Journal of Machine Learning Research x (2018) y-z Submitted 00/00; Published 00/00

Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model

Sophie Burkhardt BURKHARDT@INFORMATIK.UNI-MAINZ.DE
University of Mainz
Department of Computer Science
Staudingerweg 9
55128 Mainz, Germany

Stefan Kramer KRAMER@INFORMATIK.UNI-MAINZ.DE
University of Mainz
Department of Computer Science
Staudingerweg 9
55128 Mainz, Germany

Editor:

Abstract

Recent work on variational autoencoders (VAEs) has enabled the development of generative topic models using neural networks. Topic models based on latent Dirichlet allocation (LDA) successfully use the Dirichlet distribution as a prior for the topic and word distributions to enforce sparseness. However, there is a trade-off between sparsity and smoothness in Dirichlet distributions. Sparsity is important for a low reconstruction error during training of the autoencoder, whereas smoothness enables generalization and leads to a better log-likelihood of the test data. Both of these properties are encoded in the Dirichlet parameter vector. By rewriting this parameter vector into a product of a sparse binary vector and a smoothness vector, we decouple the two properties, leading to a model that features both a competitive topic coherence and a high log-likelihood. Efficient training is enabled using rejection sampling variational inference for the reparameterization of the Dirichlet distribution. Our experiments show that our method is competitive with other recent VAE topic models.

Keywords: Variational autoencoders, Topic models, Dirichlet distribution, Reparameterization, Generative models

1. Introduction

Generative models have been successfully employed for the modeling of text data since they allow to learn a model of the data and thus are highly interpretable. The arguably most successful model of this kind has been latent Dirichlet allocation (LDA, Blei et al., 2003), which represents each document as a mixture of different topics. This model is typically trained using either Markov chain Monte Carlo (Griffiths and Steyvers, 2004) or variational Bayes (Blei et al., 2003). Recently, several approaches were put forward to train LDA using variational autoencoder neural networks (Miao et al., 2016; Srivastava and Sutton, 2017), thus making them more easily adaptable and extendable by using advances from deep learning research.

Using a Dirichlet distribution as a prior for the latent variables has advantages and disadvantages. The biggest advantage is the sparsity that it encourages in the latent variables. However, we find that the desired sparsity is at odds with the generalization performance of the model. This was previously noticed by Wang and Blei (2009) who tried to decouple sparsity and smoothness in the discrete hierarchical Dirichlet process. In the case of variational autoencoders (VAEs), the problem can be analyzed in terms of the KL-divergence between two Dirichlet distributions, which diverges if the prior parameter is close to zero. Therefore, we are faced with a trade-off: Either the model is sparse and achieves a good reconstruction error, but has a high KL-divergence term and thus a low log-likelihood of the model, or, it avoids low settings of the prior and thereby achieves a better log-likelihood at the cost of a higher reconstruction error. A similar observation was made by Srivastava and Sutton (2018) who observe that a slow minimization of the KL-divergence leads to better topics. In this paper we will discuss how this trade-off can be avoided to achieve both sparsity and smoothness in a variational autoencoder Dirichlet topic model. l

One remaining problem with the training of generative models using variational autoencoders is the need for a differentiable non-centered reparameterization function (Kingma and Welling, 2014a) for the distribution of the latent variables. Unfortunately such a function is not available for the most widely used probability distribution in topic modeling, the Dirichlet distribution. A model that does not directly reparameterize the Dirichlet distribution is ProdLDA (Srivastava and Sutton, 2017) which employs a Laplace approximation for the Dirichlet distribution, thus enabling the training of a Dirichlet variational autoencoder. Other recent solutions include implicit reparameterization gradients (Figurnov et al., 2018), using an approximation for the inverse CDF (Joo et al., 2019) and using the Weibull distribution as a replacement (Zhang et al., 2018). In this work we will instead apply rejection sampling variational inference (Naesseth et al., 2017) to autoencoders.

The contributions of our paper are as follows:

1. We introduce a Dirichlet variational autoencoder based on RSVI (DVAE, see Figure 1a), which is as powerful and efficient as the Gaussian variational autoencoders except it uses the Dirichlet distribution as a prior on the latent variables.

2. We identify the trade-off between sparsity and smoothness as the main factor that prevents successful training of DVAE topic models and identify the KL-divergence between two Dirichlet distributions as the crucial factor influencing model performance.

3. We provide an adapted neural network architecture to decouple sparsity and smoothness.

4. We show that our Dirichlet variational autoencoder has an improved topic coherence, whereas the adapted sparse Dirichlet variational autoencoder has a competitive perplexity.

5. We propose to use the new topic redundancy measure to obtain further information on topic quality when topic coherence scores are high.

6. We perform an extensive experimental comparison with state-of-the-art VAE topic models to show that our model achieves the highest topic coherence.

2. Related Work

Variational autoencoders (VAEs) were first introduced by Kingma and Welling (2014b) and Rezende et al. (2014). These type of deep neural networks combine ideas from approximate Bayesian inference and deep neural networks, yielding generative models that can be trained by neural networks. Models are trained by stochastic backpropagation using the reparameterization trick. This trick allows the gradient to be backpropagated through the stochastic latent variables efficiently if a reparameterization is available as is the case for the Gaussian distribution.

Topic models based on latent Dirichlet allocation (LDA) (Blei et al., 2003) are directed generative models, which makes it possible to train such topic models using variational autoencoders. The first such neural variational document model (NVDM) was developed by Miao et al. (2016). This model successfully uses the idea of variational autoencoders to train a topic model with a Gaussian prior for the latent variables. This work was extended by Miao et al. (2017), who introduce Gaussian stick breaking priors in a model which is able to let the number of topics grow dynamically.

Another model using Dirichlet stick-breaking priors is SB-VAE (Nalisnick and Smyth, 2017), however, it requires Taylor expansion to compute the KL-divergence and a Kumaraswamy distribution to approximate the posterior. Since it uses a nonparametric stick-breaking prior instead of a parametric Dirichlet it is not directly comparable to our model.

LDA topic models use a Dirichlet distribution as a prior for the latent variables since it promotes sparsity and leads to more interpretable topics. Therefore, Srivastava and Sutton (2017) propose a VAE topic model called ProdLDA that employs a Laplace approximation for modeling a Dirichlet prior of the latent variables. As their results show, while the perplexity is not as good as that of previous models, the topic coherence is significantly improved over Gaussian VAEs. They also show that the Dirichlet prior leads to more sparseness in the document-topic distributions. This work is extended by Srivastava and Sutton (2018) who introduce a hierarchical VAE topic model. Their analysis shows that batch normalization leads to a slower minimization of the KL-divergence which improves the topic quality.

The reason that VAEs cannot be directly trained using a Dirichlet prior is that there is no simple variable transformation for the Dirichlet distribution. Whereas a Gaussian variable $z \sim N(\mu, \sigma^2)$ can be reparameterized as $z = \mu + \epsilon\sigma, \epsilon \sim N(0, 1)$ thus allowing the gradient to be backpropagated through the latent variable z, there is no such reparameterization for the Dirichlet distribution. However, there is a method to solve this problem which is based on rejection samplers (Naesseth et al., 2017), called RSVI. The Dirichlet distribution can be represented by gamma distributions for which there exists an efficient rejection sampler. The proposal function of this rejection sampler can be used for reparameterization with one minor difference: A correction term has to be added to the gradient that accounts for the difference between the proposal function and the true gamma distribution. In this paper we show how this elegant method can be used for efficient and effective training of Dirichlet variational autoencoders.

A different approach was taken by Zhang et al. (2018), who used the Weibull distribution as an approximation to the gamma distribution. This takes advantage of the fact that the Weibull and Gamma distributions have similar probability density functions and the Weibull distribution is reparameterizable. Zhang et al. (2018) also compare their model with a model trained using RSVI. Overall, their model structure is different however, with a hierarchical generative distribution and a hybrid training algorithm.

Other recent solutions include implicit reparameterization gradients by Figurnov et al. (2018) who introduce a trick to avoid inverting the CDF by differentiating it instead. Furthermore, Joo et al. (2019) use an approximation for the inverse CDF of the gamma distribution. We compare to both of these methods (see Section 5.3 for details).

Our work draws significant inspiration from Wang and Blei (2009). This work is concerned with the HDP-LDA topic model which corresponds to a hierarchical Dirichlet process (HDP) applied to text, resulting in the nonparametric version of latent Dirichlet allocation (Blei et al., 2003). The model represents topics by multinomial distributions over words $\beta \sim Dir(\gamma)$, where β is the parameter vector of the multinomial distribution which is drawn from a Dirichlet distribution with parameter γ. The decoupling of sparsity and smoothness is applied to these probability distributions β, such that first the words are identified that contribute to the topic and then the Dirichlet distribution is only over the sparse word vector, i.e. a reduced probability simplex. The sparsity is modeled by assuming the topics are generated as $\beta \sim Dir(b \cdot \gamma)$, where b is a vector of Bernoulli variables with a Beta prior. In contrast, we apply the decoupling to the document-topic distributions instead of the word-topic distributions.

