Deep Latent Dirichlet Allocation with Topic-Layer-Adaptive Stochastic Gradient Riemannian MCMC

Yulai Cong [1] **Bo Chen** [1] **Hongwei Liu** [1] **Mingyuan Zhou** [2]

Abstract

It is challenging to develop stochastic gradient based scalable inference for deep discrete latent variable models (LVMs), due to the difficulties in not only computing the gradients, but also adapting the step sizes to different latent factors and hidden layers. For the Poisson gamma belief network (PGBN), a recently proposed deep discrete LVM, we derive an alternative representation that is referred to as deep latent Dirichlet allocation (DLDA). Exploiting data augmentation and marginalization techniques, we derive a block-diagonal Fisher information matrix and its inverse for the simplex-constrained global model parameters of DLDA. Exploiting that Fisher information matrix with stochastic gradient MCMC, we present topic-layer-adaptive stochastic gradient Riemannian (TLASGR) MCMC that jointly learns simplex-constrained global parameters across all layers and topics, with topic and layer specific learning rates. State-of-the-art results are demonstrated on big data sets.

1. Introduction

The increasing amount and complexity of data call for large-capacity models, such as deep discrete latent variable models (LVMs) for unsupervised data analysis (Hinton et al., 2006; Bengio et al., 2007; Srivastava et al., 2013; Ranganath et al., 2015; Zhou et al., 2016a), and scalable inference methods, such as stochastic gradient Markov chain Monte Carlo (SG-MCMC) that provides posterior samples in a non-batch learning setting (Welling & Teh, 2011; Patterson & Teh, 2013; Ma et al., 2015). Unfortunately, most deep LVMs, such as deep belief network (DBN) (Hinton et al., 2006) and deep Boltzmann machines (DBM) (Salakhutdinov & Hinton, 2009), use greedy layerwise training, without a principled way to jointly learn multilayers in an unsupervised manner (Bengio et al., 2007). While SG-MCMC has recently been successfully applied to several "shallow" LVMs, such as mixture models (Welling & Teh, 2011) and mixed-membership models (Patterson & Teh, 2013), it has been rarely applied to "deep" ones, probably due to the lack of understanding on how to jointly learn the latent variables of different layers and adjust the layer and topic specific learning rates in a non-batch learning setting.

To investigate scalable SG-MCMC inference for deep LVMs, we focus our study on the recently proposed Poisson gamma belief network (PGBN), whose hidden layers are parameterized with gamma distributed hidden units and connected with Dirichlet distributed basis vectors (Zhou et al., 2016a). The PGBN is capable of extracting topics from a text corpus at multiple layers and outperforms a large number of topic modeling algorithms. However, the PGBN is currently trained with a batch Gibbs sampler that is not scalable to big data. In this paper, we focus on developing scalable multilayer joint inference for the PGBN.

We will show that scalable multilayer joint inference of the PGBN could be facilitated by its Fisher information matrix (FIM) (Amari, 1998; Girolami & Calderhead, 2011; Pascanu & Bengio, 2013), which, although seemingly impossible to derive and challenging to work with due to the need to compute the expectations over trigamma functions, is readily available under an alternative representation of the PGBN, referred to as deep latent Dirichlet allocation (DLDA). DLDA, derived by exploiting data augmentation and marginalization techniques on the PGBN, can be considered as a multilayer generalization of latent Dirichlet allocation (LDA) (Blei et al., 2003). Following a general framework for SG-MCMC (Ma et al., 2015), the block diagonal structure of the FIM of DLDA makes it be easily inverted to precondition the mini-batch based noisy gradients to exploit the second-order local curvature information, leading to topic-layer-adaptive step sizes based on the Riemannian manifold and the same asymptotic performance as a natural gradient based batch-learning algorithm (Amari, 1998; Pascanu & Bengio, 2013). To the best of our knowl-

[1]National Laboratory of Radar Signal Processing, Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an, China. [2]McCombs School of Business, The University of Texas at Austin, Austin, TX 78712, USA. Correspondence to: Bo Chen <bchen@mail.xidian.edu.cn>, Mingyuan Zhou <mingyuan.zhou@mccombs.utexas.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

edge, this is the first time that the FIM of a deep LVM is shown to have an analytical and practical form. How we derive the FIM for the PGBN using data augmentation and marginalization techniques in this paper may serve as an example to help derive the FIMs for other deep LVMs.

Besides presenting the analytical FIM of the PGBN, important for the marriage of a deep LVM and SG-MCMC, we make another contribution in showing how to facilitate SG-MCMC for an LVM equipped with simplex-constrained model parameters $\boldsymbol{\phi}_k = (\phi_{1k}, \dots, \phi_{Vk})^T$, which means $\sum_{v=1}^{V} \phi_{vk} = 1$ and $\phi_{vk} \in \mathbb{R}_+$, where $\mathbb{R}_+ := \{x, x \geq 0\}$, by using a reduced-mean simplex parameterization together with a fast sampling procedure recently introduced in Cong et al. (2017). Unlike other simplex parameterizations, the reduced-mean one does not make heuristic pseudolikelihood assumptions. Though it has previously been deemed unsound, it is successfully integrated into our SG-MCMC framework to deliver state-of-the-art results. Exploiting the analytical FIM of DLDA and novel inference for simplex-constrained parameters under a general SG-MCMC framework (Ma et al., 2015), we present topic-layer-adaptive stochastic gradient Riemannian (TLASGR) MCMC for DLDA, which automatically adjusts the learning rates of global model parameters across all layers and topics, without the need to set the same learning rate for all that is commonly used in practice due to the difficulty in identifying an appropriate combination of the learning rates for different layers and topics.

2. PGBN and SG-MCMC

The generative model of the Poisson gamma belief network (PGBN) (Zhou et al., 2016a) with L hidden layers, from top to bottom, is expressed as

$$\begin{aligned}
&\boldsymbol{\theta}_j^{(L)} \sim \mathrm{Gam}\left(\boldsymbol{r}, 1/c_j^{(L+1)}\right), \\
&\cdots \\
&\boldsymbol{\theta}_j^{(l)} \sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_j^{(l+1)}, 1/c_j^{(l+1)}\right), \\
&\cdots \\
&\boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right),\ \boldsymbol{\theta}_j^{(1)} \sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, \frac{p_j^{(2)}}{1-p_j^{(2)}}\right),
\end{aligned} \tag{1}$$

where the j^{th} observed or latent V-dimensional count vectors $\boldsymbol{x}_j^{(1)} \in \mathbb{Z}^V$, where $\mathbb{Z} := \{0, 1, \dots\}$, are factorized under the Poisson (Pois) likelihood; the hidden units $\boldsymbol{\theta}_j^{(l)} \in \mathbb{R}_+^{K_l}$ of layer l are factorized under the gamma (Gam) likelihood into the product of the basis vector matrix $\boldsymbol{\Phi}^{(l)} = \left(\boldsymbol{\phi}_1^{(l)}, \dots, \boldsymbol{\phi}_{K_l}^{(l)}\right) \in \mathbb{R}_+^{K_{l-1} \times K_l}$ and the hidden units of the next layer, where $\boldsymbol{\phi}_k^{(l)} \sim \mathrm{Dir}\left(\eta^{(l)} \mathbf{1}_{K_{l-1}}\right)$ are Dirichlet (Dir) distributed and $\mathbf{1}_{K_{l-1}}$ is a K_{l-1}-dimensional vector of all ones; the gamma shape parameters $\boldsymbol{r} = (r_1, \cdots, r_{K_L})^T$ at the top layer are shared across all j; $\{1/c_j^{(l)}\}_{3,L+1}$ are gamma scale parameters, where $c_j^{(l)} \sim \mathrm{Gam}(e_0, 1/f_0)$, and $c_j^{(2)} := \left(1 - p_j^{(2)}\right)/p_j^{(2)}$, where $p_j^{(2)} \sim \mathrm{Beta}(a_0, b_0)$ are introduced to help reduce the dependencies between $\theta_{jk}^{(1)}$ and $c_j^{(2)}$. The PGBN in (1) can be further extended under the Bernoulli-Poisson link as $\boldsymbol{b}_j^{(1)} = \mathbf{1}\left(\boldsymbol{x}_j^{(1)} > 0\right)$, and under the Poisson randomized gamma link as $\boldsymbol{y}_j^{(1)} \sim \mathrm{Gam}\left(\boldsymbol{x}_j^{(1)}, 1/a_j\right)$, where $a_j \sim \mathrm{Gam}(e_0, 1/f_0)$.

The PGBN infers a multilayer deep representation of the data, whose inferred basis vectors $\boldsymbol{\phi}_k^{(l)}$ at hidden layer l can be directly visualized as $\left[\prod_{t=1}^{l-1} \boldsymbol{\Phi}^{(t)}\right] \boldsymbol{\phi}_k^{(l)}$, which are their projections into the V-dimensional probability simplex. The information of the whole data set is compressed by the PGBN into the inferred sparse network $\{\boldsymbol{\Phi}^{(1)}, \dots, \boldsymbol{\Phi}^{(L)}\}$, where $\phi_{k_1 k_2}^{(l)}$ indicates the connection strength between node (basis vector) k_1 of layer $l-1$ and node k_2 of layer l. Moreover, the network structure can be inferred from the data by combining the gamma-negative binomial process of Zhou & Carin (2015) with a greedy layer-wise training strategy. Extensive experiments in Zhou et al. (2016a) show that the PGBN can extract basis vectors that are very specific/abstract in the bottom layer and become increasingly more general when moving upwards from the bottom to top hidden layers, and the K_1 hidden units $\boldsymbol{\theta}_j^{(1)}$ in the first hidden layer, which are unsupervisedly extracted and regularized with the deep network, are well suited for out-of-sample prediction and being used as features for classification.

Despite all these attractive model properties, the current inference of the PGBN relies on an upward-downward Gibbs sampler that requires processing all data in each iteration and hence often does not scale well to big data unless with parallel computing. To make its inference scalable to allow processing a large amount of data sufficiently fast on a regular personal computer, we resort to SG-MCMC that subsamples the data and utilizes stochastic gradients in each MCMC iteration to generate posterior samples for globally shared model parameters. Let us denote the posterior of model parameters $\boldsymbol{z}$ given the data $X = \{\boldsymbol{x}_j\}_{1,J}$ as $p(\boldsymbol{z} \,|X) \propto e^{-H(\boldsymbol{z})}$, with potential function $H(\boldsymbol{z}) = -\ln p(\boldsymbol{z}) - \sum_j \ln p(\boldsymbol{x}_j \,|\boldsymbol{z})$. As in Theorem 1 of Ma et al. (2015), $p(\boldsymbol{z} \,|X)$ is the stationary distribution of the dynamics defined by the stochastic differential equation (SDE) $d\boldsymbol{z} = \boldsymbol{f}(\boldsymbol{z})\, dt + \sqrt{2\mathbf{D}(\boldsymbol{z})} d\mathbf{W}(t)$, if the deterministic drift $\boldsymbol{f}(\boldsymbol{z})$ is restricted to the form

$$\boldsymbol{f}(\boldsymbol{z}) = -\left[\mathbf{D}(\boldsymbol{z}) + \mathbf{Q}(\boldsymbol{z})\right] \nabla H(\boldsymbol{z}) + \Gamma(\boldsymbol{z}), \tag{2}$$

$$\Gamma_i(\boldsymbol{z}) = \sum_j \frac{\partial}{\partial z_j} \left[\mathbf{D}_{ij}(\boldsymbol{z}) + \mathbf{Q}_{ij}(\boldsymbol{z})\right], \tag{3}$$

where $\mathbf{D}(\boldsymbol{z})$ is a positive semidefinite diffusion matrix, $\mathbf{W}(t)$ is a Wiener process, $\mathbf{Q}(\boldsymbol{z})$ is a skew-symmetric curl matrix, and $\Gamma_i(\boldsymbol{z})$ is the ith element of the compensation vector $\Gamma(\boldsymbol{z})$. Thus one has a mini-batch update rule as

$$\begin{aligned}
\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \varepsilon_t \Big\{ -\left[\mathbf{D}(\boldsymbol{z}_t) + \mathbf{Q}(\boldsymbol{z}_t)\right] \nabla \tilde{H}(\boldsymbol{z}_t) + \Gamma(\boldsymbol{z}_t) \Big\} \\
+ \mathcal{N}\Big(\mathbf{0}, \varepsilon_t \left[2\mathbf{D}(\boldsymbol{z}_t) - \varepsilon_t \hat{\mathbf{B}}_t\right]\Big),
\end{aligned} \tag{4}$$

where ε_t denotes step sizes, $\tilde{H}(\boldsymbol{z}) = -\ln p(\boldsymbol{z}) - \rho \sum_{\boldsymbol{x}\in\tilde{X}} \ln p(\boldsymbol{x}\,|\,\boldsymbol{z})$, $\tilde{X}$ the mini-batch, ρ the ratio of the dataset size $|X|$ to the mini-batch size $|\tilde{X}|$, and $\hat{\mathbf{B}}_t$ an estimate of the stochastic gradient noise variance satisfying a positive definite constraint as $2\mathbf{D}(\boldsymbol{z}_t) - \varepsilon_t \hat{\mathbf{B}}_t \succ \mathbf{0}$.

