Variational Bayes for generic topic models

Gregor Heinrich[1] and Michael Goesele[2]

[1] Fraunhofer IGD and University of Leipzig
[2] TU Darmstadt

Abstract. The article contributes a derivation of variational Bayes for a large class of topic models by generalising from the well-known model of latent Dirichlet allocation. For an abstraction of these models as systems of interconnected mixtures, variational update equations are obtained, leading to inference algorithms for models that so far have used Gibbs sampling exclusively.

1 Introduction

Topic models (TMs) are a set of unsupervised learning models used in many areas of artificial intelligence: In text mining, they allow retrieval and automatic thesaurus generation; computer vision uses TMs for image classification and content based retrieval; in bioinformatics they are the basis for protein relationship models etc.

In all of these cases, TMs learn latent variables from co-occurrences of features in data. Following the seminal model of latent Dirichlet allocation (LDA [6]), this is done efficiently according to a model that exploits the conjugacy of Dirichlet and multinomial probability distributions. Although the original work by Blei et al. [6] has shown the applicability of variational Bayes (VB) for TMs with impressive results, inference especially in more complex models has not adopted this technique but remains the domain of Gibbs sampling (e.g., [12,9,8]).

In this article, we explore variational Bayes for TMs in general rather than specific for some given model. We start with an overview of TMs and specify general properties (Sec. 2). Using these properties, we develop a generic approach to VB that can be applied to a large class of models (Sec. 3). We verify the variational algorithms on real data and several models (Sec. 4). This paper is therefore the VB counterpart to [7].

2 Topic models

We characterise topic models as a form of discrete mixture models. Mixture models approximate complex distributions by a convex sum of component distributions, $p(x) = \sum_{k=1}^{K} p(x|z{=}k)p(z{=}k)$, where $p(z{=}k)$ is the weight of a component with index k and distribution $p(x|z{=}k)$.

Latent Dirichlet allocation as the simplest TM can be considered a mixture model with two interrelated mixtures: It represents documents m as mixtures of latent variables z with components $\vec{\vartheta}_m = p(z|m)$ and latent topics z as mixtures of words w with

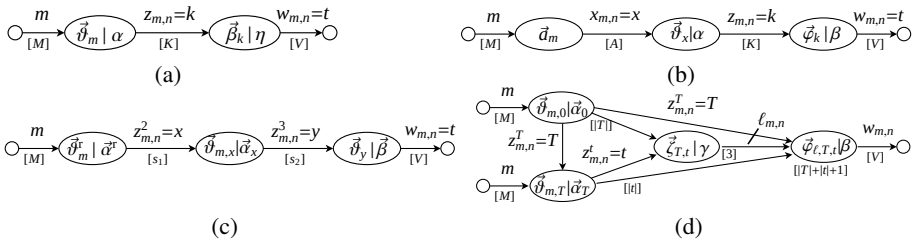

Fig. 1. Dependencies of mixture levels (ellipses) via discrete variables (arrows) in examples from literature: (a) latent Dirichlet allocation [6], (b) author–topic model (ATM [12], using observed parameters $\vec{a}_m$ to label documents, see end of Sec. 3), (c) 4-level pachinko allocation (PAM [9], models semantic structure with a hierarchy of topics $\vec{\vartheta}_m, \vec{\vartheta}_{m,x}, \vec{\vartheta}_y$), (d) hierarchical pachinko allocation (hPAM [8], topic hierarchy; complex mixture structure).

components $\vec{\beta}_k = p(w|z{=}k)$ and component weights $\vec{\vartheta}_m$, leading to a distribution over words w of $p(w|m) = \sum_{k=1}^{K} \vartheta_{m,k}\,\beta_{k,w}$.[3] The corresponding generative process is illustrative: For each text document m, a multinomial distribution $\vec{\vartheta}_m$ is drawn from a Dirichlet prior $\mathrm{Dir}(\vec{\vartheta}_m|\alpha)$ with hyperparameter α. For each word token $w_{m,n}$ of that document, a topic $z_{m,n}{=}k$ is drawn from the document multinomial $\vec{\vartheta}_m$ and finally the word observation w drawn from a topic-specific multinomial over terms $\vec{\beta}_k$. Pursuing a Bayesian strategy with parameters handled as random variables, the topic-specific multinomial is itself drawn from another Dirichlet, $\mathrm{Dir}(\vec{\beta}_k|\eta)$, similar to the document multinomial.

