Black Box Variational Inference on Latent Dirichlet Allocation

Aodong Li
Computer Science Department
New York University
New York, NY 10003
al5350@nyu.edu

Abstract

Latend Dirichlet Allocation (LDA) is an important member of topic models. It is used to model the generation process of a document. However, due to the triple-level hierarchy, the inference in LDA is typically hard. To tackle this problem, various special-purpose inference algorithms have been proposed. Black Box Variational Inference (BBVI)[6], on the other hand, is an instance of general-purpose inference algorithms. Although it is introduced as a generic algorithm to simplify the inference problem, BBVI has never been used for LDA. In this work, we apply BBVI to LDA and compare it with vanilla variational inference. Experiments show that BBVI achieves a slightly lower log likelihood than the vanilla variational inference.[1]

1 Introduction

Latend Dirichlet Allocation (LDA)[1] is a generative probabilistic model of corpus. But it is not limited to texts. It also generalizes into domains including content-based image retrieval, collaborative filtering, and bioinformatics. Such generalization is based on the property that a collection of data has latent components to enable clustering. LDA, however, goes beyond the clustering property and fits a generative model that also generalizes to unseen data, by which it is superior to its predecessors, e.g., probabilistic Latent Semantic Indexing (pLSI)[4]. Wide as the applicable domains are, in this paper, we use the language of texts to illustrate the ideas for convenience.

The graphical representation of LDA is illustrated in Figure 1 and the generation of a document can be seen as follows. For each document d, we first sample its topic distribution θ_d, the parameter of a multinomial distribution, from a Dirichlet distribution with parameter α,

$$\theta_d \sim \text{Dir}(\theta|\alpha).$$

Note that θ_d is a valid parameter for multinomial distribution because $\sum_j^k \theta_{d,j} = 1$. Then for each word n in document d, we sample its topic $z_{d,n}$ from a multinomial distribution with parameter θ_d,

$$z_{d,n} \sim \text{Mult}(z|\theta_d).$$

Finally, the corresponding word $w_{d,n}$ is sampled from a multinomial distribution specific to topic $z_{d,n}$,

$$w_{d,n} \sim \text{Mult}(w|\beta_{z_{d,n}})$$

where $\beta_{z_{d,n}}$ is the parameter of a multinomial distribution over the vocabulary.

The motivation of LDA comes from generalizing the "bag-of-words" assumption for a document and the "bag-of-documents" assumption for a corpora. These assumptions ignore the order of words in a

[1] Code can be found at https://github.com/aodongli/Black-Box-Variational-Inference-on-Latent-Dirichlet-Allocation

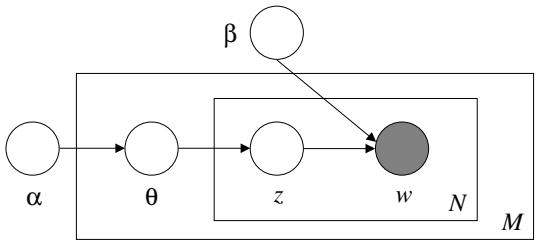

Figure 1: Graphical representation of LDA. The corpus has M documents and each document involves N words.

document, and the order of documents in a corpora. Remarkably, order-ignorant assumptions coincide with the more general notion of *exchangeability* in the language of probability theory, by which we are able to utilize the probability tools to better solve the problem. Among the probability theorems, de Finetti's representation theorem[2] says that a collection of exchangeable random variables has a representation as a mixture distribution – in general an infinite distribution. The theorem essentially dictates a conditionally independent and identically distributed assumption over the random variables we want to model. This insight reflects the design of probability of a sequence of words and topics in LDA, which can be treated as an infinite mixture over an underlying set of topic distributions governed by a multinomial parameter:

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(w_n|z_n) p(z_n|\theta) \right) d\theta.$$

Elegant as LDA is, its triple-level structure posits some difficulty on the inference of the parameters. Thus a lot of special-purpose inference algorithms have been proposed to tackle this problem, e.g., collapsed Gibbs sampling[5], collapsed variational inference[7], and stochastic collapsed variational Bayesian inference[3], etc. On the other hand, Black Box Variational Inference (BBVI)[6] is proposed as a generic inference algorithm and has not been used for LDA. So in this paper, we apply BBVI to LDA. The contributions include

- We re-derive the vanilla variational inference algorithm on LDA in a general-to-specific way by the power of exponential families[8].
- We derive the inference formula of LDA under the framework of BBVI.
- We implement it with an extension of Adagrad algorithm[9].