3. Background

3.1 LDA

Please note that the notation in this subsection follows the established conventions in topic modeling work. The rest of the paper may use the same variable names in different contexts.

Topic models based on latent Dirichlet allocation (LDA) describe each document in a collection of text documents as a mixture of topics. This means that each document d is associated with a document-topic distribution $\theta_d \sim Dir(\alpha)$ with a Dirichlet prior. For each word w_i, a topic z_i is drawn from θ_d and the word w_i is drawn from a topic-word distribution ϕ_{z_i} associated with the topic z_i. The joint distribution of the latent variables z, θ and ϕ is then optimized to maximize the marginal log-likelihood of the data D under the given model.

$$L(D|\theta, \phi, z) = \sum_{d \in D} \log p(d|z, \theta, \phi) \tag{1}$$

3.2 Variational autoencoders

The notation for variational autoencoders (VAEs) unfortunately does not correspond well to the common notation in the context of LDA. In VAEs the document-topic distribution is generated by the encoder network parameterized by ϕ, whereas the topic-word distribution is given through the decoder network parameterized by θ. Since this paper mainly focuses on VAEs, we adopt the notation common to VAEs here, although it is exactly reversed compared to the topic model notation.

Following Kingma and Welling (2014b), we define a generative model with a latent variable z and an observed input x that are generated from a generative distribution $p_\theta(z)p_\theta(x|z)$

with parameters θ. This generative model corresponds to the reconstruction/decoder part of the VAE, whereas the encoder corresponds to the variational approximation $q_\phi(z|x)$ parameterized by ϕ.

VAEs are generative models similar to LDA topic models. Accordingly, the input data x is assumed to be generated from a conditional distribution $p_\theta(x|z)$. The latent topic vectors z are generated from a prior distribution $p_\theta(z)$. Generally, it is assumed that the integral of the marginal likelihood $p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$ is intractable as well as the posterior and the integrals required for mean-field variational Bayes. In the case of LDA, it is possible to train the model using mean-field variational Bayes, however, this restricts the model and does not easily allow for certain modifications. For example, using VAEs, it is easily possible to lift the restriction of the mixture model by removing the constraint on the parameters θ, which makes the model more expressive (Srivastava and Sutton, 2017).

The marginal likelihood of the model is given by

$$\log p(x) = \sum_{i=1}^{N} \log p(x_i) =$$

$$\sum_{i=1}^{N} \left(KL(q(z_i|x_i)||p(z_i|x_i))) + L(\theta, \phi; x_i) \right),$$

where N is the number of documents in the dataset.

Since the KL-divergence is always positive, the second term is the variational lower bound $\mathcal{L}$, which can be written as

$$\begin{aligned}
&\log p(x_i) \geq \mathcal{L}(\theta, \phi; x_i) = \\
&E_{q(z|x_i)}[-\log q(z|x_i)] + E_{q(z|x_i)}[\log p(x_i, z)] = \\
&- KL(q(z_i|x_i)||p(z_i))) + E_{q(z_i|x_i)}[\log p(x_i|z_i)],
\end{aligned}$$

where the first term in the second line corresponds to the entropy and the second term in the last line corresponds to the negative reconstruction error, whereas the KL-divergence acts as a regularizer that keeps the variational distribution close to the prior. To form an estimate of the ELBO, we need to sample z from q_ϕ and calculate the ELBO by accumulating the contributions of the samples:

$$\tilde{\mathcal{L}}(\theta, \phi; x_i) = \frac{1}{L} \sum_{l=1}^{L} (\log p(x_i, z_{i,l}) - \log q(z_{i,l}|x_i)), \tag{2}$$

where L denotes the number of samples.

3.3 Rejection sampling variational inference

The idea of rejection sampling variational inference (RSVI) is to use the proposal function of a rejection sampler as the reparameterization function. This makes it possible to use probability distributions that do not have non-central differentiable reparameterizations such as the Dirichlet distribution. It also utilizes statistical research into efficient rejection samplers.

The Dirichlet distribution can be simulated using gamma random variables since if $z_i \sim \Gamma(\alpha_i)$ then $\tilde{z}_{1:K} = \frac{z_{1:K}}{\sum_i z_i} \sim Dirichlet(\alpha_{1:K})$. This is relevant because for the gamma distribution there exists an efficient rejection sampler:

$$z = h_\Gamma(\epsilon, \alpha) := (\alpha - \frac{1}{3})(1 + \frac{\epsilon}{\sqrt{9\alpha - 3}})^3, \epsilon \sim N(0,1) \tag{3}$$

However, using this proposal function is not equivalent to using a gamma distribution because some samples are rejected. Therefore we need the distribution of an accepted sample $\epsilon \sim s(\epsilon)$, which is obtained by marginalizing over the uniform variable u of the rejection sampler.

$$\pi(\epsilon; \theta) = \int \pi(\epsilon, u; \theta) du = s(\epsilon) \frac{q(h(\epsilon, \theta))}{M_\theta r(h(\epsilon, \theta))}, \tag{4}$$

where r is the proposal function for the rejection sampler, $z = h(\epsilon, \theta), \epsilon \sim s(\epsilon)$ is the reparameterization of the proposal distribution and M_θ is a constant used in the rejection sampler. Using this, the ELBO can be rewritten as

$$\begin{gathered} E_{q(z|x_i)}[-\log q(z|x_i)] + E_{q(z|x_i)}[\log p(x_i, z)] = \\ E_{q(z|x_i)}[-\log q(z|x_i)] + E_{\pi(\epsilon;\theta)}[\log p(x_i, h(\epsilon, \theta))] \end{gathered}$$

As Naesseth et al. (2017) show, the gradient can then be decomposed into three parts:

$$\nabla_\theta L(\theta) = g_{\text{rep}} + g_{\text{cor}} + \nabla_\theta E_{q(z|x_i)}[-\log q(z|x_i)] \tag{5}$$

where g_{rep} and g_{cor} for the case of a one-sample Monte Carlo estimator are given as

$$g_{\text{rep}} = \nabla_z \log p(x_i, z) \nabla_\theta h(\epsilon, \theta) \tag{6}$$

$$g_{\text{cor}} = \log p(x_i, z) \nabla_\theta \log \frac{q(h(\epsilon, \theta))}{r(h(\epsilon, \theta))} \tag{7}$$

and the entropy can be calculated analytically. g_{rep} corresponds to the gradient assuming that the proposal is exact and always accepted and g_{cor} corresponds to a correction part of the gradient that accounts for not using an exact proposal.

3.4 Shape augmentation for gamma distribution

The rejection sampler for the gamma distribution has higher acceptance rates for higher values of the parameter α. Naesseth et al. (2017) use this fact to augment the shape of the gamma distribution and thereby achieve a lower variance gradient. As they show, a $\Gamma(\alpha, 1)$ distributed variable z can be expressed as $z = \tilde{z} \prod_{i=1}^{B} u_i^{\frac{1}{\alpha+i-1}}$ for a positive integer B and i.i.d. uniform random variables u, where $\tilde{z} \sim \Gamma(\alpha + B, 1)$. This makes it possible to use the rejection sampler for $\alpha < 1$ and also decreases the correction term for increasing $\alpha + B$.

4. Method

4.1 Sparsity and Smoothness

The trade-off between sparsity and smoothness is a known issue in the topic modeling literature (Wang and Blei, 2009). Aspects of it were also discussed in recent literature by Srivastava and Sutton (2017, 2018). According to them, batch normalization and dropout slow down the minimization of the KL-divergence in the beginning of training. This has the effect of increasing the smoothness and thereby increases the topic coherence because the model generalizes better. A fast minimization puts a greater emphasis on sparseness which ensures a lower perplexity, but is prone to overfitting. A minimization that is too fast leads to maximal smoothness which means the Dirichlet parameter is equal to the prior. In this case no meaningful topics can be learned since the latent variables are only sampling noise. This problem is known as component collapsing.

But influencing the training process through e.g. batch normalization or dropout is only one way to control this trade-off. By decoupling both aspects explicitly we can allow both, a good generalization performance and a good fit to the data at hand. To do this, first, the topics are selected that represent the current document. Second, a distribution is sampled for the selected topics. The details are described in the next section.

4.2 DVAE Sparse

A schematic view of our methods is depicted in Figures 1a and 1c. We will now describe the sparse Dirichlet variational autoencoder (DVAE Sparse) depicted in Figure 1c. The DVAE model works analogously except that the vector b is set to a vector of only ones. The model equation is given by $p(x, z, \theta) = p(x|z, \theta) \cdot p(z|\theta) \cdot p(\theta)$. The input x is transformed to the Dirichlet parameter α and a sparsity parameter b by a neural network, the latent variables z are sampled from $Dir(b \cdot \alpha)$ and the decoder network reconstructs the original input. This is in contrast to the Gaussian autoencoder shown in Figure 1b, where the latent variables are drawn from a Gaussian distribution. To be able to backpropagate the gradients through the latent variables z, we use rejection sampling variational inference.