Generic TMs. As generalisations of LDA, topic models can be seen as a powerful yet flexible framework to model complex relationships in data that are based on only two modelling assumptions: (1) TMs are structured into Dirichlet–multinomial mixture "levels" to learn discrete latent variables (in LDA: z) and multinomial parameters (in LDA: β and ϑ). And (2) these levels are coupled via the values of discrete variables, similar to the coupling in LDA between ϑ and β via z.

More specifically, topic models form graphs of mixture levels with sets of multinomial components as nodes connected by discrete random values as directed edges. Conditioned on discrete inputs, each mixture level chooses one of its components to generate discrete output propagated to the next level(s), until one or more final levels produce observable discrete data. For some examples from literature, corresponding "mixture networks" are shown in Fig. 1, including the variant of observed multinomial parameters substituting the Dirichlet prior, which will be discussed further below.

For the following derivations, we introduce sets of discrete variables X, multinomial parameters Θ and Dirichlet hyperparameters A as model-wide quantities, and the corresponding level-specific quanitities $X^\ell, \Theta^\ell, A^\ell$ where superscript ℓ indicates the mixture level. The constraint of connecting different mixture levels (ellipses in Fig. 1) via discrete variables (arrows in Fig. 1) can be expressed by an operator $\uparrow x^\ell$ that yields all parent variables of a mixture level $\ell \in L$ generating variable x^ℓ. Here x^ℓ can refer to

[3] In example models, we use the symbols from the original literature.

specific tokens $\uparrow x_i^\ell$ or configurations $\uparrow X^\ell$. Based on this and the definitions of the multinomial and Dirichlet distributions, the joint likelihood of any TM is:

$$p(X, \Theta|A) = \prod_{\ell\in L} p(X^\ell, \Theta^\ell|A^\ell, \uparrow X^\ell) = \prod_{\ell\in L}\left[\prod_i \mathrm{Mult}(x_i|\Theta, \uparrow x_i)\prod_k \mathrm{Dir}(\vec{\vartheta}_k|A, \uparrow X)\right]^{[\ell]} \quad (1)$$

$$= \prod_{\ell\in L}\left[\prod_i \vartheta_{k_i,x_i}\prod_k \frac{\Gamma(\sum_t \alpha_{j,t})}{\prod_t \Gamma(\alpha_{j,t})}\prod_t \vartheta_{k,t}^{\alpha_{j,t}-1}\right]^{[\ell]};\ k_i^\ell = g^\ell(\uparrow x_i^\ell, i),\ j^\ell = f^\ell(k^\ell)$$

$$= \prod_{\ell\in L}\left[\prod_k \frac{1}{\Delta(\vec{\alpha}_j)}\prod_t \vartheta_{k,t}^{n_{k,t}+\alpha_{j,t}-1}\right]^{[\ell]};\quad n_{k,t}^\ell = \left[\sum_i \delta(k_i-k)\,\delta(x_i-t)\right]^{[\ell]}. \quad (2)$$

In this equation, some further notation is introduced: We use brackets $[\cdot]^{[\ell]}$ to indicate that the contained quantities are specific to level ℓ. Moreover, the mappings from parent variables to component indices k_i are expressed by (level-specific) $k_i = g(\uparrow x_i, i)$, and $n_{k,t}^\ell$ is the number of times that a configuration $\{\uparrow x_i, i\}$ for level ℓ lead to component k^ℓ. Further, models are allowed to group components by providing group-specific hyperparameters $\vec{\alpha}_j$ with mapping $j = f(k)$. Finally, $\Delta(\vec{\alpha})$ is the normalisation function of the Dirichlet distribution, a K-dimensional beta function: $\Delta(\vec{\alpha}) \triangleq \prod_t \Gamma(\alpha_t)/\Gamma(\sum_t \alpha_t)$.

