} \delta_{\eta_k} \\ Z_k^{(m)} &\sim \mathrm{Gamma}(\beta p_k, \mathrm{e}^{-w_k^{(m)}}) \\ w^{(m)} &\overset{iid}{\sim} \mathrm{GP}(\mathbf{m}, \mathbf{K}). \end{aligned} \tag{10}$$

Recall that $p_k := V_k \prod_{j=1}^{k-1} (1 - V_j)$. A proof that the normalizing constant is almost surely finite is given in the appendix. We note that we have suppressed the location ℓ_k of atoms in G'_m, since these are no longer relevant after its construction.

This representation arises as follows. We construct the distribution G_m^{DP} in equation (8) using gamma-distributed random variables, as in equation (4). Next, observe that the exponential scaling term in (8) can be absorbed within the gamma distribution; recall that a random variable $x \sim \mathrm{Gamma}(a, 1)$ that is scaled by $b > 0$ to produce $y := bx$ results in a new random variable $y \sim \mathrm{Gamma}(a, b^{-1})$.

For the topic model, the data-generating distribution $f(\cdot)$ is multinomial, and drawing an observation proceeds as for the HDP,

$$X_n^{(m)} \sim \text{Mult}(\eta_{C_n^{(m)}}), \quad C_n^{(m)} \sim \sum_{k=1}^{\infty} \frac{Z_k^{(m)}}{\sum_{j=1}^{\infty} Z_j^{(m)}} \delta_k \,. \tag{11}$$

The latent variable $C_n^{(m)}$ gives the index of the topic associated with observation $X_n^{(m)}$. Figure 1 shows a graphical representation of the DILN topic model.

2.4 The Covariance Structure of DILN

The two-parameter gamma representation of DILN permits simple calculation of the central moments and covariance between components prior to normalization. In the following calculations, we assume that the mean function $\mathbf{m} = 0$ and let $k_{ij} = \mathbf{K}(\ell_i, \ell_j)$. Conditioned on parameters $\{\beta, p_i, \mathbf{K}\}$, the expectation, variance and covariance of $Z_i^{(m)}$ and $Z_j^{(m)}$ are

$$\begin{aligned} \mathbb{E}\left[Z_i^{(m)} | \cdot\right] &= \beta p_i \mathrm{e}^{\frac{1}{2}k_{ii}}, \qquad (12)\\ \mathbb{V}\left[Z_i^{(m)} | \cdot\right] &= \beta p_i \mathrm{e}^{2k_{ii}} + \beta^2 p_i^2 \mathrm{e}^{k_{ii}} \left(\mathrm{e}^{k_{ii}} - 1\right), \\ \text{Cov}\left[Z_i^{(m)}, Z_j^{(m)} | \cdot\right] &= \beta^2 p_i p_j \mathrm{e}^{\frac{1}{2}(k_{ii}+k_{jj})} \left(\mathrm{e}^{k_{ij}} - 1\right). \end{aligned}$$

We derived these properties by integrating out the vector $w_i^{(m)}$ using the law of iterated expectation. We observe that the covariance is similar to the un-normalized logistic normal (Aitchison, 1982), but with the additional term $\beta^2 p_i p_j$. In general, these p_i terms show how sparsity is enforced by the top-level DP, since both the expectation and variance terms go to zero as i increases. These values with $\{p_i\}$ integrated out are omitted for space.

The available covariance structure depends on the kernel. For example, when a Gaussian kernel is used, negative covariance is not achievable since $k_{ij} \geq 0$. In the next section, we will propose learning the kernel values directly, which will result in a simpler algorithm, and remove any restrictions on the kernel values imposed by a specific function.

3 Variational Inference for DILN

The central computational problem in Bayesian nonparametric mixed-membership modeling is posterior inference. The exact posterior—the conditional distribution of the top and lower-level parameters given a set of grouped data—is not tractable to compute. For HDP-based models, several approximation methods have been developed (Teh *et al.*, 2007; Liang *et al.*, 2007; Teh *et al.*, 2009).

