Kernel Topic Models

Philipp Hennig[1,2,3] **David Stern**[3] **Ralf Herbrich**[3,*] **Thore Graepel**[3]
[1]Cavendish Laboratory Cambridge, UK
[2]Max Planck Institute for Intelligent Systems Tübingen, Germany
[3]Microsoft Research Cambridge, UK

Abstract

Latent Dirichlet Allocation models discrete data as a mixture of discrete distributions, using Dirichlet beliefs over the mixture weights. We study a variation of this concept, in which the documents' mixture weight beliefs are replaced with squashed Gaussian distributions. This allows documents to be associated with elements of a Hilbert space, admitting *kernel topic models (KTM)*, modelling temporal, spatial, hierarchical, social and other structure between documents. The main challenge is efficient approximate inference on the latent Gaussian. We present an approximate algorithm cast around a Laplace approximation in a transformed basis. The KTM can also be interpreted as a type of Gaussian process latent variable model, or as a topic model conditional on document features, uncovering links between earlier work in these areas.

1 Introduction

Latent Dirichlet Allocation (LDA) [Blei et al., 2003] is a generative model for datasets comprising collections of discrete samples. Each collection is assumed to be generated from a mixture of discrete distributions, such that both the belief over the discrete distributions and over the mixture weights are Dirichlet. Text documents constitute the most popular domain with this anatomy: Each document in a corpus, treated as a "bag of words" (i.e. ignoring word order), is one collection of (discrete) words, and the mixture components are interpreted as *topics*. Thus, each document exhibits several topics to varying degree, with each word in the document sampled from one specific topic.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.
* RH is with Facebook Inc., Menlo Park, California, USA

Real documents do not exist void of context. They are products of their authors, time, and place. Electronic communication has intensified this truism, and online corpora are now invariably accompanied by copious amounts of meta-data. The identity of the author may be augmented by additional knowledge about their location in a social graph, autobiographic information, and many more. Such features convey semantic information: Topic popularity varies between West and East, conservatives and progressives, rich and poor, scientists and celebrities, young and old, contemporaries and forebears.

In its standard form, LDA can not take advantage of such metadata; but extensions proposed by several authors have addressed certain types of meta-structure. Dynamic development of topics over sequential sets of documents was considered by Blei and Lafferty [2006], Wang and McCallum [2006] and Wang et al. [2009]. Both Mimno and McCallum [2008] and Zhu and Xing [2010] considered a more general description of topics in terms of a linear function in a latent real vector space, linked to the topic dimension through the softmax function. These works differ in their details (some assume the topics stay constant over time while their distribution changes, others that the topics themselves change. Words may be assumed to generate features, or the other way round), but are linked by their common use of *Gaussian* random variables to describe dynamics or regress on document features. They also all use maximum likelihood, or maximum a-posteriori inference to fit regression weights where they exist.

This work generalizes these approaches by replacing real-valued features with elements of a Hilbert space, and point estimates with Gaussian process measures (Figure 1). The resulting *kernel topic model* provides an expressive framework for the inclusion of virtually all types of metadata in the semantic description of topical data, and allows a rich description of nonlinear topic dynamics. The main mathematical challenge is that inference on the latent Gaussian belief is not analytically tractable. We address this through a numerically lightweight Laplace approximation for Dirichlet

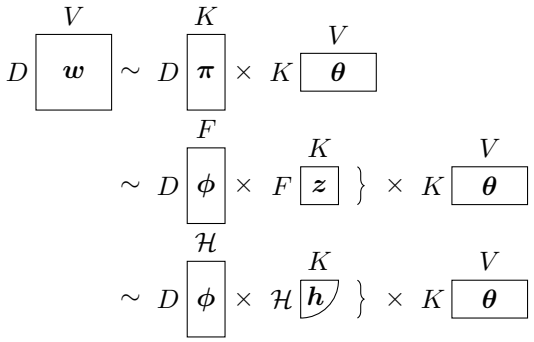

Figure 1: Dimensionality-reduction view of topic models. **Top:** LDA describes D documents containing words from a vocabulary of size V in terms of K topics. **Middle:** Dirichlet multinomial regression portrays the documents in terms of F features, which generate the topics through a linear map. **Bottom:** The kernel topic model replaces the features with coordinates of a Hilbert space $\mathcal{H}$, and the linear map with a nonlinear one. The curly brace denotes a softmax-projection from $\mathbb{R}^K$ to the $[0,1]^K$ simplex.

distributions in the softmax basis, extending on a note by MacKay [1998]. As a side effect, this approximation also admits a particularly efficient implementation of Bayesian inference on linear latent models, such as the one introduced by Mimno and McCallum [2008]. The kernel topic model links topic modelling and Gaussian process latent variable models, effectively casting LDA as a likelihood for generalised Gaussian process models. The price of the increased modelling flexibility is a comparably high computational cost – cubic in the number of documents.

2 Methods

2.1 Model

We consider a corpus of D documents. Document d contains I_d words $w_{di} \in \{1,\dots,V\}, d \in \{1,\dots,D\}, i \in \{1,\dots,I_d\}$ from a vocabulary of size V. Additional aspects of d are described by features $\boldsymbol{\phi}_d \in \mathcal{H}$ in a Hilbert space $\mathcal{H}$. In other words, the dataset consists of pairs $(\boldsymbol{w}_d, \boldsymbol{\phi}_d) \in \{1,\dots,V\}^{I_d} \times \mathcal{H}$.

