Adversarial Learning for Topic Models

Tomonari Masada[1][0000−0002−8358−3699] and Atsuhiro Takasu[2]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, Japan
masada@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takasu@nii.ac.jp

Abstract. This paper proposes adversarial learning for topic models. Adversarial learning we consider here is a method of density ratio estimation using a neural network called discriminator. In generative adversarial networks (GANs), we train discriminator for estimating the density ratio between the true data distribution and the model. Also in variational inference (VI) for Bayesian models, we can train discriminator for estimating the density ratio between the approximate posterior distribution and the prior distribution. The adversarial learning in VI enables us to adopt implicit distribution as approximate posterior. This paper proposes an adversarial learning for latent Dirichlet allocation (LDA) to improve the expressiveness of the approximate posterior. Our experimental results showed that the quality of extracted topics was improved in terms of test perplexity. However, it was regrettably observed that high probability words in each extracted topic were less interpretable than those obtained by collapsed Gibbs sampling for LDA, though it was also observed that our proposal works in document expansion.

Keywords: Topic models · Adversarial learning · Variational inference.

1 Introduction

This paper proposes adversarial learning for topic models. Topic modeling [3, 2] is widely used as text analysis method and can extract diverse topics latent in a given document set. Topic modeling represents each latent topic as a probability distribution defined over vocabulary words. By visualizing such word probability distributions as word clouds (cf. Fig. 2 and Fig. 3), we can intuitively grasp what semantic contents are discussed in the given document set by going through the word clouds. Further, topic modeling can also provide topic proportions for each document [2, Figure 1]. Therefore, after finding favorite topics by going through the word clouds, we may retrieve the documents where the proportions of our favorite topics are larger than other topics and read only those documents carefully. In this way topic modeling mitigates the burden imposed on us by the excessive diversity of contents delivered by a huge set of documents. Topic modeling can provide us a bird's-eye view over the diverse document contents.

Since topic modeling considered in this paper is a Bayesian modeling, we need to infer posterior distribution. However, the true posterior distribution is

intractable. We can approximate the true posterior either by drawing many samples with MCMC techniques or by seeking a tractable approximate distribution as a surrogate of the true posterior. In this paper we consider the latter way of posterior inference and propose a GAN-like [7] adversarial learning for approximating the true posterior in variational inference (VI) for topic models.

The adversarial learning we consider here is a density ratio estimation using a neural network called discriminator [7]. In VI we approximate the true posterior by maximizing the evidence lower bound (ELBO). This maximization requires estimation of the KL-divergence from the approximate posterior distribution to the prior distribution.[1] When we choose an approximate posterior whose density function is given explicitly, the KL-divergence can be easily estimated. However, the expressiveness of the approximate posterior is then limited. This paper proposes to use implicit distribution, i.e., the probability distribution whose log-likelihood function is not explicitly given [10, 13], for approximating the true posterior. Even when we adopt implicit distribution, we can estimate the KL-divergence by using adversarial learning.

This paper provides an adversarial learning for latent Dirichlet allocation (LDA) [3]. Our aim is to improve the expressiveness of approximate posterior for LDA. In VI for LDA we consider the log evidence for each document d:

$$\log p(\boldsymbol{x}_d; \boldsymbol{\Phi}) = \int p(\boldsymbol{x}_d|\boldsymbol{\theta}_d; \boldsymbol{\Phi}) p(\boldsymbol{\theta}_d) d\boldsymbol{\theta}_d \tag{1}$$

where $\boldsymbol{x}_d$ is the observed word count vector of document d, $\boldsymbol{\theta}_d$ is the document-wise parameter vector of categorical distribution over K latent topics, and $\boldsymbol{\Phi}$ denotes the free parameters for modeling topic-wise probability distributions over V vocabulary words. The main task we consider is to approximate the true posterior $p(\boldsymbol{\theta}_d|\boldsymbol{x}_d)$ by a surrogate distribution $q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)$. This approximation is equivalent to the maximization of the following ELBO $\mathcal{L}(\boldsymbol{\theta}_d, \boldsymbol{\Phi})$:

$$\mathcal{L}(\boldsymbol{\theta}_d, \boldsymbol{\Phi}) = \mathbb{E}_{q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)}[\log p(\boldsymbol{x}_d|\boldsymbol{\theta}_d; \boldsymbol{\Phi})] - \mathrm{KL}(q(\boldsymbol{\theta}_d|\boldsymbol{x}_d) \| p(\boldsymbol{\theta}_d)) \leq \log p(\boldsymbol{x}_d; \boldsymbol{\Phi}) \tag{2}$$

The estimation of $\boldsymbol{\Phi}$ can be performed by maximizing the log-likelihood term $\mathbb{E}_{q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)}[\log p(\boldsymbol{x}_d|\boldsymbol{\theta}_d; \boldsymbol{\Phi})]$ in Eq. (2) when $\boldsymbol{\theta}_d$ is given. Our proposal concerns the approximation of the KL-divergence in Eq. (2). We propose to use implicit distribution as $q(\theta_d|\boldsymbol{x}_d)$, which is represented by a neural network called *encoder*, and make $q(\theta_d|\boldsymbol{x}_d)$ more expressive than when we use explicit distribution. However, this makes the estimation of the KL-divergence in Eq. (2) intractable. Therefore, we provide an adversarial learning. The proposed adversarial learning employs a neural network called *discriminator*, with which we can obtain an approximation of the KL-divergence from $q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)$ to $p(\boldsymbol{\theta}_d)$ in Eq. (2).