3 Variational Bayes for topic models

As in many latent-variable models, determining the posterior distribution $p(H, \Theta|V) = p(V, H, \Theta) \,/\, \sum_H \int p(V, H, \Theta)\,\mathrm{d}\Theta$ with hidden and visible variables $\{H, V\} = X$, is intractable in TMs because of excessive dependencies between the sets of latent variables H and parameters Θ in the marginal likelihood $p(V) = \sum_H \int p(V, H, \Theta)\,\mathrm{d}\Theta$ in the denominator. Variational Bayes [2] is an approximative inference technique that relaxes the structure of $p(H, \Theta|V)$ by a simpler variational distribution $q(H, \Theta|\Psi, \Xi)$ conditioned on sets of free variational parameters Ψ and Ξ to be estimated in lieu of H and Θ. Minimizing the Kullback-Leibler divergence of the distribution q to the true posterior can be shown to be equivalent to maximising a lower bound on the log marginal likelihood:

$$\begin{aligned}\log p(V) &\geq \log p(V) - \mathrm{KL}\{q(H, \Theta) \parallel p(H, \Theta|V)\}\\ &= \left\langle \log p(V, H, \Theta)\right\rangle_{q(H,\Theta)} + \mathrm{H}\{q(H, \Theta)\} \triangleq \mathcal{F}\{q(H, \Theta)\}\end{aligned} \quad (3)$$

with entropy $\mathrm{H}\{\cdot\}$. $\mathcal{F}\{q(H, \Theta)\}$ is the (negative) variational free energy – the quantity to be optimised using an EM-like algorithm that alternates between (E) maximising $\mathcal{F}$ w.r.t. the variational parameters to pull the lower bound towards the marginal likelihood and (M) maximising $\mathcal{F}$ w.r.t. the true parameters to raise the marginal likelihood.

Mean-field approximation. Following the variational mean field approach [2], in the LDA model the variational distribution consists of fully factorised Dirichlet and multinomial distributions [6]:[4]

$$q(\vec{z}, \beta, \vartheta|\varphi, \lambda, \gamma) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} \mathrm{Mult}(z_{m,n}|\vec{\varphi}_{m,n}) \cdot \prod_{k=1}^{K} \mathrm{Dir}(\vec{\beta}_k|\vec{\lambda}_k)\prod_{m=1}^{M} \mathrm{Dir}(\vec{\vartheta}_m|\vec{\gamma}_m)\,. \quad (4)$$

[4] In [6] this refers to the smoothed version; it is described in more detail in [5].

In [6], this approach proved very successful, which raises the question how it can be transferred to more generic TMs. Our approach is to view Eq. 4 as a special case of a more generic variational structure that captures dependencies $\uparrow X$ between multiple hidden mixture levels and includes LDA for the case of one hidden level ($H = \{\vec{z}\}$):

$$q(H, \Theta|\Psi, \Xi) = \prod_{\ell\in H}\left[\prod_i \mathrm{Mult}(x_i|\vec{\psi}_i, \uparrow x_i)\right]^{[\ell]} \prod_{\ell\in L}\left[\prod_k \mathrm{Dir}(\vec{\vartheta}_k|\vec{\xi}_k, \uparrow X)\right]^{[\ell]} , \qquad (5)$$

where $\ell \in H$ refers to all levels that produce hidden variables. In the following, we assume that the indicator i is identical for all levels ℓ, e.g., words in documents $i^\ell = i \equiv (m, n)$. Further, tokens i in the corpus can be grouped into terms v and (observable) document-specific term frequencies $n_{m,v}$ introduced. We use shorthand $u = (m, v)$ to refer to specific unique tokens or document–term pairs.