We construct a topic model conditional on the observable features of the documents, using the following generative process for the vector $\boldsymbol{w}_d$ from K topics:

- For each topic $k \in \{1,\dots,K\}$, generate a discrete probability distribution with parameters $\boldsymbol{\theta}_k \in [0,1]^V$ over the vocabulary of size V by sampling from a Dirichlet distribution with parameter vector $\boldsymbol{\beta}_k$ (Γ denotes the Gamma function):

$$p(\boldsymbol{\theta}_k \,|\, \boldsymbol{\beta}_k) = \mathcal{D}(\boldsymbol{\theta}_k; \boldsymbol{\beta}_k) = \frac{\Gamma\left(\sum_v^V \beta_{kv}\right)}{\prod_v^V \Gamma(\beta_{kv})} \prod_k \theta_{kv}^{\beta_{kv}-1}. \tag{1}$$

- Independently sample K functions $h_k(\boldsymbol{\phi}) : \mathcal{H} \to \mathbb{R}$ from the Hilbert space of real-valued functions over $\mathcal{H}$, by sampling from Gaussian process priors with mean functions $\mu_k(\boldsymbol{\phi}_d)$ and covariance functions $\Sigma_k(\boldsymbol{\phi}_d, \boldsymbol{\phi}_{d'})$, induced by (potentially topic-specific) kernels η_k:

$$p(h_k \,|\, \mu_k, \Sigma_k) = \mathcal{GP}(h_k; \mu_k, \Sigma_k^2) \tag{2}$$

- For each document d with features $\boldsymbol{\phi}_d \in \mathbb{R}^F$,
  - Draw a latent variable $\boldsymbol{y}_d$ by evaluating $\boldsymbol{h}(\boldsymbol{\phi}_d)$ and adding Gaussian noise of standard deviation τ:

$$p(\boldsymbol{y}_d \,|\, \boldsymbol{h}, \tau, \boldsymbol{\phi}_d) = \prod_k \mathcal{N}(\boldsymbol{y}_{dk}; h_k(\boldsymbol{\phi}_d), \tau^2) \tag{3}$$

  - Define the topic proportions $\boldsymbol{\pi}_d = \sigma(\boldsymbol{y}) \in [0,1]^K$ where σ is the softmax function

$$\sigma_k(\boldsymbol{y}) = \frac{\exp(y_k)}{\sum_\ell^K \exp(y_\ell)} \tag{4}$$

  - For each of I_d words
    - draw a topic c_{di} from the discrete distribution defined by $\boldsymbol{\pi}_d$:

$$p(c_{di} = k \,|\, \boldsymbol{\pi}_d) = \pi_{dk} \tag{5}$$

    - draw word w_{di} from the discrete distribution of topic c_{di}:

$$p(w_{di} = v \,|\, c_{di}, \boldsymbol{\Theta}) = \theta_{c_{di} v} \tag{6}$$

The directed graphical model in Figure 2, left, sheds light on the dependency structure of this generative model. If we replace everything to the left of $\boldsymbol{\pi}_d$ in that figure by a single Dirichlet parameter vector $\boldsymbol{\alpha}$ (identical for all d), then the parts shown to the right of and including the node $\boldsymbol{\pi}$ correspond to the traditional LDA model [Blei et al., 2003] (Figure 2, right). On the other hand, we can identify the parts to the left of (and excluding) $\boldsymbol{\pi}$ as a case of Gaussian process regression. It is the connection between these two parts that makes the model challenging, and approximate inference will in fact separate in this way.

In passing, we note a connection to the *correlated topic model* [Blei and Lafferty, 2007], which shares everything to the right of and including $\boldsymbol{y}$ in Figure 2, but not the regression element to its left. Instead, that

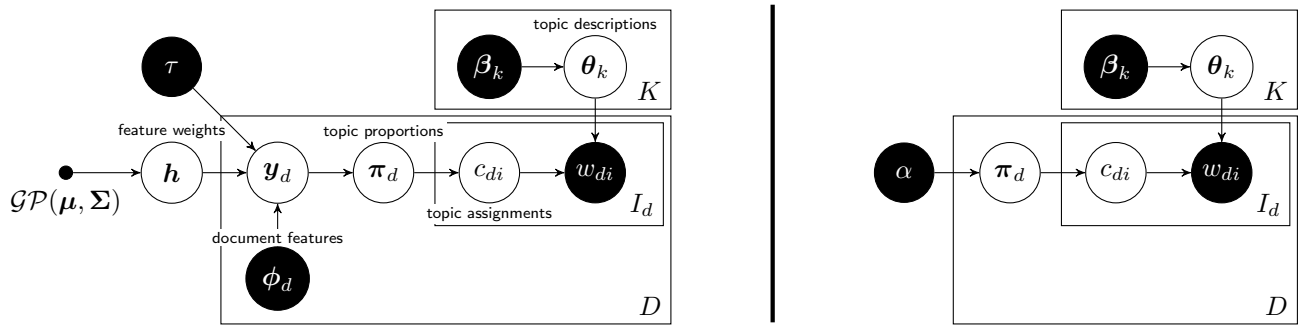

Figure 2: **Left:** Directed graphical model of the kernel topic model. Some variables labeled for clarity. **Right:** latent Dirichlet allocation. The models are identical to the right of and including $\boldsymbol{\pi}$.

model focusses on estimating the correlation between topics, which here is replaced by a simpler, diagonal covariance. Introducing correlations between topics is possible in our model using an approach analogous to the cited work (maximum likelihood estimation on the covariance structure), but left out here for clarity.