2 Variational Inference on LDA

Different from the original derivation in [1], we derive the variational inference algorithm for LDA in a general-to-specific way, in which we utilize the representation of exponential families.

2.1 Generic Mean-Field Variational Inference

We first derive the lower bound of the log-likelihood,

$$\begin{aligned}
\ln p(X) &= \ln \int_z p(X, Z) dz \\
&= \ln \int_z p(X, Z) \frac{q(Z)}{q(Z)} dz \\
&= \ln \mathop{\mathrm{E}}_{q} \left[\frac{p(X, Z)}{q(Z)} \right] \\
&\geq \mathop{\mathrm{E}}_{q} \left[\ln \frac{p(X, Z)}{q(Z)} \right] \\
&= \mathop{\mathrm{E}}_{q} \left[\ln p(X, Z) - \ln q(Z) \right] \\
&= \mathcal{L}(q)
\end{aligned}$$

where the inequality comes from the concavity of $\ln(\cdot)$. The lower bound $\mathcal{L}(q)$ is called evidence lower bound (ELBO) and $q(Z)$ is the variational distribution.

For a successful derivation, we use mean-field approximation for $q(Z) = \prod_{i=1}^{m} q_i(Z_i)$ or it is equivalent to say that the latent variational variables factorize. If we concentrate on one particular component Z_j, then $\mathcal{L}(q)$ can be written as

$$\begin{aligned}\mathcal{L}(q_j) &= \int_{z_j} q_j(z_j) \mathop{\mathrm{E}}_{q_{\neq j}}[\ln p(X, Z)] dz_j - \sum_{i=1}^{m} \mathop{\mathrm{E}}_{q_i}[\ln q_i(Z_i)] \\ &= \int_{z_j} q_j(z_j) \mathop{\mathrm{E}}_{q_{\neq j}}[\ln p(X, Z)] dz_j - \mathop{\mathrm{E}}_{q_j}[\ln q_j(Z_j)] + \text{const.}\end{aligned}$$

Furthermore, let $\ln \widetilde{p}(X, Z_j) = \mathrm{E}_{q_{\neq j}}[\ln p(X, Z)]$, then $\widetilde{p}(X, Z_j)$ is a pseudo distribution for $p(X, Z_j)$ in that $\widetilde{p}(X, Z_j) = \exp\left(\int q_{\neq j}(Z_{\neq j}) \ln p(X, Z) dZ_{\neq j}\right)$. Also we have

$$\mathcal{L}(q_j) = \int_{z_j} q_j(z_j) \ln \left[\frac{\widetilde{p}(X, Z_j)}{q_j(Z_j)}\right] dz_j + \text{const},$$

which is the same as $-\mathbb{KL}(\widetilde{p}(X, Z_j)||q_j(Z_j))$. Thus to maximize $\mathcal{L}(q)$, $\mathbb{KL}(\widetilde{p}(X, Z_j)||q_j(Z_j))$ should be minimized for every $q_j(Z_j)$. Thus it provides an update rule

$$\ln q_j^*(Z_j) = \widetilde{p}(X, Z_j) = \mathop{\mathrm{E}}_{q_{\neq j}}[\ln p(X, Z)], \tag{1}$$

which is an iterative algorithm for all $q_j^*(Z_j), j = 1, ..., m$.