As shown in Ma et al. (2015), stochastic gradient Riemannian Langevin dynamics (SGRLD) of Patterson & Teh (2013) is a special case with $\mathbf{D}(\boldsymbol{z}) = \mathbf{G}(\boldsymbol{z})^{-1}, \mathbf{Q}(\boldsymbol{z}) = \mathbf{0}, \hat{\mathbf{B}}_t = \mathbf{0}$, where $\mathbf{G}(\boldsymbol{z})$ denotes the Fisher information matrix (FIM). SGRLD is designed to solve the inference on the probability simplex, where four different parameterizations of the simplex-constrained basis vectors are discussed, including reduced-mean, expanded-mean, reduced-natural, and expanded-natural. Here, we consider both expanded-mean, previously shown to provide the best overall results, and reduced-mean, which, although discarded in Patterson & Teh (2013) due to its unstable gradients, is used in this paper to produce state-of-the-art results.

Let us denote $\boldsymbol{\phi}_k \in \mathbb{R}_+^V$ as a vector on the probability simplex, $\hat{\boldsymbol{\phi}}_k \in \mathbb{R}_+^V$ as a nonnegative vector, and $\boldsymbol{\varphi}_k \in \mathbb{R}_+^{V-1}$ as a nonnegative vector constrained with $\varphi_{\cdot k} := \sum_{v=1}^{V-1} \varphi_{vk} \le 1$. For convenience, the symbol "$\cdot$" will denote the operation of summing over the corresponding index. We use $\left(\hat{\phi}_{1k}, \cdots, \hat{\phi}_{Vk}\right)^T / \sum_v \hat{\phi}_{vk}$ as an expanded-mean parameterization of $\boldsymbol{\phi}_k$ and $\left(\varphi_{1k}, \cdots, \varphi_{(V-1)k}, 1 - \sum_{v<V} \varphi_{vk}\right)^T$ as a reduced-mean parametrization of $\boldsymbol{\phi}_k$. SGRLD focuses on a single-layer model with a multinomial likelihood $\boldsymbol{n}_k \sim \mathrm{Mult}(n_{\cdot k}, \boldsymbol{\phi}_k)$ and a Dirichlet distributed prior $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\eta \mathbf{1}_V)$. For inference, it adopts the expanded-mean parameterization of $\boldsymbol{\phi}_k$ and makes a heuristic assumption that $n_{\cdot k} \sim \mathrm{Pois}(\hat{\phi}_{\cdot k})$. While that heuristic pseudolikelihood assumption of SGRLD is neither supported by the original generative model nor rigorously justified in theory, it converts a Dirichlet-multinomial model into a gamma-Poisson one, allowing a simple sampling equation for $\hat{\boldsymbol{\phi}}_k$ as

$$\begin{aligned}(\hat{\boldsymbol{\phi}}_k)_{t+1} = \Big| (\hat{\boldsymbol{\phi}}_k)_t + \varepsilon_t \Big[(\boldsymbol{n}_k + \eta) - (n_{\cdot k} + \hat{\phi}_{\cdot k})(\boldsymbol{\phi}_k)_t \Big] \\ + \mathcal{N}\Big(0, 2\varepsilon_t \mathrm{diag}\Big[(\hat{\boldsymbol{\phi}}_k)_t\Big]\Big) \Big|, \end{aligned} \tag{5}$$

where the absolute operation $|\cdot|$ is used to ensure positive-valued $\hat{\boldsymbol{\phi}}_k$. Below we show how to eliminate that heuristic assumption by parameterizing $\boldsymbol{\phi}_k$ with reduced-mean, and develop efficient SG-MCMC for the PGBN, which reduces to LDA when the number of hidden layers is one.

3. Deep Latent Dirichlet Allocation

While the original construction of PGBN in (1) makes it seemingly impossible to compute the FIM, as shown in Appendix A, we find that, by exploiting data augmentation and marginalization techniques, the PGBN generative model can be rewritten under an alternative representation that marginalizes out all the gamma distributed hidden units, as shown in the following Lemma, where $\mathrm{Log}(\cdot)$ denotes the logarithmic distribution (Johnson et al., 1997), $m \sim \mathrm{SumLog}(x,p)$ represents the sum-logarithmic distribution generated with $m = \sum_{i=1}^{x} u_i, u_i \sim \mathrm{Log}(p)$ (Zhou et al., 2016b). The proof is deferred to the Appendix.

Lemma 3.1. *Denote* $q_j^{(l+1)} = \ln\left(1 + q_j^{(l)}/c_j^{(l+1)}\right)$ *for* $l = 1, \ldots, L$, *where* $q_j^{(1)} := 1$, *which means* $q_j^{(l+1)} = \ln\left(1 + \frac{1}{c_j^{(l+1)}} \ln\left\{1 + \frac{1}{c_j^{(l)}} \ln\left[1 + \cdots \ln\left(1 + \frac{1}{c_j^{(2)}}\right)\right]\right\}\right)$. *With* $p_j^{(l)} := 1 - e^{-q_j^{(l)}}$ *and* $\tilde{p} := q_\cdot^{(L+1)}/(c_0 + q_\cdot^{(L+1)})$, *one may re-express the hierarchical model of the PGBN as deep latent Dirichlet allocation (DLDA) as*

$$\begin{aligned}
&x_{k\cdot}^{(L+1)} \sim \mathrm{Log}(\tilde{p}), K_L \sim \mathrm{Pois}[-\gamma_0 \ln(1-\tilde{p})], \\
&X^{(L+1)} = \textstyle\sum_{k=1}^{K_L} x_{k\cdot}^{(L+1)} \delta_{\boldsymbol{\phi}_k^{(L)}}, \\
&\left(x_{vj}^{(L+1)}\right)_j \sim \mathrm{Mult}\left[x_{v\cdot}^{(L+1)}, \left(q_j^{(L+1)}\right)_j / q_\cdot^{(L+1)}\right], \\
&m_{vj}^{(L)(L+1)} \sim \mathrm{SumLog}(x_{vj}^{(L+1)}, p_j^{(L+1)}), \\
&\cdots \\
&x_{vj}^{(l)} = \textstyle\sum_{k=1}^{K_l} x_{vkj}^{(l)}, \left(x_{vkj}^{(l)}\right)_v \sim \mathrm{Mult}\left(m_{kj}^{(l)(l+1)}, \boldsymbol{\phi}_k^{(l)}\right), \\
&m_{vj}^{(l-1)(l)} \sim \mathrm{SumLog}(x_{vj}^{(l)}, p_j^{(l)}), \\
&\cdots \\
&x_{vj}^{(1)} = \textstyle\sum_{k=1}^{K_1} x_{vkj}^{(1)}, \left(x_{vkj}^{(1)}\right)_v \sim \mathrm{Mult}\left(m_{kj}^{(1)(2)}, \boldsymbol{\phi}_k^{(1)}\right).
\end{aligned} \tag{6}$$

Note that the equations in the first four lines of (6) precisely represent a random count matrix generated from a gamma-negative binomial process that can also be generated from

$$\begin{aligned}
&x_{kj}^{(L+1)} \sim \mathrm{Pois}\left(r_k q_j^{(L+1)}\right), r_k \sim \mathrm{Gam}(\gamma_0/K, 1/c_0), \\
&m_{kj}^{(L)(L+1)} \sim \mathrm{SumLog}\left(x_{kj}^{(L+1)}, p_j^{(L+1)}\right)
\end{aligned} \tag{7}$$

by letting $K \to \infty$ (Zhou et al., 2016b). When $L = 1$, the PGBN whose $(r_k, \boldsymbol{\phi}_k)$ are the points of a gamma process reduces to the gamma-negative binomial process PFA of Zhou & Carin (2015), whose alternative representation is provided in Corollary D.1 in the Appendix. Note that how we re-express the PGBN as DLDA is related to how Schein et al. (2016) re-express their Poisson–gamma dynamic systems into an alternative representation that facilitates inference.

DLDA, designed to infer a multilayer representation of observed or latent high-dimensional sparse count vectors, constrains all the basis vectors of different layers to probability simplices. It is clear from (6) that a data point backpropagates its counts through the network one layer at a time via a sum-logarithmic distribution to *enlarge* each element of a K_l-dimensional count vector, a multinomial distribution to *partition* that enlarged count vector into a $K_{l-1} \times K_l$ count matrix, and then a row-sum operation to *aggregate* that latent count matrix into a K_{l-1}-dimensional count vector, where $K_0 := V$ is the feature dimension. Below we show that such an alternative representation that repeats the *enlarge-partition-augment* operation brings significant benefits when it comes to deriving SG-MCMC inference with preconditioned gradients.

3.1. Fisher Information Matrix of DLDA

In deep LVMs, whose parameters of different layers are often highly correlated to each other, it is often difficult to tune the step sizes of different layers together and hence one often chooses to train an unsupervised deep model in a greedy layer-wise manner (Bengio et al., 2007), which is a sensible but not optimal training strategy. To address that issue, we resort to the inverse of the FIM that is widely used to precondition the gradients to adjust the learning rates of different model parameters (Amari, 1998; Pascanu & Bengio, 2013; Ma et al., 2015; Li et al., 2016). However, it is often difficult to compute the FIMs of deep LVMs as

$$\mathbf{G}(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{\Omega}|\boldsymbol{z}}\left[-\frac{\partial^2}{\partial \boldsymbol{z}^2}\ln p(\boldsymbol{\Omega}\,|\boldsymbol{z})\right], \tag{8}$$

where $\boldsymbol{z}$ denotes the set of all global variables and $\boldsymbol{\Omega}$ is the set of all observed and local variables.

Although deriving the FIM for the PGBN generative model shown in (1) seems impossible, we find it to be straightforward under the alternative DLDA representation in (6). Since the likelihood of (6) is fully factorized with respect to the global parameters $\boldsymbol{z}$, *i.e.,* $\boldsymbol{\phi}_k^{(l)}$ and $\boldsymbol{r}$, one may readily show the FIM $\mathbf{G}(\boldsymbol{z})$ of (6) has a block diagonal form as

$$\operatorname{diag}\left[\mathbf{I}\left(\boldsymbol{\varphi}_1^{(1)}\right), \cdots, \mathbf{I}\left(\boldsymbol{\varphi}_{K_L}^{(L)}\right), \mathbf{I}(\boldsymbol{r})\right]; \tag{9}$$

with the likelihood $\left(x_{vkj}^{(l)}\right)_v \sim \mathrm{Mult}\left(m_{kj}^{(l)(l+1)}, \boldsymbol{\phi}_k^{(l)}\right)$ and the reduced-mean parameterization, we have

$$\begin{aligned}\mathbf{I}\left(\boldsymbol{\varphi}_k^{(l)}\right) &= -\mathbb{E}\left[\tfrac{\partial^2}{\partial \boldsymbol{\varphi}_k^{(l)2}}\ln\left(\prod_j \mathrm{Mult}\left[(x_{vkj}^{(l)})_v; m_{kj}^{(l)(l+1)}, \boldsymbol{\phi}_k^{(l)}\right]\right)\right]\\ &= M_k^{(l)}\left[\operatorname{diag}\left(1/\boldsymbol{\varphi}_k^{(l)}\right) + \mathbf{1}\mathbf{1}^T/(1-\varphi_{\cdot k}^{(l)})\right],\end{aligned} \tag{10}$$

where $M_k^{(l)} := \mathbb{E}\left[m_{k\cdot}^{(l)(l+1)}\right] = \mathbb{E}\left[x_{\cdot k\cdot}^{(l)}\right]$. Similarly, with the likelihood $x_{kj}^{(L+1)} \sim \mathrm{Pois}(r_k q_j^{(L+1)})$, we have

$$\mathbf{I}(\boldsymbol{r}) = M^{(L+1)}\operatorname{diag}(1/\boldsymbol{r}), \tag{11}$$

where $M^{(L+1)} := \mathbb{E}\left[q_\cdot^{(L+1)}\right]$.

The block diagonal structure of the FIM of DLDA makes it computationally appealing to apply its inverse for preconditioning. Under the framework suggested by (4), we adopt the similar settings used in SGRLD (Patterson & Teh, 2013) that lets $\mathbf{D}(\boldsymbol{z}) = \mathbf{G}(\boldsymbol{z})^{-1}$, $\mathbf{Q}(\boldsymbol{z}) = \mathbf{0}$, and $\hat{\mathbf{B}}_t = \mathbf{0}$. While other more sophisticated settings described in Ma et al. (2015), including as special examples stochastic gradient Hamiltonian Monte Carlo in Chen et al. (2014) and stochastic gradient thermostats in Ding et al. (2014), may be used to further improve the performance, we choose this specific one to make a direct comparison with SGRLD.

By substituting the FIM $\mathbf{G}(\boldsymbol{z})$ and the adopted settings into (4), it is apparent that we only need to choose a single step size ε_t, relying on the FIM to automatically adjust the relatively learning rates for different parameters across all layers and topics. Moreover, the block-diagonal structure of $\mathbf{G}(\boldsymbol{z})$ will be carried over to its inverse $\mathbf{D}(\boldsymbol{z})$, making it simple to perform updating using (4), as described below.