3 Inference

Ample expertise has accumulated in the literature, on inference for both LDA, and Gaussian processes given (approximately) Gaussian likelihoods. What is missing is a connection between the two paradigms. This link is the main contribution of this paper. To clarify the setting, however, we give a very brief introduction to the two sub-systems in this section, then derive the link – the *Laplace bridge* – in Section 3.3.

3.1 Semi-Collapsed Variational Inference

Broadly speaking, there are two popular methods for inference in LDA: variational inference [Blei et al., 2003] and collapsed Gibbs sampling [Griffiths and Steyvers, 2004]. Gibbs samples come from the exact posterior, but provide no analytic form for the beliefs. Since our extension benefits from such forms, we opt for a variational approximation. Standard inference in LDA [Blei et al., 2003, Blei and Lafferty, 2009, Hoffman et al., 2010] uses a fully factorized approximate distribution, but Teh et al. [2007] showed that this Ansatz entails an unnecessarily loose bounds and slow convergence. To mitigate this problem, latent variables should be integrated out wherever possible. Since we require explicit forms for the per-document topic distributions $\boldsymbol{\pi}_d$, we can not integrate out this variable, but we *can* collapse the bound on the per-topic distributions $\boldsymbol{\theta}$. This amounts to an adaptation to Teh et al.'s work, which we do not dwell on here for brevity. The bottom line is that it is possible to construct a variational bound that, given a Dirichlet prior

$$p(\boldsymbol{\pi}_d \,|\, \boldsymbol{\alpha}_d) = \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d) \tag{7}$$

on π_d, assigns approximate Dirichlet "posterior" beliefs

$$p(\boldsymbol{\pi}_d \,|\, \boldsymbol{\alpha_d}, \boldsymbol{w}_d) = \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d + \boldsymbol{\nu}_d) \tag{8}$$

with a vector $\boldsymbol{\nu}_d \in \mathbb{R}^K$ of pseudo-counts. At the LDA end of the divide between Gaussian regression and LDA, we thus require a Dirichlet belief.

3.2 Gaussian Process Regression

For the moment, assume there be some isomorphism $\mathfrak{L}$ between K-dimensional Dirichlet distributions and K approximately independent Gaussian ones (to be developed in Section 3.3).

$$\mathfrak{L}: \quad \mathcal{D}(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_d) \leftrightarrow \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{y}_d; \mu_{dk}, \sigma^2_{dk}) \tag{9}$$

This transformation provides approximate Gaussian *messages* from $\boldsymbol{\pi}_d$ to $\boldsymbol{y}_d$ in the graph of Figure 2. With these messages, Gaussian process inference over the Hilbert space $\mathcal{H}$ becomes a known problem, and we can implement an approximate Gaussian process latent inference algorithm: For every topic k, the posterior belief over the function $h_k(\boldsymbol{\phi}_*)$ at the Hilbert location $\boldsymbol{\phi}_*$ is the product of the Gaussian process prior and the D approximately independent Gaussian messages $p(\boldsymbol{y}_d \,|\, h(\phi_d), \boldsymbol{W}, \boldsymbol{\Theta}) = \mathcal{N}(\mu_{kd}; h_k(\phi_d), \tau^2 + \sigma^2_{kd})$. We subsume the means of these messages into a vector $\boldsymbol{\mu}_k$ and their variances into a diagonal matrix $\Sigma_k = \mathrm{diag}(\tau^2 + \sigma^2_{dk})$, which allows us to write the mean and marginal variance functions of the posterior Gaussian process as

$$\begin{aligned} \mathsf{E}[h_*] &= \eta_k(\boldsymbol{\phi}_*, \boldsymbol{\Phi})^\intercal H^{-1}(H^{-1} + \tilde{\Sigma}_k^{-1})^{-1}\tilde{\Sigma}_k^{-1}\tilde{\boldsymbol{\mu}}_k \\ \mathsf{V}[h_*] &= \eta_k(\boldsymbol{\phi}_*, \boldsymbol{\phi_*}) - \eta_k(\boldsymbol{\phi}_*, \boldsymbol{\Phi})(H + \tilde{\Sigma})^{-1}\eta_k(\boldsymbol{\Phi}, \boldsymbol{\phi}_*) \end{aligned} \tag{10}$$

writing the message *precisions* (inverse variances) as $\zeta_d = (\sigma_d^2 + \tau^2)^{-1}$, we construct a matrix $\tilde{S} = \mathrm{diag}(\boldsymbol{\zeta})$

and message precision adjusted means $\tilde{\boldsymbol{\nu}} = \tilde{S}\tilde{\boldsymbol{\mu}}$. Using this notation, the implementation of iterative Gaussian process inference from approximate Gaussian messages contained in Section 3.6.3, in particular Algorithms 3.5 and 3.6 in Rasmussen and Williams [2006] can be used almost without changes.

In Gaussian process generalised regression, the hyperparameters (kernel parameters and observation noise) are usually estimated by evidence maximisation ("type-II maximum likelihood"). In our case, the unknown function is h, the data is $\boldsymbol{w}$ and let the hyperparameters be ξ. Evidence maximisation would amount to optimising $p(\boldsymbol{w} \,|\, \xi) = \int p(\boldsymbol{w} \,|\, f, \xi) p(f \,|\, \xi) \,\mathrm{d}f$. However, in our case, there is an approximate inference algorithm separating the Gaussian process regression from the observed data, so this kind of optimisation has exceedingly high computational cost (each evaluation of $p(\boldsymbol{w} \,|\, \xi)$ involves running the LDA part of Section 3.1 to convergence). Instead, a much cheaper, if less effective, method is to maximise $p(\boldsymbol{y} \,|\, \xi)$, where $\boldsymbol{y}$ are the estimated per-document topic distributions. Defining the matrix $B = \boldsymbol{I} + \tilde{S}^{1/2} K \tilde{S}^{1/2}$, a simple algebraic argument similar to the one in Rasmussen and Williams [2006], Section 3.6.3, gives the log evidence

$$\log Z = \frac{1}{2}\left[\log|\tilde{S}| - \log|B| - \tilde{\mu}^\intercal \tilde{S}^{1/2} B^{-1} \tilde{S}^{1/2} \tilde{\mu}\right] \tag{11}$$

which is numerically stable (because all eigenvalues of B are larger than 1), and can be implemented efficiently. Derivatives of $\log Z$ with respect to the kernel parameters, required for efficient optimisation, are straightforward to calculate using linear algebra identities.