In this section and the appendix, we derive a mean-field variational inference algorithm (Jordan *et al.*, 1999) for approximate posterior inference of a DILN mixed-membership model. We focus on topic modeling but note that our algorithm can be applied (with little modification) to any DILN mixed-membership model. Further, since the HDP is an instance of a DILN model, this algorithm provides a new inference method for HDP mixed-membership models using the gamma process representation.

Variational methods for Bayesian inference attempt to minimize the Kullback-Leibler divergence between a distribution over the hidden variables (indexed by variational parameters) and the true posterior. In a DILN topic model, the hidden variables are gamma variables $Z_k^{(m)}$, topic indexes $C_n^{(m)}$, GP draws $w_k^{(m)}$, topic distributions η_k, top-level proportions V_k, concentration parameters α and β, and GP parameters $\mathbf{m}$ and $\mathbf{K}$. Under the mean-field assumption—where the variational distribution is fully factorized—the variational distribution is

$$\begin{aligned} Q := \prod_{m=1}^{M} \prod_{n=1}^{N_m} \prod_{k=1}^{T} q(Z_k^{(m)}) q(C_n^{(m)}) q(w_k^{(m)}) q(\eta_k) q(V_k) \\ \times\, q(\alpha) q(\beta) q(\mathbf{m}) q(\mathbf{K}), \end{aligned} \tag{13}$$

where the components are defined as,

$$\begin{aligned} q(C_n^{(m)}) &= \text{Multinomial}(C_n^{(m)} | \phi_n^{(m)}) \\ q(Z_k^{(m)}) &= \text{Gamma}(Z_k^{(m)} | a_k^{(m)}, b_k^{(m)}) \\ q(w_k^{(m)}) &= \text{Normal}(w_k^{(m)} | \mu_k^{(m)}, v_k^{(m)}) \\ q(\eta_k) &= \text{Dirichlet}(\eta_k | \gamma'_{k,1}, \dots, \gamma'_{k,D}) \\ q(V_k) &= \delta_{V_k} \\ q(\mathbf{m}) q(\mathbf{K}) &= \delta_{\mathbf{m}} \cdot \delta_{\mathbf{K}} \\ q(\alpha) q(\beta) &= \delta_{\alpha} \cdot \delta_{\beta}\,. \end{aligned} \tag{14}$$

Note that we truncate the number of components to T in the top-level Dirichlet process (Blei & Jordan, 2005).[2] The truncation level T should be set larger than the total number of topics expected to be used by the data. The variational approximation will then prefer a distribution on topics that is sparse. We contrast this with the CTM and other finite topic models, which fit a pre-specified number of topics to the data, and potentially overfit if that number is too large.

Further note that we have selected several delta functions as variational distributions. In the case of V_k and β, we have followed Liang *et al.* (2007) in doing this for tractability. In the case of α, we have done this for simplicity. We have also selected delta functions for $\mathbf{m}$ and $\mathbf{K}$ for simple updates, which we discuss further in the next section. We observe that most parameters—and all document level parameters—have functional q distributions.

[2] Kurihara *et al.* (2006) show how infinite-dimensional objective functions can be defined for variational inference, however the conditions for this are not met by the DILN model as represented here.

Corpus	# training	# testing	vocabulary size	# total words
Huffington Post	3000	1000	6313	660,000
New York Times	5000	2000	3012	720,000
Science	5000	2000	4403	1,380,000
Wikipedia	5000	2000	6131	1,770,000

Table 1: Data sets considered for experiments. Five training/testing sets were constructed by selecting the number of documents shown for each corpus from larger data sets.

With the variational family defined, approximate inference proceeds by optimizing the variational objective function with respect to the variational parameters. The objective is a lower bound of the marginal likelihood of the data that is expressed as an expectation over q and is equivalent, up to a constant, to the negative KL divergence between the variational family and the true posterior. Given this objective, the variational parameters associated with each factor are iteratively optimized, forming a coordinate ascent optimization algorithm on the objective. The updates and objective are given in the appendix. Here, we discuss two characteristics of the algorithm: the relationship between DILN and HDP inference, and the fitting of the latent kernel matrix.