The evaluation experiment showed that the latent topics extracted by the proposed method were better than those extracted by collapsed Gibbs sampling (CGS) [8] for LDA in terms of test perplexity. However, it was regrettably discovered that our method had a drawback in interpretability. To apply the proposed

[1] We do not consider the *joint contrastive* form of ELBO [10] in this paper.

adversarial learning, we need to collapse discrete latent variables of LDA as described in [18, Section 3.1]. The same collapsing is also performed in other work [5]. However, it was suspected that this collapsing led to a poor interpretability of extracted topics. The high probability words in each topic only gave a vague intuition about the corresponding semantic content, though those obtained by CGS for LDA were easy to interpret. However, it was also observed that our proposal worked in document expansion even when topic words were of poor quality. In sum, our contributions are:

- We propose an adversarial learning for LDA using discriminator network,
- We show that the proposed adversarial learning achieves better topic extraction in terms of test perplexity, and
- We observed that the collapsing of discrete variables seemingly led to poor interpretability of topic words.
- We also observed that our method worked in document expansion.

2 Method

2.1 Adversarial Learning in VI for LDA

Adversarial learning we consider here is a method of density ratio estimation using a discriminator network. The density ratio we estimate is $\frac{q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)}{p(\boldsymbol{\theta}_d)}$, which is used for approximating the KL-divergence $\mathrm{KL}(q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)\|p(\boldsymbol{\theta}_d))$ in Eq. (2). In this paper we propose to represent $q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)$ with an encoder network $\boldsymbol{\theta}_d = g(\boldsymbol{x}_d, \boldsymbol{\epsilon})$, whose input is the concatenation of a word count vector $\boldsymbol{x}_d$ and a noise vector $\boldsymbol{\epsilon}$ drawn from the standard multivariate Gaussian distribution, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_e)$. The dimension e of $\boldsymbol{\epsilon}$ is determined independently of the number of latent topics K in LDA and also of the vocabulary size V. The output of the encoder then has randomness thanks to the noise vector. However, the density function cannot be explicitly given for the distribution of the encoder output. Therefore, we provide an adversarial learning, where we use a discriminator network $r(\boldsymbol{\theta}_d, \boldsymbol{x}_d)$ to estimate the density ratio between $q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)$ and $p(\boldsymbol{\theta}_d)$ [10, 13]. By maximizing the following objective function with respect to the discriminator parameters, we obtain an approximation of the logarithmic density ratio $\log \frac{q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)}{p(\boldsymbol{\theta}_d)}$ as $r(\boldsymbol{\theta}_d, \boldsymbol{x}_d)$:

$$\ell(r) = \sum_{d=1}^{D} \Big[\mathbb{E}_{\boldsymbol{\theta}_d \sim p(\boldsymbol{\theta}_d)} \log \big(1 - \sigma(r(\boldsymbol{x}_d, \boldsymbol{\theta}_d))\big) + \mathbb{E}_{\boldsymbol{\theta}_d \sim q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)} \log \big(\sigma(r(\boldsymbol{x}_d, \boldsymbol{\theta}_d))\big) \Big] \tag{3}$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the standard sigmoid function. We assume that the prior $p(\boldsymbol{\theta}_d)$ is an isotropic Gaussian distribution, whose mean and standard deviation parameters are estimated by empirical Bayes approach.

It should be noted that, for the optimal discriminator r^* obtained by maximizing $\ell(r)$ in Eq. (3), it holds that $\sigma(r^*) = \frac{1}{1+\exp(-\log(q/p))} = \frac{q}{p+q}$, which

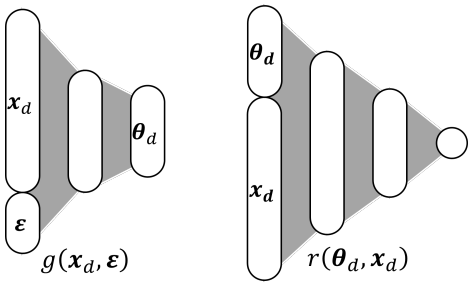

Fig. 1. Schematic depiction of the encoder network (left) and the discriminator network (right) in AVB-LDA. The softmaxed encoder output $\boldsymbol{\theta}_d$ is the parameter vector of document-wise categorical distribution over topics. The discriminator output is an approximation of the logarithmic density ratio between $q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)$ and $p(\boldsymbol{\theta}_d)$. See Eq. (4).

corresponds to the optimal discriminator D^* in GANs [7, Proposition 1]. Therefore, precisely speaking, it is an abuse of terminology to call r discriminator, because r is not equal to the discriminator D in GANs. However, we think that it is harmless, because r corresponds D through the mapping $D = \sigma(r)$.