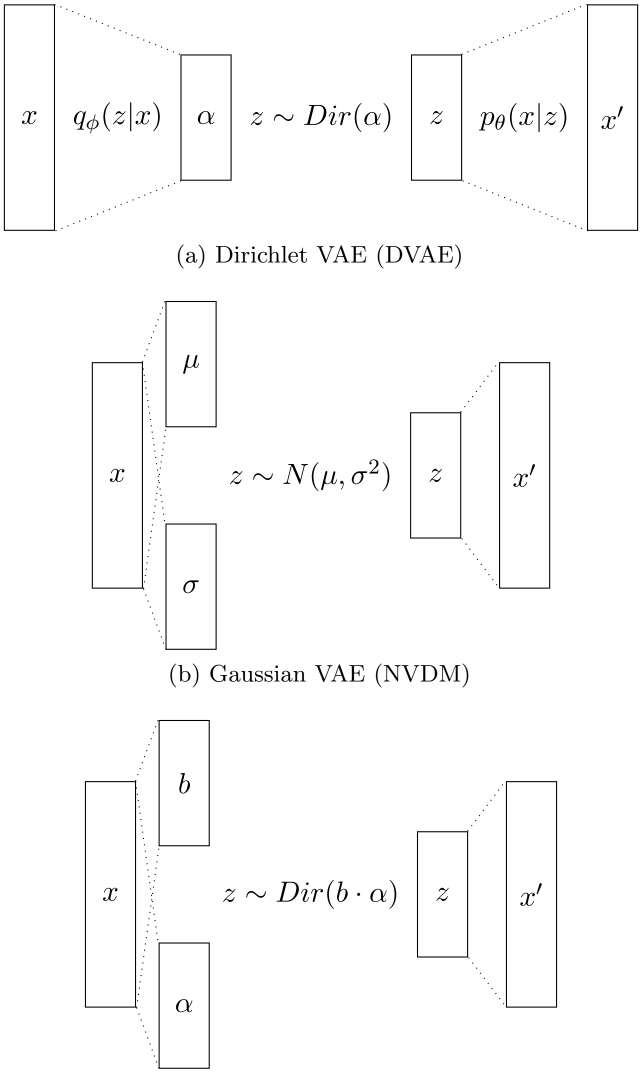
The parameter b is inspired by Wang and Blei (2009), who use a Bernoulli distributed variable. In contrast, we simply use a sigmoid function and subsequently round the result to zero or one. The distribution $Dir(b \cdot \alpha)$ is a degenerate Dirichlet distribution over the sub-simplex specified by b. If the vector b only consists of ones, this reduces to the normal Dirichlet distribution $Dir(\alpha)$.

$$b = l_2(\lambda) = \text{round}(\frac{1}{1 + e^{-\lambda}}) \tag{8}$$

We optimize α in an unconstrained domain by letting

$$\alpha = \log(1 + \exp(a)), \tag{9}$$

where $a = l_1(\lambda)$ is a linear transformation of λ and λ is the transformation of the input through the neural network $\lambda = f_{\text{MLP}}(x)$.



(a) Dirichlet VAE (DVAE)

(b) Gaussian VAE (NVDM)

(c) Sparse Dirichlet VAE (DVAE Sparse)

Figure 1: Schematic figure of Dirichlet VAE, Gaussian VAE and the proposed Sparse Dirichlet VAE. For the Dirichlet VAE the input is transformed by a neural network to yield parameter α. The latent variables z are subsequently sampled from $Dir(\alpha)$. Finally the input is reconstructed as x'. The Gaussian VAE has parameters μ and σ instead and uses a Gaussian distribution.

The ELBO is given by:

$$L(\theta, \phi; x_i) = \tag{10}$$

$$-KL(q(z|x_i)||p(z))) + E_{q(z|x_i)}[\log p(x_i|z)] = \tag{11}$$

$$-KL(q(z|x_i)||p(z))) + E_{\pi(\epsilon;\phi)}[\log p(x_i|h_\Gamma(\epsilon, \phi))] \tag{12}$$

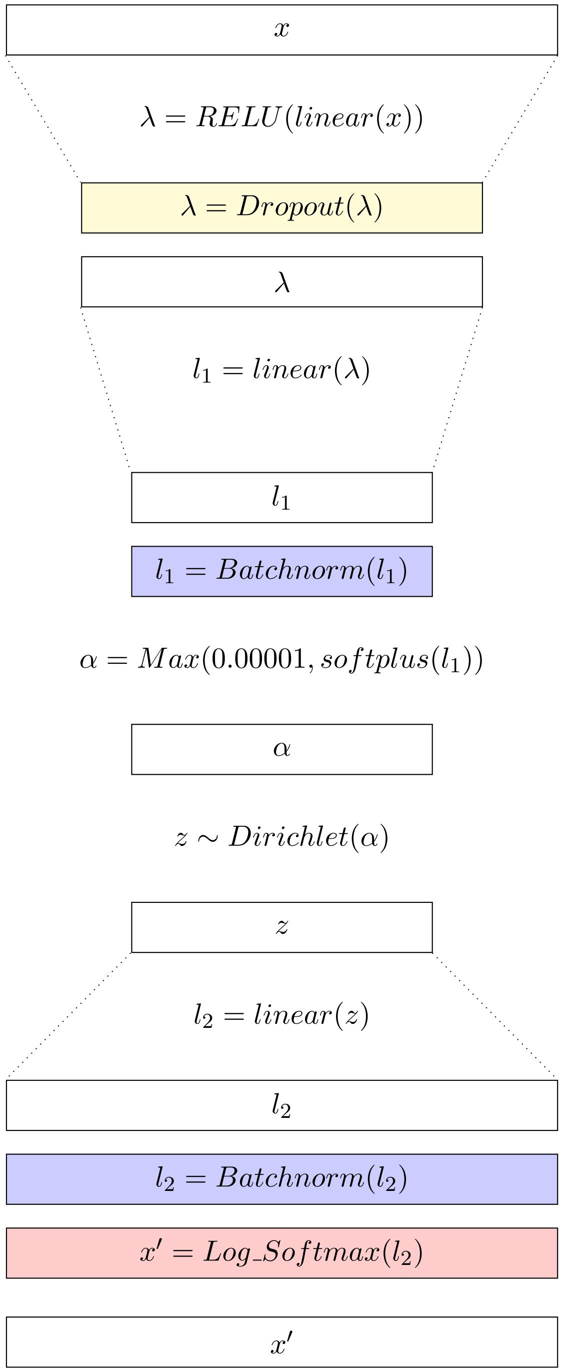


Figure 2: Illustration of the neural network architecture used for DVAE, ProdLDA, the implicit reparameterization gradient method, the Weibull VAE method and the inverse CDF gradient method. The architecture was proposed by Srivastava and Sutton (2017) because it allows to train topic models with a high learning rate, thus avoiding local minima and converging quickly.

Here, the expectation representing the negative reconstruction error is rephrased with respect to the distribution of the accepted sample π and the latent variables z are reparameterized by $h_\Gamma(\epsilon, \phi)$ given in Equation 3. $\pi(\epsilon; \phi)$, the distribution of the accepted sample ϵ, is obtained by marginalizing over the uniform variable u. Here, we use the parameter ϕ since our encoder network is parameterized by ϕ.

$$\pi(\epsilon; \phi) = \int \pi(\epsilon, u; \phi) du = s(\epsilon) \frac{q(h_\Gamma(\epsilon, \phi))}{r(h_\Gamma(\epsilon, \phi))} \tag{13}$$

The gradient of the ELBO for the generative/decoder network is given as follows:

$$\nabla_\theta L(\theta, \phi; x_i) = \nabla_\theta E_{q(z|x_i)}[\log p(x_i|z)] \tag{14}$$

This corresponds to the logarithm of the reconstruction probability. Gradient backpropagation for this term is unproblematic since the samples used to estimate L depend on q which is parameterized by ϕ. Therefore we do not need to take care of reparameterization for this term. The gradients of the ELBO for the the variational/encoder network are as follows:

$$\nabla_\phi L(\theta, \phi; x_i) = \nabla_\phi - KL(q(z|x_i)||p(z))) + \tag{15}$$

$$\nabla_\phi E_{\pi(\epsilon;\theta)}[\log p(x_i|h_\Gamma(\epsilon, \phi))] \tag{16}$$