2.2 Dirichlet distribution in exponential families

For Dirichlet distribution

$$\text{Dir}(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1},$$

it can be written in exponential family as

$$\exp\left\{\sum_{i=1}^{k} (\alpha_i - 1) \ln \theta_i - \sum_{i=1}^{k} \ln \Gamma(\alpha_i) + \ln \Gamma\left(\sum_{i=1}^{k} \alpha_i\right)\right\}.$$

The sufficient statistic is $T(\theta_i) = \ln \theta_i$. So by the property of the log normalizer in exponential family distribution, we conclude

$$\mathop{\mathrm{E}}_{p(\theta_i|\alpha)}[\ln \theta_i] = \Psi(\alpha_i) - \Psi\left(\sum_{i=1}^{k} \alpha_i\right)$$

where Ψ is di-gamma function and $\Psi(\alpha_i) = \frac{\Gamma'(\alpha_i)}{\Gamma(\alpha_i)}$, which is the derivative of log gamma function.

2.3 Mean-Field Variational Inference on LDA

For LDA, we want to infer $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$ given a document $\mathbf{w}$. We want to remove the problematic connections between β and θ. By mean-field approximation, we use $q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{i=1}^{N} q(z_i|\phi_i)$ where γ and ϕ are the variational parameters controlling the distribution.

First we derive $q(\theta|\gamma)$. The right hand side of (1) is

$$\begin{aligned}&\mathop{\mathrm{E}}_{q(Z|\phi)}[\ln p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] \\ =&\mathop{\mathrm{E}}_{q(Z|\phi)}[\ln p(\theta|\mathbf{z}, \alpha) + \ln p(\mathbf{z}, \mathbf{w}|\beta)] \\ =&\mathop{\mathrm{E}}_{q(Z|\phi)}[\ln p(\theta|\mathbf{z}, \alpha)] + \text{const} \\ =&h(\theta) + T(\theta)^\top \mathop{\mathrm{E}}_{q(Z|\phi)}[\eta_g(\mathbf{z}, \alpha)] + \text{const},\end{aligned}$$

where we used the exponential family representation of $p(\theta|\mathbf{z},\alpha) \propto h(\theta)\exp\left(T(\theta)^\top \eta_g(\mathbf{z},\alpha)\right)$ and did not write the normalizer explicitly. Also in exponential family representation, the left hand side of (1) is

$$h(\theta) + T(\theta)^\top \eta_g(\gamma) - A_g(\gamma).$$

Then make both sides equal and omit the normalizer, we have

$$\eta_g(\gamma) = \mathop{\mathrm{E}}_{q(Z|\phi)}\left[\eta_g(\mathbf{z},\alpha)\right]. \tag{2}$$

Then in a similar fashion, we can derive $q(z_i|\phi_i)$ in a parameterizing form

$$\eta_l(\phi_i) = \mathop{\mathrm{E}}_{q(\theta|\gamma)}\left[\eta_l(w_i,\beta,\theta)\right]. \tag{3}$$

Now we consider the posterior probability of $p(z_{d,n}=j|\theta_d, w_{d,n}, \beta)$ in exponential family form, which can be expressed as

$$\begin{aligned} p(z_{d,n}=j|\theta_d,w_{d,n},\beta) &\propto p(z_{d,n}=j|\theta_d)p(w_{d,n}|z_{d,n}=j,\beta) \\ &= \text{Mult}(\theta_{d,j})\text{Mult}(\beta_{j,w_{d,n}}) \\ &\propto \exp\left(\left(\ln\theta_{d,j} + \ln\beta_{j,w_{d,n}}\right)\cdot 1\right). \end{aligned}$$