With $r(\boldsymbol{\theta}_d, \boldsymbol{x}_d)$, the ELBO in Eq. (2) can be rewritten as

$$\mathcal{L}(g, \boldsymbol{\Phi}) = \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_e)}[\log p(\boldsymbol{x}_d | g(\boldsymbol{x}_d, \boldsymbol{\epsilon}); \boldsymbol{\Phi}) - r(g(\boldsymbol{x}_d, \boldsymbol{\epsilon}), \boldsymbol{x}_d)] \tag{4}$$

We update the parameters of the encoder network $g(\boldsymbol{x}_d, \boldsymbol{\epsilon})$ and the model parameters $\boldsymbol{\Phi}$ by maximizing $\mathcal{L}(g, \boldsymbol{\Phi})$ in Eq. (4) for a fixed $r(\boldsymbol{\theta}_d, \boldsymbol{x}_d)$. The expectation with respect to $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_e)$ is approximated by Monte Carlo integration.

We call this inference *adversarial variational Bayes (AVB)* by following [13]. Our AVB for LDA uses two multilayer perceptrons (MLPs), i.e., one for encoder and another for discriminator, as depicted in Fig. 1. The number of hidden layers of each MLP was set differently for each data set. The experiment showed that two hidden layers were enough to achieve any good evaluation result and that the results were not improved by increasing the number of hidden layers. We next discuss how the topic-wise word probabilities $\boldsymbol{\Phi}$ are modeled.

2.2 Topic-Wise Word Probabilities

In the generative story of the original LDA [3] we first draw a topic for each word token from the document-wise categorical distribution over topics, i.e., Categorical($\boldsymbol{\theta}_d$). The elements $(\theta_{d,1}, \ldots, \theta_{d,K})$ of $\boldsymbol{\theta}_d$ are the topic proportions in document d. We then draw a word from the topic-wise categorical distribution over words, which corresponds to the drawn topic k, i.e., Categorical($\boldsymbol{\phi}_k$). The elements $(\phi_{k,1}, \ldots, \phi_{k,V})$ of the parameter vector $\boldsymbol{\phi}_k$ are the word probabilities in topic k. This generative story assigns each word token to a topic randomly chosen based on the document-wise topic proportions. Therefore, we introduce the discrete latent variable $z_{d,i}$ indicating to what topic the i-th word token in document d is assigned for formulating the joint distribution of LDA [3, Eq. (2)].

Our AVB for LDA, however, collapses the discrete latent variables $\boldsymbol{z}_d$ in the same manner as [5, 18] to reduce computational burden in VI. We thus obtain the log-likelihood $\log p(\boldsymbol{x}_d|\boldsymbol{\theta}_d;\boldsymbol{\Phi})$, which is expressed by using continuous parameters $\boldsymbol{\Phi} = \{\boldsymbol{B}, \boldsymbol{b}_0\}$, as follows:

$$\log p(\boldsymbol{x}_d|\boldsymbol{\theta}_d;\boldsymbol{\Phi}) = \boldsymbol{x}_d^\top \mathrm{LogSoftmax}(\boldsymbol{B}\boldsymbol{\theta}_d + \boldsymbol{b}_0) \tag{5}$$

where $\boldsymbol{x}_d$ is the word count vector of document d. The parameter vector $\boldsymbol{\theta}_d$ is affine-transformed with the $V \times K$ matrix $\boldsymbol{B}$ and the V dimensional vector $\boldsymbol{b}_0$. The computation $\boldsymbol{B}\boldsymbol{\theta}_d + \boldsymbol{b}_0$ in Eq. (5) can be regarded as forward pass of a single-layer neural network. We call this network *decoder*. By applying the log-softmax function to the decoder output, we obtain the logarithmic word probabilities in document d. The log-likelihood of document d is then obtained as the inner product of the word count vector $\boldsymbol{x}_d$ with the logarithmic word probability vector as in Eq. (5). We use no prior distribution for word probabilities in our LDA. However, the V-dimensional bias vector $\boldsymbol{b}_0$ plays a similar role to smoothing parameter [4], because it depends on no particular topics. We train the decoder also by maximizing the ELBO in Eq. (4). While we tested decoders having hidden layers accompanied with nonlinear activation function in the evaluation experiment, the results were not improved. Therefore, word probabilities are modeled in this simple way.

The pseudo code of the proposed adversarial learning for LDA, abbreviated as AVB-LDA, is given in Algorithm 1, where C is the mini-batch size, and M is the number of iterations for the training of discriminator. While $M = 1$ works in many cases, we may obtain better results by setting $M > 1$. In our experiment, the initialization method of the parameters of the encoder g and the discriminator r was chosen from among Xavier uniform, Xavier normal, Kaiming uniform, and Kaiming normal [6, 9]. The activation function was chosen from among ReLU and LeakyReLU. For the parameters of the decoder, $\boldsymbol{\Phi}$ was initialized by standard normal random numbers and $\boldsymbol{b}_0$ by zeros.