3.2. Inference on the Probability Simplex

As discussed in Section 2, to sample simplex-constrained model parameters for a Dirichlet-multinomial model, the SGRLD of Patterson & Teh (2013) adopts the expanded-mean parameterization of simplex-constrained vectors and makes a pseudolikelihood assumption to simplify the derivation of update equations. In this paper, without replying on that pseudolikelihood assumption, we choose to use the reduced-mean parameterization of simplex-constrained vectors, despite being considered as an unsound choice in Patterson & Teh (2013). In the following discussion, we omit the layer-index superscript $^{(l)}$ for simplicity.

With the multinomial likelihood in (6) and the Dirichlet-multinomial conjugacy, the conditional posterior of $\boldsymbol{\phi}_k$ can be expressed as $(\boldsymbol{\phi}_k\,|-) \sim \mathrm{Dir}(x_{1k\cdot}+\eta, \ldots, x_{Vk\cdot}+\eta)$. Taking the gradient with respect to $\boldsymbol{\varphi}_k \in \mathbb{R}_+^{V-1}$ on the summation of the negative log-likelihood of a mini-batch $\tilde{X}$ scaled by $\rho = |X|/|\tilde{X}|$ and the negative log-likelihood of the Dirichlet prior, we have

$$\nabla_{\boldsymbol{\varphi}_k}\left[-\tilde{H}(\boldsymbol{\varphi}_k)\right] = \frac{\rho\bar{\boldsymbol{x}}_{:k\cdot}+\eta-1}{\boldsymbol{\varphi}_k} - \frac{\rho\tilde{x}_{Vk\cdot}+\eta-1}{1-\varphi_{\cdot k}}, \tag{12}$$

where $\tilde{x}_{vk\cdot} = \sum_{j:x_j\in\tilde{X}} x_{vkj}$ and $\bar{\boldsymbol{x}}_{:k\cdot} := (\tilde{x}_{1k\cdot}, \ldots, \tilde{x}_{(V-1)k\cdot})^T$. Note the gradient in (12) becomes unstable when some components of $\boldsymbol{\varphi}_k$ approach zeros, a key reason that this approach is mentioned but not further pursued in Patterson & Teh (2013).

However, after preconditioning the noisy gradient with the inverse of the FIM, it is intriguing to find out that the stability issue completely disappears. More specifically, by plugging both the FIM in (10) and noisy gradient in (12) into the SG-MCMC update in (4), a noisy estimate of the deterministic drift defined in (2) obtained using the current mini-batch can be expressed as

$$\begin{aligned}&\mathbf{I}(\boldsymbol{\varphi}_k)^{-1}\nabla_{\boldsymbol{\varphi}_k}\left[-\tilde{H}(\boldsymbol{\varphi}_k)\right] + \Gamma(\boldsymbol{\varphi}_k)\\ &= M_k^{-1}\left[(\rho\bar{\boldsymbol{x}}_{:k\cdot}+\eta) - (\rho\tilde{x}_{\cdot k\cdot}+\eta V)\,\boldsymbol{\varphi}_k\right],\end{aligned} \tag{13}$$

where $\Gamma(\boldsymbol{\varphi}_k) = M_k^{-1}[1 - V\boldsymbol{\varphi}_k]$ according to (3), as derived in detail in Appendix B. Consequently, with $[\cdot]_\triangle$ denoting the constraint that $\varphi_{vk} \geq 0$ and $\sum_{v=1}^{V-1} \varphi_{vk} \leq 1$, using (4), the sampling of $\boldsymbol{\varphi}_k$ becomes

$$(\boldsymbol{\varphi}_k)_{t+1} = \Big[(\boldsymbol{\varphi}_k)_t + \tfrac{\varepsilon_t}{M_k}\left[(\rho\bar{\boldsymbol{x}}_{:k\cdot} + \eta) - (\rho\tilde{x}_{\cdot k\cdot} + \eta V)(\boldsymbol{\varphi}_k)_t\right] + \mathcal{N}\Big(\mathbf{0}, \tfrac{2\varepsilon_t}{M_k}\Big[\mathrm{diag}\,(\boldsymbol{\varphi}_k)_t - (\boldsymbol{\varphi}_k)_t\,(\boldsymbol{\varphi}_k)_t^T\Big]\Big)\Big]_{\triangle}. \tag{14}$$

Even without the $[\cdot]_{\triangle}$ constraint, the multivariate normal (MVN) simulation in (14), although easy to interpret and numerically stable, is computationally expensive if the Cholesky decomposition, with $\mathcal{O}((V-1)^3)$ complexity (Golub & Van Loan, 2012), is adopted directly. Fortunately, using Theorem 2 of Cong et al. (2017), the special structure of its covariance matrix allows an equivalent but substantially more efficient simulation of $\mathcal{O}(V)$ complexity by transforming a random variable drawn from a related MVN that has a diagonal covariance matrix. More specifically, the sampling of (14) can be efficiently realized in a V-dimensional space as

$$(\boldsymbol{\phi}_k)_{t+1} = \Bigg[(\boldsymbol{\phi}_k)_t + \frac{\varepsilon_t}{M_k}\left[(\rho\tilde{\boldsymbol{x}}_{:k\cdot} + \eta) - (\rho\tilde{x}_{\cdot k\cdot} + \eta V)(\boldsymbol{\phi}_k)_t\right] + \mathcal{N}\left(\mathbf{0}, \frac{2\varepsilon_t}{M_k}\mathrm{diag}\,(\boldsymbol{\phi}_k)_t\right)\Bigg]_{\angle}, \tag{15}$$

where $[\cdot]_{\angle}$ denotes the simplex constraint that $\phi_{vk} \geq 0$ and $\sum_{v=1}^{V} \phi_{vk} = 1$. More details on simulating (14) and (15) can be found in Examples 1-3 of Cong et al. (2017).

Similarly, with the gamma-Poisson construction in (7), we have $\Gamma_k(\boldsymbol{r}) = 1/M^{(L+1)}$, as in Appendix B, and

$$\nabla_{r_k}\left[-\tilde{H}(\boldsymbol{r})\right] = r_k^{-1}\left(\rho\tilde{x}_{k\cdot}^{(L+1)} + \frac{\gamma_0}{K_L} - 1\right) - \left(c_0 + \rho\tilde{q}_{\cdot}^{(L+1)}\right), \tag{16}$$

which also becomes unstable if r_k approaches zero. Substituting (16) and (11) into (4) leads to

$$\boldsymbol{r}_{t+1} = \Bigg|\boldsymbol{r}_t + \frac{\varepsilon_t}{M^{(L+1)}}\left[\left(\rho\tilde{\boldsymbol{x}}_{:\cdot}^{(L+1)} + \frac{\gamma_0}{K_L}\right) - \boldsymbol{r}_t\left(c_0 + \rho\tilde{q}_{\cdot}^{(L+1)}\right)\right] + \mathcal{N}\left(\mathbf{0}, \frac{2\varepsilon_t}{M^{(L+1)}}\mathrm{diag}\,(\boldsymbol{r}_t)\right)\Bigg|, \tag{17}$$

for which there is no stability issue.

3.3. Topic-Layer-Adaptive Stochastic Gradient Riemannian MCMC

Note that $M^{(L+1)}$ and $M_k^{(l)}$ for $l \in \{1, \ldots, L\}$, appearing as denominates in (17) and (15), respectively, are expectations that need to be approximately calculated. We update them using annealed weighting (Polatkan et al., 2015) as

$$M_k^{(l)} = \left(1 - \varepsilon_t'\right) M_k^{(l)} + \varepsilon_t' \rho E\left[\tilde{x}_{\cdot k\cdot}^{(l)}\right], \tag{18}$$

$$M^{(L+1)} = \left(1 - \varepsilon_t'\right) M^{(L+1)} + \varepsilon_t' \rho E\left[\tilde{q}_{\cdot}^{(L+1)}\right], \tag{19}$$

where $E[\cdot]$ denotes averaging over the collected MCMC samples. For simplicity, we set $\varepsilon_t' = \varepsilon_t$ in this paper, which is found to work well in practice.

Algorithm 1 TLASGR MCMC for DLDA (PGBN).

Input: Data mini-batches;
Output: Global parameters of DLDA (PGBN).
1: **for** $t = 1, 2, \cdots$ **do**
2: /* *Collect local information*
3: Upward-downward Gibbs sampling (Zhou et al., 2016a) on the t^{th} mini-batch for $\tilde{\boldsymbol{x}}_{:k\cdot}$, $\tilde{x}_{\cdot k\cdot}$, $\tilde{\boldsymbol{x}}_{:\cdot}^{(L+1)}$, and $\tilde{q}_{\cdot}^{(L+1)}$;
4: /* *Update global parameters*
5: **for** $l = 1, \cdots, L$ and $k = 1, \cdots, K_l$ **do**
6: Update $M_k^{(l)}$ with (18); then $\boldsymbol{\phi}_k^{(l)}$ with (15);
7: **end for**
8: Update $M^{(L+1)}$ with (19) and then $\boldsymbol{r}$ with (17).
9: **end for**

Note that as in (15) and (17), instead of having a single learning rate for all layers and topics, a common practice due to the difficulty to adapt the step sizes to different layers and/or topics, the proposed inference employs topic-layer-adaptive learning rates as $\varepsilon_t/M_k^{(l)}$, where $M_k^{(L+1)} := M^{(L+1)}$, adapting a single step size ε_t to different topics and layers by multiplying it with the weights $1/M_k^{(l)}$ for $l \in \{1, \ldots, L\}$ and $k \in \{1, \ldots, K_l\}$. We refer to the proposed inference algorithm with adaptive learning rates as topic-layer-adaptive stochastic gradient Riemannian (TLASGR) MCMC, as summarized in Algorithm 1 that is simple to implement.

4. Related Work

Both LDA (Blei et al., 2003) and the related Poisson factor analysis (PFA) (Zhou et al., 2012) are equipped with scalable inference, such as stochastic variational inference (SVI) (Hoffman et al., 2010; Mimno et al., 2012) and SGRLD (Patterson & Teh, 2013). However, both are shallow LVMs whose modeling capacities are often insufficient for big and complex data. The deep Poisson factor models of Gan et al. (2015) and Henao et al. (2015) are proposed to generalize PFA with deep structures, but both of them only explore the deep information in binary topic usage patterns instead of the full connection weights that are used in the PGBN. The proposed DLDA shares some similarities with the pachinko allocation model of Li & McCallum (2006) in that they both adopt layered construction and use Dirichlet distributed topics. Ranganath et al. (2015) propose deep exponential family (DEF), which differs from the PGBN in connecting adjacent layers via the gamma rate parameters and using black-box variational inference (BBVI) (Ranganath et al., 2014).

Some commonly used neural networks, such as deep belief network (DBN) (Hinton et al., 2006) and deep Boltzmann machines (DBM) (Salakhutdinov & Hinton, 2009), have also been modified for text analysis (Hinton & Salakhutdinov, 2009; Larochelle & Lauly, 2012; Srivastava et al., 2013). Although they may work well for certain text analysis tasks, they are not naturally designed for count data and often yield latent structures that are not readily interpretable.

The neural variational document model (NVDM) of Miao et al. (2016), even though using deep neural networks in its variational auto-encoder (VAE) (Kingma & Welling, 2013), still relies on a single-layer model for data generalization.

Generally speaking, it is challenging to develop an efficient and principled multilayer joint learning algorithm for deep LVMs. Scalable variational inference, such as BBVI, often makes the restrictive mean-field assumption. Neural variational inference and learning (NVIL) relies on variance reduction techniques that are often difficult to be generalized for discrete LVMs (Mnih & Gregor, 2014; Rezende et al., 2014). When a SG-MCMC algorithm is used, a single learning rate is often applied for different variables across all layers (Welling & Teh, 2011; Neal et al., 2011; Chen et al., 2014; Ding et al., 2014). It is possible to improve SG-MCMC by adjusting its noisy gradients with some stochastic optimization technique, such as Adagrad (Duchi et al., 2011), Adadelta (Zeiler, 2012), Adam (Kingma & Ba, 2014), and RMSprop (Tieleman & Hinton, 2012). For example, Li et al. (2016) show that preconditioning the gradients with diagonal approximated FIM improves SG-MCMC in both training speed and predictive accuracy for supervised learning where gradients are easy to calculate. Other efforts exploiting similar preconditioning idea focus on shallow and/or binary models (Mimno et al., 2012; Patterson & Teh, 2013; Grosse & Salakhutdinov, 2015; Song et al., 2016; Simsekli et al., 2016), and it is unclear how that idea can be extended to deep LVMs whose gradients and FIM maybe difficult to approximate.

5. Experiment results

We present experimental results on three benchmark corpora: 20Newsgroups (20News), Reuters Corpus Volume I (RCV1) that is moderately large, and Wikipedia (Wiki) that is huge. 20News consists of 18,845 documents with a vocabulary size of 2,000, partitioned into 11,315 training documents and 7,531 test ones. RCV1 consists of 804,414 documents with a vocabulary size of 10,000, where 10,000 documents are randomly selected for testing. Wiki consists of 10 million documents randomly downloaded from *Wikipedia* using the scripts provided in Hoffman et al. (2010); as in Hoffman et al. (2010), Gan et al. (2015), and Henao et al. (2015), we use a vocabulary with 7,702 words and randomly select 1,000 documents for testing. To make a fair comparison, these corpora, including the training/testing partitions, are set to be the same as those in Gan et al. (2015) and Henao et al. (2015). To be consistent with the settings of Gan et al. (2015) and Henao et al. (2015), no precautions are taken in the scripts for *Wikipedia* to prevent a testing document from being downloaded into a mini-batch for training.