Thus the natural parameters are

$$\eta_l(w_{d,n},\beta,\theta_d) = \ln\theta_{d,j} + \ln\beta_{j,w_{d,n}}.$$

Based on (3), if we take multinomial distribution over z, then

$$\begin{aligned} \ln(\phi_{d,n}^j) &= \mathop{\mathrm{E}}_{q(\theta_d|\gamma_d)}\left[\ln\theta_{d,j} + \ln\beta_{j,w_{d,n}}\right] \\ &= \ln\beta_{j,w_{d,n}} + \Psi(\gamma_{d,j}) - \Psi\left(\sum_{i=1}^k \gamma_{d,i}\right), \end{aligned}$$

where we used the property of Dirichlet distribution. Then rewrite them and we get

$$\phi_{d,n}^j \propto \beta_{j,w_{d,n}} \exp\left(\Psi(\gamma_{d,j}) - \Psi\left(\sum_{i=1}^k \gamma_{d,i}\right)\right). \tag{4}$$

On the other hand, we consider the posterior probability of $p(\theta_d|\alpha,\mathbf{z}_d,\mathbf{w}_d)$ in exponential family form, which can be expressed

$$\begin{aligned} p(\theta_d|\alpha,\mathbf{z}_d,\mathbf{w}_d) \propto& p(\theta_d|\alpha)p(\mathbf{z}_d|\theta_d) \\ =& \text{Dir}(\alpha)\prod_{n=1}^N \text{Mult}(z_{d,n}|\theta_d) \\ \propto& \exp\left(\sum_{j=1}^k(\alpha_j-1)\ln\theta_{d,j} + \sum_{j=1}^k\sum_{n=1}^N \delta(z_{d,n},j)\ln\theta_{d,j}\right) \\ =& \exp\left(\sum_{j=1}^k\left((\alpha_j-1) + \sum_{n=1}^N\delta(z_{d,n},j)\right)\ln\theta_{d,j}\right). \end{aligned}$$

By focusing on what we are interested in, we can see

$$\eta_g(\alpha,\mathbf{z}_d) = \left(\alpha_1 - 1 + \sum_{n=1}^N \delta(z_{d,n},1), ..., \alpha_k - 1 + \sum_{n=1}^N \delta(z_{d,n},k)\right).$$

Similarly, in terms of natural parameter, based on (2),

$$\begin{aligned}\eta_g(\gamma_d) &= \mathop{\mathrm{E}}_{q(z_{d,n}|\phi_{d,n})}\left[\eta_g(\alpha, \mathbf{z}_d)\right] \\ &= \mathop{\mathrm{E}}_{q(z_{d,n}|\phi_{d,n})}\left[\left(\alpha_1 - 1 + \sum_{n=1}^{N}\delta(z_{d,n},1), ..., \alpha_k - 1 + \sum_{n=1}^{N}\delta(z_{d,n},k)\right)\right] \\ &= \left(\alpha_1 - 1 + \sum_{n=1}^{N}\delta(z_{d,n},1)\phi_{d,n}^1, ..., \alpha_k - 1 + \sum_{n=1}^{N}\delta(z_{d,n},k)\phi_{d,n}^k\right).\end{aligned}$$

Using the fact of $\eta_g(x) = x - 1$, the variational parameter is

$$\gamma_d = \left(\alpha_1 + \sum_{n=1}^{N}\delta(z_{d,n},1)\phi_{d,n}^1, ..., \alpha_k + \sum_{n=1}^{N}\delta(z_{d,n},k)\phi_{d,n}^k\right). \tag{5}$$

In summary, the inference algorithm is a coordinate ascent over (4) and (5).

3 Black Box Variational Inference on LDA

Given certain conditions (Robbins-Monro conditions) on learning rate, stochastic optimization algorithms are guaranteed to converge to a maximum of an objective function $f(x)$:

$$x_{t+1} \leftarrow x_t + \rho_t h_t(x_t)$$

where $h_t(x_t)$ is a realization of random variable $H(x)$ whose expectation is the gradient of $f(x)$[6].

Given variational distribution $q(z|\lambda)$, by [6] the gradient of ELBO can be derived to be

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathop{\mathrm{E}}_{q(z|\lambda)}\left[\nabla_\lambda \ln q(z|\lambda)(\ln p(x,z) - \ln q(z|\lambda))\right]. \tag{6}$$

Thus we are able to perform inference using gradient ascent. The gradient can be estimated by Monte Carlo method. Here comes the definition of black box, we only require the model $\ln p(x,z)$ is evaluable and then we are able to do inference in a model-agnostic way.