3.3 The Laplace Bridge

To link these two parts of the inference, we must connect the Dirichlet belief on $\boldsymbol{\pi}_d$ and the Gaussian domain required for $\boldsymbol{y}_d$. Since $\sigma(\boldsymbol{y}_d) = \boldsymbol{\pi}_d$, this task amounts to an uncertain form of logistic regression, in the sense that discrete samples c_{dn} from the distribution defined by $\boldsymbol{\pi}_d$ are replaced by probabilistic beliefs over c_{dn}. Our solution to this problem is to construct a Laplace approximation to Dirichlet distributions *in the softmax basis*, in which these distributions can be approximated by Gaussians much better than in the popular simplex basis.

MacKay [1998] showed that, because the softmax function has a Jacobian proportional to $\prod_k \pi_k$, a basis change from probabilities $\boldsymbol{\pi}$ to real numbers $\boldsymbol{y} = \sigma^{-1}(\boldsymbol{\pi})$ gives the Dirichlet a new parametric form

$$\mathcal{D}_y(\boldsymbol{\pi}(\boldsymbol{y}); \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_k^K \alpha_k\right)}{\prod_k^K \Gamma(\alpha_k)} \prod_k^K \pi_k^{\alpha_k} g(\mathbf{1}^\intercal \boldsymbol{y}) \tag{12}$$

$g(\mathbf{1}^\intercal \boldsymbol{y})$ is an arbitrary normalisable measure, required to ensure integrability by restricting the sum of the elements of $\boldsymbol{y}$ ($\mathbf{1}$ is the vector $[1, 1, 1, \dots]$). In this basis, the Dirichlet lacks the -1 terms in the exponents present in the standard representation, and thus does not diverge for $|x| \to \infty$ and $\alpha_i < 1$. It is also a unimodal distribution whose mode at $\boldsymbol{\pi}(\boldsymbol{y}) = \boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|$ now falls together with its mean. These aspects allow a good quality Laplace approximation.

For numerical convenience, we choose (like MacKay)

$$g = \exp\left(-\frac{\epsilon}{2}(\mathbf{1}^\intercal \boldsymbol{y})^2\right). \tag{13}$$

MacKay shows the Hessian of the logarithm of this distribution has elements

$$L_{k\ell}(\boldsymbol{y}) = \frac{\partial^2 \mathcal{D}(\boldsymbol{y})}{\partial y_k \partial y_l} = \hat{\alpha}\left(\delta_{k\ell}\pi_k - \pi_k \pi_\ell\right) + \epsilon(\mathbf{1}\mathbf{1}^\intercal)_{k\ell} \tag{14}$$

(using Kronecker's δ, and $\hat{\alpha} = \sum_k \alpha_k$. The ϵ stems from Eq. (13)). To construct a Laplace approximation of the Dirichlet in the form of a multivariate Gaussian $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ (deviating from MacKay's derivations from here on), we identify the mean $\boldsymbol{\mu}$ with the mode of the distribution,

$$\mu_k = \log \alpha_k - \frac{1}{K}\sum_{\ell=1}^K \log \alpha_\ell, \tag{15}$$

and the negative logarithm of its Hessian with $\boldsymbol{\Sigma}$. To gain a sparse approximation, we analytically invert the Hessian. To do so, we introduce the rectangular matrix $\boldsymbol{X} \in \mathbb{R}^{K \times 2}$ with elements $X_{ku} = \hat{\pi}_k \delta_{1u} + \mathbf{1}_k \delta_{2u}$ and the square matrices $\boldsymbol{A} \in \mathbb{R}^{K \times K}$ and $\boldsymbol{B} \in \mathbb{R}^{2 \times 2}$

$$\boldsymbol{A} = \mathrm{diag}(\boldsymbol{\alpha}) \qquad \text{and} \qquad \boldsymbol{B} = \begin{pmatrix} -\hat{\alpha} & 0 \\ 0 & \epsilon \end{pmatrix} \tag{16}$$

which allows us to write $\boldsymbol{L} = \boldsymbol{A} + \boldsymbol{X}\boldsymbol{B}\boldsymbol{X}^\intercal$. Both $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal with strictly positive diagonal elements, and thus invertible. Hence we can use the matrix inversion lemma, which exposes an analytically invertible 2×2 Schur complement and thus easily yields the inverse of the Hessian

$$L^{-1}_{k\ell} = \delta_{k\ell}\frac{1}{\alpha_k} - \frac{1}{K}\left[\frac{1}{\alpha_k} + \frac{1}{\alpha_\ell} - \frac{1}{K}\left(\frac{1}{\epsilon} + \sum_u^K \frac{1}{\alpha_u}\right)\right] \tag{17}$$

because this inverse is defined for all positive values of ϵ, we can safely take the limit of $\epsilon \to \infty$, i.e. $g(x) \to \delta(x)$, to the Dirac point distribution. Note that the off-diagonal elements of this matrix are suppressed with $\mathcal{O}(1/K)$, so for large K, the belief is approximately independent, with element-wise variances

$$\Sigma_{kk} = \frac{1}{\alpha_k}\left(1 - \frac{2}{K}\right) + \frac{1}{K^2}\sum_\ell^K \frac{1}{\alpha_\ell}. \tag{18}$$