Variational DILN vs HDP. The variational inference algorithm derived in the appendix is similar to one that can be derived for the HDP using the representation discussed in Section 2.1. The difference lies in the update for $q(Z_k^{(m)})$. We give these updates for DILN below to facilitate discussion. Let N_m be the number of observations (e.g., words) in group m. The update for $q(Z_k^{(m)})$ is

$$\begin{aligned} a_k^{(m)} &= \beta p_k + \sum_{n=1}^{N_m} \phi_n^{(m)}(k), \\ b_k^{(m)} &= \mathbb{E}_Q[\exp\{-w_k^{(m)}\}] + \frac{N_m}{\xi_m}, \end{aligned} \tag{15}$$

where ξ_m is an auxiliary variable used in a first-order Taylor expansion discussed in the appendix. The update for $a_k^{(m)}$ contains the prior from the top-level DP, and the expected number of words in document m drawn from topic k. The parameter $b^{(m)}$ distinguishes DILN from the HDP.

A new inference algorithm arises for the HDP when the first term in the update for $b_k^{(m)}$ is set equal to one. In contrast, the first term for DILN is the expectation of $\exp\{-w_k^{(m)}\}$, which is the term introduced in Section 2.3 to model covariance between components. Including or excluding this parameter allows one to switch between variational inference for DILN and the HDP.

Kernel Learning. We optimize the kernel directly, rather than optimize locations in a latent space through a kernel function. This leads to the analytical update,

$$\mathbf{K} = \frac{1}{M}\sum_{m=1}^{M}\left\{(\mu^{(m)} - \mathbf{m})(\mu^{(m)} - \mathbf{m})^T + \mathrm{diag}(v^{(m)})\right\},$$

where $\mathbf{m}$ is the mean of the Gaussian process, $\mu^{(m)}$ is the variational mean of the log-normal vector of document m and $v^{(m)}$ is its variance (see the appendix). Hence, the update for $\mathbf{K}$ is approximately the covariance of these log-normal vectors.[3]

We follow Lanckriet *et al.* (2002) who motivated this approach in a similar situation by noting that any positive definite matrix is guaranteed to have some implicit mapping into a Hilbert space, and therefore can be called a kernel matrix. We note that the gram matrix, $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$, can be taken to be an eigendecomposition of $\mathbf{K}$, or $\mathbf{\Phi} := U\Lambda^{1/2}$, in which case the function mapping ℓ_k from its latent space into an inner product embedding space $\mathbf{\Phi}(\ell_k)$ is the kth row of $U\Lambda^{1/2}$; alternatively ℓ_k can be taken to be $\mathbf{\Phi}(\ell_k)$.

Finally, we note that updating $\mathbf{K}$ as above gives the optimal positive definite kernel matrix with respect to optimizing the lower bound in (19). We assume that a base measure L_0 can be defined such that the measure induced by this implicit kernel mapping leads to the updates given in the previous section when $q(\mathbf{K}) = \delta_{\mathbf{K}}$.

4 Experiments

We evaluate the performance of DILN as a topic modeling prior and compare with the HDP and CTM. We perform experiments on four text corpora: the *Huffington Post*, the *New York Times*, *Science* and *Wikipedia*. Each corpus was divided into five training and testing groups selected from a larger set of documents. See Table 1.

Experimental settings. We trained all models using variational inference; for the CTM, this is the algorithm given in Blei & Lafferty (2007); for the HDP, we use the inference method that arises as a special case of DILN, discussed in Section 3. (Therefore, the benefit of learning correlation structure in DILN is especially highlighted in the

[3] We considered working directly with the location space through a Gaussian kernel function. However, the resulting gradient algorithm was unnecessarily complicated, and was significantly slower due to a large number of matrix inversions.