We consider two related performance measures. The first one is the commonly-used per-heldout-word perplexity calculated as follows: for each test document, we randomly select 80% of the word tokens to sample the local variables specific to the document, under the global model parameters of each MCMC iteration; after the burn-in period, we accumulate the first layer's Poisson rates in each collected MCMC sample; in the end, we normalize these accumulated Poisson rates to calculate the perplexity using the remaining 20% word tokens. Similar evaluation methods have been widely used, $e.g.$, in Wallach et al. (2009), Paisley et al. (2011), and Zhou & Carin (2015). Although a good measure for overall performance, the per-heldout-word perplexity, calculated based on multiple collected MCMC samples of global parameters, may not be ideal to check the performance in real time to assess how efficient an iterative algorithm improves its performance as time increases. Therefore, we slightly modify it to provide a *point* per-heldout-word perplexity calculated based on only the global parameters of the most recent MCMC sample. For simplicity, we refer to (point) per-heldout-word perplexity as (point) perplexity.

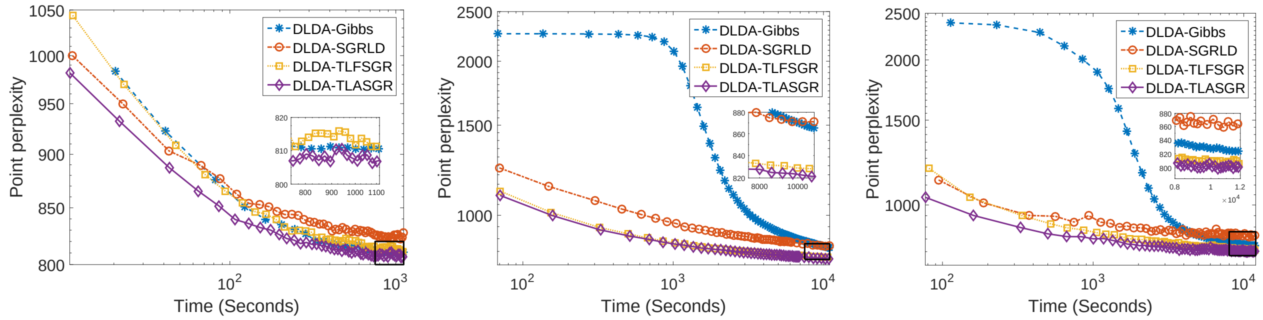
For comparison, we consider LDA of Blei et al. (2003), focused topic model (FTM) of Williamson et al. (2010), replicated softmax (RSM) of Hinton & Salakhutdinov (2009), nested Hierarchical Dirichlet process (nHDP) of Paisley et al. (2015), DPFA of Gan et al. (2015), and DPFM of Henao et al. (2015). For these methods, the perplexity results are taken from Gan et al. (2015) and Henao et al. (2015). For the proposed algorithms, we set the mini-batch size as 200, and use as burn-in 2000 mini-batches for both 20News and RCV1 and 3500 mini-batches for Wiki. We collect 1500 samples to calculate perplexity. For point perplexity, given the global parameters of an MCMC sample, we sample the local variables with 600 iterations and collect one sample every two iterations during the last 400 iterations. The hyperparameters of DLDA are set as: $\eta^{(l)} = 1/K_l$, $a_0 = b_0 = 0.01$, and $\gamma_0 = c_0 = e_0 = f_0 = 1$. Note $\eta^{(l)}$ and K_l are set similar to that of DPFM for fair comparisons, while other hyperparameters follow Zhou et al. (2016a).

To demonstrate the advantages of using the reduced-mean simplex parameterization and inverting the FIM for preconditioning to obtain topic-layer-adaptive learning rates, we consider four different inference methods:

1) TLASGR: topic-layer-adaptive stochastic gradient Riemannian MCMC for DLDA, as described in Algorithm 1.

2) TLFSGR: topic-layer-fixed stochastic gradient Riemannian MCMC for DLDA that replaces the adaptive learning rates $\varepsilon_t/M_k^{(l)}$ of TLASGR with $\varepsilon_t/(\sum_{k=1}^{K_1} M_k^{(1)}/K_1)$.

3) SGRLD: updating each $\phi_k^{(l)}$ under the expanded-mean parameterization as in (5), served as a good scalable baseline for comparison since it was shown in Patterson & Teh (2013) to perform significantly better than SVI. It differs from TLFSGR mainly in using a different parameterization for



(a) A single-layer DLDA on 20News (b) DLDA of size 128-64 on RCV1 (c) DLDA of size 128-64 on Wiki

Figure 1. Plot of point perplexity as a function of time. (a) 20News with a single-layer DLDA with 128 topics. (b) RCV1 with a two-layer DLDA with 128 and 64 topics in the first and second layers, respectively. (c) Wiki with a two-layer DLDA, with 128 and 64 topics in the first and second layers, respectively. Note a small subset of 10^6 documents from Wiki is used for demonstration.

Table 1. Per-heldout-word perplexities on 20 News, RCV1 and Wiki. For models except DLDA, the results are taken from Gan et al. (2015) and Henao et al. (2015). Note that for Wiki, DPFM with MCMC infers the global parameters on a subset of the corpus with 3,000 MCMC iterations.

Model	Method	Size	20 News	RCV1	Wiki
DLDA	TLASGR	128-64-32	757	815	**786**
DLDA	TLASGR	128-64	758	817	787
DLDA	TLASGR	128	770	823	802
DLDA	TLFSGR	128-64-32	760	817	789
DLDA	TLFSGR	128-64	759	819	791
DLDA	TLFSGR	128	772	829	804
DLDA	SGRLD	128-64-32	775	827	792
DLDA	SGRLD	128-64	770	823	792
DLDA	SGRLD	128	777	829	803
DLDA	Gibbs	128-64-32	**752**	**802**	—
DLDA	Gibbs	128-64	**754**	**804**	—
DLDA	Gibbs	128	768	818	—
DPFM	SVI	128-64	818	961	791
DPFM	MCMC	128-64	780	908	**783**
DPFA-SBN	SGNHT	128-64-32	827	1143	876
DPFA-RBM	SGNHT	128-64-32	896	920	942
nHDP	SVI	(10,10,5)	889	1041	932
LDA	Gibbs	128	893	1179	1059
FTM	Gibbs	128	887	1155	991
RSM	CD5	128	877	1171	1001

$\phi_k^{(l)}$ and adding a pseudolikelihood assumption.

4) Gibbs: the upward-downward Gibbs sampler in Zhou et al. (2016a).

Both TLASGR and TLFSGR differ from SGRLD mainly in how the global parameters $\phi_k^{(l)}$ are updated. While TLASGR uses topic-layer-adaptive learning rates, both TLFSGR and SGRLD use a single learning rate, a common practice due to the difficulty of tuning the step sizes across layers and topics. We keep the same stepsize schedule of $\varepsilon_t = a(1+t/b)^{-c}$ as in Patterson & Teh (2013) and Ma et al. (2015).

Let us first examine how various inference algorithms perform on 20News with a single-layer DLDA of size 128, which can be considered as a topic model that imposes an asymmetric prior on a document's proportion over these 128 topics. Under this setting, as shown in Fig. 1(a), TLFSGR clearly outperforms SGRLD in providing lower point perplexities as time progresses, which is not surprising as under the reduced-mean simplex parameterization, to derive its sampling equations, TLFSGR does not rely on a pseudolikelihood assumption that is adopted by SGRLD in its expanded-mean simplex parameterization. Moreover, TLASGR is found to further improve TLFSGR, suggesting that even for a single-layer model, replacing a fixed learning rate as $\varepsilon_t/(\sum_{k=1}^{K_1} M_k^{(1)}/K_1)$ with topic-adaptive learning rates as $\varepsilon_t/M_k^{(1)}$ could further improve the performance.

Let us then examine how these algorithms perform on two larger corpora—RCV1 and Wiki—with a two-layer DLDA of size 128-64, which improves the single-layer one by capturing the co-occurrence patterns between the topics of the first layer with those of the second layer (Zhou et al., 2016a). As show in Figs. 1(b) and 1(c), it is clear that the proposed TLASGR performs the best for the two-layer DLDA and consistently outperforms TLFSGR as time progresses. In comparison, SGRLD quickly improves its performance as a function of time in the beginning but its point perplexity remains higher even after 10,000 seconds, whereas Gibbs sampling slowly improves its performance as a function of time in the beginning but moves its point perplexity closer and closer to that of TLASGR as time progresses.

Note that for 20News, the point perplexity of the mini-batch based TLASGR quickly decreases as time increases, while that of Gibbs sampling decreases relatively slowly. That discrepancy of convergence rate as a function of time becomes much more evident for both RCV1 and Wiki, as shown in Figs. 1(b) and 1(c). This is expected as both RCV1 and Wiki are much larger corpora, for which a mini-batch based inference algorithm can already make significant progress in learning the global model parameters, before a batch-learning Gibbs sampler finishes a single iteration that needs to go through all documents.

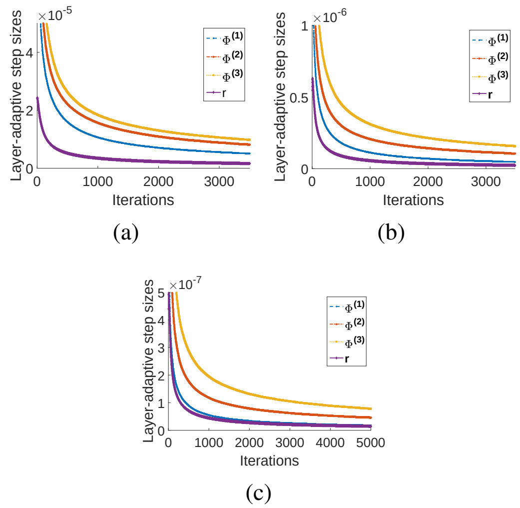


Figure 2. Topic-layer-adaptive learning rates inferred with a three-layer DLDA of size 128-64-32. (a) 20News. (b) RCV1. (c) Wiki. Note the layer-adaptive learning rate for layer l is obtained by averaging over the topic-layer-adaptive learning rates of all $\phi_k^{(l)}$ for $k = 1, \ldots, K_l$.

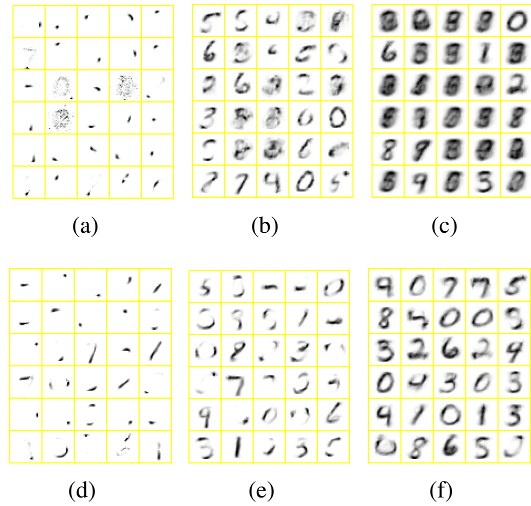


Figure 3. Learned dictionary atoms on MNIST digits with a three-layer GBN of size 128-64-32 after one full epoch. Shown in (a)-(c) are example atoms for $\phi_k^{(1)}$, $\mathbf{\Phi}^{(1)}\phi_k^{(2)}$, and $\mathbf{\Phi}^{(1)}\mathbf{\Phi}^{(2)}\phi_k^{(3)}$, respectively, learned with TLFSGR, and shown in (d)-(f) are example ones learned with TLASGR.

To illustrate the working mechanism of TLASGR, we show how its inferred learning rates are adapted to different layers in Fig. 2. By contrast, TLFSGR admits a fixed learning rate that leads to worse performance. Several interesting observations can be made for TLASGR from Figs. 2(a)-2(c): 1) for $\mathbf{\Phi}^{(l)}$, higher layers prefer larger step sizes, which may be explained by the *enlarge-partition-augment* data generating mechanism of DLDA; 2) larger datasets prefer slower learning rates, reflected by the scales of the vertical axes; 3) and the relative learning rates between different layers vary across different datasets.

To further verify the excellent performance of DLDA inferred with TLASGR, we compare a wide variety of models and inference algorithms in Table 1. For 20News and RCV1, DLDA with Gibbs sampling performs the best in terms of perplexity and exhibits a clear trend of improvement as the number of hidden layers increases. For Wiki, a single iteration of the DLDA Gibbs sampler on the full corpus is so expensive in both time and memory that its performance is not reported. For DLDA on 20News and RCV1, TLASGR only slightly underperforms Gibbs sampling, and the performance degradation from Gibbs sampling to TLASGR is significantly smaller than that from MCMC to SVI for DPFM. That relative small degradation caused by changing from Gibbs sampling to the mini-batch based TLASGR could be attributed to the Fisher efficiency brought by the FIM. Generally speaking, in comparison to SGRLD, TLASGR brings a clear boost in performance, which is particularly evident for a deeper DLDA, and TLASGR consistently outperforms TLFSGR that does not adapt its learning rates to different topics and layers.