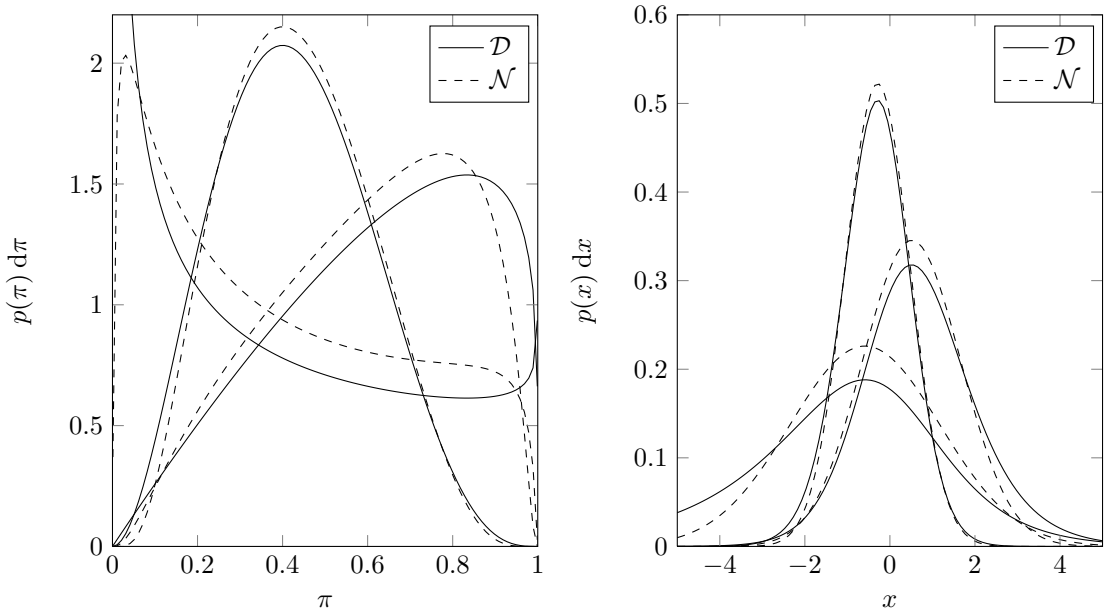

Figure 3: Laplace approximations between Gaussians and Dirichlets. **Left:** Simplex basis. **Right:** Softmax basis. The parameter choices for the Beta distributions (special 1D case of the Dirichlet) are $(a, b) = (2, 1.2); (0.5, 0.9); (3, 4)$. Under the Laplace approximation, these correspond to one-dimensional Gaussian parameters $(\mu, \sigma^2) = (0.5, 1.3); (-0.6, 3.1); (-0.3, 0.6)$. Note that the Laplace approximation matches modes and means in the softmax basis, but not in the simplex basis.

(This map is only valid for $K > 2$. In the 2D-case, a special, much simpler solution can be derived by mapping directly to the real line. See also Figure 3). It is not hard to invert this $\boldsymbol{\alpha} \twoheadrightarrow (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ map from Dirichlet to Gaussian parameters, giving

$$\alpha_k = \frac{1}{\Sigma_{kk}} \left(1 - \frac{2}{K} + \frac{e^{-\mu_k}}{K^2} \sum_{\ell}^{K} e^{-\mu_\ell} \right) \; \forall \; k = 1, \ldots, K \tag{19}$$

Figure 3 gives an intuition for the quality and defects of this approximation in the 2D case. The approximation is very good for large entries in $\boldsymbol{\alpha}$, but retains good quality even for $\alpha < 1$, which is important for topic models, where the prior is often sparse.

While it has previously been investigated in MacKay [1998], the use of this approximation here differs considerably from the setting studied in the cited paper (which dealt with evidence estimation in neural networks). Its use here amounts to the following:

- Some unobserved process with known parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$ generates data as follows:
  - Sample $\boldsymbol{x} \in \mathbb{R}^K \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathcal{N}(0; \mathbf{1}^\intercal \boldsymbol{x}, \epsilon^2)$
  - Map $\boldsymbol{\pi} = \sigma(\boldsymbol{x})$
  - Sample data c from $p(c = k \,|\, \boldsymbol{\pi}) = \boldsymbol{\pi}$
- The inference method tries to infer $\boldsymbol{x}$ thus:
  - Use the Laplace map to gain a Dirichlet belief on $\boldsymbol{\pi}$ from the Gaussian prior (15)
  - Update this belief using the data (which is trivial, due to the Dirichlet's conjugacy to the Multinomial distribution)
  - Use the Laplace map in the opposite direction, to get a Gaussian belief on $\mathbb{R}^k$, claim the resulting belief to be an approximate posterior on $\boldsymbol{x}$

Figure 4 compares this approximate scheme to an asymptotically exact Markov Chain Monte Carlo scheme (the particular MCMC method chosen for this task is elliptical slice sampling [Murray et al., 2010], which has the advantage of having no free parameters). The figure shows the 2-norm error of a point estimate for $\boldsymbol{x}$ returned by the two methods (solid lines) and error estimates constructed from the algorithms' results. For the MCMC sampler, these two estimates are the sample mean and (unbiased) sample covariance. For the Laplace approximations, the two estimates are the mean and standard deviation of the approximate Gaussian belief. The prior mean and covariance were sampled, for each experiment separately, from the standard Gaussian and the standard inverse Wishart distribution, respectively. The number of dimensions was set to $K = 10$. Note that the Laplace bridge does not show any discernible bias or

over-convergence. Its only two apparent drawbacks are its relatively bad fit for $\alpha \rightarrowtriangle 0$ and that covariance can not be captured by the Dirichlet.