Topic field. The dependency between mixture levels, $\uparrow x_u^\ell$, can be expressed by the likelihood of a particular configuration of hidden variables $\vec{x}_u{=}\vec{t} \triangleq \{x_u^\ell{=}t^\ell\}_{\ell\in H}$ under the variational distribution: $\psi_{u,\vec{t}} = q(\vec{x}_u{=}\vec{t}|\Psi)$. The complete structure $\underline{\psi}_u$ (the joint distribution over all $\ell \in H$ with $\underline{\Psi} = \{\underline{\psi}_u\}_{\forall u}$) is a multi-way array of likelihoods for all latent configurations of token u with as many index dimensions as there are dependent variables. For instance, Fig. 1 reveals that LDA has one hidden variable with dimension K while PAM has two with dimensions $s_1 \times s_2$. Because of its interpretation as a mean field of topic states in the model, we refer to $\underline{\psi}_u$ as a "topic field" (in underline notation).

We further define $\psi^\ell_{u,k,t}$ as the likelihood of configuration (k^ℓ, t^ℓ) for document–term pair u. This "marginal" of $\underline{\psi}_u$ depends on the mappings between parent variables $\uparrow x_u$ and components k on each level. To obtain $\psi^\ell_{u,k,t}$, the topic field $\underline{\psi}_u$ is summed over all descendant paths that $x_u{=}t$ causes and the ancestor paths that can cause $k = g(\uparrow x_u, u)$ on level ℓ according to the generative process:

$$\psi^\ell_{u,k,t} = \sum_{\{\vec{t}^\ell_{\mathrm{A}}, \vec{t}^\ell_{\mathrm{D}}\}} \psi_{u;(\vec{t}^\ell_{\mathrm{A}}, k^\ell, t^\ell, \vec{t}^\ell_{\mathrm{D}})} \; ; \quad \vec{t}^\ell_{\mathrm{A}} = \text{path causing } k^\ell \; , \vec{t}^\ell_{\mathrm{D}} = \text{path caused by } t^\ell \; . \qquad (6)$$

Descendant paths $\vec{t}^\ell_{\mathrm{D}}$ of t^ℓ are obtained via recursion of $k = g(\uparrow x^d_u, u)$ over ℓ's descendant levels d. Assuming bijective $g(\cdot)$ as in the TMs in Fig. 1, the ancestor paths $\vec{t}^\ell_{\mathrm{A}}$ that correspond to components in parents leading to k^ℓ are obtained via $(\uparrow x^a_u, u) = g^{-1}(k)$ on ℓ's ancestor levels a recursively. Each pair $\{\vec{t}^\ell_{\mathrm{A}}, \vec{t}^\ell_{\mathrm{D}}\}$ corresponds to one element in $\underline{\psi}_u$ per $\{k^\ell, t^\ell\}$ at index vector $\vec{t} = (\vec{t}^\ell_{\mathrm{A}}, k^\ell, t^\ell, \vec{t}^\ell_{\mathrm{D}})$.

Free energy. Using Eqs. 2, 3, 5 and 6, the free energy of the generic model becomes:

$$\begin{aligned}\mathcal{F} = &\sum_{\ell\in L}\left[\sum_k \log\Delta(\vec{\xi}_k) - \log\Delta(\vec{\alpha}_j) + \sum_t \left((\textstyle\sum_u n_u\psi_{u,k,t}) + \alpha_{j,t} - \xi_{k,t}\right)\cdot\mu_t(\vec{\xi}_k)\right]^{[\ell]} \\ &- \sum_u n_u \sum_{\vec{t}} \psi_{u,\vec{t}}\log\psi_{u,\vec{t}} = \sum_{\ell\in L}\mathcal{F}^\ell + \mathrm{H}\{\Psi\}\,, \end{aligned} \qquad (7)$$