MNIST. To further illustrate the advantages of using the inverse of the FIM for preconditioning in a deep generative model, and to visualize the benefits of automatically adjusting the relative learning rates of different hidden layers, we apply a three-layer Poisson randomized gamma gamma belief network (PRG-GBN) (Zhou et al., 2016a) to 60,000 MNIST digits and present the learned dictionary atoms after one full epoch, as shown in Fig. 3. It is clear that, with topic-layer-adaptive learning rates, which are made possible by utilizing the FIM, TLASGR provides more effective mini-batch based stochastic updates to allow better information propagation between different hidden layers, extracting more interpretable features at multiple layers.

6. Conclusions

For scalable multilayer joint inference of the Poisson gamma belief network (PGBN), we introduce an alternative representation of the PGBN, which is referred to as deep latent Dirichlet allocation (DLDA) that can be considered as a multilayer generalization of latent Dirichlet allocation. We show how to reparameterize the simplex constrained basis vectors, derive a block-diagonal Fisher information matrix (FIM), and efficiently compute the inverse of the FIM, leading to a stochastic gradient MCMC algorithm referred to as topic-layer-adaptive stochastic gradient Riemannian (TLASGR) MCMC. The proposed TLASGR-MCMC is able to jointly learn the parameters of different layers with topic-layer-adaptive step sizes, which makes DLDA (PGBN) much more practical in a big data setting. Compelling experimental results on large text corpora and the MNIST dataset demonstrated the advantages of TLASGR-MCMC.

Acknowledgements

Bo Chen thanks the support of the Thousand Young Talent Program of China, NSFC (61372132), and NDPR-9140A07010115DZ01019. Hongwei Liu thanks the support of NSFC for Distinguished Young Scholars (61525105).

References

Amari, S. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. Greedy layer-wise training of deep networks. In *NIPS*, pp. 153–160, 2007.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

Chen, T., Fox, E. B., and Guestrin, C. Stochastic gradient Hamiltonian Monte Carlo. In *ICML*, pp. 1683–1691, 2014.

Cong, Y., Chen, B., and Zhou, M. Fast simulation of hyperplane-truncated multivariate normal distributions. *Bayesian Analysis Advance Publication*, 2017.

Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D., and Neven, H. Bayesian sampling using stochastic gradient thermostats. In *NIPS*, pp. 3203–3211, 2014.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12(Jul):2121–2159, 2011.

Gan, Z., Chen, C., Henao, R., Carlson, D., and Carin, L. Scalable deep Poisson factor analysis for topic modeling. In *ICML*, pp. 1823–1832, 2015.

Girolami, M. and Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *JRSS-B*, 73(2):123–214, 2011.

Golub, Gene H and Van Loan, Charles F. *Matrix computations*, volume 3. JHU Press, 2012.

Grosse, R. B. and Salakhutdinov, R. Scaling up natural gradient by sparsely factorizing the inverse Fisher matrix. In *ICML*, pp. 2304–2313, 2015.

Henao, R., Gan, Z., Lu, J., and Carin, L. Deep Poisson factor modeling. In *NIPS*, pp. 2782–2790, 2015.

Hinton, G. E. and Salakhutdinov, R. R. Replicated softmax: an undirected topic model. In *NIPS*, pp. 1607–1614, 2009.

Hinton, G. E., Osindero, S., and Teh, Y. W. A fast learning algorithm for deep belief nets. *Neural Computation*, 18 (7):1527–1554, 2006.

Hoffman, M. D., Bach, F. R., and Blei, D. M. Online learning for latent Dirichlet allocation. In *NIPS*, pp. 856–864, 2010.

Johnson, N. L., Kotz, S., and Balakrishnan, N. *Discrete Multivariate Distributions*, volume 165. Wiley New York, 1997.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Kingma, Diederik P and Welling, Max. Auto-encoding variational Bayes. In *ICLR*, number 2014, 2013.

Larochelle, H. and Lauly, S. A neural autoregressive topic model. In *NIPS*, 2012.

Li, C., Chen, C., Carlson, D., and Carin, L. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. *AAAI*, 2016.

Li, W. and McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, pp. 577–584, 2006.

Ma, Y., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *NIPS*, pp. 2899–2907, 2015.

Miao, Y., Yu, L., and Blunsom, P. Neural variational inference for text processing. In *ICML*, 2016.

Mimno, D., Hoffman, M. D., and Blei, D. M. Sparse stochastic inference for latent Dirichlet allocation. In *ICML*, pp. 362 – 365, 2012.

Mnih, A. and Gregor, K. Neural variational inference and learning in belief networks. 2014.

Neal, R. M. et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

Paisley, J., Wang, C., and Blei, D. The discrete infinite logistic normal distribution for mixed-membership modeling. In *AISTATS*, 2011.

Paisley, J., Wang, C., Blei, D. M., and Jordan, M. I. Nested hierarchical dirichlet processes. *PAMI*, 37(2):256–270, 2015.

Pascanu, R. and Bengio, Y. Revisiting natural gradient for deep networks. *arXiv:1301.3584*, 2013.

Patterson, S. and Teh, Y. W. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *NIPS*, pp. 3102–3110, 2013.

Polatkan, G., Zhou, M., Carin, L., Blei, D., and Daubechies, I. A Bayesian nonparametric approach to image super-resolution. *PAMI*, 37(2):346–358, 2015.

Ranganath, R., Gerrish, S., and Blei, D. M. Black box variational inference. In *AISTATS*, 2014.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. Deep exponential families. In *AISTATS*, 2015.

Rezende, Danilo J, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pp. 1278–1286, 2014.

Salakhutdinov, R. and Hinton, G. E. Deep Boltzmann machines. In *AISTATS*, volume 1, pp. 3, 2009.

Schein, A., Zhou, M., and Wallach, H. Poisson–gamma dynamical systems. In *NIPS*, pp. 5006–5014, 2016.

Simsekli, U., Badeau, R., Cemgil, A. T., and G., Richard. Stochastic quasi-Newton Langevin Monte Carlo. In *ICML*, 2016.

Song, Z., Henao, R., Carlson, D., and Carin, L. Learning sigmoid belief networks via Monte Carlo expectation maximization. In *AISTATS*, pp. 1347–1355, 2016.

Srivastava, N., Salakhutdinov, R. R., and Hinton, G. E. Modeling documents with deep Boltzmann machines. In *UAI*, 2013.

Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2), 2012.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. Evaluation methods for topic models. In *ICML*, 2009.

Welling, M. and Teh, Y.-W. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, pp. 681–688, 2011.

Williamson, S., Wang, C., Heller, K. A., and Blei, D. M. The IBP compound Dirichlet process and its application to focused topic modeling. In *ICML*, pp. 1151–1158, 2010.

Zeiler, M. D. Adadelta: an adaptive learning rate method. *arXiv:1212.5701*, 2012.

Zhou, M. and Carin, L. Negative binomial process count and mixture modeling. *PAMI*, 37(2):307–320, 2015.

Zhou, M., Hannah, L., Dunson, D. B., and Carin, L. Beta-negative binomial process and Poisson factor analysis. In *AISTATS*, pp. 1462–1471, 2012.

Zhou, M., Cong, Y., and Chen, B. Augmentable gamma belief networks. *Journal of Machine Learning Research*, 17(163):1–44, 2016a.

Zhou, M., Padilla, O., and Scott, J. G. Priors for random count matrices derived from a family of negative binomial processes. *J. Amer. Statist. Assoc.*, 111(515):1144–1156, 2016b.

Supplementary Material for Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC

Yulai Cong, Bo Chen, Hongwei Liu, and Mingyuan Zhou

A. Naive derivation of the Fisher information matrix of the Poisson gamma belief network

For simplicity, we take for example a two-layer Poisson gamma belief network (PGBN), expressed as

$$\begin{gathered}\boldsymbol{\theta}_j^{(2)} \sim \operatorname{Gam}\left(\boldsymbol{r}, 1/c_j^{(3)}\right), \\ \boldsymbol{x}_j^{(1)} \sim \operatorname{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right), \boldsymbol{\theta}_j^{(1)} \sim \operatorname{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, \tfrac{p_j^{(2)}}{1-p_j^{(2)}}\right),\end{gathered} \tag{20}$$

and focus on a specific element $\boldsymbol{\Phi}_{vk}^{(2)}$ only.

With the definition in (8), it is straight to show that the $\boldsymbol{\Phi}^{(2)}$-relevant part in $\ln p\left(\boldsymbol{\Omega}\,|\boldsymbol{z}\right)$ is

$$\sum\nolimits_{vj}\left[\boldsymbol{\Phi}_{v:}^{(2)}\boldsymbol{\theta}_{:j}^{(2)}\ln\left(c_j^{(2)}\boldsymbol{\theta}_{vj}^{(1)}\right) - \ln\Gamma\left(\boldsymbol{\Phi}_{v:}^{(2)}\boldsymbol{\theta}_{:j}^{(2)}\right)\right]. \tag{21}$$

Accordingly, for $\boldsymbol{\Phi}_{vk}^{(2)}$, we have

$$\mathbb{E}\left[-\frac{\partial^2}{\partial[\boldsymbol{\Phi}_{vk}^{(2)}]^2}\ln p\left(\boldsymbol{\Omega}\,|\boldsymbol{z}\right)\right] = \mathbb{E}\left[\sum_j \psi'\left(\boldsymbol{\Phi}_{v:}^{(2)}\boldsymbol{\theta}_{:j}^{(2)}\right)\left[\boldsymbol{\theta}_{:j}^{(2)}\right]^2\right], \tag{22}$$

where $\psi'(\cdot)$ is the trigamma function. This expectation involving the trigamma function is difficult to calculate.

B. Derivation of the $\Gamma(\cdot)$ functions in Section 3.2

With $\mathbf{D}(\boldsymbol{z}) = \mathbf{G}(\boldsymbol{z})^{-1}, \mathbf{Q}(\boldsymbol{z}) = \mathbf{0}$, and the block-diagonal Fisher information matrix (FIM) $\mathbf{G}(\boldsymbol{z})$ in (9), it is straight to show that $\frac{\partial}{\partial\boldsymbol{\varphi}_k}\left[\mathbf{D}(\boldsymbol{z}) + \mathbf{Q}(\boldsymbol{z})\right]$ is non-zero only in the $\boldsymbol{\varphi}_k$-related block $\mathbf{I}(\boldsymbol{\varphi}_k)$ in (10). Therefore, we focus on this block and have

$$\Gamma_v(\boldsymbol{\varphi}_k) = \sum\nolimits_u \frac{\partial}{\partial\varphi_{uk}}\left[\mathbf{I}_{vu}^{-1}(\boldsymbol{\varphi}_k)\right], \tag{23}$$

where $\mathbf{I}^{-1}(\boldsymbol{\varphi}_k) = M_k^{-1}\left[\operatorname{diag}(\boldsymbol{\varphi}) - \boldsymbol{\varphi}\boldsymbol{\varphi}^T\right]$. Accordingly, we have

$$\begin{aligned}\Gamma_v(\boldsymbol{\varphi}_k) &= M_k^{-1}\sum\nolimits_u \frac{\partial}{\partial\varphi_{uk}}\left[\delta_{u=v}\varphi_{uk} - \varphi_{vk}\varphi_{uk}\right] \\ &= M_k^{-1}(1 - V\varphi_{vk}).\end{aligned} \tag{24}$$

Since $\mathbf{G}(\boldsymbol{z})$ is block-diagonal with its $\boldsymbol{r}$-relevant block being $\mathbf{I}(\boldsymbol{r}) = M^{(L+1)}\operatorname{diag}(1/\boldsymbol{r})$, according to (3), it is straightforward to show that

$$\begin{aligned}\Gamma_k(\boldsymbol{r}) &= \sum\nolimits_u \frac{\partial}{\partial r_u}\left[\mathbf{I}_{ku}^{-1}(\boldsymbol{r})\right], \\ &= \sum\nolimits_u \frac{\partial}{\partial r_u}\left[\delta_{u=k}\frac{r_u}{M^{(L+1)}}\right], \\ &= 1/M^{(L+1)}.\end{aligned} \tag{25}$$

C. Proof of Lemma 3.1

Note that the counts in $x_{vj}^{(l)} \sim \operatorname{Pois}\big(q_j^{(l)}\sum_{k=1}^{K_l}\phi_{vk}^{(l)}\theta_{kj}^{(l)}\big)$ can be augmented as

$$\begin{aligned}x_{vj}^{(l)} &= \sum\nolimits_{k=1}^{K_l} x_{vkj}^{(l)}, \\ x_{vkj}^{(l)} &\sim \operatorname{Pois}\big(q_j^{(l)}\phi_{vk}^{(l)}\theta_{kj}^{(l)}\big),\end{aligned} \tag{26}$$

which, according to Lemma 4.1 of Zhou et al. (2012), can be equivalently expressed as

$$\begin{aligned}\left(x_{vkj}^{(l)}\right)_v &\sim \operatorname{Mult}\left(m_{kj}^{(l)(l+1)}, \boldsymbol{\phi}_k^{(l)}\right), \\ m_{kj}^{(l)(l+1)} &\sim \operatorname{Pois}\left(q_j^{(l)}\theta_{kj}^{(l)}\right),\end{aligned} \tag{27}$$

where $m_{kj}^{(l)(l+1)} := \sum_{v=1}^{K_{l-1}} x_{vkj}^{(l)}$. Marginalizing out $\theta_{vj}^{(l)} \sim \operatorname{Gam}\left(\sum_{k=1}^{K_{l+1}}\boldsymbol{\phi}_{vk}^{(l+1)}\boldsymbol{\theta}_{kj}^{(l+1)}, 1/c_j^{(l+1)}\right)$ from (27) leads to

$$m_{vj}^{(l)(l+1)} \sim \operatorname{NB}\left(\sum\nolimits_{k=1}^{K_{l+1}}\boldsymbol{\phi}_{vk}^{(l+1)}\boldsymbol{\theta}_{kj}^{(l+1)},\ p_j^{(l+1)}\right), \tag{28}$$

which can be augmented as

$$\begin{aligned}m_{vj}^{(l)(l+1)} &\sim \operatorname{SumLog}(x_{vj}^{(l+1)}, p_j^{(l+1)}), \\ x_{vj}^{(l+1)} &\sim \operatorname{Pois}\left(q_j^{(l+1)}\sum\nolimits_{k=1}^{K_{l+1}}\boldsymbol{\phi}_{vk}^{(l+1)}\boldsymbol{\theta}_{kj}^{(l+1)}\right).\end{aligned} \tag{29}$$

When $l = L$, we have

$$m_{kj}^{(L)(L+1)} \sim \operatorname{NB}(r_k, p_j^{(L+1)}), \tag{30}$$

marginalizing the gamma process $G \sim \operatorname{GaP}(G_0, 1/c_0)$ from which leads to a gamma-negative binomial process random count matrix, as expressed in the first two lines of (6).