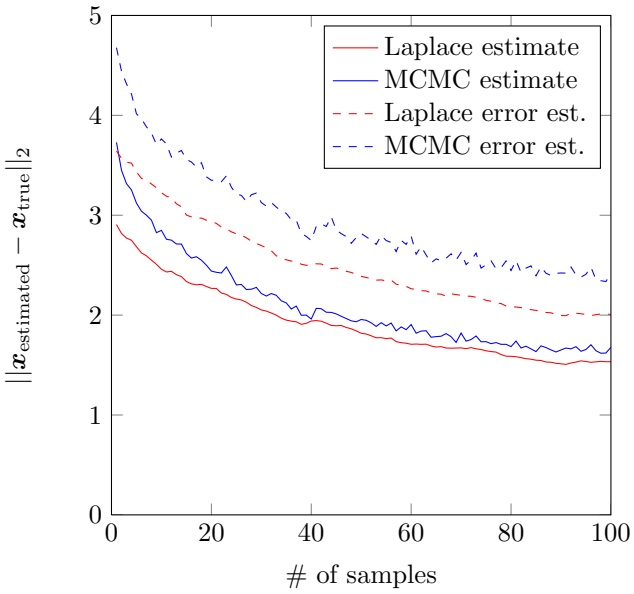

Figure 4: Convergence behaviour of approximate inference using the Laplace bridge compared to MCMC inference. Solid lines represent deviation of mean estimate (sample mean for MCMC) from ground truth, dashed lines the error estimate of the inference algorithm (one standard deviation). Both methods were initialised with a prior of $\mu = 0, \sigma = 1$. Plots are averages over 12 independent experiments.

3.4 The Wider View

Within the wider context of unsupervised learning methods, the kernel topic model establishes a connection between conditional topic models and Gaussian process latent variable models (GPLVM) [Lawrence, 2004]. GPLVMs learn mappings from data-space to a lower-dimensional space, assuming the generative model for the data in the latent space is a Gaussian process. In the case of the kernel topic model, the variational part of the inference learns a mapping from the V-dimensional space of documents defined by their words to the K-dimensional space of topics defined by their topics (Figure 1), where the documents' topics are assumed to be generated by a Gaussian process. However, in GPLVMs the map between data and their low-dimensional representation is usually assumed to be generated by another Gaussian process. In the kernel topic model, the lower dimensional distributions are discrete, and sampled from Dirichlet distributions. The kernel topic model thus performs Gaussian process regression, under a "latent Dirichlet likelihood".

4 Experiments

4.1 Euclidean and Discrete Spaces

We compare the kernel topic model to its conceptually closest competitor, the Dirichlet-Multinomial Regression (DMR) model by Mimno and McCallum [2008], which was, in the cited work, shown to give superior results to a number of other models, such as topics through time [Wang and McCallum, 2006] and the author topic model [Rosen-Zvi et al., 2004]. The dataset consists of the annual State Of The Union addresses by US presidents to the joint chambers of Congress, annotated with both the speaker's identity and the year of delivery. This dataset is interesting because it combines continuous features (time) with 44 discrete ones (author identity) and thus falls outside of the descriptive power of time drift models like the one by Wang et al. [2009]. All models used $K = 10$ topics.

For the linear model of DMR, we represented time using 100 radial basis functions spaced evenly through the time period from years 1790 to 2011, each with a width of 5 years, and used 44 binary author indicator features. For the kernel topic model, we used a rational quadratic kernel [Matérn, 1960, Rasmussen and Williams, 2006] on the space of time and author identity, assigning a distance between documents linear in time (initially using the same scale of 5 years), with an additional constant term if the authors of two documents are not the same. The rational quadratic kernel is equivalent to an infinite scale mixture of square exponential kernels: It assigns nonzero mass to functions with a range of length scales, while the the square exponential (for which the radial basis functions of the linear model are a finite-dimensional approximation) can only construct functions of a single length scale. So the kernel model is strictly more expressive than the linear model in this case. In addition, the evidence maximisation description as introduced in Section 3.2 allows an optimisation of the kernel parameters during training. For DMR, this would amount to optimising the feature set, rather than the feature parameters, which is more difficult to do efficiently.

Figure 5 shows the consequences of this additional expressive power: The kernel model captures interesting detail in the development of American interior and foreign policy, including long-term developments like the industrial revolution (bright red topic at bottom of plot) and faster developments like the Spanish-American war (light blue, top).