Here, we need to reparameterize the latent variables z. The gradient of the KL-divergence will be discussed in Section 4.3 and the rest of the gradient is given analogously to Equation 5. Note that we now use the expectation of the conditional distribution $p(x_i|z)$ instead of the joint distribution $p(x_i, z)$, since we write the ELBO in a different way which is more common for VAEs. The gradient based on a one-sample Monte Carlo estimator is given as:

$$\nabla_\phi L(\theta, \phi; x_i) = \nabla_\phi - KL(q(z|x_i)||p(z))) + g^\phi_{\text{rep}} + g^\phi_{\text{cor}} \tag{17}$$

$$g^\phi_{\text{rep}} = \nabla_z \log p(x_i|z) \nabla_\phi h(\epsilon, \phi)$$

$$g^\phi_{\text{cor}} = \log p(x_i|z) \nabla_\phi \log \frac{q(h(\epsilon, \phi))}{r(h(\epsilon, \phi))}$$

To understand the step from Equation 16 to Equation 17, we refer to the derivation by Naesseth et al. (2017):

$$E_{\pi(\epsilon;\theta)}[f(h(\epsilon, \phi))] = \tag{18}$$

$$\int s(\epsilon) \nabla_\phi \left(f(h(\epsilon, \phi)) \frac{q(h(\epsilon, \phi); \phi)}{r(h(\epsilon, \phi); \phi)} \right) d\epsilon = \tag{19}$$

$$\int s(\epsilon) \frac{q(h(\epsilon, \phi); \phi)}{r(h(\epsilon, \phi); \phi)} \nabla_\phi f(h(\epsilon, \phi)) d\epsilon + \tag{20}$$

$$\int s(\epsilon) f(h(\epsilon, \phi)) \nabla_\phi \left(\frac{q(h(\epsilon, \phi); \phi)}{r(h(\epsilon, \phi); \phi)} \right) d\epsilon = \tag{21}$$

$$E_{\pi(\epsilon;\theta)}[\nabla_\phi f(h(\epsilon, \phi))] + \tag{22}$$

$$E_{\pi(\epsilon;\theta)}[f(h(\epsilon, \phi)) \nabla_\phi \log \frac{q(h(\epsilon, \phi); \phi)}{r(h(\epsilon, \phi); \phi)}], \tag{23}$$

where in the last step the log-derivative trick was used and $f(h(\epsilon, \phi))$ in our case corresponds to $\log p(x_i|h(\epsilon, \phi))$.

4.3 The KL-divergence

The KL-divergence can be calculated analytically, meaning that we do not have to use any samples and thus do not need to take care of reparameterization in the gradient back propagation for this term. $p(z)$ is a Dirichlet with prior α^0, whereas $q(z|x)$ is the Dirichlet parameterized by α as given by the variational network parameterized by ϕ.

$$\begin{aligned} KL[q(z|x)||p(z)] = &\log(\Gamma(\sum_k \alpha_k)) - \log(\Gamma(\sum_k \alpha_k^0))+ \\ &\sum_k \log\Gamma(\alpha_k^0) - \sum_k \log\Gamma(\alpha_k)+ \\ &\sum_k (\alpha_k - \alpha_k^0)(\Psi(\alpha) - \Psi(\sum_k \alpha_k)), \end{aligned}$$

where Ψ is the digamma function.

In our experiments we found that the analytical calculation of the KL-divergence, even though it is theoretically more appealing, does not lead to optimal results. We therefore resort to using a simple sampling approximation of the KL-divergence by calculating the KL-divergence as $KL[q(z|x)||p(z)] = q(z|x)\log\frac{q(z|x)}{p(z)}$. Note that test results are calculated using the analytical KL-divergence to enable a fair comparison. Sampling the KL-divergence leads to more stable gradients, however, it introduces a second correction term for the gradient of the encoder network, since we are now using the rejection sampler again to produce z.

We therefore rewrite the ELBO as:

$$\begin{aligned} &L(\theta, \phi; x_i) = \\ &- KL(q(z|x_i)||p(z))) + E_{q(z|x_i)}[\log p(x_i|z)] = \\ &E_{\pi(\epsilon;\phi)}[\log\frac{p(h_\Gamma(\epsilon,\phi))}{q(h_\Gamma(\epsilon,\phi)|x)}] + E_{\pi(\epsilon;\phi)}[\log p(x_i|h_\Gamma(\epsilon,\phi))] \end{aligned}$$

The gradient of the KL-divergence is now given as

$$\nabla_\phi KL[q(h_\Gamma(\epsilon,\phi))|x)||p(h_\Gamma(\epsilon,\phi))] = \nabla_\phi \log\frac{p(h_\Gamma(\epsilon,\phi))}{q(h_\Gamma(\epsilon,\phi)|x)} + g^\phi_{\text{kl-cor}}, \tag{24}$$

where $g^\phi_{\text{kl-cor}}$ is defined as:

$$g^\phi_{\text{kl-cor}} = \log\frac{p(h_\Gamma(\epsilon,\phi))}{q(h_\Gamma(\epsilon,\phi)|x)}\nabla_\phi \log\frac{q(h_\Gamma(\epsilon,\phi))}{r(h_\Gamma(\epsilon,\phi))} \tag{25}$$

The log fraction in this correction term is the same as the one for g^ϕ_{cor}, however, it is weighted by the KL-divergence instead of the reconstruction probability. The derivation is again done using Equation 23 except now $f(h(\epsilon,\phi))$ corresponds to $\log\frac{p(z)}{q(h(\epsilon,\phi)|x)}$. Note, that the two terms in Equation 24 correspond to a reparameterization part and a score function part. If the proposal is equal to the distribution and always accepted this reduces to the reparameterization gradient, whereas if the proposal is always rejected, this reduces to the score function gradient, also known as REINFORCE[1].

[1] https://casmls.github.io/general/2017/04/25/rsvi.html

4.4 Neural network architecture

The architecture of the neural network for the DVAE variational autoencoder is illustrated in Figure 2.

1. Encoder: The input is transformed using a RELU-layer with dropout. The result is linearly transformed and batch normalization is applied. Batch normalization is crucial for the convergence of the model. VAEs are prone to so-called component collapsing, a well-known problem (Bowman et al., 2016). Using batch normalization in combination with a relatively high learning rate (10^{-2}–10^{-4}), helps to avoid component collapsing and leads to a fast convergence (Srivastava and Sutton, 2017). A softplus transformation is applied because the Dirichlet parameter needs to be positive. Furthermore, a minimum value of 0.00001 is used for the Dirichlet parameter α which is necessary for stability of the gradients. The latent variables z are sampled from the Dirichlet distribution.

2. Decoder: The decoder is structured as follows: The latent variables z are transformed linearly. Again, batch normalization is applied. The final result is obtained through a log-softmax transformation. Note that the decoder weights are not restricted to be on the probability simplex as would be necessary if a Dirichlet prior was put on the topic-word distributions. This lifts the restriction of the mixture model as proposed by Srivastava and Sutton (2017) (The same was done in the NVDM model by Miao et al. (2016), although it was not explicitly discussed.). This is why in our experiments we compare to the ProdLDA model of Srivastava and Sutton (2017) and not to AVITM, which puts the simplex constraint on the decoder weights. Nevertheless, we can use the decoder weights to infer our topics since it is sufficient to rank the words according to the respective weights for each topic. The weights do not have be given a probabilistic interpretation as in LDA models. The words with the largest weights for each latent topic, are usually representative for that topic. It was shown by Srivastava and Sutton (2017) and is confirmed in our experiments that the non-mixture version yields lower perplexity but higher topic coherence. The reason for this is suspected to be the lifting of the mixture model contraint which makes the model more expressive.

4.5 Component collapsing

As mentioned in the previous section, VAEs often suffer from the well-known issue of component collapsing (Bowman et al., 2016). This is the problem of a local minimum that is usually encountered early on in the training process and which the model cannot escape as the training continues. It occurs because it is often easier for the model to minimize the KL-divergence than to minimize the reconstruction error. This leads to parameters that are equal to the prior and thus, many or all topics are the same. This means that there are a lot of redundant topics, but the perplexity is relatively low. To prevent this from happening, a common method is annealing of the KL-divergence (Bowman et al., 2016) to slowly introduce it into the loss function. Other methods include batch normalization and dropout (Srivastava and Sutton, 2017, 2018). In our experiments, we found that batch normalization had the largest effect. None of the implemented Dirichlet autoencoders learn meaningful topics without batch normalization. NVDM does not employ batch normalization, but has to use a lower learning rate to prevent component collapsing.

5. Experiments

5.1 Experimental settings

The number of neurons was set to 100 for all hidden layers in the neural network models. We used a single sample for all methods. A batch size of 200 was used for the training of all models. Training was monitored on a validation set with early stopping and a look-ahead of 30 iterations. We used the Adam optimizer for training. For NVDM each training iteration spans ten iterations for the decoder and ten iterations for the encoder. DVAE and ProdLDA optimize decoder and encoder jointly. Since we noticed that DVAE has no difference in performance between using the correction part of the gradient when shape augmentation is used, we omitted that part in the experiments to save unnecessary computation. Similar to Srivastava and Sutton (2017), we employ batch normalization to improve the convergence of our model. The learning rate is optimized for all neural network methods using Bayesian optimization.