D. Corollary D.1

Corollary D.1. *The gamma-negative binomial process PFA can be equivalently expressed as*

$$\begin{aligned}\ell_{k\cdot} &\sim \operatorname{Log}\left(\frac{q_\cdot}{c_0 + q_\cdot}\right),\ K \sim \operatorname{Pois}\left(\gamma_0\ln\frac{c_0 + q_\cdot}{c_0}\right), \\ \mathcal{L} &= \sum\nolimits_{k=1}^{K}\ell_{k\cdot}\delta_{\boldsymbol{\phi}_k}, \\ \left(\ell_{kj}\right)_j &\sim \operatorname{Mult}\left[\ell_{k\cdot}, \left(q_j\right)_j/q_\cdot\right], \\ m_{kj} &\sim \operatorname{SumLog}(\ell_{kj}, p_j) \\ x_{vj} &= \sum\nolimits_{k=1}^{K} x_{vkj},\ \left(x_{vkj}\right)_v \sim \operatorname{Mult}\left(m_{kj}, \boldsymbol{\phi}_k\right).\end{aligned} \tag{31}$$

E. Visualizations of the extracted topics and networks

In the following, we provide some example results, obtained using DLDA where $[K_1, K_2, K_3] = [128, 64, 32]$ and $\eta^{(l)} = 1/K_l$ for the lth layer, on extracting multilayer representations/topics from both the RCV1 and Wiki corpora. Clearly interpretable results, which are similar to those reported in Zhou et al. (2016a) and hence omitted here for brevity, are also extracted from the 20Newsgroups corpus.

E.1. RCV1

Following the visualization techniques in Zhou et al. (2016a), we plot 54 example topics of layer one in Figure 4, the top 30 topics of layer two in Figure 5, and the top 30 topics of layer three in Figure 6. Figure 4 clearly shows that the topics of layer one are very specific. For example, topics 41, 71 and 62 in the first row are about "Germany," "Polish," and "France," respectively; topics 53 and 54 in the second row are about "airline" and "European union," respectively; and topics 85 and 36 in the third row are about "ship & island" and "comput & techn," respectively. By contrast, the topics of layers two and three, shown in Figures 5 and 6, respectively, are increasingly more general. Such topics can be better interpreted via the following informative tree structured visualizations. Note that a tree defined in this paper allows a child node of a layer to be connected to more than one parent node of the adjacent higher layer.

Shown in Figure 7 is a [10, 3, 1] tree rooted at node 4 of the top layer on "bonds, rates, & credit markets." Apparently, the topics become more and more specific when moving from top to bottom following the branches. For example, the root node splits into three nodes from layers three to two, which focus differently on "treasury bill," "dollar rate," and "bond, credit, & debt," respectively. When moving from layers two to one, all three topics in layer two splits into multiple ones that is clearly more specific. For example, topics 1, 17, and 87 are about "months," "loan & credit," and "bond & pay," respectively.

Shown in Figure 8 is another analogous tree rooted at node 24 of layer three. It is clear that, as the nodes of this tree, topics 55, 38, 34, and 30 of layer two are mainly about "Germany," "France," "airline," and "labor union," respectively. Moreover, these four topics at layer two are all connected to topic 8 of layer one, which is very specific on "office meetings."

To understand the relationships and distinctions between different trees, we construct subnetworks as shown in Figures 9-10. Figure 9 clearly shows that all three trees, rooted at nodes 16, 10, and 17 of layer three, respectively, are highly related to topic 3 of layer two on "low & expect". However, the two trees rooted at node 10 and 17, respectively, both have their own specificities. For example, topic 52 of layer two on "wall street," is unique to node 10 of layer three, and topic 35 of layer two on "India" is unique to node 17 of layer three. Similar phenomena can also be observed from another subnetwork on "China," shown in Figure 10, where both nodes of the top layer are connected to topic 19 of layer two on "corp & techn," topic 36 on "China," and topic 12 on "profit & expect." Though related to each other, the tree rooted at node 18 of the top layer is also strongly connected to topic 31 on "project" and topic 34 on "airline," whereas the other one focuses differently on topic 49 on "car & Korea" and topic 44 on "growth rate".

E.2. Wiki

What follows are analogous figures illustrating some interpretable topics extracted from Wiki.

Figures 11-13 show the top example topics at layers one, two, and three, respectively. It is obvious that topics of layer one are specific, such as topic 31 on "university & research," topic 72 on "news, magazine, & times," topic 83 on "military & army," topic 74 on "police, crime & prison," topic 75 on "birds & species," topic 36 on "British & England," and so on. By contrast, when going to higher layers, topics become more general, as shown in Figures 12 and 13. To better illustrate topics of higher layers, we explicitly show their hierarchical structures via the following trees and subnetworks.

Figure 14 shows a tree rooted at node 1 of layer three on "music & song," whose topics at layer two are about "song & band" and "music, piano, & theatre," respectively. Figure 15 demonstrates another tree consisting of topic 9 of layer two on "London & British," topic 50 on "church & Catholic," and topic 25 on "king & prince," which is mainly about "United Kingdom." Given in Figure 16 is another tree on "art & museum," where the left side is about "art" while the right is on "history & building." These trees are all clearly interpretable.

In the subnetwork shown in Figure 17, all three trees are related to topic 9 of layer two on "London, British, & Sir." But they focus differently on topic 16 of layer two on "Irish Americans," topic 24 on "life, birth, education, career, family, & death," and topic 25 on "king & prince," respectively. Similar phenomena can also be observed in Figure 18. Both trees are related to topic 52 of layer two on "ship" and topic 49 on "air," but the left one is about various means of transportation and communication while the right one is about various components of "war." Figure 19 shows another subnetwork on "team & race," where three trees, all include topic 6 of layer two, focus differently on "goals, clubs, & league," "world cup," and "rank, first, second, & third," respectively.

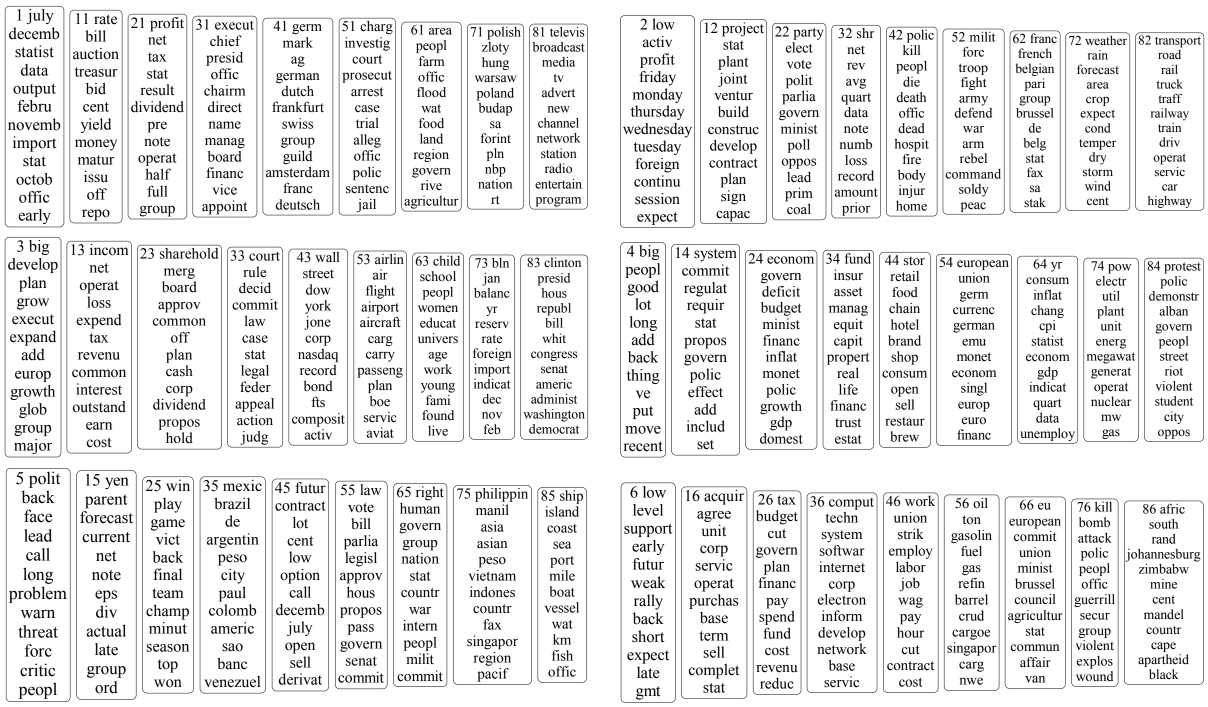


Figure 4. Example topics of layer one of DLDA trained with TLASGR MCMC on RCV1.

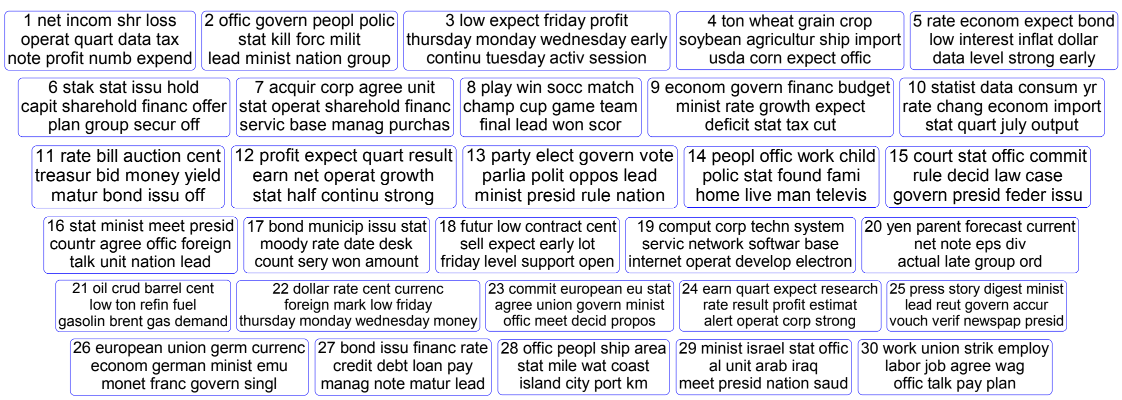


Figure 5. The top 30 topics of layer two of DLDA trained with TLASGR MCMC on RCV1.

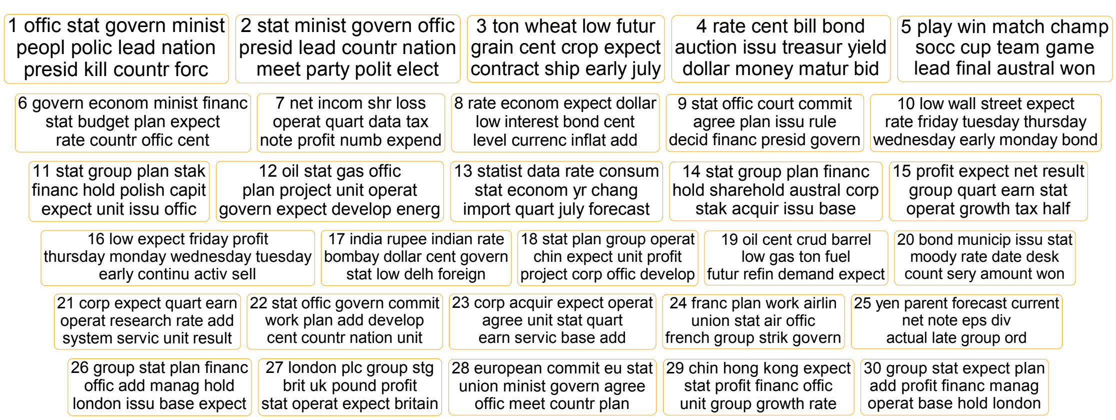


Figure 6. The top 30 topics of layer three of DLDA trained with TLASGR MCMC on RCV1.

