Topic Model with meta-data via SGRLD

Peshal Agarwal
13817472

Yash Travadi
13817748

Asim Unmesh
13907167

Javed Akhtar
13817324

1 Problem Description and Motivation

In this project we have surveyed Stochastic Gradient Langevin Dynamics, its variants and its application to Dirichlet-Multinomial Regression (DMR) model. Monte-Carlo approaches have been extensively used for Bayesian inference, and proven to be superior over many methods. However, MCMC procedures generally have a random-walk behaviour and thus leads to slow convergence. We explore an old but classic concept of Physics called the Langevin Dynamics derived from Hamiltonian mechanics and Reimannian geometry which have been recently applied to utilize the gradient information making it computationally less expensive particularly for large datasets. This technique provide fast convergence to the true posterior without over-fitting to the data.

2 Literature Review

In the modern scenario, there is an increasing demand for scalable MCMC methods for probabilistic modeling and inference over large scale datasets. A lot of recent work has been done focused on this aspect. Several modification, improvements and novel ideas have been proposed, which we surveyed in this project. The main focus is to propose a stochastic mini-batch versions of the MCMC algorithms to overcome the computationally intensive framework of the traditional methods. In 2010, Neal[2], proposed a new class of algorithm called Hamiltonian Dynamics having its origin in Statistical Physics. Inspired by this work, Welling and Teh[4], applied this technique to modify the update equation of parameters in the stochastic gradient step, by adding a noise term, which is known as Stochastic Gradient Langevin Dynamics(SGLD). They hence proposed an algorithm to perform MCMC which is similar to Stochastic Gradient Descent, where each update to the parameters uses only a subset of the data, instead to using the entire dataset. Stochastic Gradient Descent has made a huge impact on reducing learning time for non-Bayesian algorithms and SGLD extends its use for Bayesian learning as well. However, SGLD has its limitations as it couldn't be applied when concerned parameters lie in some constrained sets, for example in a probability simplex. To address this specific issue, i.e. application of SGLD when the parameters lie in a probability simplex, Patterson and Teh[3] (2013), proposed a new method called Stochastic Gradient Riemannian Langevin dynamic(SGRLD) and demonstrated its application in topic modelling using Latent Dirichlet Allocation(LDA) in an online mini-batch setting. There are many variants of LDA where application of SGRLD can be investigated. One of them is Dirichlet -multinomial Regression (DMR) for topic model with metadata. Below, we give an intelligible description of each of the SGLD, SGRLD and DMR models.

2.1 Stochastic Gradient Langevin Dynamics

SGLD is basically a combination of two well known class of algorithm i.e Stochastic Optimization and Langevin Dynamics. Langevin Dynamics injects noise into the gradient descent parameter updates. This allows the algorithm to explore the whole posterior rather than just converging to a MAP estimate.

Let θ denote parameter vector, $p(\theta)$ a prior distribution and $p(x|\theta)$, likelihood. Let $X = \{x_i\}_{i=1}^N$ be data. Then posterior is given by

$$p(\theta|X) \propto p(\theta) \prod_{i=1}^{N} p(x_i|\theta) \tag{1}$$

Stochastic Optimization: It is one of the mostly widely used class of algorithm for machine learning often used to scale up algorithms to extend their use to big data. It processes a mini-batch of dataset in each iteration of parameter update, similar to working in an online setting. These mini-batch parameter updates are equal to the batch-updates in expectation.
In this framework, at each iteration t, a subset of n data items $X_t = \{x_{t1}, ..., x_{tn}\}$ is given, and the parameters are updated as follows:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta)\right) \tag{2}$$

where the step sizes satisfy the following condition:

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty \tag{3}$$

The above constraints on step size are required to ensure convergence to a local maxima. The first condition on step size ensures that the parameters will reach high probability regions irrespective of where it was initialized, while the second condition ensures its convergence to the mode. To ensure that step size decreases with time, Welling and Teh[4] suggest to set them as $\epsilon_t = a(b + t)^{-\gamma}$, where $\gamma \in (0.5, 1]$

Langevin Dynamics: This method of optimization injects a Gaussian noise into the parameter updates of gradient descent steps. The injected noise is a zero mean Gaussian with variance equal to twice the step size for that iteration. Since the updates are stochastic, this method gives a distribution over the parameters unlike gradient descent which only gives a point estimate of the parameters. The updates are given as follows:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \sum_{i=1}^{n} \nabla \log p(x_t|\theta)\right) + \eta_t \tag{4}$$

where $\eta_t \sim \mathcal{N}(0, \epsilon_t)$

Stochastic Gradient Langevin Dynamics: Combining the Stochastic Optimization and Langevin Dynamics, Welling and Teh[4] propose the SGLD update as follows:

$$\Delta\theta_t = \frac{\epsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{N}{n}\sum_{i=1}^{n} \nabla \log p(x_{ti}|\theta_t)\right) + \eta_t \tag{4}$$

where $\eta_t \sim \mathcal{N}(0, \epsilon_t)$ and step size decreases as $\epsilon_t = a(b + t)^{-\gamma}$, where $\gamma \in (0.5, 1]$.
Initially when the gradients are large, the former term of the update is dominating and the algorithm is said to be in the Stochastic Optimization stage. Eventually as the gradients become small, the noise term dominates

and the algorithm is said to have transitioned into Langevin Dynamics stage for sufficiently large t. Once the algorithm has entered the Langevin Dynamics phase, it starts to generate samples from a distribution that is "close" to the posterior. It does allow to capture uncertainty in the estimates of parameters in a Bayesian manner. Another advantage of this algorithm is that the MH rejection rate goes to zero asymptotically as t increases, which is favourable for scalability. A detailed discussion on the convergence analysis of this algorithm can be found in Welling and Teh[4].