5.2 Datasets

- `20news`: We used the same version of this dataset as Srivastava and Sutton (2017). It is a widely used text dataset consisting of news articles that are grouped into 20 different categories. It has 11,000 training instances and a 2,000 word vocabulary.

- `NIPS`: The NIPS dataset was retrieved in a preprocessed format from the UCI Machine Learning Repository (Perrone et al., 2016). It consist of 5,812 documents and has 11,463 features. It contains text data from NIPS conference papers published between 1987 and 2015. Compared to the other datasets, individual documents are large. 1,000 documents were separated as a test set.

- `KOS`: The 3,430 blog entries of this dataset were originally extracted from `http://www.dailykos.com/`, the dataset is available in the UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/datasets/Bag+of+Words`. The number of features is 6,906.

- `Rcv1`[2]: This corpus contains 810,000 documents of Reuters news stories. We separated 10,000 documents as a testset and pruned the vocabulary to 10,000 words after stopword removal following Miao et al. (2016).

5.3 Methods

5.3.1 Dirichlet Autoencoders

We compare to four different autoencoders that all have exactly the same architecture and hyperparameters as DVAE. They are only different with respect to the modeling of the Dirichlet prior on the latent variables.

[2] `http://trec.nist.gov/data/reuters/reuters.html`

1. ProdLDA: This method uses a Laplace approximation to approximate the Dirichlet prior. We use the implementation provided by the authors[3].

2. Inverse CDF (Joo et al., 2019): This method builds on previous work by Knowles (2015). If $X \sim \Gamma(\alpha, \beta)$ and $F(x; \alpha, \beta)$ is a CDF of the random variable X, the inverse CDF can be approximated by $F^{-1}(u; \alpha, \beta) \approx \beta^{-1}(u\alpha\Gamma(\alpha))^{1/\alpha}$, where $u \sim$ Uniform(0,1). The approximated inverse CDF can then be used as a reparameterization function.

3. Implicit reparameterization gradients (Figurnov et al., 2018): This method introduces a trick to avoid inverting the CDF by differentiating it instead. It is integrated into the tensorflow library and can be used out-of-the-box.

4. Weibull autoencoder (Zhang et al., 2018): This method uses the Weibull distribution instead of the Gamma distribution because it has a similar PDF and the Weibull distribution can be reparameterized easily. The Weibull PDF is given as

$$P(x|k, \lambda_{\text{weibull}}) = \frac{k}{\lambda_{\text{weibull}}^k} x^{k-1} \exp\left(x/\lambda_{\text{weibull}}\right)^k \tag{26}$$

   The reparameterization for $x \sim Weibull(k, \lambda_{\text{weibull}})$ is

$$x = \lambda(-\ln(1-\epsilon))^{1/k}, \epsilon \sim \text{Uniform(0,1)} \tag{27}$$

   and the KL-divergence from the gamma distribution is analytically computed as

$$KL(Weibull(k, \lambda_{\text{weibull}})||Gamma(\alpha, \beta)) =$$
$$\alpha \ln \lambda_{\text{weibull}} - \frac{\gamma\alpha}{k} - \ln k - \beta\lambda_{\text{weibull}}\Gamma\left(1 + \frac{1}{k}\right) + \gamma + 1 + \alpha \ln \beta - \ln \Gamma(\alpha).$$

   The parameters λ_{weibull} and k are computed through the neural network as

$$\lambda_{\text{weibull}} = \log(1 + \exp(b)), k = \log(1 + \exp(c)), \tag{28}$$

   where $b = l_2(\lambda)$ and $c = l_3(\lambda)$, where l_2 and l_3 are linear transformations of λ which is the same as in Equation 9.

Additionally, k is restricted to be greater or equal to 0.1 because for values of k that are lower than that the method becomes unstable.

5.3.2 Other comparison methods

Additionally, we compare to two other algorithms:

1. NVDM: This is the Gaussian VAE topic model. We used the code provided by the authors[4].

2. SCVB: As a representative of the original LDA models, we compare to SCVB (Foulds et al., 2013), a stochastic variant of collapsed variational Bayes. This method is trained online using minibatches in the same way as the neural networks. Therefore the method is applicable to large datasets. We used our own Java reimplementation.

[3] `https://github.com/akashgit/autoencoding_vi_for_topic_models`

[4] `https://github.com/ysmiao/nvdm`

5.4 Evaluation

All models are evaluated on three measures, perplexity, topic redundancy and topic coherence. The per-word perplexity is calculated from the ELBO as $\exp(-\frac{1}{N_w}\sum_d^D \log p(d))$, where N_w is the number of words in the corpus. We use the analytical KL-divergence for all models to calculate the perplexity to allow a fair comparison.

The topic coherence is calculated following Srivastava and Sutton (2017) using the normalized pointwise mutual information (NPMI), averaged over all pairs of words of all topics, where the NPMI is set to zero for the case that a word pair does not occur together. We calculate the topic coherence on the test set using the 10 first words of each topic, i.e. the 10 words with the highest probability for that topic. The NPMI for topic t is given as follows:

$$\mathrm{NPMI}(t) = \sum_{j=2}^{N}\sum_{i=1}^{j-1} \frac{\log \frac{P(w_j,w_i)}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)}, \tag{29}$$

where N is the number of words in topic t, w_i is the ith word of topic t and $P(w_i, w_j)$ is the probability of words w_i and w_j occurring together in the test set, which is approximated by counting the number of documents where both words appear together and dividing the result by the total number of documents.

The topic redundancy corresponds to the average probability of each word to occur in one of the other topics of the same model. The redundancy for topic k is thus given as

$$R(k) = \frac{1}{K-1}\sum_{i=1}^{N}\sum_{j\neq k} P(w_{ik}, j), \tag{30}$$

where $P(w_{ik}, j)$ is one if the ith word of topic k, w_{ik}, occurs in topic j and otherwise zero, and $K-1$ is the number of topics excluding the current topic.

5.5 Experimental results

5.5.1 Perplexity

We compare our models, DVAE and DVAE Sparse, to the Laplace approximation *ProdLDA*, the Gaussian VAE *NVDM* and the online LDA model *SCVB* in Tables 1 and 2. Additional results are in the Appendix in Tables 4 and 5. The overall best perplexity is achieved by NVDM as can be seen from the average ranks in each table, whereas the Weibull VAE has the worst perplexity, closely followed by DVAE. Implicit gradients VAE and inverse CDF VAE are worse than NVDM and DVAE Sparse, but a little bit better than DVAE and Weibull VAE. The reason for the high perplexity of DVAE is the KL-divergence part of the ELBO (Equation 12). The KL-divergence between two Dirichlet distributions can have extreme values as shown in Figure 3. If the Dirichlet parameter is close to zero, the KL-divergence and therefore also the perplexity is high. As Tables 1 and 2 show, the DVAE Sparse model successfully circumvents this problem by decoupling sparsity and smoothness. The perplexity of DVAE Sparse, while not as good as that of NVDM, is mostly better or at least comparable to that of the ProdLDA model.

dataset	DVAE	DVAE Sp.	Impl.	Inv.	Weibull	ProdLDA	NVDM	SCVB
	Perplexity							
					50 topics			
KOS	7872	2408	2894	2598	7795	2811	**2097**	2365
NIPS	2779	2197	2327	2141	3.55E+52	2303	**1933**	3973
20news	5009	1066	1842	1533	5209	1128	**813**	817
Rcv1	5885	1464	1635	1547	2344	2265	**907**	2034
					200 topics			
KOS	2.59E+05	2397	3238	3233	3.11E+04	3161	**2216**	2222
NIPS	3110	2152	2250	2135	8.50E+167	2532	**1957**	5787
20news	4.93E+05	1075	3021	3033	1.82E+08	1271	**875**	951
Rcv1	8.66E+04	1614	1471	1.07E+10	4426	2559	**891**	2779
av. rank	7.125	2.875	4.750	4.375	7.375	4.500	**1.000**	4.000
	Topic Coherence							
					50 topics			
KOS	0.194	0.050	0.131	0.130	0.067	**0.218**	0.070	0.145
NIPS	**0.313**	0.295	0.276	0.224	0.167	0.293	0.070	0.116
20news	**0.324**	0.248	0.322	0.304	0.271	0.223	0.140	0.244
Rcv1	**0.381**	0.293	0.340	0.162	0.216	0.140	0.110	0.239
					200 topics			
KOS	**0.212**	0.072	0.118	0.076	0.086	0.071	0.060	0.107
NIPS	**0.280**	0.234	0.227	0.208	0.176	0.277	0.060	0.105
20news	**0.307**	0.174	0.298	0.281	0.258	0.150	0.140	0.204
Rcv1	**0.373**	0.115	0.237	0.138	0.069	0.370	0.051	0.263
av. rank	1.125	4.875	3.000	4.625	5.375	4.500	7.750	4.750
	Topic Redundancy							
					50 topics			
KOS	0.017	0.014	0.024	**0.011**	0.020	0.040	0.012	0.110
NIPS	0.016	0.013	0.008	**0.003**	0.123	0.019	0.003	0.088
20news	0.026	0.025	0.022	0.018	0.039	0.020	**0.007**	0.037
Rcv1	0.009	0.269	0.002	**0.001**	0.095	0.017	0.002	0.039
					200 topics			
KOS	0.026	0.013	0.016	0.010	0.020	**0.006**	0.010	0.088
NIPS	0.021	0.010	0.012	0.008	0.022	0.018	**0.003**	0.058
20news	0.030	0.016	0.030	0.027	0.044	0.013	**0.011**	0.046
Rcv1	0.007	0.017	0.104	0.817	0.668	0.253	**0.004**	0.024
av. rank	4.875	4.125	4.375	2.750	6.875	4.375	1.625	7.000