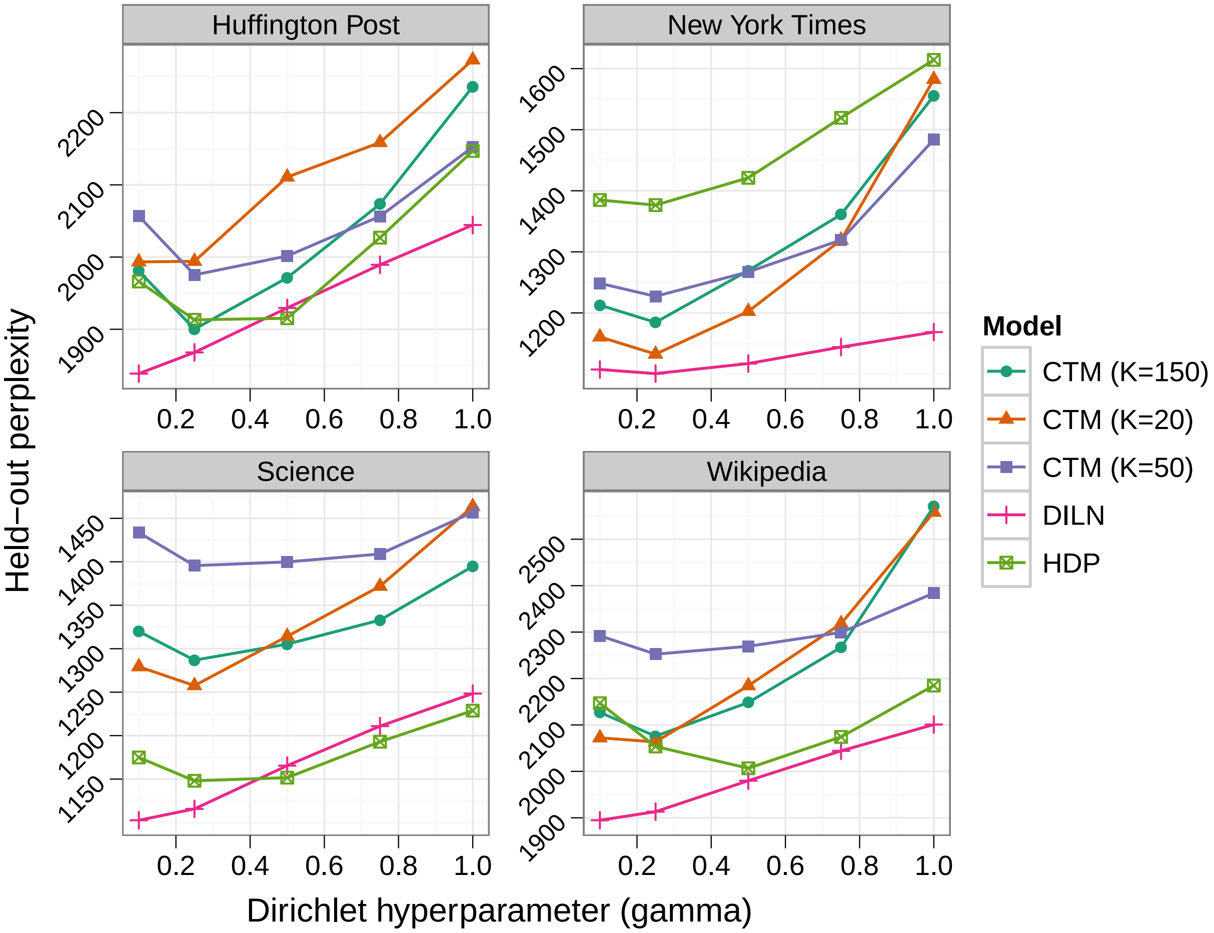


Figure 2: Perplexity results for four text corpora and averaged over five training/testing sets. For a fixed Dirichlet hyperparameter, the DILN topic model typically achieves better perplexity than both the HDP and CTM models. In all corpora, for some value of the hyperparameter, the DILN topic model achieves the best perplexity overall.

comparison with the HDP.) For the DILN and HDP models, we truncate the top level stick-breaking construction to $T = 200$ components, while we consider $K = 20, 50, 150$ topics for the CTM. Other values of K were also considered, and results for the three values shown here are representative of the performance observed for the CTM.

We initialize all models in the same way; we first cluster the empirical word distributions of each document with three iterations of k-means using the L_1 distance measure, and reorder these topics by size according to the indicators produced by k-means. We then scale the k-means centroids and add a small constant plus noise to smooth the initialization of the Dirichlet q distribution of each topic. All other parameters are initialized to values that result in a uniform distribution on these topics. Variational inference was terminated when the fractional change in the lower bound fell below 10^{-3}. In addition, we ran each algorithm using five different topic Dirichlet parameter settings: $\gamma = 0.1, 0.25, 0.5, 0.75, 1.0$.