where $\mu_t(\vec{\xi}) \triangleq \Psi(\xi_t) - \Psi(\sum_t \xi_t) = \left\langle \log\vec{\vartheta}|\vec{\xi}\right\rangle_{\mathrm{Dir}(\vec{\vartheta}|\vec{\xi})} = \nabla_t \log\Delta(\vec{\xi})$, and $\Psi(\xi) \triangleq \mathrm{d}/\mathrm{d}x \log\Gamma(\xi)$ is the digamma function.[5]

[5] Note the distinction between the function $\Psi(\cdot)$ and quantity Ψ.

Variational E-steps. In the E-step of each model, the variational distributions for the joint multinomial $\underline{\psi}_u$ for each token (its topic field) and the Dirichlet parameters $\vec{\xi}_k^{\,\ell}$ on each level need to be estimated. The updates can be derived from the generic Eq. 7 by setting derivatives with respect to the variational parameters to zero, which yields:[6]

$$\psi_{u,\vec{t}} \propto \exp\left(\textstyle\sum_{\ell\in L}\left[\mu_t(\vec{\xi}_k)\right]^{[\ell]}\right) , \tag{8}$$

$$\xi^{\ell}_{k,t} = \left[(\textstyle\sum_u n_u \psi_{u,k,t}) + \alpha_{j,t}\right]^{[\ell]} \tag{9}$$

where the sum $\sum_u n_u \psi^{\ell}_{u,k,t}$ for level ℓ can be interpreted as the expected counts $\langle n^{\ell}_{k,t}\rangle_q$ of co-occurrence of the value pair (k^{ℓ}, t^{ℓ}). The result in Eqs. 8 and 9 perfectly generalises that for LDA in [5].

M-steps. In the M-step of each model, the Dirichlet hyperparameters $\vec{\alpha}^{\ell}_j$ (or scalar α^{ℓ}) are calculated from the variational expectations of the log model parameters $\langle \log \vartheta_{k,t}\rangle_q = \mu_t(\vec{\xi}_k)$, which can be done at mixture level (Eq. 9 has no reference to $\vec{\alpha}^{\ell}_j$ across levels).

Each estimator for $\vec{\alpha}_j$ (omitting level ℓ) should "see" only the expected parameters $\mu_t(\vec{\xi}_k)$ of the K_j components associated with its group $j = f(k)$. We assume that components be associated a priori (e.g., PAM in Fig. 1c has $\vec{\vartheta}_{m,x} \sim \mathrm{Dir}(\vec{\alpha}_x)$) and K_j is known. Then the Dirichlet ML parameter estimation procedure given in [6,10] can be used in modified form. It is based on Newton's method with the Dirichlet log likelihood function f as well as its gradient and Hessian elements g_t and h_{tu}:

$$\begin{aligned}
f(\vec{\alpha}_j) &= -K_j \log\Delta(\vec{\alpha}_j) + \textstyle\sum_t(\alpha_{j,t}-1)\sum_{\{k:f(k)=j\}}\mu_t(\vec{\xi}_k)\\
g_t(\vec{\alpha}_j) &= -K_j\,\mu_t(\vec{\alpha}_j) + \textstyle\sum_{\{k:f(k)=j\}}\mu_t(\vec{\xi}_k)\\
h_{tu}(\vec{\alpha}_j) &= -K_j\Psi'(\textstyle\sum_s\alpha_{j,s}) + \delta(t-u)K_j\Psi'(\alpha_{j,t}) = z + \delta(t-u)h_{tt}\\
\alpha_{j,t} &\leftarrow \alpha_{j,t} - (\underline{H}^{-1}\vec{g})_t = \alpha_{j,t} - h_{tt}^{-1}\left(g_t - (\textstyle\sum_s g_s h_{ss}^{-1})\Big/(z^{-1}+\sum_s h_{ss}^{-1})\right) .
\end{aligned} \tag{10}$$