Table 1: Perplexity, topic coherence and topic redundancy for **Dirichlet prior 0.02**.

dataset	DVAE	DVAE Sp.	Impl.	Inv.	Weibull	ProdLDA	NVDM	SCVB
	Perplexity							
					50 topics			
KOS	4190	2314	2399	2242	4274	2555	**2097**	2579
NIPS	2686	2185	2263	2064	1.83E+141	2250	**1933**	4290
20news	2158	1080	2505	2443	1546	1041	**813**	903
Rcv1	3449	1646	1506	1355	1815	2165	**907**	1986
					200 topics			
KOS	2.06E+04	2758	2713	3807	1.21E+12	3127	**2216**	2500
NIPS	2744	2072	2076	**1936**	1.70E+249	2422	1957	5911
20news	1.71E+04	1100	2257	1678	1.13E+09	1294	875	**874**
Rcv1	1.01E+04	1196	1356	1675	2463	2146	**891**	3317
av. rank	6.875	3.250	4.500	3.625	7.000	4.750	1.250	4.750
	Topic Coherence							
					50 topics			
KOS	0.185	0.117	0.134	0.127	0.051	**0.220**	0.070	0.137
NIPS	**0.313**	0.284	0.284	0.233	0.306	0.312	0.070	0.111
20news	**0.337**	0.248	0.318	0.258	0.243	0.286	0.140	0.234
Rcv1	**0.389**	0.345	0.325	0.083	0.311	0.194	0.110	0.249
					200 topics			
KOS	**0.204**	0.104	0.113	0.137	0.133	0.140	0.060	0.110
NIPS	**0.281**	0.243	0.240	0.208	0.186	0.224	0.060	0.101
20news	**0.303**	0.222	0.302	0.270	0.264	0.216	0.140	0.188
Rcv1	**0.364**	0.121	0.321	0.187	0.147	0.200	0.051	0.257
av. rank	1.125	4.875	3.125	4.875	5.125	3.500	7.750	5.625
	Topic Redundancy							
					50 topics			
KOS	0.021	0.024	0.010	**0.007**	0.016	0.055	0.012	0.111
NIPS	0.014	0.014	0.009	**0.002**	0.075	0.018	0.003	0.086
20news	0.027	0.035	0.033	0.063	0.032	0.027	**0.007**	0.035
Rcv1	0.007	0.011	0.005	0.004	0.012	0.023	**0.002**	0.036
					200 topics			
KOS	0.030	0.033	0.019	0.031	0.032	0.014	**0.010**	0.084
NIPS	0.018	0.010	0.009	0.003	0.033	0.014	**0.003**	0.062
20news	0.024	0.033	0.039	0.034	0.050	0.022	**0.011**	0.065
Rcv1	0.008	0.023	0.017	0.023	0.262	0.019	**0.004**	0.021
av. rank	3.750	5.500	3.500	3.750	6.125	4.500	1.375	7.500

Table 2: Perplexity, topic coherence and topic redundancy for **Dirichlet prior 0.1**.

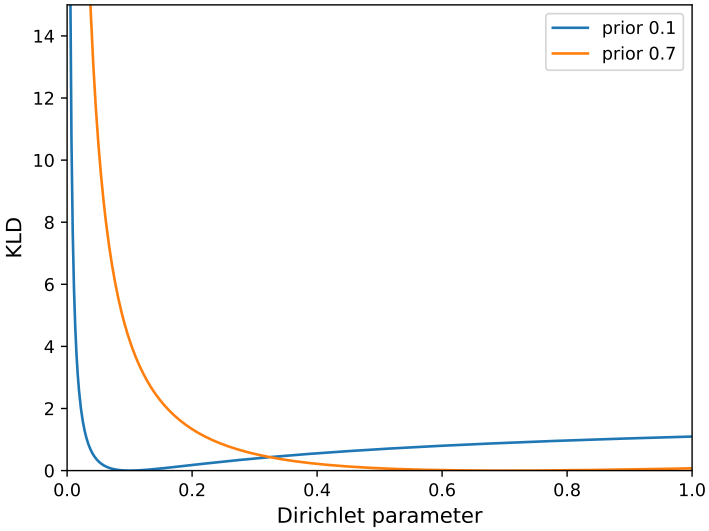
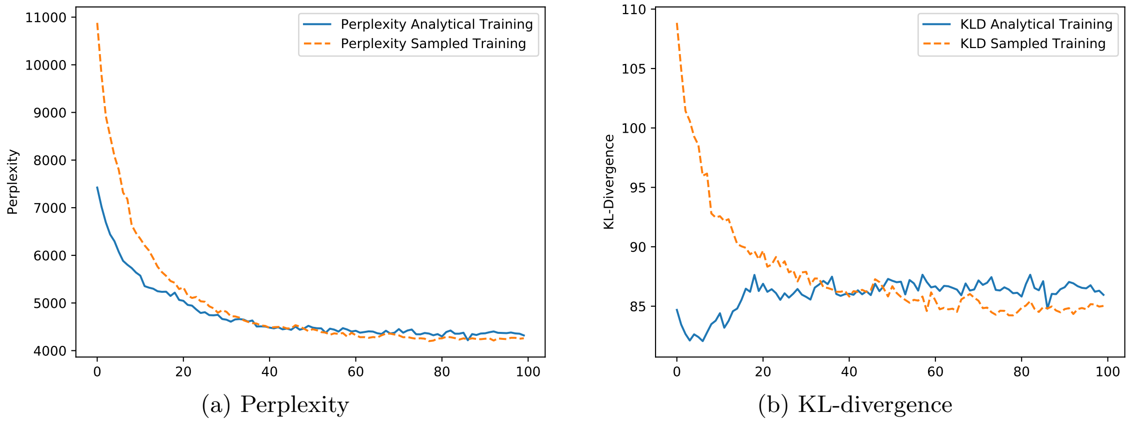


Figure 3: This plot shows the KL-divergence between two Dirichlet distributions with identical parameters, where only one element of one parameter vector is varied between zero and one. All other elements are fixed to 0.1 (blue line) or 0.7 (orange line). We can see that the KL-divergence diverges for small parameter values.

5.5.2 Topic coherence

In terms of topic coherence (see Tables 1 and 2, middle section, as well as additional results in the Appendix), DVAE has consistently high values as compared to all other methods with the best average rank for all settings of the Dirichlet parameter. A slight decrease in the average rank (1.125,1.125,1.250,1.750) can be observed for increasing values of the Dirichlet parameter which indicates that DVAE is especially good at modeling the sparseness implied by small parameter settings. DVAE Sparse is slightly worse than ProdLDA, but generally still better than NVDM and SCVB in terms of the average rank in topic coherence. This shows a trade-off between perplexity and topic coherence which can be influenced through the sparseness of the model. While DVAE Sparse has a far better perplexity than DVAE, its topic coherence is generally worse. While increased sparseness may thus lead to a lower topic coherence, this is not the only indicator for the topic quality (see Section 5.5.3).

Taking a closer look at the topic outputs, it is not always the case that a higher topic coherence corresponds to better topics. In Table 3 we listed all the topics that ProdLDA learned on the KOS dataset that have the words “iraq” and “war” in the most frequent words. This concerns ten out of fifty topics. ProdLDA often learns topics of relatively high quality that are redundant. This way, it is able to achieve a high topic coherence, but the diversity of the topics is not accounted for by this measure. DVAE in contrast learned only one topic containing the words “iraq” and “war” (see also Table 3), whereas DVAE Sparse only learned five topics containing these two terms. While NVDM only learned a single



(a) Perplexity

(b) KL-divergence

Figure 4: This figure compares the performance of DVAE trained with the analytical KL-divergence with the performance of DVAE trained with the sampled KL-divergence on the 20news dataset using a Dirichlet prior of 0.02. Reported is the perplexity using the analytical KL-divergence as well as the analytical KL-divergence. We can see that the final perplexity and KL-divergence for the sampled version are lower even though the convergence is slower in the beginning.

topic on the Iraq war, the quality of this topic is subjectively bad as none of the other words are closely related to the war. Nevertheless, ProdLDA has a topic coherence of 0.22 and DVAE and DVAE Sparse only reach 0.19 and 0.12, respectively. This shows that, while topic coherence is considered to be closely related to human judgement about topic quality (Lau et al., 2014), it is not a good indicator of topic quality in the case where topic redundancy is a problem.