Testing. We use a set-up for testing similar to one used by Asuncion *et al.* (2009). We randomly partition each test document $\mathbf{X}$ into two halves, $\mathbf{X}'$ and $\mathbf{X}''$. The first half of each document is used to learn the document-specific variational distributions. This includes the q distributions for $Z_{1:T}$, μ and v, and involves the values $\mathbf{m}$, $\mathbf{K}$, β, $V_{1:T}$ and $q(\eta_{1:T})$ learned in training. The second half of the testing document is then used for prediction.

These predictions are made by approximating the conditional marginal probability,

$$p(\mathbf{X}''|\mathbf{X}') = \qquad (16)$$

$$\int_{\Omega_{\boldsymbol{\eta},\boldsymbol{Z}}} \prod_{n=1}^{N} \left\{ \sum_{k=1}^{T} p(X''_n|\eta_k) p(C''_n = k|Z_{1:T}) \right\} dQ(\boldsymbol{Z}) dQ(\boldsymbol{\eta}),$$

where N is the number of observations constituting $\mathbf{X}''$,

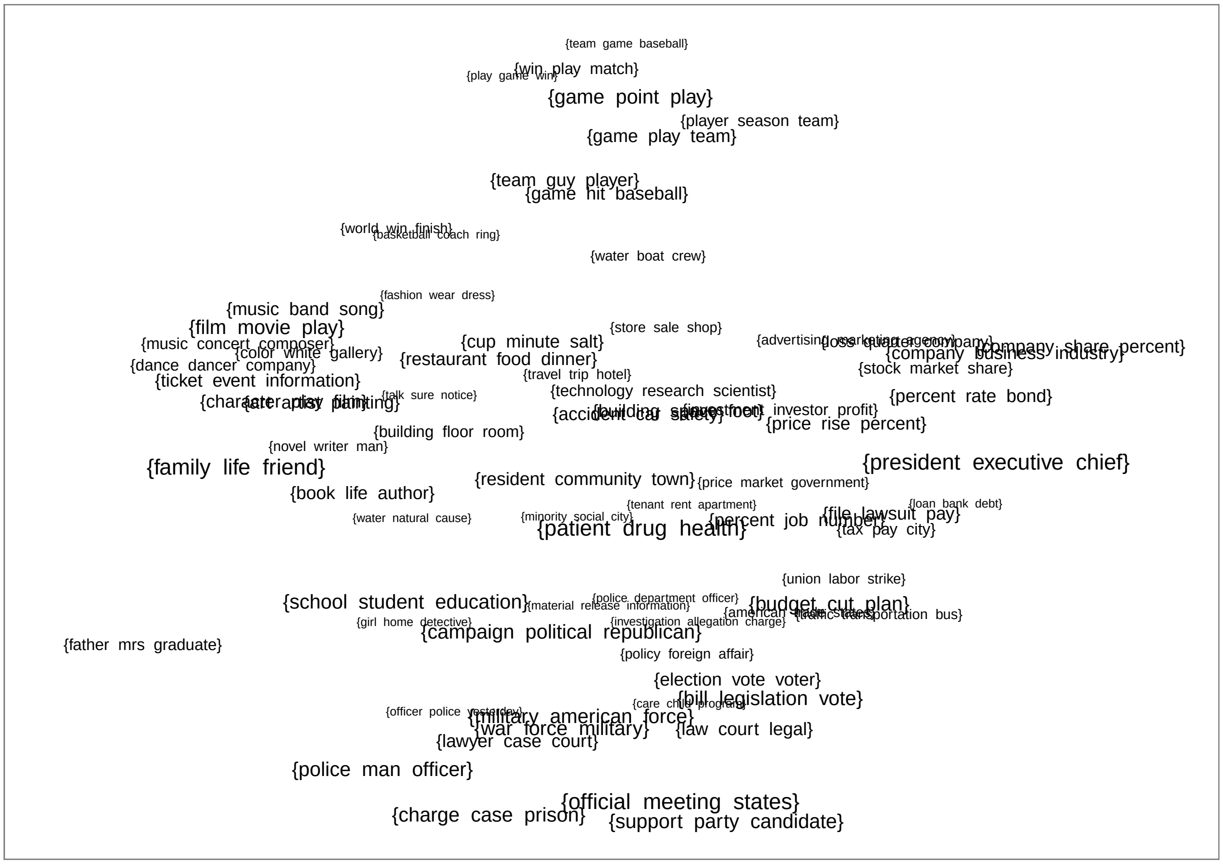


Figure 3: A visualization of words from the New York Times corpus with locations determined by the learned kernel. Each set represents a topic and contains its three most probable words, with size proportional to topic probability. Topics that are close together in the plot are conditionally correlated—given the values of the other topic proportions, two topics that are close together will tend to co-occur. (See the text for details.)

C''_n is the latent indicator associated with the nth word in $\mathbf{X}''$, and $\boldsymbol{\eta} := \eta_{1:T}$ and $\boldsymbol{Z} := Z_{1:T}$. Since this integral is intractable, we sample iid values from the factorized forms of $Q(Z_{1:T})$ and $Q(\eta_{1:T})$ for approximation. We note that the information regarding the document's correlation structure contained in μ and v can be found in $Q(Z_{1:T})$.

This approximation of the marginal likelihood is then used to compute the average per-word perplexity for the second half of the test document,

$$\text{perplexity} = \exp\left\{\frac{-\ln p(\mathbf{X}''|\mathbf{X}')}{N}\right\}. \qquad (17)$$

Lower perplexity indicates better performance.