Figure 6 shows the development of the perplexity score [Rosen-Zvi et al., 2004] of the two models, on the training set, during training on three different datasets (see caption for details on datasets). (The vocabulary size

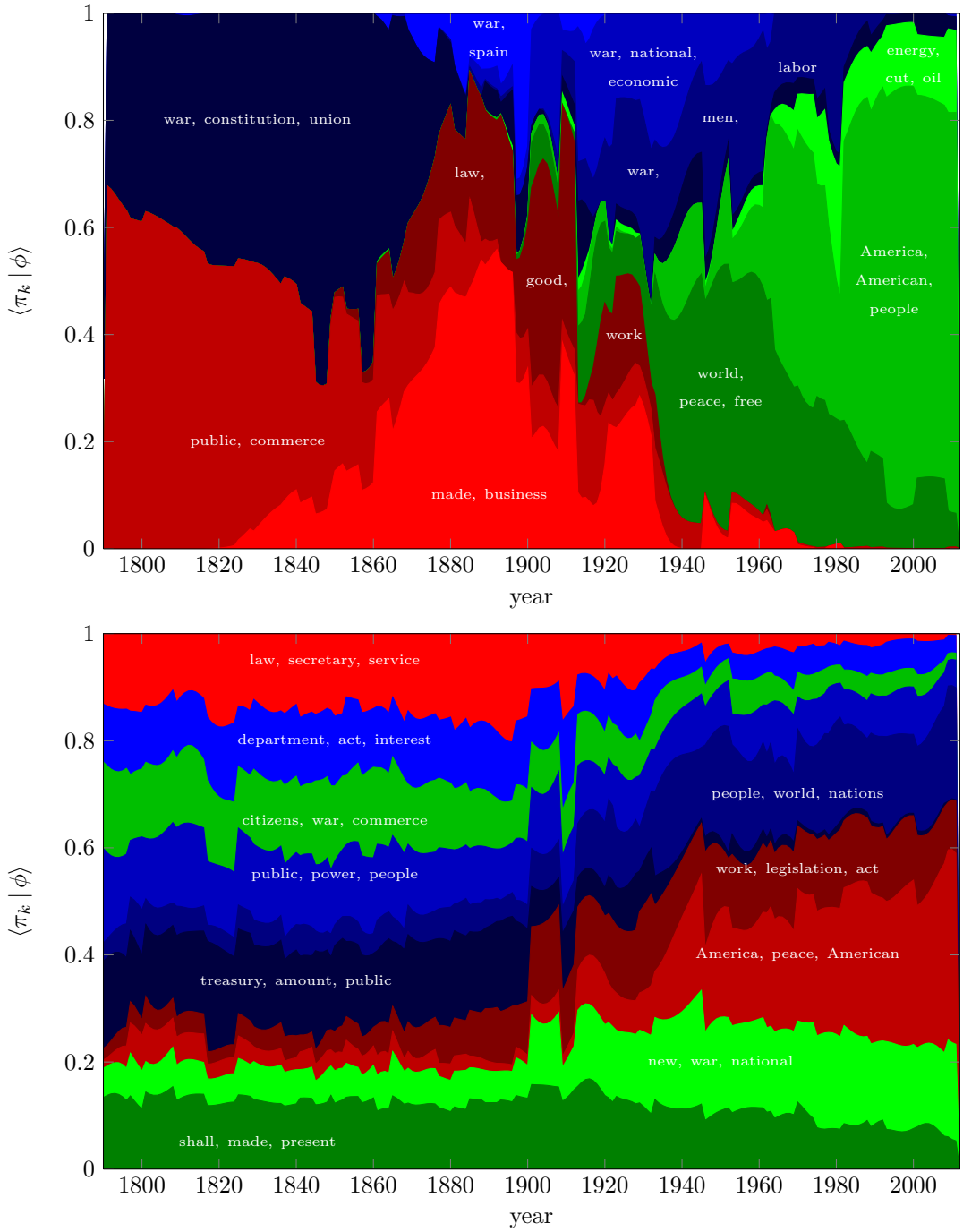

Figure 5: Inferred topic distribution of State Of The Union addresses by US American presidents. **Top:** kernel topic model using a rational quadratic kernel on the 45 dimensional space of authors and time; **Bottom:** Linear model using 100 radial basis functions in time and 44 binary author features. To generate this plot, either model was used to predict the topic distributions at the given date, conditioned on the author being the president in office at that time.

for this dataset is $V = 5000$, so the initial perplexity is 5000.) Optimisation of kernel hyperparameters was performed every tenth variational loop, and is visible as discrete steps in the plots when it has non-negligible effect, thus also giving an intuition for the model performance without hyper-optimisation.

The kernel topic model converges about as fast as the DMR, but achieves a final score about 12% below that of DMR. The two methods' runtimes are roughly com-

parable on our datasets: Both models share the LDA part. In the regression part, DMR requires numerical optimisation of the feature weights, while the kernel topic model requires inverting a large matrix.

4.2 Topics on Graphs

The kernel view on topic models also allows a relatively elegant treatment of non-Euclidean feature spaces. As an example, we construct a topic model on a graph. For our experiment, $D = 318$ documents were taken from Wikipedia's "list of probability topics[1]". We construct a positive definite kernel by embedding the documents in the $\mathbb{R}^D$ Euclidean vector space, setting

$$k(d_1, d_2) = s \exp(-\frac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{x}_2)^\intercal \boldsymbol{S}(\boldsymbol{x}_1 - \boldsymbol{x}_2)) \quad (20)$$

where the vector elements $x_{d,j}$ are the shortest distances, on the graph of links between documents, from document d to document i (links are interpreted as undirected edges, documents not linked by any path are assigned infinite distance), s and $\boldsymbol{S} = \mathrm{diag}_i(S_i)$ are parameters. Of course it is possible to define corresponding linear features, but the kernel view arguably allows a more natural way of deriving such measures.