Comparing the implicit gradients VAE, inverse CDF VAE and Weibull VAE in term of topic coherence, Tables 1 and 2 show that the implicit gradients method achieves the second best topic coherence overall and is only worse than DVAE. The inverse CDF and Weibull methods however, show a topic coherence that is in general comparable to the DVAE Sparse method. Judging from the topic coherence results, the approximations that are made by the Weibull and inverse CDF autoencoders are not as good as the implicit gradients and RSVI methods. We can therefore conclude that the RSVI and implicit gradients methods are preferable for Dirichlet parameter settings < 1 as are typical in topic modeling. The difference between the DVAE and the implicit gradients VAE method is minimal according to Figurnov et al. (2018), however, the used Dirichlet parameter is not reported since it was optimized during training. We therefore hypothesize that our improved performance is due to the sampled KL-divergence as well as a better approximation for low settings of the Dirichlet parameter.

5.5.3 Topic Redundancy

To quantify the redundancy of the topics we also reported the average probability of each word to occur in one of the other topics of the same model. The results reported in Tables 1

	DVAE
1	iraq hussein war destruction weapons saddam bush bin mass ladens
	DVAE Sparse
1	iraq war iraqi soldiers military sunni pleasure forces insurgency kurdish
2	war iraq baghdad juan cole soldiers helicopter iraqi killed troops
3	occupation license war iraq reid freeway iraqi ministries forces idiot
4	iraq bush war president administration billion house fiscal bushs cost
5	iraq weapons administration intelligence war bush saddam united clarke destruction
	ProdLDA
1	war iraq iraqi military baghdad juan administration bush mortar gulf
2	war iraq baghdad juan iraqi cole moqtada insurgency sunni soldiers
3	war fatalities authorities cow uniform barnes iraq administration ashamed mps
4	war iraq jury emotional assertions ice melanie film rationing studies
5	war iraq baghdad iraqi uniform prisoners military occupation saddam noble
6	iraq war iraqi moqtada kurdish baghdad shiite sunni kurds shia
7	bush patriot war aclu counterterrorism oversight iraq provisions allday sunset
8	iraq war interrogation pentagon apples ghraib abu guantanamo intelligence weapons
9	iraq war iraqi juan sistani baghdad kufa duration forces cole
10	bush gwb contracts ownership london defence iraq amazingly war infamous
	NVDM
1	senate chemical war races voted iraq barbero return qaqaa materiel

Table 3: 10 out of 50 topics for ProdLDA are about the Iraq war on the KOS dataset with a Dirichlet prior of 0.1. The topic coherence is 0.22 for this run, since topic coherence does not account for the diversity of topics. The topic coherence for DVAE Sparse is only 0.12, however, the topic quality is higher as can be seen from the topic redundancy score and qualitatively from this example. For DVAE, only one topic is about the Iraq war and the topic redundancy score is lower than that of ProdLDA and DVAE Sparse while the topic coherence is still high.

and 2 show that the lowest redundancy scores were achieved by NVDM, however, NVDM has a far worse topic coherence. In our experience the overall topic quality of all other models was mostly better than that of NVDM. If the topic coherence is low, however, the redundancy score is not that meaningful, since a low redundancy can also be expected from a model that assigns words to topics randomly. If a model achieves a high topic coherence as can be seen in Table 5 for the implicit gradients method with the Rcv1 dataset and 200 topics, this can still be associated with a very hight topic redundancy (0.45 in this case) which points to component collapsing and devalues the high topic coherence score. DVAE almost always has a low topic redundancy except for a prior of 0.6 and 200 topics on the KOS dataset. The topic coherence is also low in this case. We can therefore conclude that topic redundancy is an important measure if the topic coherence is also high. This is because it helps to differentiate models with a diverse set of high quality topics and models where component collapsing led to a single topic being repeated several times.

6. Discussion

DVAE, DVAE Sparse, implicit gradients VAE, inverse CDF VAE and Weibull VAE are most closely related to the ProdLDA model. All of these models can be trained with relatively high learning rates (10^{-3}–10^{-2}) as compared to NVDM, which is trained with much lower learning rates (around 10^{-5}). This speeds up the training process considerably. The efficiency of ProdLDA, DVAE and DVAE Sparse and the other Dirichlet autoencoders is comparable as long as the correction term is disregarded for the DVAE models. This can safely be done as long as the shape augmentation parameter is chosen high enough (≥ 5). The training efficiency might also be affected by other architectural decisions that warrant further investigation in the future. NVDM trains the encoder and decoder networks separately, whereas DVAE and ProdLDA update both networks at once. DVAE and ProdLDA use dropout and batch normalization, which enable the high learning rates. All these details need to be taken into consideration and their potential influence will be investigated in future work.

All models with a Dirichlet latent parameter vector are characterized by a relatively high topic coherence as compared to the Gaussian model NVDM. This shows that it is worth considering to use the Dirichlet distribution in cases where interpretable topics are important. Nevertheless, topic coherence alone is not a good indicator of topic quality and should be viewed in conjunction with our newly introduced measure of topic redundancy. This is especially important in VAE topic models due to the problem of component collapsing, which is more prevalent in VAEs than in other LDA topic models.

Comparing our two models DVAE and DVAE Sparse, DVAE clearly leads to topics of higher quality. The perplexity, however, is bad in comparison. According to our analysis this is due to the KL-divergence term in the ELBO that leads to high values when the parameter is predicted to be close to zero. By modeling the sparseness in a separate vector, this problem is alleviated, leading to lower perplexities. DVAE Sparse therefore serves as a test to show that our original hypothesis is correct, namely that the coupling of sparsity and smoothness is responsible for the high perplexity scores. Furthermore, this shows that perplexity scores are not always meaningful when comparing different topic models, as in our experiments the topic quality has nothing to do with the perplexity a model achieves.

The KL-divergence can be sampled or calculated analytically. In our experiments we chose to sample the KL-divergence during training. The comparison methods all use analytical KL-divergence terms. For the KL-divergence of the Dirichlet distribution it can make a big difference whether the KL-divergence is sampled or not. In the case where the parameter is close to zero, the analytical KL-divergence takes on very large values (see Figure 3), whereas in this case the sampled KL-divergence is often negative in fact. This can lead to much lower perplexity scores when the sampled KL-divergence is used. Therefore, we chose to calculate the testing perplexity using the analytical KL-divergence. (In the other case, when the parameter is larger, the analytical KL-divergence is often lower than the sampled KL-divergence, although the difference is not as big as in the first case.) This stability issue has a large influence on the perplexity and our sparse method avoids it to a certain extent, but not completely. A related method avoids this by preconditioning the gradient in a Hamiltonian MCMC method with the inverse of the Fisher information matrix (Cong et al., 2017; Patterson and Teh, 2013). According to Cong et al. (2017), this makes the stability problem completely disappear. How such a method could be applied to the variational autoencoder however, remains open.

Our experiments show that DVAE is superior to other state-of-the-art topic models in terms of topic coherence. This further underlines the argument by Srivastava and Sutton (2017) that the sparsity induced by the Dirichlet indeed leads to higher topic coherence. The comparison with other Dirichlet autoencoders shows that while our version of the RSVI autoencoder works best, the implicit gradients VAE is the most competitive method, whereas the inverse CDF and Weibull autoencoders are even worse than the Laplace approximation of ProdLDA in terms of topic coherence.

We discussed the trade-off between sparsity and smoothness in Section 4.1. This trade-off can be influenced using the DVAE Sparse model with the additional Bernoulli variable. Furthermore, the sampling of the KL-divergence is an important factor for increasing the topic coherence. As shown in Figure 4, while the convergence of perplexity and KL-divergence seems to be slower in the beginning, this may lead to a lower perplexity and KL-divergence in the final stages of training. For the shown comparison the final topic coherence score was 0.33 for the training method with the analytical KL-divergence, whereas training with the sampled KL-divergence led to a topic coherence of 0.35 after 100 iterations of training. This effect is similar to the effect of batch normalization as reported by Srivastava and Sutton (2018) who did a similar experiment. They show that batch normalization slows down the convergence of the KL-divergence in the beginning of training. This reduces the component collapsing and enables the learning of more meaningful topics.