Scalar α (without grouping) is found accordingly via the symmetric Dirichlet:

$$\begin{aligned}
f &= -K[T\log\Gamma(\alpha) - \log\Gamma(T\alpha)] + (\alpha-1)s_\alpha\ , \quad s_\alpha = \textstyle\sum_{k=1}^{K}\sum_{t=1}^{T}\mu_t(\vec{\xi}_k)\\
g &= KT[\Psi(T\alpha) - \Psi(\alpha) + s_\alpha]\ , \quad h = KT[T\Psi'(T\alpha) - \Psi'(\alpha)]\\
\alpha &\leftarrow \alpha - gh^{-1}\ .
\end{aligned} \tag{11}$$

Variants. As an alternative to Bayesian estimation of all mixture level parameters, for some mixture levels ML point estimates may be used that are computationally less expensive (e.g., unsmoothed LDA [6]). By applying ML only to levels without document-specific components, the generative process for unseen documents is retained. The E-step with ML levels has a simplified Eq. 8, and ML parameters ϑ^c are estimated in the M-step (instead of hyperparameters):

$$\psi_{u,\vec{t}} \propto \exp\left(\textstyle\sum_{\ell\in L\setminus c}\left[\mu_t(\vec{\xi}_k)\right]^{[\ell]}\right)\cdot\vartheta^c_{k,t}\ , \quad \vartheta^c_{k,t} = \langle n^c_{k,t}\rangle_q \Big/ \langle n^c_k\rangle_q \propto \textstyle\sum_u n_u\psi^c_{u,k,t}\ . \tag{12}$$

[6] In Eq. 8 we assume that $t^{\ell}{=}v$ on final mixture level(s) ("leaves"), which ties observed terms v to the latent structure. For "root" levels where component indices are observed, $\mu_t(\vec{\xi}_k)$ in Eq. 8 can be replaced by $\Psi(\xi_{k,t})$.

Moreover, as an extension to the framework specified in Sec. 2, it is straightforward to introduce observed parameters that for instance can represent labels, as in the author–topic model, cf. Fig 1. In the free energy in Eq. 7, the term with $\mu_t(\vec{\xi}_k)$ is replaced by $(\sum_u n_u \psi_{u,k,t}) \log \vartheta_{k,t}$, and consequently, Eq. 8 takes the form of Eq. 12 (left), as well.

Other variants like specific distributions for priors (e.g., logistic-normal to model topic correlation [4] and non-parametric approaches [14]) and observations (e.g., Gaussian components to model continuous data [1]), will not be covered here.

Algorithm structure. The complete variational EM algorithm alternates between the variational E-step and M-step until the variational free energy $\mathcal{F}$ converges at an optimum. At convergence, the estimated document and topic multinomials can be obtained via the variational expectation $\log \hat{\vartheta}_{k,t} = \mu_t(\vec{\xi}_k)$. Initialisation plays an important role to avoid local optima, and a common approach is to initialise topic distributions with observed data, possibly using several such initialisations concurrently. The actual variational EM loop can be outlined in its generic form as follows:

1. Repeat E-step loop until convergence w.r.t. variational parameters:
   1. For each observed unique token u:
      1. For each configuration $\vec{t}$: calculate var. multinomial $\psi_{u,\vec{t}}$ (Eq. 8 or 12 left).
      2. For each (k,t) on each level ℓ: calculate var. Dirichlet parameters $\xi^{\ell}_{k,t}$ based on topic field marginals $\psi^{\ell}_{u,k,t}$ (Eqs. 6 and 9), which can be done differentially: $\xi^{\ell}_{k,t} \leftarrow \xi^{\ell}_{k,t} + n_u \Delta\psi^{\ell}_{u,k,t}$ with $\Delta\psi^{\ell}_{u,k,t}$ the change of $\psi^{\ell}_{u,k,t}$.
   2. Finish variational E-step if free energy $\mathcal{F}$ (Eq. 7) converged.
2. Perform M-step:
   1. For each j on each level ℓ: calculate hyperparameter $\alpha^{\ell}_{j,t}$ (Eqs. 10 or 11), inner iteration loop over t.
   2. For each (k,t) in point-estimated nodes ℓ: estimate $\vartheta^{\ell}_{k,t}$ (Eq. 12 right).
3. Finish variational EM loop if free energy $\mathcal{F}$ (Eq. 7) converged.