3 Experiment

3.1 Document Sets for Evaluation

In the evaluation experiment we used three data sets, whose specifications are given in Table 1, where D_{train} and D_{test} are the numbers of documents in training and test sets, respectively, and V is the vocabulary size. The first data set, denoted by NIPS, is the set of NIPS full papers obtained from UCI Bag of Words Data Set.[2] The second data set is a subset of the questions from the StackOverflow data set available at Kaggle.[3] We denote this document set by STOF. The third one, a set of New York Times articles also obtained from UCI Bag of Words Data Set, is denoted by NYT. Each document set is split into

[2] https://archive.ics.uci.edu/ml/datasets/bag+of+words
[3] https://www.kaggle.com/stackoverflow/rquestions

Algorithm 1 Adversarial variational Bayes for LDA

procedure AVB-LDA($\mathcal{D}, r, g, \boldsymbol{\Phi}, C, M$)
  repeat
    Sample C items $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(C)}$ from the training set $\mathcal{D}$ to make a mini-batch.
    for $m = 1$ to M **do**
      ▷ discriminator update
      Sample C noise vectors $\boldsymbol{\epsilon}_{(1)}, \ldots, \boldsymbol{\epsilon}_{(C)}$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
      Compute $g(\boldsymbol{x}_{(c)}, \boldsymbol{\epsilon}_{(c)})$ for each $c \in \{1, \ldots, C\}$.
      Sample C document-wise topic proportions $\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(C)}$ from $p(\boldsymbol{\theta})$.
      Maximize ℓ in Eq. (3) with respect to the parameters of r.
    ▷ encoder update
    Sample C noise vectors $\boldsymbol{\epsilon}_{(1)}, \ldots, \boldsymbol{\epsilon}_{(C)}$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
    Compute $g(\boldsymbol{x}_{(c)}, \boldsymbol{\epsilon}_{(c)})$ for each $c \in \{1, \ldots, C\}$.
    Maximize $\mathcal{L}$ in Eq. (4) with respect to the parameters of g.
    ▷ decoder update
    Sample C noise vectors $\boldsymbol{\epsilon}_{(1)}, \ldots, \boldsymbol{\epsilon}_{(C)}$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
    Compute $g(\boldsymbol{x}_{(c)}, \boldsymbol{\epsilon}_{(c)})$ for each $c \in \{1, \ldots, C\}$.
    Maximize $\mathcal{L}$ in Eq. (4) with respect to $\boldsymbol{\Phi}$.
    ▷ prior update
    Sample C noise vectors $\boldsymbol{\epsilon}_{(1)}, \ldots, \boldsymbol{\epsilon}_{(C)}$ from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
    Compute $g(\boldsymbol{x}_{(c)}, \boldsymbol{\epsilon}_{(c)})$ for each $c \in \{1, \ldots, C\}$.
    Maximize $\mathcal{L}$ in Eq. (4) with respect to the parameters of the prior.
    ▷ postprocessing
    Adjust learning rate.
  until change in parameters is negligible

training and test sets as given in Table 1. Based on the perplexity computed over training set, we tuned free parameters for each compared method. The final evaluation is performed in terms of the perplexity computed over test set. The perplexity is defined as the exponential of the negative log likelihood of the corpus, where the log likelihood is normalized by the number of word tokens.

3.2 Compared Methods

We compared the proposed adversarial learning for LDA, denoted by AVB-LDA, to the following three methods.

The first compared method is collapsed Gibbs sampling (CGS) for LDA [8]. This choice aims to compare our proposal to the vanilla topic modeling. CGS is a sampling-based posterior inference and is thus time-consuming. However, it is known that CGS often gives better test perplexity than VI [1]. In CGS we have two free parameters, i.e., the hyperparameter α of the symmetric Dirichlet prior distribution for document-wise topic categorical distributions and the hyperparameter β of that for topic-wise word categorical distributions. We tuned these two hyperparameters by grid search [1] based on training set perplexity. We call this method CGS-LDA.

Table 1. Specifications of data sets used in comparison experiment

NIPS			STOF			NYT		
D_{train}	D_{test}	V	D_{train}	D_{test}	V	D_{train}	D_{test}	V
1,050	225	12,419	20,486	2,561	13,184	239,785	29,968	102,660