Results and Discussion. Figure 2 contains testing results for the four corpora. We see that, in general, DILN outperforms both the HDP and CTM in terms of perplexity. Given the difference between DILN and HDP inference algorithms discussed in Section 3, this also shows the effectiveness of the kernel. Both DILN and HDP used significantly fewer topics than the truncation level $T = 200$; in general, between 50 and 100 topics were required. As expected, the CTM results were not sparse in their topic usage. Computation time for DILN and the HDP was comparable; both required on the order of one minute per iteration. Depending on the truncation level, the CTM was slightly to significantly faster than both DILN and HDP.

Topic models are often used to summarize and explore a collection. The correlation structure learned by DILN can benefit this task. In Figure 3, we represent the topics learned by the DILN prior in a latent space using the top three most probable words. For this figure, we used multidimensional scaling to project the inverse of the expected kernel to two dimensions. Two topics are close together in the plot if, conditioned on the other topic proportions components, their topic proportions are correlated. For example, "police man officer" is more conditionally correlated with "charge case prison" than "game point play."

Appendix

Proof of Almost Sure Finiteness of $\sum_{i=1}^{\infty} Z_i \mathrm{e}^{w_i}$

The normalizing constant is $S := \sum_{i=1}^{\infty} Z_i \mathrm{e}^{w_i}$ prior to absorbing the scaling factor within the gamma distribution. We show that this value is finite almost surely provided that the Gaussian process has bounded mean and covariance functions, and therefore the normalization of the DILN is well-defined. Let $S_T := \sum_{i=1}^{T} Z_i \mathrm{e}^{w_i}$. It follows that $S_1 \leq \cdots \leq S_T \leq \cdots \leq S$ and $S = \lim_{T\to\infty} S_T$. To prove that S is finite almost surely, we only need to prove that $\mathbb{E}[S]$ is finite. From the monotone convergence theorem, $\mathbb{E}[S] = \lim_{T\to\infty} \mathbb{E}[S_T]$. Since

$$\begin{aligned} \mathbb{E}[S_T] &= \textstyle\sum_{i=1}^{T} \mathbb{E}[Z_i]\mathbb{E}[\mathrm{e}^{w_i}] \\ &\leq \mathrm{e}^{\max_i(\mu_i + \frac{1}{2}\sigma_i^2)} \textstyle\sum_{i=1}^{T} \mathbb{E}[Z_i], \end{aligned} \quad (18)$$

$\mathbb{E}[S] \leq \beta \mathrm{e}^{\max_i(\mu_i + \frac{1}{2}\sigma_i^2)}$ and S is finite almost surely.