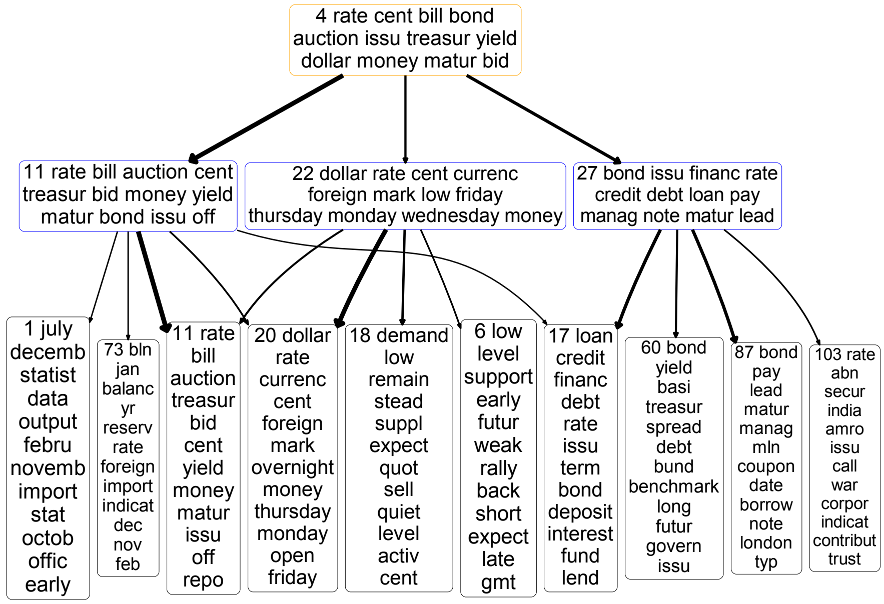


Figure 7. A $[10, 3, 1]$ tree on "bonds, rates, & credit markets" that includes all the lower-layer nodes (directly or indirectly) linked with non-negligible weights to node 4 of the top layer, taken from the full $[128, 64, 32]$ DLDA network trained with TLASGR MCMC on the 794,414 training documents of the RCV1 corpus, with $\eta^{(l)} = 1/K_l$ for the lth layer. A line from node k at layer l to node k' at layer $l-1$ indicates that $\mathbf{\Phi}^{(l)}(k', k) > 5/K_{l-1}$, with the width of the line proportional to $\sqrt{\mathbf{\Phi}^{(l)}(k', k)}$. For each node, the rank (in terms of popularity) at the corresponding layer and the top 12 words of the corresponding topic are displayed inside the text box, where the text font size monotonically decreases as the popularity of the node decreases, and the outside border of the text box is colored as orange, blue, or black if the node is at layer three, two, or one, respectively.

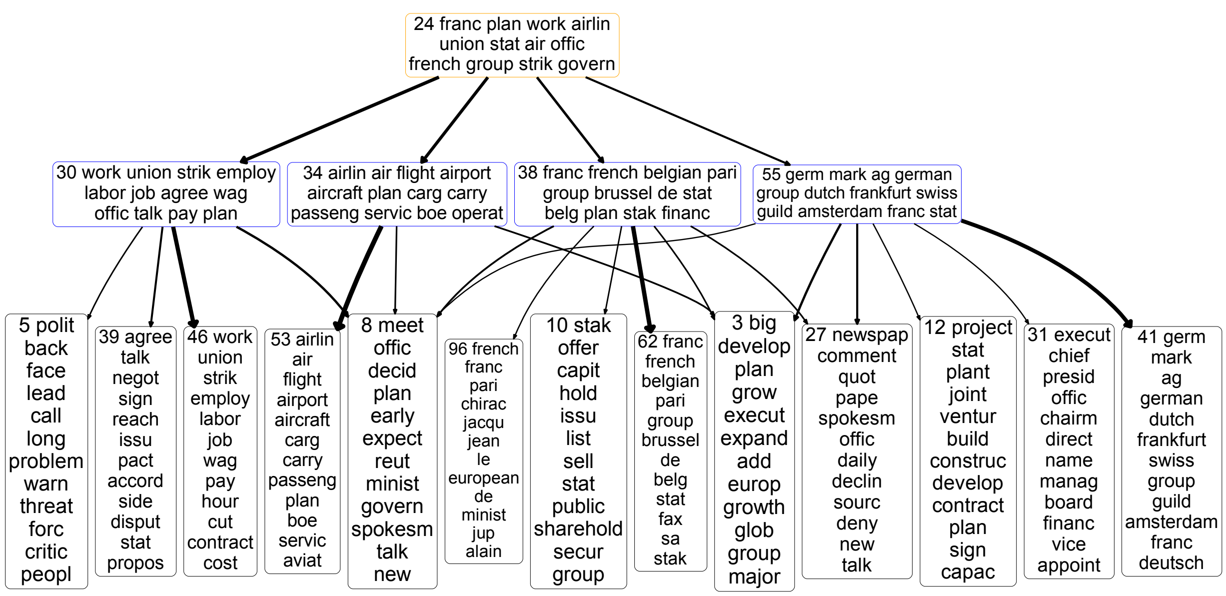


Figure 8. Analogous plots to Figure 7 for a tree rooted at node 24 on "France, Germany, & airline" from RCV1.

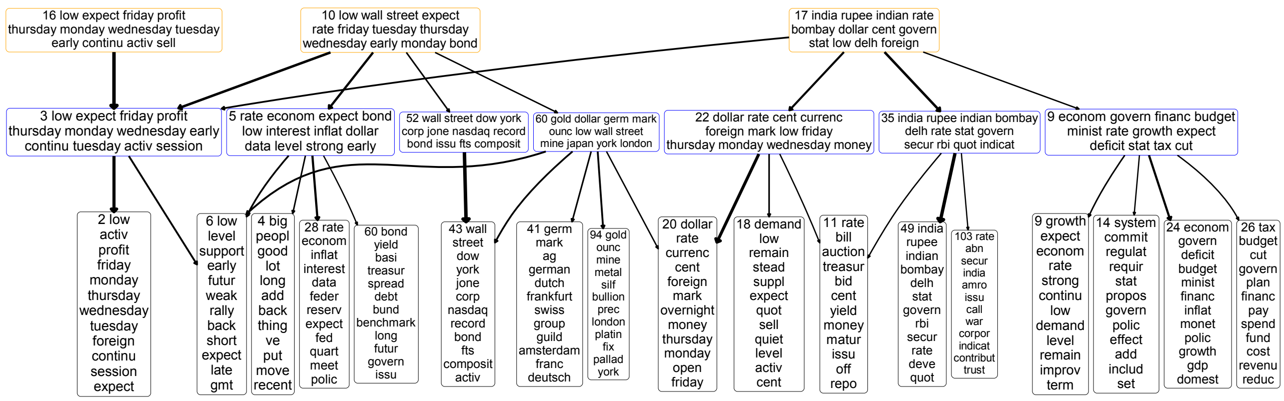


Figure 9. Analogous plots to Figure 7 for a subnetwork related to “low & expect” from RCV1, consisting of three trees rooted at nodes 16, 10 and 17, respectively, of layer three.

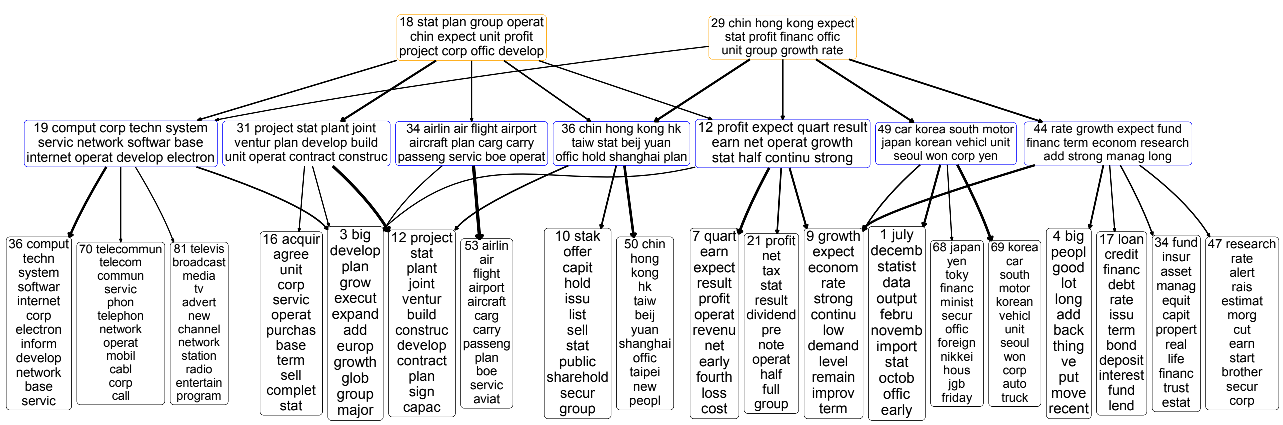


Figure 10. Analogous plots to Figure 7 for a subnetwork on “China” from RCV1, consisting of two trees rooted at nodes 18 and 29, respectively, of layer three.

1 district province village population rural county central department region families located town
11 john david paul michael peter tom jones mark jack joe martin bill
21 back get tells time takes help finds plot gets goes friends home
31 university research professor science society journal studies international press history scientific sciences
41 party election liberal elections conservative parliament labour elected assembly democratic member seats
51 art museum gallery arts work artist artists works painting exhibition design collection
61 library published book edition first books chapter text new history written collection
71 music piano musical songs folk major string minor concert performed instruments instrument
81 age population average years income living families males females people older size
2 born career life years school early worked university first year age occupation
12 time system different two level term common possible based process single number
22 born ireland american irish refer actor english dublin writer british people author
32 season league team played games game career home professional major year signed
42 best directed director actor cast production festival drama released language won produced
52 social theory human self people life world nature knowledge philosophy concept view
62 republican democratic elected governor state senate house district served election party representatives
72 news magazine times newspaper daily press editor published jan post oct nov
82 japan japanese china chinese asia lee han south peoples world republic east

3 time new first made years early two continued left due came end
13 march january november october december april february september june july august accessed
23 name known list first named names present called history see originally current
33 years high large low number population due early million period found formation
43 italy european italian van dutch netherlands germany france europe spain belgium greece
53 church catholic bishop saint roman pope churches cathedral holy parish john chapel
63 act state government federal public law constitution commission bill section laws authority
73 house historic national listed places built home added located property private story
83 military service army forces force navy naval training officer commander command operations
4 species family genus found plant tree described plants known leaves description marine
14 john william james george thomas robert charles smith henry richard joseph edward
24 said stated people times saying wrote told asked made called described just
34 new york city jersey boston massachusetts american times connecticut philadelphia pennsylvania washington
44 land people first native territory region colony settlement established river population history
54 university college students education student school science faculty arts degree program year
64 king prince kingdom emperor iii empire duke royal kings queen princess son
74 police crime prison death murder arrested trial killed found investigation criminal case
84 week day year festival held events event days annual sunday saturday ceremony

5 first two second three four third year time five six last fourth
15 development community program organization members public association international national project foundation work
25 century ancient history period early roman found known stone modern greek late
35 president minister general secretary office council government prime member appointed committee executive
45 island islands sea bay coast port ocean pacific north south tropical atlantic
55 war army division general regiment infantry corps brigade battle cavalry world artillery
65 south australia africa new wales african sydney victoria western cape coast dir
75 birds species fish bird large female small eggs male long common body
85 god christian religious bible jesus religion life faith spiritual gods spirit holy
6 river lake park mountain area valley forest national mountains located mount creek
16 north west east south poland county central eastern western northern village district
26 style like made popular known early called first famous fashion time modern
36 london british england royal scotland britain scottish english oxford great edinburgh cambridge
46 series show season guest appeared shows comedy voice played original character star
56 battle attack war forces killed army men soldiers troops captured force attacks
66 american washington america war african first history americans johnson president black civil
76 coach game season conference team head yards bowl yard field college record
86 bank million financial tax market money price income value economic billion cost

Figure 11. Example topics of layer one of DLDA trained with TLASGR MCMC on Wiki.

1 song released band music single track records songs first rock recorded live
2 species family genus found plant tree known described plants white leaves black
3 district village province county population rural north west central east south region
4 two system group time high number first water process different form energy
5 goals clubs years league club season cup team first played division career
6 score team first round final world won win points cup group second
7 world team race won cup time event mens place national summer rank
8 university research new college science professor american school born society studies york
9 london british john sir england william henry lord son died royal thomas
10 government political war president march state international national group october january party
11 united states act law government state public court new federal national first
12 city district town village population province area india county rural located indian
13 back best time new first get two family father life series tells
14 river lake park area mountain valley north south forest national water located
15 company business group new million companies inc market products first services corporation
16 born american ireland irish refer actor english john british dublin writer people
17 party election liberal member elected democratic president elections conservative minister council general
18 series best show new born director life directed actor first production theatre
19 university college school development community education national program public international association students
20 data system information software systems computer windows based code company available network
21 series show season first new week year day september march two october
22 social new people first book time theory language life human work world
23 city century district name town area population village history known province church
24 born years life career family school first son time died early married
25 king son prince kingdom emperor family iii empire name died first death
26 season league team played new games game career first home year york
27 published first book works german work university english french literature poetry press
28 war army battle general division regiment infantry military corps brigade world forces
29 engine car two first new cars air model design speed built ford
30 said time first two new made day police september october june july

Figure 12. The top 30 topics of layer two of DLDA trained with TLASGR MCMC on Wiki.