5 Conclusion

We have presented the kernel topic model, allowing nonparametric regression of topics on document metadata of various kinds. The model is a combination of Gaussian process regression and latent Dirichlet allocation; these two conditionally independent parts are linked efficiently through a lightweight Laplace approximation. Inference in the kernel topic model is cubic in the number of documents. In large corpora, this can compare unfavourably to other feature-based topic models, but it offers superior power of expression for small and medium-sized corpora, where (approximate) analytic Gaussian process inference can even be faster than EM optimization of point estimates. An elegant side-effect of the Laplace approximation, which we have only touched upon marginally, is that it replaces the point estimates of earlier approaches with a full Bayesian belief. This means that topics can be predicted with uncertainty, and that hyperparameters of the model can be inferred consistently, using higher order maximum likelihood (maximum evidence) optimisation. A strength of the kernel formulation is its applicability to non-Euclidean feature spaces. Although not detailed this paper, this may connect our work to special types of topic models, e.g. citation [Daumé III, 2009], network [Chang and Blei, 2009], and multilingual models [Mimno et al., 2009].

[1] http://en.wikipedia.org/wiki/List_of_probability_topics

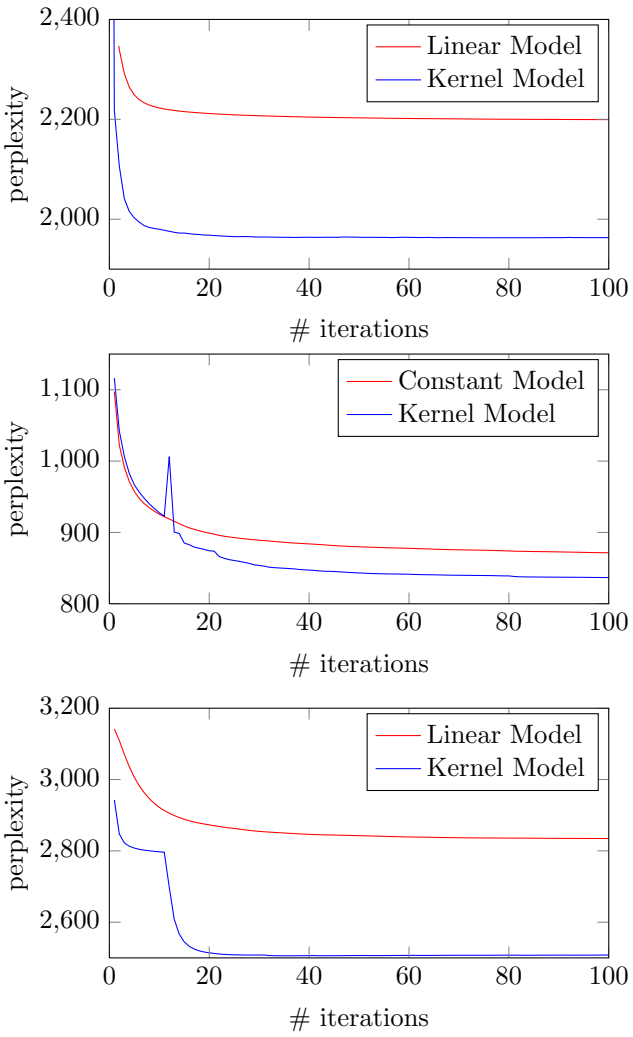

Figure 6: Perplexity of kernel topic model (blue) and linear maximum-likelihood Gaussian model or a constant LDA model (red). KTM *hyper*-parameters were optimised after every 10 iterations (the kernel regression itself is updated after every document inference). **Top:** State Of The Union dataset. Here, the hyperparameters happened to be chosen well, optimising them had negligible effect on perplexity. **Middle:** Wiki documents (Section 4.2). Note the spike in the perplexity of the kernel model in the latter plot, caused by the optimisation of hyperparameters – since the optimisation is not performed directly on the word level, the topic model crosses over into a more perplexed state at this point, but this subsequently allows a better representation. **Bottom:** NIPS dataset [Globerson et al., 2007], again showing considerable improvement after hyperparameter optimisation.

Acknowledgements

We would like to thank David MacKay and David Knowles for helpful discussions and comments.

References

D.M. Blei and J.D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

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

D.M. Blei and J.D. Lafferty. Topic models. In A. Srivastava and M. Sahami, editors, *Text Mining: Theory and Application*. Taylor & Francis, 2009.

D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, 2009.

H. Daumé III. Markov random topic fields. In *Int. Joint Conference on Natural Language Processing*, pages 293–296. Association for Computational Linguistics, 2009.

A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean Embedding of Co-occurrence Data. *Journal of Machine Learning Research*, 8:2265–2295, 2007.

T.L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228, 2004.

M. Hoffman, D. Blei, and F. Bach. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23, pages 856–864, 2010.

N.D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, volume 16, 2004.

D.J.C. MacKay. Choice of basis for Laplace approximation. *Machine Learning*, 33(1):77–86, 1998.

B. Matérn. Spatial variation. *Meddelanden från statens Skogsforskningsinstitut*, 49(5), 1960.

D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, 2008.

D. Mimno, H.M. Wallach, J. Naradowsky, D.A. Smith, and A. McCallum. Polylingual topic models. In *Conference on Empirical Methods in Natural Language Processing: Volume 2*, pages 880–889. Association for Computational Linguistics, 2009.

I. Murray, R.P. Adams, and D.J.C. MacKay. Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

Y.W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 19:1353, 2007.

C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proc. of Uncertainty in Artificial Intelligence*. Citeseer, 2009.

X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, KDD '06, pages 424–433, 2006.

J. Zhu and E.P. Xing. Conditional topic random fields. In *International Conference on Machine Learning*, 2010.