The Inference Algorithm for DILN

We present the variational inference algorithm for the DILN topic model. The variational lower bound is calculated by taking the following expectations with respect to the set of variational parameters, denoted $\mathbf{\Psi}$,

$$\begin{aligned} \mathcal{L}(\mathbf{X}, \mathbf{\Psi}) = & \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T} \phi_n^{(m)}(k)\mathbb{E}_Q[\ln p(X_n^{(m)}|\eta_k)] \\ & + \sum_{m=1}^{M}\sum_{n=1}^{N_m}\sum_{k=1}^{T} \phi_n^{(m)}(k)\mathbb{E}_Q[\ln p(C_n^{(m)} = k|Z_{1:T}^{(m)})] \\ & + \sum_{m=1}^{M}\sum_{k=1}^{T} \mathbb{E}_Q[\ln p(Z_k^{(m)}|\beta p_k, w_k^{(m)})] \\ & + \sum_{k=1}^{T} \mathbb{E}_Q[\ln p(\eta_k|\gamma)] + \sum_{k=1}^{T} \mathbb{E}_Q[\ln p(V_k|\alpha)] \\ & + \sum_{m=1}^{M} \mathbb{E}_Q[\ln p(w^{(m)}|\mathbf{m}, \mathbf{K})] \\ & + \mathbb{E}_Q[\ln p(\alpha)] + \mathbb{E}_Q[\ln p(\beta)] + \mathbb{E}_Q[\ln p(\mathbf{m})] \\ & + \mathbb{E}_Q[\ln p(\mathbf{K})] - \mathbb{E}_Q[\ln Q]. \end{aligned} \quad (19)$$

We optimize this lower bound using coordinate ascent. We present document and corpus level parameter updates below using the following definitions:

$$\begin{aligned} \mathbb{E}_Q[Z_k^{(m)}] &= a_k^{(m)}/b_k^{(m)} \quad (20) \\ \mathbb{E}_Q[\ln Z_k^{(m)}] &= \psi(a_k^{(m)}) - \ln b_k^{(m)} \\ \mathbb{E}_Q[\exp\{-w_k^{(m)}\}] &= \exp\left\{-\mu_k^{(m)} + \frac{1}{2}v_k^{(m)}\right\} \\ \mathbb{E}_Q[\ln \eta_{k,X_n^{(m)}}] &= \psi(\gamma'_{k,X_n^{(m)}}) - \psi(\textstyle\sum_d \gamma'_{k,d}) \end{aligned}$$

The symbol $\psi(\cdot)$ represents the digamma function.

Document-Level Parameters

Update $q(C_n^{(m)})$. For $k = 1, \ldots, T$ topics (21)

$$\phi_n^{(m)}(k) \propto \exp\left\{\mathbb{E}_Q[\ln \eta_{k,X_n^{(m)}}] + \mathbb{E}_Q[\ln Z_k^{(m)}]\right\}.$$

Update $q(Z_k^{(m)})$. This update is given in the text. The term $\mathbb{E}_Q[\ln p(C_n^{(m)} = k|Z_{1:T}^{(m)})]$ in the lower bound requires an approximation. We use a first-order Taylor expansion on the following intractable expectation,

$$-\mathbb{E}_Q\left[\ln \sum_{k=1}^{T} Z_k^{(m)}\right] \geq -\ln \xi_m - \frac{\sum_{k=1}^{T}\mathbb{E}_Q[Z_k^{(m)}] - \xi_m}{\xi_m}.$$

This approximation introduces an auxiliary parameter ξ_m for each group-level distribution. The update for ξ_m is: $\xi_m = \sum_{k=1}^{T}\mathbb{E}_Q[Z_k^{(m)}]$. Analytical updates for $a_k^{(m)}$ and $b_k^{(m)}$ result, and inference was approximately five times faster than the model with ξ pre-optimized since this approach required gradient methods.