In practice, similar to [5], this algorithm can be modified by separating levels with document-specific variational parameters $\Xi^{\ell,m}$ and such with corpus-wide parameters $\Xi^{\ell,*}$. This allows a separate E-step loop for each document m that updates $\underline{\psi}_u$ and $\Xi^{\ell,m}$ with $\Xi^{\ell,*}$ fixed. Parameters $\Xi^{\ell,*}$ are updated afterwards from changes $\Delta\psi^{\ell}_{u,k,t}$ cumulated in the document-specific loops, and their contribution added to $\mathcal{F}$.

4 Experimental verification

In this section, we present initial validation results based on the algorithm in Sec. 3.

Setting. We chose models from Fig. 1, LDA, ATM and PAM, and investigated two versions of each: an unsmoothed version that performs ML estimation of the final mixture level (using Eq. 12) and a smoothed version that places variational distributions over all parameters (using Eq. 8). Except for the component grouping in PAM ($\vec{\vartheta}_{m,x}$ have vector hyperparameter $\vec{\alpha}_x$), we used scalar hyperparameters. As a base-line, we used Gibbs sampling implementations of the corresponding models. Two criteria are immediately useful: the ability to generalise to test data V' given the model parameters Θ, and the convergence time (assuming single-threaded operation). For the

Model:		LDA			ATM			PAM		
Dimensions {A,B}:		$K = \{25, 100\}$			$K = \{25, 100\}$			$s_{1,2} = \{(5, 10), (25, 25)\}$		
Method:		GS	VB_{ML}	VB	GS	VB_{ML}	VB	GS	VB_{ML}	VB
Convergence time [h]	A	0.39	0.83	0.91	0.73	1.62	1.79	0.5	1.25	1.27
	B	1.92	3.75	4.29	3.66	7.59	8.1	5.61	14.86	16.06
Iteration time [sec]	A	4.01	157.3	164.2	6.89	254.3	257.8	5.44	205.1	207.9
	B	16.11	643.3	671.0	29.95	1139.2	1166.9	53.15	2058.2	2065.1
Iterations	A	350	19	20	380	23	25	330	22	22
	B	430	21	23	440	24	25	380	26	28
Perplexity	A	1787.7	1918.5	1906.0	1860.4	1935.2	1922.8	2053.8	2103.0	2115.1
	B	1613.9	1677.6	1660.2	1630.6	1704.0	1701.9	1909.2	1980.5	1972.6

Fig. 2. Results of VB and Gibbs experiments.

first criterion, because of its frequent usage with topic models we use the perplexity, the inverse geometric mean of the likelihood of test data tokens given the model: $\mathcal{P}(V') = \exp(-\sum_u n_u \log p(v'_u|\Theta')/W')$ where Θ' are the parameters fitted to the test data V' with W' tokens. The log likelihood of test tokens $\log p(v'_u|\Theta')$ is obtained by (1) running the inference algorithms on the test data, which yields Ξ' and consequently Θ', and (2) marginalising all hidden variables h'_u in the likelihood $p(v'_u|h'_u, \Theta') = \prod_{\ell \in L} [\vartheta_{k,t}]^{[\ell]}$.[7] The experiments were performed on the NIPS corpus [11] with $M = 1740$ documents (174 held-out), $V = 13649$ terms, $W = 2301375$ tokens, and $A = 2037$ authors.