1 music song released band single first new songs records track rock recorded
2 county river north road united south park new city states area station
3 goals years clubs league club team season cup first played world division
4 ship air first two new station united line engine states car company
5 district village province county population rural north west central east south region
6 first series new time man two best game back character story book
7 city district area village town river population province north south county island
8 born american ireland irish john british party english refer actor william london
9 party government election president national state first new united political member minister
10 series best new show born first life director directed actor american john
11 war army ship air united new military first general battle states march
12 two first time name group known like called system new number made
13 team first world new won score time final two race cup points
14 city century town village area north castle church south west name district
15 first new church book university life published name century two time work
16 species family genus found plant known tree white new described plants name
17 new united states american state county university york democratic republican first john
18 new university book published american first books york born life work art
19 king church son first prince family kingdom died name century death emperor
20 system data new information united states based first company time software systems
21 university new united states company research college international school national first development
22 season team league game new played first games coach state year career
23 species family genus found plant known tree described new plants white south
24 company new united states business group first million york companies american based
25 team world race won cup time event place national mens first games
26 station network channel news series new show first stations fox time september
27 university new research born music first college school american science professor united
28 county age city population school united states town area years average american
29 school high schools students education university state district college public year new
30 art museum new work arts gallery works artist artists first painting design

Figure 13. The top 30 topics of layer three of DLDA trained with TLASGR MCMC on Wiki.

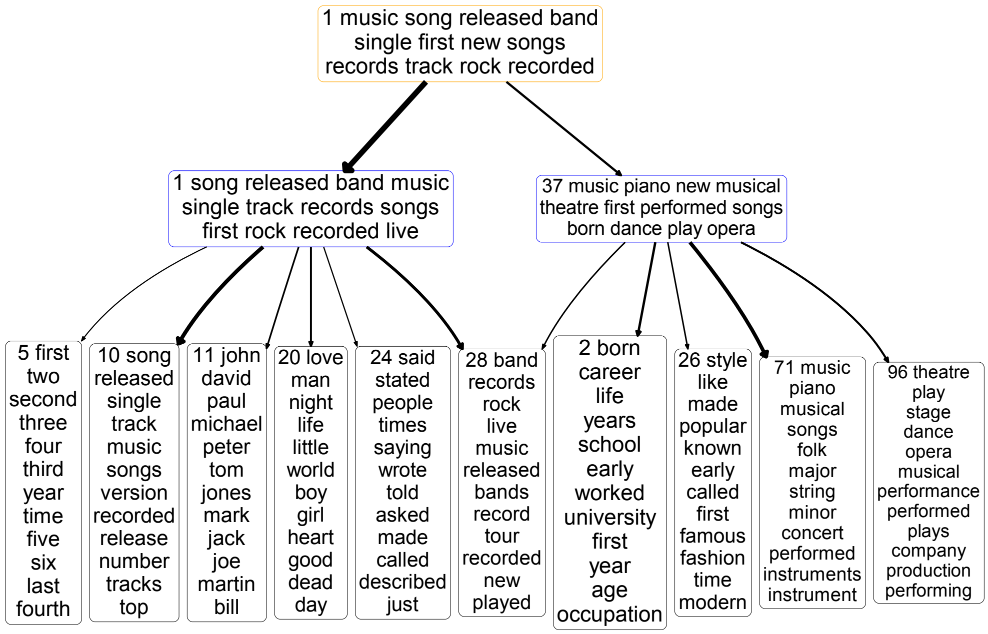


Figure 14. Analogous plots to Figure 7 for a tree rooted at node 1 on "music & song" from Wiki.

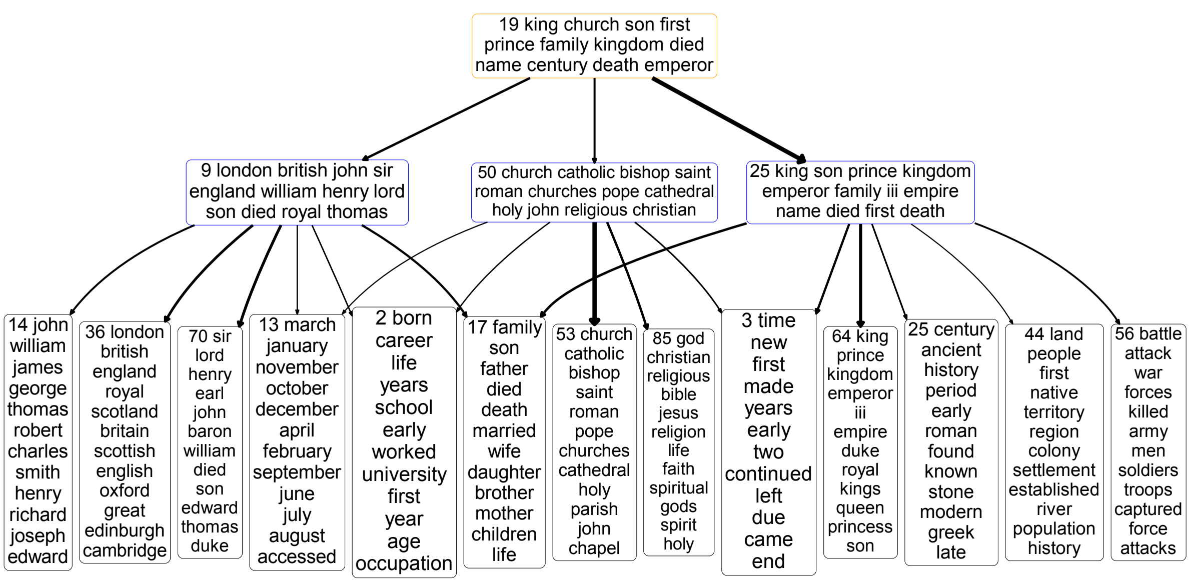


Figure 15. Analogous plots to Figure 7 for a tree rooted at node 19 on "United Kingdom" from Wiki.

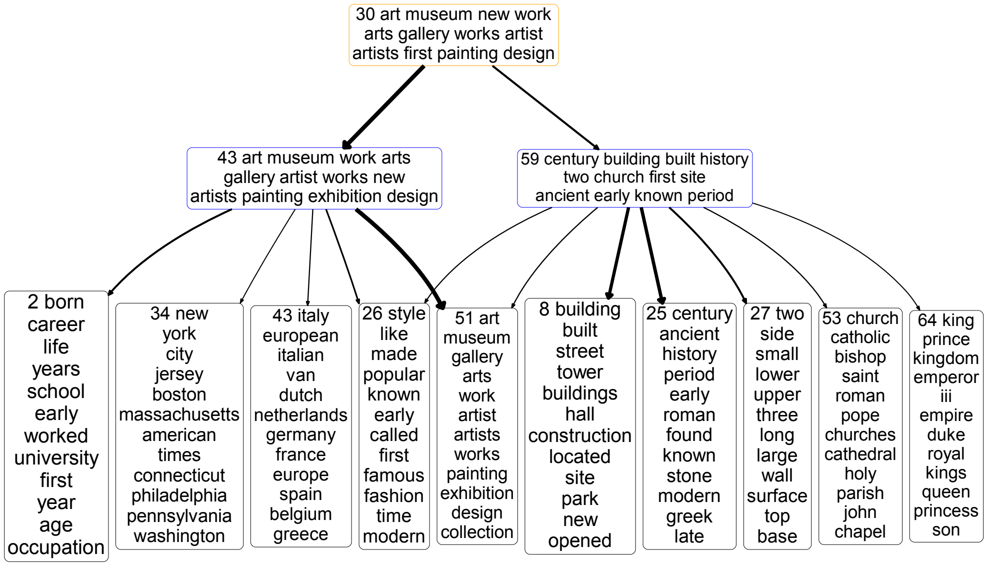


Figure 16. Analogous plots to Figure 7 for a tree rooted at node 30 on "art & museum" from Wiki.

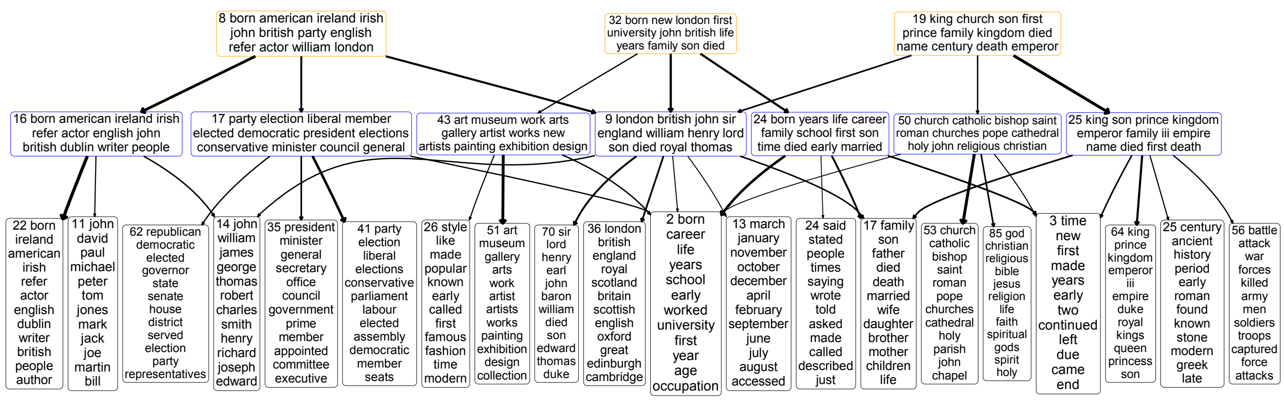


Figure 17. Analogous plots to Figure 7 for a subnetwork on "British" from Wiki, consisting of three trees rooted at nodes 8, 32, and 19, respectively, of layer three.

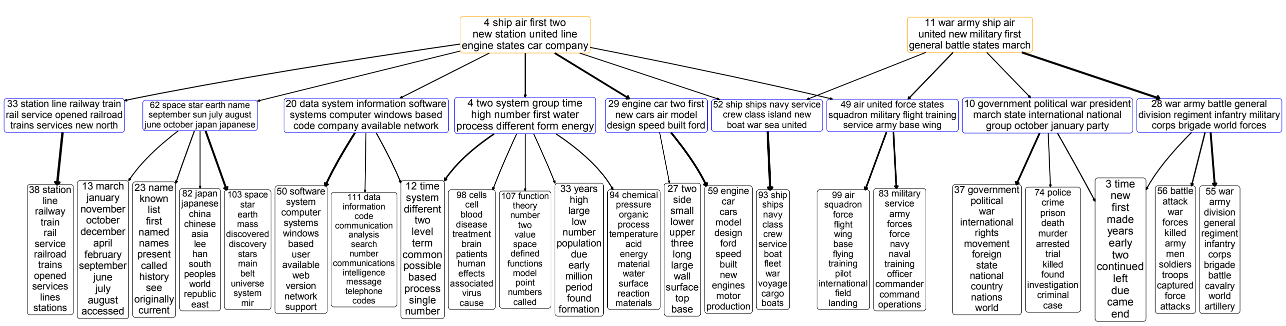


Figure 18. Analogous plots to Figure 7 for a subnetwork on "ship & air" from Wiki, consisting of two trees rooted at nodes 4 and 11, respectively, of layer three.

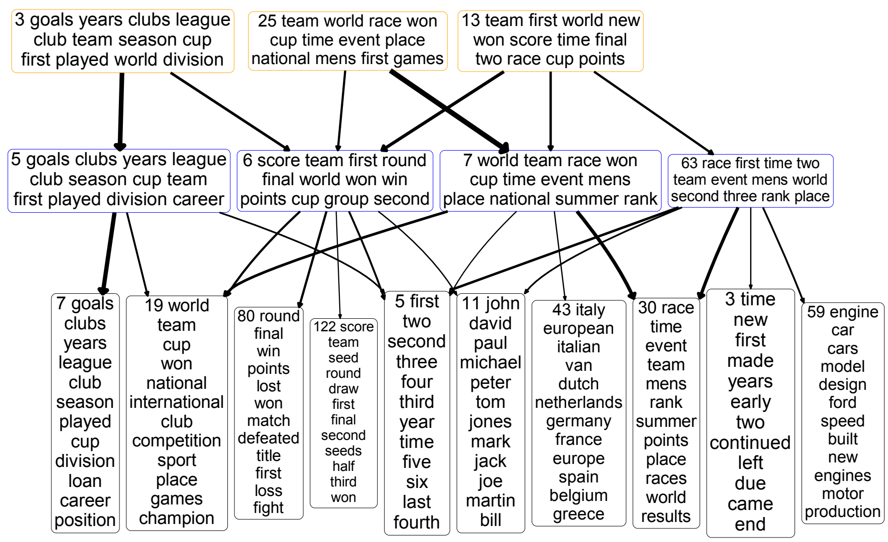


Figure 19. Analogous plots to Figure 7 for a subnetwork on "team & race" from Wiki, consisting of three trees rooted at nodes 3, 25, and 13, respectively, of layer three.