However, one of the shortage is that the estimator of (6) is that it is of high variance and it is hard to compute the gradients for all variational parameters at one time.

To tackle this, one variance reduction technique is to Rao-Blackwellize the estimator. By the implication of Rao-Blackwell theorem on mean-field approximation, all we need to do is to integrate out some irrelevant variables[6],

$$\mathrm{E}[J(X,Y)|X] = \frac{\int J(x,y)p(x)p(y)dy}{\int p(x)p(y)dy} = \mathop{\mathrm{E}}_{p(y)}[J(X,Y)].$$

With this implication, first we derive the gradient of γ and integrate out ϕ,

$$\begin{aligned}\nabla_\gamma \mathcal{L}(\gamma) &= \mathop{\mathrm{E}}_{q(\theta_d|\gamma)}\left[\nabla_\gamma \ln q(\theta_d|\gamma) \mathop{\mathrm{E}}_{q(\mathbf{z}_d|\phi_d)}\left[\ln p(\theta_d, \mathbf{z}_d, \mathbf{w}_d|\alpha,\beta) - \ln q(\theta_d|\gamma) - \ln q(\mathbf{z}_d|\phi_d)\right]\right] \\ &= \mathop{\mathrm{E}}_{q(\theta_d|\gamma)}\left[\nabla_\gamma \ln q(\theta_d|\gamma)\left(\ln p(\theta_d|\alpha) + \mathop{\mathrm{E}}_{q(\mathbf{z}_d|\phi_d)}\left[\ln p(\mathbf{z}_d|\theta_d)\right] - \ln q(\theta_d|\gamma) + C\right)\right] \\ &= \mathop{\mathrm{E}}_{q(\theta_d|\gamma)}\left[\nabla_\gamma \ln q(\theta_d|\gamma)\left(\ln p(\theta_d|\alpha) + \mathop{\mathrm{E}}_{q(\mathbf{z}_d|\phi_d)}\left[\ln p(\mathbf{z}_d|\theta_d)\right] - \ln q(\theta_d|\gamma)\right)\right],\end{aligned}$$

in which for each term we have

$$\begin{aligned}\nabla_\gamma \ln q(\theta_d|\gamma) &= \ln \theta_d - \Psi(\gamma) + \Psi\left(\sum_{i=1}^{k}\gamma_i\right), \\ \ln p(\theta_d|\alpha) &= \ln \mathrm{Dir}(\theta_d|\alpha), \\ \ln q(\theta_d|\gamma) &= \ln \mathrm{Dir}(\theta_d|\gamma),\end{aligned}$$

and $\mathrm{E}_{q(\mathbf{z}_d|\phi_d)}\left[\ln p(\mathbf{z}_d|\theta_d)\right]$ can be evaluated by Monte-Carlo estimation (please see code for explicit representation).

Similarly, we derive the gradient of ϕ and integrate out γ,

$$\nabla_{\phi_{d,n}} = \mathop{\mathrm{E}}_{q(z_{d,n}|\phi_{d,n})}\left[\nabla_{\phi_{d,n}} \ln q(z_{d,n}|\phi_{d,n}) \left(\ln p(w_{d,n}|z_{d,n},\beta) - \ln q(z_{d,n}|\phi_{d,n}) + \mathop{\mathrm{E}}_{q(\theta_d|\gamma)}\left[\ln p(z_{d,n}|\theta_d)\right]\right)\right],$$

where we used the property of Dirichlet distribution and each term in the formula is

$$\begin{aligned}
\nabla_{\phi_{d,n}} \ln q(z_{d,n}|\phi_{d,n}) &= \begin{cases} 0, & z_{d,n} \neq j, \\ \frac{1}{\phi^j_{d,n}}, & z_{d,n} = j, \end{cases} \\
\ln p(w_{d,n}|z_{d,n},\beta) &= \ln \beta_{z_{d,n},w_{d,n}}, \\
\ln q(z_{d,n}|\phi_{d,n}) &= \ln \phi^{z_{d,n}}_{d,n},
\end{aligned}$$

and $\mathrm{E}_{q(\theta_d|\gamma)}\left[\ln p(z_{d,n}|\theta_d)\right]$ can be evaluated by Monte-Carlo estimation (please see code for explicit representation).