Results. The results of the experiments are shown in Fig. 2. It turns out that generally the VB algorithms were able to achieve perplexity reductions in the range of their Gibbs counterparts, which verifies the approach taken. Further, the full VB approaches tend to yield slightly improved perplexity reductions compared to the ML versions. However, these first VB results were consistently weaker compared to the baselines. This may be due to adverse initialisation of variational distributions, causing VB algorithms to become trapped at local optima. It may alternatively be a systematic issue due to the correlation between Ψ and Ξ assumed independent in Eq. 5, a fact that has motivated the collapsed variant of variational Bayes in [13]. Considering the second evaluation criterion, the results show that the current VB implementations generally converge less than half as fast as the corresponding Gibbs samplers. This is why currently work is undertaken in the direction of code optimisation, including parallelisation for multikernel CPUs, which, opposed to (collapsed) Gibbs samplers, is straightforward for VB.

5 Conclusions

We have derived variational Bayes algorithms for a large class of topic models by generalising from the well-known model of latent Dirichlet allocation. By an abstraction of these models as systems of interconnected mixture levels, we could obtain variational update equations in a generic way, which are the basis for an algorithm, that can be easily applied to specific topic models. Finally, we have applied the algorithm to a couple of example models, verifying the general applicability of the approach. So far, especially more complex topic models have predominantly used inference based on Gibbs sampling. Therefore, this paper is a step towards exploring the possibility of variational

[7] In contrast to [12], we also used this method to determine ATM perplexity (from the $\vec{\varphi}_k$).

approaches. However, what can be drawn as a conclusion from the experimental study in this paper, more work remains to be done in order to make VB algorithms as effective and efficient as their Gibbs counterparts.

Related work. Beside the relation to the original LDA model [6,5], especially the proposed representation of topic models as networks of mixture levels makes work on discrete DAG models relevant: In [3], a variational approach for structure learning in DAGs is provided with an alternative derivation based on exponential families leading to a structure similar to the topic field. They do not discuss mapping of components or hyperparameters and restrict their implementations to structure learning in graphs bipartite between hidden and observed nodes. Also, the authors of [9] present their pachinko allocation models as DAGs, but formulate inference based on Gibbs sampling. In contrast to this, the novelty of the work presented here is that it unifies the theory of topic models in general including labels, the option of point estimates and component grouping for variational Bayes, giving empirical results for real-world topic models.

Future work will optimise the current implementations with respect to efficiency in order to improve the experimental results presented here, and an important aspect is to develop parallel algorithms for the models at hand. Another research direction is the extension of the framework of generic topic models, especially taking into consideration the variants of mixture levels outlined in Sec. 3. Finally, we will investigate a generalisation of collapsed variational Bayes [13].

References

1. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *JMLR*, 3(6):1107–1136, 2003.
2. M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
3. M. J. Beal and Z. Ghahramani. Variational bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1:793–832, 2006.
4. D. Blei and J. Lafferty. A correlated topic model of science. *AOAS*, 1:17–35, 2007.
5. D. Blei, A. Ng, and M. Jordan. Hierarchical Bayesian models for applications in information retrieval. *Bayesian Statistics*, 7:25–44, 2003.
6. D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.
7. G. Heinrich. A generic approach to topic models. In *ECML/PKDD*, 2009.
8. W. Li, D. Blei, and A. McCallum. Mixtures of hierarchical topics with pachinko allocation. In *ICML*, 2007.
9. W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML*, 2006.
10. T. Minka. Estimating a Dirichlet distribution. Web, 2003.
11. NIPS corpus. `http://www.cs.toronto.edu/~roweis/data.html`.
12. M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *ACM SIGKDD*, 2004.
13. Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *NIPS*, volume 19, 2007.
14. Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. Technical Report 653, Department of Statistics, University of California at Berkeley, 2004.