The second compared method is the adversarial learning for document modeling, not for topic modeling, This choice aims to compare the effectiveness of adversarial learning for topic modeling to that for plain document modeling. By plain document modeling we mean a modeling of documents simply by mapping them into some lower-dimensional space. We refer to this method as AVB-DM. AVB-DM uses the following three MLPs. The first MLP is encoder with a single hidden layer for mapping document vectors to their lower-dimensional representation. The second one is decoder with a single hidden layer for mapping the encoded representation to the reconstructed document vector. The dimension of encoded representations is set to K, i.e., the number of topics in CGS-LDA and AVB-LDA. The input of the encoder is a concatenation of a word count vector $\boldsymbol{x}_d$ and a noise vector $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_e)$ as described in [13]. The dimension e of the noise vector was tuned based on training set perplexity. Since the encoder represents an implicit distribution, we adopt an adversarial learning. Therefore, AVB-DM uses the third MLP as discriminator in a similar manner to our proposal. The number of the hidden layers of discriminator was set to two only for NIPS data set and one for the other data sets. Since AVB-DM is not a topic model, it does not provide latent topics as categorical distributions over vocabulary words. AVB-DM provides no straightforward ways to obtain diverse document contents as word lists, each possibly visualized as word clouds. Therefore, it can be said that AVB-DM is inferior to AVB-LDA with respect to the interpretability of analysis results.

The third compared method is the variational autoencoder (VAE) [11] for document modeling. This method, denoted by VAE-DM, is simpler than AVB-DM, because no implicit distribution is used for approximating the true posterior. The aim of this choice is to clarify what we lose by discarding adversarial learning. In VAE-DM we consider here the encoder MLP maps each document vector $\boldsymbol{x}_d$ to a concatenation of the mean parameter vector $\boldsymbol{\mu}_d$ and the standard deviation parameter vector $\boldsymbol{\sigma}_d$ of diagonal Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_d, \mathrm{diag}(\boldsymbol{\sigma}_d^2))$, where $\mathrm{diag}(\boldsymbol{\sigma}_d)$ is the diagonal matrix whose diagonal elements are $\boldsymbol{\sigma}_d^2$. The dimension of the diagonal Gaussian distribution is set to K. We can draw samples from this diagonal Gaussian by using reparameterization trick [16, 19]. That is, the samples from $q(\boldsymbol{z}_d|\boldsymbol{x}_d)$ are obtained as $\boldsymbol{\epsilon} \odot \boldsymbol{\sigma}_d + \boldsymbol{\mu}_d$, where $\odot$ is element-wise product and $\boldsymbol{\epsilon}$ is drawn from the standard multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$. The number of the hidden layers of the encoder was set to two only for STOF data set and one for the other data sets. There is no difference between AVB-DM and VAE-DM with respect to the decoder MLP, which maps the encoded representation to a reconstruction of the input document vector. The number of the hidden layers of the decoder was set to one for all data sets. We made the mini-batch size for VAE-DM larger than AVB-DM and AVB-LDA

for obtaining better training set perplexity. While the VAE-DM we consider here is almost the same with neural variational document model (NVDM) [14], we applied dropout to the input vectors in every mini-batch, because perplexity was improved greatly for all data sets.

The compared methods were implemented in PyTorch[4] except CGS-LDA, which uses no neural networks. For AVB-LDA, the number of the hidden layers of the encoder was set to two only for STOF data set and one for the other data sets. Further, the number of the hidden layers of the discriminator was set to one only for NYT data set and two for the other data sets. The number of the hidden layers were determined based on training set perplexity for all cases and for all compared methods. While we could compare AVB-LDA to the VAE for LDA, it was already proposed by Srivastava et al. [18] and was compared to CGS-LDA in their paper. Therefore, we did not consider the VAE for LDA and avoided redundant experiments by comparing our method to CGS-LDA.

3.3 Evaluation Results

Table 2 contains all evaluation results in terms of test set perplexity, where we also present the optimal settings obtained based on training set perplexity. $h_{\mathrm{Enc},1}$ means the size of the first hidden layer of the encoder, and η_{Dis} the initial learning rate of the discriminator. As Table 2 shows, AVB-LDA achieved the best test perplexity among the compared methods. However, AVB-DM gave the second best results for all data sets. Therefore, we can say that adversarial learning works both for topic modeling and for plain document modeling. VAE-DM led to a comparable result only for NIPS data set, and the perplexity of CGS-LDA exhibits a tendency similar to that of VAE-DM. It can be concluded that AVB-LDA is the best choice if our aim is to achieve excellent test perplexity.

However, by going through the topic words, i.e., high probability words in each extracted topics, it was revealed that better test perplexity did not necessarily mean better interpretability as shown in Fig. 2 and Fig. 3, where we depict topic words as word clouds. In Fig. 2 the top and bottom panels present high probability words in three among 100 topics extracted by CGS-LDA and AVB-LDA, respectively, from NIPS data set. Based on the words appearing in the top three panels, it is relatively easy to guess the corresponding contents, i.e., machine learning inspired by physics concepts (the left panel), motion capturing (the center panel), and fundamentals in machine learning (the right panel). However, the topic words extracted by AVB-LDA, shown in the bottom three panels, provide only a weak clue to their corresponding semantic contents. Many of the topic words extracted by AVB-LDA are the words that are rarely used when compared with those extracted by CGS-LDA. Fig. 3 provides the topic words extracted from NYT data set by CGS-LDA on the top panels and those by AVB-LDA on the bottom panels. We can observe that many of the topic words extracted by AVB-LDA are named-entity annotation, which has the zzz