Then we can do the inference by iteratively updating those parameters.

4 Experiments

In this section, we describe the dataset we used and the experiments we conducted to compare the performance of different variational inference algorithms.

4.1 Dataset

20 Newsgroups dataset[2] is used to evaluate the performances of different algorithms. The dataset consists of 20 different classes. For this task, we basically do not distinguish classes and treat the inference and learning in an unsupervised way.

We used the pre-split training set for parameter learning. But for the inference task, we only used 100 documents of the test set. In summary, we used 11269 documents for training and 100 documents for comparisons of inference algorithms.

4.2 Training

We build the model as described above and estimate the parameters α and β by variational EM algorithm[1]. In E step, we update the variational latent parameters by the formulas in the previous section. In M step, we maximize the ELBO with variational latent variables fixed. The implementation uses the standard LDA code[3].

We use randomly initialized parameters and set the number of topics to be 20. The whole training set is employed to estimate the parameters. After training we save the parameters and keep them fixed during inference.

4.3 Inference

We compared two versions of BBVI with vanilla variational inference (VI) on LDA. Specifically, we did BBVI with stochastic gradient descent (SGD) and Adagrad[9] repectively. The gradients are estimated by 10 samples. For BBVI with SGD, the learning rate is set to be 1e-4, while for BBVI with Adagrad, the initial learning rate is set to be 1e-1 and is adjusted automatically for each parameter during update.

Inference is performed on 100 test documents. The results are plotted in Figure 2. VI in general behaves better than BBVI because the estimated gradient incorporates noise, leading to a noisy estimate. But it is worth noting that BBVI with Adagrad behaves much worse than BBVI with SGD, which should be investigated in the future.

[2] http://qwone.com/∼jason/20Newsgroups/

[3] http://www.cs.columbia.edu/∼blei/lda-c/

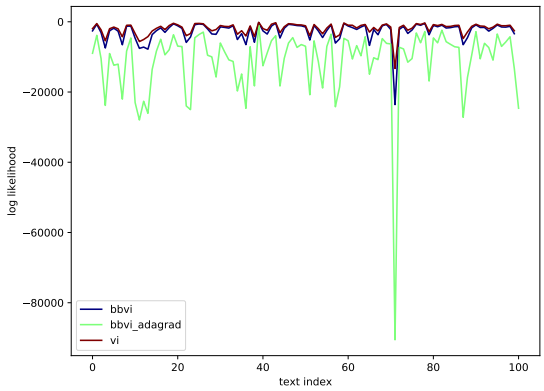

Figure 2: Log likelihood on testset documents.

5 Future work

First, the reason why Adagrad behaves much worse than SGD should be investigated. In addition, more systematic comparison of algorithmic efficiency is worth being conducted because some random variables are hard to sample. Finally, comparison with other sampling methods, like Gibbs sampling, should be conducted.

6 Conclusion

In this project, we apply Black Box Variational Inference to Latent Dirichlet Allocation. The derivation work of BBVI is greatly reduced compared to vanilla variational inference. Experiments show BBVI with SGD achieves a slightly lower log likelihood than VI, but BBVI with Adagrad has much lower log likelihood, which should be investigated.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Bruno De Finetti. *Theory of probability: a critical introductory treatment*, volume 6. John Wiley & Sons, 2017.

[3] 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*, pages 446–454. ACM, 2013.

[4] Thomas Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Forum*, volume 51, pages 211–218. ACM, 2017.

[5] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM, 2008.

[6] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[7] Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2007.

[8] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[9] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.