When talking about a trade-off, this implies that there is a continuum of models between extreme sparseness and extreme smoothness. This can in principle be explored using our DVAE Sparse model. While at the moment, the parameter for our Bernoulli variables is $p = 0.5$, this parameter could also be set to other values and thereby increase or decrease the sparseness of the model. On the one hand, a value of $p = 0$ would mean that the vector b will always consist of only ones. This would reduce the model to the original DVAE model, which is a smooth model. On the other hand, for $p = 1$ the vector b would consist of only zeros, which entails maximal sparseness and only topics with zero values, preventing any topics from being learned. Intermediary values can be used to adjust the degree of sparseness and interpolate between the sparse and the smooth model.

7. Conclusion

We have introduced a variational autoencoder with a Dirichlet prior called DVAE. A trade-off between sparsity and smoothness was identified, leading to high perplexity as well as topic coherence scores. To test our hypothesis about the reason for the high perplexity scores, a new model called DVAE Sparse was introduced. DVAE Sparse decouples sparsity and smoothness and thus enables lower perplexity scores on the one hand while on the other hand the topic coherence is still competitive. The topic quality was also measured in terms of topic redundancy which is a good indicator for component collapsing as opposed to topic coherence. Overall, this shows that, despite the bad perplexity scores, DVAE is an effective model. An extensive comparison with all known training methods for Dirichlet VAEs shows that our DVAE using the RSVI method achieves the highest topic coherence scores.

To conclude, we think that this work is an important step towards making VAEs with Dirichlet priors more widely used. In the future we would like to extend our model to work with hierarchical Dirichlet priors.

Acknowledgements

We thank the anonymous reviewers for their helpful comments.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

Yulai Cong, Bo Chen, Hongwei Liu, and Mingyuan Zhou. Deep latent Dirichlet allocation with topic-layer-adaptive stochastic gradient Riemannian MCMC. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 864–873, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 439–450, 2018. URL `http://papers.nips.cc/paper/7326-implicit-reparameterization-gradients`.

James Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 446–454, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. In *Proc. of the National Academy of Sciences of the United States of America*, volume 101, pages 5228–5235. National Acad Sciences, 2004.

Weonyoung Joo, Wonsung Lee, Sungrae Park, and Il-Chul Moon. Dirichlet variational autoencoder. *CoRR*, abs/1901.02739, 2019. URL `http://arxiv.org/abs/1901.02739`.

Diederik Kingma and Max Welling. Efficient gradient-based inference through transformations between bayes nets and neural nets. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1782–1790, Bejing, China, 22–24 Jun 2014a. PMLR.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014b.

David A Knowles. Stochastic gradient variational bayes for gamma approximating distributions. *arXiv*, page 1509.01631, 2015. URL `https://arxiv.org/abs/1509.01631`.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539. The Association for Computer Linguistics, 2014.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA, 20–22 Jun 2016. PMLR.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 489–498, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

Eric Nalisnick and Padhraic Smyth. Stick-breaking variational autoencoders. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Sam Patterson and Yee Whye Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, pages 3102–3110, 2013.

V. Perrone, P. A. Jenkins, D. Spano, and Y. W. Teh. Poisson random fields for dynamic feature models. ArXiv e-prints: 1611.07460, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Akash Srivastava and Charles A. Sutton. Variational inference in pachinko allocation machines. *CoRR*, abs/1804.07944, 2018. URL `http://arxiv.org/abs/1804.07944`.

Chong Wang and David M Blei. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. In *Advances in Neural Information Processing Systems*, pages 1982–1989, 2009.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *International Conference on Learning Representations*, 2018.

Appendix A. Additional results

dataset	DVAE	DVAE Sp.	Impl.	Inv.	Weibull	ProdLDA	NVDM	SCVB
	Perplexity							
					50 topics			
KOS	2858	2322	2232	**2056**	2923	2537	2097	2501
NIPS	2666	2142	2223	4858	1.46E+26	2239	**1933**	8137
20news	1345	1030	1678	1484	1212	1076	**813**	961
Rcv1	2542	1234	1458	1480	1397	1788	**907**	1990
					200 topics			
KOS	4503	2443	2611	2395	4839	3739	**2216**	2404
NIPS	2551	2043	1983	2155	inf	2313	**1957**	6627
20news	2869	1123	1734	2120	4.66E+04	1496	**875**	925
Rcv1	2913	1052	1337	1519	1613	2206	**891**	4573
av. rank	6.625	2.875	4.125	4.375	6.625	5.125	1.125	5.125
	Topic Coherence							
					50 topics			
KOS	0.202	0.151	0.132	0.135	0.146	**0.232**	0.070	0.147
NIPS	**0.330**	0.263	0.284	0.144	0.120	0.281	0.070	0.093
20news	**0.342**	0.265	0.280	0.276	0.243	0.291	0.140	0.218
Rcv1	**0.402**	0.121	0.326	0.150	0.308	0.094	0.110	0.246
					200 topics			
KOS	**0.218**	0.070	0.143	0.098	0.152	0.124	0.060	0.114
NIPS	**0.295**	0.213	0.236	0.244	0.088	0.260	0.060	0.074
20news	**0.334**	0.197	0.300	0.260	0.215	0.297	0.140	0.183
Rcv1	0.286	0.113	**0.386**	0.138	0.256	0.078	0.051	0.227
av. rank	1.250	5.250	3.000	4.750	4.500	3.750	7.875	5.625
	Topic Redundancy							
					50 topics			
KOS	0.024	0.015	0.015	**0.007**	0.032	0.051	0.012	0.094
NIPS	0.018	0.012	0.010	0.030	0.068	0.014	**0.003**	0.095
20news	0.029	0.028	0.073	0.048	0.033	0.023	**0.007**	0.041
Rcv1	0.007	**0.001**	0.005	0.009	0.002	0.033	0.002	0.034
					200 topics			
KOS	0.029	**0.006**	0.051	0.047	0.013	0.081	0.010	0.168
NIPS	0.017	0.010	0.005	0.015	0.029	0.016	**0.003**	0.096
20news	0.042	0.025	0.042	0.036	0.034	0.043	**0.011**	0.151
Rcv1	0.010	**0.002**	0.098	0.015	0.025	0.024	0.004	0.025
av. rank	4.625	2.125	5.000	4.625	4.875	5.500	1.625	7.625

Table 4: Perplexity, topic coherence and topic redundancy for **Dirichlet prior 0.3**.

dataset	DVAE	DVAE Sp.	Impl.	Inv.	Weibull	ProdLDA	NVDM	SCVB
Perplexity								
50 topics								
KOS	2362	2259	2206	**1988**	2744	2623	2097	2515
NIPS	2626	2115	2219	2005	3.60E+47	2247	**1933**	4806
20news	1053	954	1531	1182	1369	1139	**813**	853
Rcv1	2154	1191	1460	7.09E+201	1294	1765	**907**	2525
200 topics								
KOS	2943	**2116**	2378	2688	4821	4263	2216	2171
NIPS	2450	1981	1971	**1848**	7.73E+112	2261	1957	1.13E+04
20news	1035	1065	2073	1357	4997	2104	**875**	943
Rcv1	2681	971	1942	7.11E+201	1433	2393	**891**	3295
av. rank	5.250	2.875	4.500	4.500	6.625	5.750	1.500	5.000
Topic Coherence								
50 topics								
KOS	**0.205**	0.192	0.144	0.128	0.057	0.160	0.070	0.139
NIPS	**0.313**	0.254	0.294	0.261	0.287	0.312	0.070	0.093
20news	**0.371**	0.250	0.303	0.245	0.206	0.324	0.140	0.214
Rcv1	**0.425**	0.285	0.327	0.130	0.294	0.260	0.110	0.230
200 topics								
KOS	0.110	0.081	0.153	**0.155**	0.141	0.137	0.060	0.123
NIPS	**0.296**	0.232	0.255	0.234	0.139	0.263	0.060	0.061
20news	**0.319**	0.235	0.285	0.261	0.192	0.313	0.140	0.180
Rcv1	0.291	0.186	**0.333**	0.165	0.274	0.077	0.051	0.196
av. rank	1.750	4.750	2.625	4.750	5.000	3.375	7.875	5.875
Topic Redundancy								
50 topics								
KOS	0.025	0.030	0.012	**0.006**	0.018	0.054	0.012	0.092
NIPS	0.015	0.012	0.010	0.004	0.081	0.020	**0.003**	0.105
20news	0.033	0.027	0.080	0.029	0.029	0.033	**0.007**	0.055
Rcv1	0.007	0.004	0.007	0.142	**0.002**	0.027	0.002	0.032
200 topics								
KOS	0.216	0.012	0.019	0.042	0.023	0.070	**0.010**	0.253
NIPS	0.014	0.013	0.009	0.008	0.079	0.014	**0.003**	0.103
20news	0.039	0.025	0.044	0.044	0.026	0.035	**0.011**	0.231
Rcv1	0.046	**0.002**	0.453	0.223	0.009	0.013	0.004	0.056
av. rank	5.500	3.000	4.750	4.250	4.125	5.375	1.500	7.500

Table 5: Perplexity, topic coherence and topic redundancy for **Dirichlet prior 0.6**.