[4] https://pytorch.org/

Table 2. Evaluations in terms of test perplexity with the optimal hyperparameters

	NIPS	STOF	NYT
VAE-DM	1494.57±2.90	884.66±2.00	2609.80±0.95
	$B = 4000$	$B = 10000$	$B = 1,000$
	$h_{\text{Enc}} = 1200$	$h_{\text{Enc},1} = 1600$	$h_{\text{Enc}} = 1000$
		$h_{\text{Enc},2} = 800$	
	$h_{\text{Dec}} = 1200$	$h_{\text{Dec}} = 800$	$h_{\text{Dec}} = 1000$
	$\eta_{\text{Enc}} = 0.001$	$\eta_{\text{Enc}} = 0.0002$	$\eta_{\text{Enc}} = 0.003$
	$\eta_{\text{Dec}} = 0.001$	$\eta_{\text{Dec}} = 0.0002$	$\eta_{\text{Dec}} = 0.01$
AVB-DM	1473.32±0.12	630.22±0.11	1794.88±0.01
	$B = 200$, $e = 300$	$B = 200$, $e = 300$	$B = 200$, $e = 200$
	$h_{\text{Enc}} = 600$	$h_{\text{Enc}} = 1600$	$h_{\text{Enc}} = 800$
	$h_{\text{Dis},1} = 600$	$h_{\text{Dis}} = 1600$	$h_{\text{Dis}} = 800$
	$h_{\text{Dis},2} = 300$		
	$h_{\text{Dec}} = 600$	$h_{\text{Dec}} = 1600$	$h_{\text{Dec}} = 800$
	$\eta_{\text{Enc}} = 0.00002$	$\eta_{\text{Enc}} = 0.0001$	$\eta_{\text{Enc}} = 0.001$
	$\eta_{\text{Dis}} = 0.003$	$\eta_{\text{Dis}} = 0.0001$	$\eta_{\text{Dis}} = 0.001$
	$\eta_{\text{Dec}} = 0.0004$	$\eta_{\text{Dec}} = 0.003$	$\eta_{\text{Dec}} = 0.003$
AVB-LDA	**1443.45±0.04**	**613.35±0.09**	**1747.42±0.01**
	$B = 200$, $e = 300$	$B = 200$, $e = 400$	$B = 200$, $e = 400$
	$h_{\text{Enc}} = 800$	$h_{\text{Enc},1} = 1600$	$h_{\text{Enc}} = 1000$
		$h_{\text{Enc},2} = 800$	
	$h_{\text{Dis},1} = 800$	$h_{\text{Dis},1} = 1600$	$h_{\text{Dis}} = 1000$
	$h_{\text{Dis},2} = 400$	$h_{\text{Dis},1} = 800$	
	$\eta_{\text{Enc}} = 0.04$	$\eta_{\text{Enc}} = 0.001$	$\eta_{\text{Enc}} = 0.001$
	$\eta_{\text{Dis}} = 0.01$	$\eta_{\text{Dis}} = 0.001$	$\eta_{\text{Dis}} = 0.001$
	$\eta_{\text{Dec}} = 0.001$	$\eta_{\text{Dec}} = 0.05$	$\eta_{\text{Dec}} = 0.1$
CGS-LDA	1524.47±0.30	843.57±0.32	3001.17±0.58
	$\alpha = 0.2$, $\beta = 0.025$	$\alpha = 0.07$, $\beta = 0.005$	$\alpha = 0.01$, $\beta = 0.01$

prefix in NYT data set. We can again say that AVB-LDA is likely to give relatively rarely used words as topic words, which makes it hard to figure out the corresponding content. If we are going to use topic modeling as described in the first paragraph of Section 1, this nature of the topic words given by AVB-LDA makes it difficult to adopt AVB-LDA.[5]

However, even when we used AVB-LDA, *document reconstruction* was successfully performed as shown in Table 3. For any unseen document d' AVB-LDA can provide its topic proportions $\boldsymbol{\theta}_{d'}$ as the softmax output of the encoder, i.e., $\boldsymbol{\theta}_{d'} = g(\boldsymbol{x}_{d'}, \boldsymbol{\epsilon})$. By passing $\boldsymbol{\theta}_{d'}$ to the decoder and sotfmax its output, we obtain the reconstructed word probabilities in d' as $\text{SoftMax}(\boldsymbol{B}\boldsymbol{\theta}_{d'} + \boldsymbol{b}_0)$. Therefore, we can obtain reconstructed word counts of d' as $n_{d'} \times \text{SoftMax}(\boldsymbol{B}\boldsymbol{\theta}_{d'} + \boldsymbol{b}_0)$, where $n_{d'}$ is the length of d'. Table 3 presents reconstructed word counts ob-

[5] Also in [18], where the collapsing of discrete variables is applied, Table 6 shows that there is a similar difference between the topic words obtained by ProdLDA and those by CGS-LDA. The words in the row labeled as ProdLDA seems more rarely used ones when compared to the words in the row labeled as LDA Collapsed Gibbs.