Update $q(w_k^{(m)})$. We use steepest ascent to jointly update the T-dimensional vectors $\mu^{(m)}$ and $v^{(m)}$. Let $\lambda_{k,m} := \mathbb{E}_Q[Z_k^{(m)}] \times \mathbb{E}_Q[\exp\{-w_k^{(m)}\}]$. The derivatives comprising the gradient vector are,

$$\begin{aligned} \frac{\partial\mathcal{L}(\cdot)}{\partial\mu_k^{(m)}} &= \lambda_{k,m} - \beta p_k - \mathbf{K}_{k,:}^{-1}(\mu^{(m)} - \mathbf{m}), \\ \frac{\partial\mathcal{L}(\cdot)}{\partial v_k^{(m)}} &= -\frac{1}{2}\left\{\lambda_{k,m} - \mathbf{K}_{k,k}^{-1} + \frac{1}{v_k^{(m)}}\right\}. \end{aligned} \quad (22)$$

Corpus-Level Parameters

Update $q(\eta_k)$. For $d = 1, \ldots, D$ vocabulary words

$$\gamma'_{k,d} = \gamma + \sum_{m,n} \phi_n^{(m)}(k)\mathbb{I}\left(X_n^{(m)} = d\right). \quad (23)$$

Update $q(V_k)$. We use steepest ascent to jointly optimize $V_1, \ldots, V_{T-1}$ (with $V_T := 1$). The gradient is

$$\begin{aligned} \frac{\partial\mathcal{L}(\cdot)}{\partial V_k} = & -\frac{\alpha - 1}{1 - V_k} + \cdots \quad (24) \\ & \beta\sum_{j=1}^{k-1}(1 - V_j)\left\{\sum_m \left(\mathbb{E}_Q[\ln Z_k^{(m)}] - \mu_k^{(m)}\right) - V_k\psi\left(\beta p_k\right)\right\} \\ & -\beta\sum_m\sum_{j>k}\frac{p_j}{1 - V_k}\left\{\mu_j^{(m)} + \psi\left(\beta p_j\right) - \mathbb{E}_Q[\ln Z_j^{(m)}]\right\}. \end{aligned}$$

Update $q(\mathbf{m})$ and $q(\mathbf{K})$. The update for the mean vector is $\mathbf{m} = \frac{1}{M}\sum_{m=1}^{M}\mu^{(m)}$. The update for $\mathbf{K}$ is in the text.

Update $q(\alpha)$ and $q(\beta)$. We omit these updates due to space constraints.

Acknowledgements

David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520, AFOSR 09NL202, the Alfred P. Sloan foundation, and a grant from Google.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B* **44**, 139–177.

Asuncion, A., Welling, M., Smyth, P. & Teh, Y. (2009). On smoothing and inference for topic models. In *Uai*.

Blei, D. & Jordan, M. (2005). Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis* **1**, 121–144.

Blei, D. & Lafferty, J. (2007). A correlated topic model of Science. *Annals of Applied Statistics* **1**, 17–35.

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

De Iorio, M., Muller, P., Rosner, G. L. & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

Dunson, D. & Park, J. (2008). Kernel stick-breaking processes. *Biometrika* **95**, 307–323.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Gelfand, A., Kottas, A. & MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

Ishwaran, H. & Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics* **30**, 269–283.

Jordan, M., Ghahramani, Z., Jaakkola, T. & Saul, L. (1999). An introduction to variational methods for graphical models. *Machine Learning* **37**, 183–233.

Kingman, J. (1993). *Poisson processes*. Oxford University Press, USA.

Kurihara, K., Welling, M. & Vlassis, N. (2006). Accelerated variational DP mixture models. In *Advances in neural information processing systems*.

Lanckriet, G., Cristianini, N., Ghaoui, L. E., Bartlett, P. & Jordan, M. (2002). Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research* , 27–72.

Lenk, P. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association* **83**, 509–516.

Liang, P., Petrov, S., Jordan, M. & Klein, D. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Emperical methods in natural language processing*.

MacEachern, S. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science* .

Rao, V. & Teh, Y. W. (2009). Spatial normalized gamma processes. In *Advances in neural information processing systems*.

Rasmussen, C. & Williams, C. (2006). *Gaussian processes for machine learning*. MIT press.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Teh, Y., Jordan, M., Beal, M. & Blei, D. (2007). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

Teh, Y., Kurihara, K. & Welling, M. (2009). Collapsed variational inference for HDP. In *Advances in neural information processing systems*.