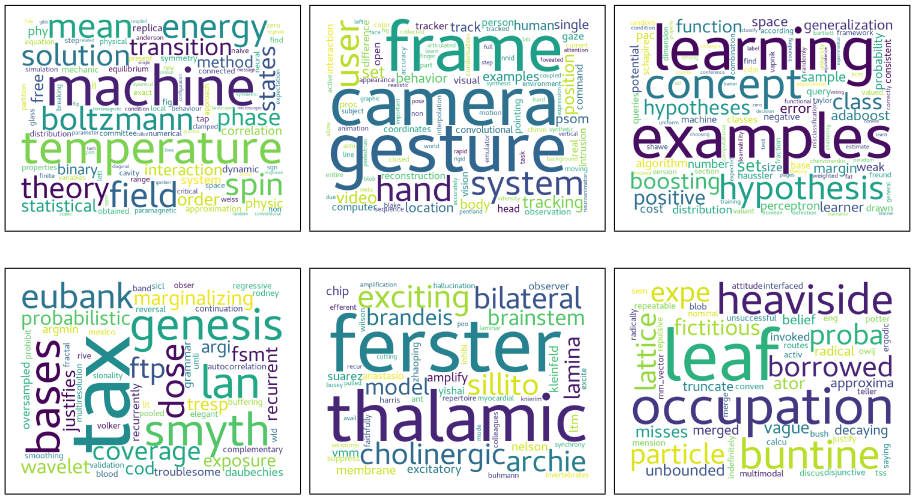

Fig. 2. High probability words extracted by CGS-LDA (the top three panels) and those by AVB-LDA (the bottom three panels) from NIPS data set.

tained by AVB-LDA for the four randomly chosen test documents in NYT data set.[6] The perplexity of each document is also given in parentheses. Words are sorted in decreasing order of their reconstructed counts. As the words of large reconstructed count show, their reconstruction counts fit well to the content of each test document. Therefore, even if AVB-LDA is not attractive for topic word visualization, it is still attractive for document reconstruction, which can be used for document-to-document similarity estimation. By regarding document reconstruction as presented in Table 3 as a kind of document expansion, we can perform information retrieval over the reconstructed documents. In this way AVB-LDA may mitigate the burden imposed on us by the excessive diversity of contents delivered by a huge set of documents. It is an interesting future research direction to evaluate AVB-LDA through document expansion in information retrieval.

4 Previous Work

The variational autoencoder (VAE) proposed by Kingma and Welling [11] has initiated a creative interaction between deep learning and Bayesian probabilistic modeling. The VAE was firstly applied to plain document modeling by Miao et al. [14] and then firstly applied to LDA by Srivastava et al. [18]. In VI presented in the original LDA paper [3], the approximate posterior for document-wise topic categorical distributions was prepared separately for each document. When compared to this, VAE for LDA has an advantage that it considers not only the document-wise structure but also the global structure of how the input vector

[6] The named-entity annotations are rewritten as proper names in Table 3.

Table 3. Reconstruction of word counts in test documents (NYT data set)

document A (perplexity = 1282.00)			document B (perplexity = 665.38)		
	actual count	reconst. count		actual count	reconst. count
genetically	2	12.11	team	5	4.61
corn	14	7.13	tournament	2	2.99
crop	1	5.28	game	4	2.87
scientist	5	5.26	Northridge	6	2.85
plant	4	3.94	scored	5	2.85
environmental	1	3.73	rebound	2	2.52
farmer	8	2.15	half	4	2.51
genes	0	2.06	loss	0	2.36
engineered	1	1.87	coach	2	2.24
modified	1	1.82	score	0	1.80
document C (perplexity = 1476.02)			**document D (perplexity = 1761.00)**		
	actual count	reconst. count		actual count	reconst. count
Elian	10	23.58	regulation	7	11.74
family	14	8.07	government	10	9.33
relatives	5	7.50	administration	8	5.62
Imigration	7	7.37	foreign	6	4.89
boy	13	7.28	trade	0	4.68
Washington	4	4.34	China	1	4.36
federal	5	3.87	rules	8	4.09
mother	0	2.87	United States	2	3.41
Cuba	7	2.76	technology	0	2.99
Miami	0	2.60	software	4	2.79

$\boldsymbol{x}_d$ provides the corresponding approximate posterior $q(\boldsymbol{\theta}_d|\boldsymbol{x}_d)$ [17, Section 3]. Adversarial learning in VI [10, 13] shares this advantage with VAE and has an additional advantage that the expressiveness of the approximate posterior is improved.

Our work is not just an application of the already proposed adversarial learning to VI for LDA. Our experiment showed that relatively simple MLPs, i.e., MLPs having at most two hidden layers, achieved better test perplexity than other methods. It also showed that AVB-LDA could not give interpretable topic words and that, however, AVB-LDA worked as document expansion. These are highly practical results, which could not be obtained if we were only talking about the theoretical possibility of the application of adversarial learning to VI for LDA. An important research direction discovered by our experiment is to modify the proposed adversarial learning for LDA so that the resulting topic words be easy to interpret.

5 Conclusions

In this paper we proposed an adversarial learning for LDA, where we used discriminator multilayer perceptron for estimating the log density ratio between the approximate posterior distribution and the prior distribution. The experimental

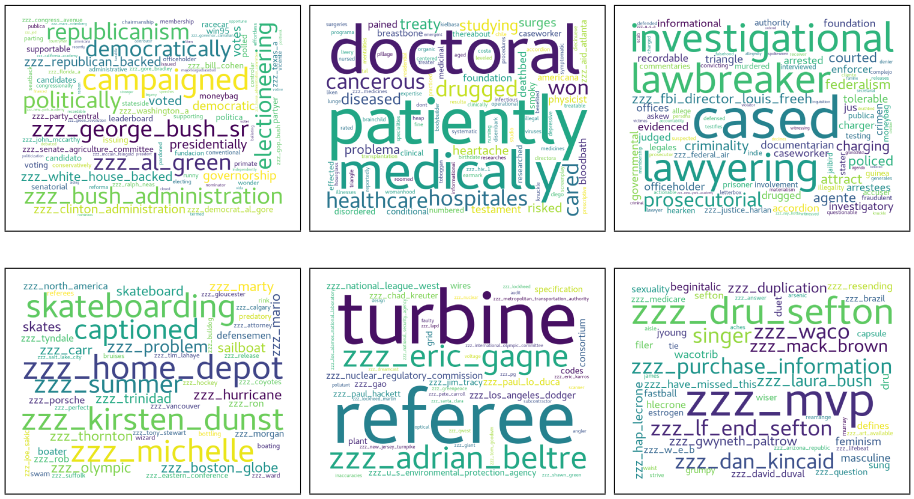

Fig. 3. High probability words extracted by CGS-LDA (the top three panels) and those by AVB-LDA (the bottom three panels) from NYT data set. Words with the zzz prefix are named-entity annotations.

results showed that our proposal AVB-LDA could achieve better test perplexity than the three compared methods, i.e., CGS-LDA, AVB-DM, and VAE-DM. However, it was observed that the high probability words in each topic obtained by AVB-LDA were harder to interpret than those obtained by CGS-LDA. While AVB-LDA worked as document expansion method, it is an important research direction to improve AVB-LDA in its interpretability. It should be noted that the proposed method is not the only way to use implicit distribution for approximating the true posterior in VI [15, 20]. Further, it is an interesting research direction to use other types of neural network, e.g., RNN [5] and CNN [12], as encoder and to provide adversarial learning for such encoders.

References

1. Asuncion, A., Welling, M., Smyth, P., Teh, Y.W.: On smoothing and inference for topic models. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. pp. 27–34. UAI '09 (2009)
2. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
4. Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. pp. 310–318. ACL '96 (1996)
5. Dieng, A.B., Wang, C., Gao, J., Paisley, J.W.: TopicRNN: A recurrent neural network with long-range semantic dependency. CoRR **abs/1611.01702** (2016), http://arxiv.org/abs/1611.01702

6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 249–256. AISTATS '10 (2010)
7. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. pp. 2672–2680. NIPS'14 (2014)
8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(suppl 1), 5228–5235 (2004)
9. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. pp. 1026–1034. ICCV '15 (2015)
10. Huszár, F.: Variational inference using implicit distributions. CoRR **abs/1702.08235** (2017), http://arxiv.org/abs/1702.08235
11. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. CoRR **abs/1312.6114** (2013), http://arxiv.org/abs/1312.6114
12. Lea, C., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks: A unified approach to action segmentation. In: Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III. pp. 47–54 (2016)
13. Mescheder, L.M., Nowozin, S., Geiger, A.: Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 2391–2400. ICML'17 (2017)
14. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. pp. 1727–1736. ICML'16 (2016)
15. Mohamed, S., Lakshminarayanan, B.: Learning in implicit generative models. CoRR **abs/1610.03483** (2016), http://arxiv.org/abs/1610.03483
16. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. pp. II–1278–II–1286. ICML'14 (2014)
17. Shu, R., Bui, H.H., Zhao, S., Kochenderfer, M.J., Ermon, S.: Amortized inference regularization. CoRR **abs/1805.08913** (2018), http://arxiv.org/abs/1805.08913
18. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. CoRR **abs/1703.01488** (2017), http://arxiv.org/abs/1703.01488
19. Titsias, M.K., Lázaro-Gredilla, M.: Doubly stochastic variational Bayes for non-conjugate inference. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. pp. II–1971–II–1980. ICML'14 (2014)
20. Uehara, M., Sato, I., Suzuki, M., Nakayama, K., Matsuo, Y.: Generative adversarial nets from a density ratio estimation perspective. CoRR **abs/1610.02920** (2016), http://arxiv.org/abs/1610.02920