2.2 Stochastic Gradient Riemannian Langevin dynamics

Stochastic Gradient Langevin Dynamics has greatly helped in building highly scalable probabilistic model. However, when it comes to constrained set of parameters it can't be applied. To address one such issue, specifically when parameter lies in a probability simplex, Patterson and Teh[3] (2013) proposed a new method combining the idea of reparameterization from Reimannian geometry and SGLD. Parameter lying on a probability simplex is given by,

$$\Delta_K = \{(\pi_1, \pi_2, ...\pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\} \subseteq \mathbb{R}^K \tag{5}$$

There are two major issues which have to be addressed and solved before applying SGLD over this parameter. First of all, the above probability simplex is compact and has boundary. Therefore occurrences of updates that brings the vector out of the simplex has to be taken care of. Secondly, in practice such parameters could be sparse as in LDA, i.e most of the entries are close to zero and masses are assigned mostly to corners and boundary of the simplex. In case of LDA gradients calculation in SGLD requires inverting the entries of π, it results in the problem of gradient being blown up.

To solve such issues, different ways of parameterizing the probability simplex are considered. It comes out that the choice of parameterization is not obvious but guided by Riemannian geometry of the simplex. The different kinds of parameterization are summarized in the figure below.

Parameterisation	Reduced-Mean	Reduced-Natural	Expanded-Mean	Expanded-Natural
θ	$\theta_k = \pi_k$	$\theta_k = \log \frac{\pi_k}{1-\sum_k^{K-1} \pi_k}$	$\pi_k = \frac{\lvert\theta_k\rvert}{\sum_{k=1} \lvert\theta_k\rvert}$	$\pi_k = \frac{e^{\theta_k}}{\sum_{k=1} e^{\theta_k}}$
$\nabla_\theta \log p(\theta \mid \mathbf{x})$	$\frac{n+\alpha}{\theta} - \mathbf{1}\frac{n_K+\alpha-1}{\pi_K}$	$n + \alpha - (n. + K\alpha)\,\pi$	$\frac{n+\alpha-1}{\theta} - \frac{n.}{\theta.} - \mathbf{1}$	$n + \alpha - n.\pi - \mathrm{e}^{\theta}$
$G(\theta)$	$n. \left(\mathrm{diag}(\theta)^{-1} + \frac{1}{1-\sum_k \theta_k}\mathbf{1}\mathbf{1}^T\right)$	$\frac{1}{n.}\left(\mathrm{diag}(\pi) - \pi\pi^T\right)$	$\mathrm{diag}\,(\theta)^{-1}$	$\mathrm{diag}\,(\mathrm{e}^{\theta})$
$G^{-1}(\theta)$	$\frac{1}{n.}\left(\mathrm{diag}(\theta) - \theta\theta^T\right)$	$n. \left(\mathrm{diag}(\pi)^{-1} + \frac{1}{1-\sum_k \pi_k}\mathbf{1}\mathbf{1}^T\right)$	$\mathrm{diag}\,(\theta)$	$\mathrm{diag}\,(\mathrm{e}^{-\theta})$
$\sum_{k=1}^{D}\left(G^{-1}\frac{\partial G}{\partial \theta_k}G^{-1}\right)_{jk}$	$K\theta_j - 1$	$\frac{1}{\pi_j^2} - \frac{K-1}{(1-\sum_k \pi_k)^2}$	-1	$\mathrm{e}^{-\theta_j}$
$\sum_{k=1}^{D}\left(G^{-1}(\theta)\right)_{jk} \mathrm{Tr}\left(G^{-1}(\theta)\frac{\partial G}{\partial \theta_k}\right)$	$K\theta_j - 1$	$\frac{1}{\pi_j^2} - \frac{K-1}{(1-\sum_k \pi_k)^2}$	-1	$\mathrm{e}^{-\theta_j}$

Figure 1: Table Courtesy Patterson and Teh [3]

Among the possible parameterization that take care of the issues, expanded mean parameterization is used for our problem of probability simplex. The preconditioning matrix $G(\theta)$ corresponding to this choice of reparameterization is diagonal which makes it computationally efficient.

Having investigated the Riemannian geometry, the Langevin dynamics is modified by incorporating a preconditionaing matrix $G(\theta)$ to overcome the limitation of isotropic proposal distribution in the case of Langevin dynamics and the new algorithm is called Riemannian Langevin dynamics. The update equation in this case becomes,

$$\theta^* = \theta + \frac{\epsilon}{2}\mu(\theta) + G^{-\frac{1}{2}}(\theta)\zeta, \quad \zeta \sim \mathcal{N}(0, \epsilon I) \tag{6}$$

where the j^{th} component of $\mu(\theta)$ is given by,

$$\mu(\theta)_j = \left(G^{-1}(\theta) \left(\nabla_\theta \log p(\theta) + \sum_{i=1}^{n} \nabla_\theta \log p(x_i|\theta) \right) \right)_j - 2 \sum_{k=1}^{D} \left(G^{-1}(\theta) \frac{\partial G(\theta)}{\partial \theta_k} G^{-1}(\theta) \right)_{jk} + \sum_{k=1}^{D} \left(G^{-1}(\theta) \right)_{jk} Tr \left(G^{-1}(\theta) \frac{\partial G(\theta)}{\partial \theta_K} \right) \quad (7)$$

The first term in (7) gives the natural gradient of the log posterior. Note that standard gradient gives the direction of steepest ascent in Euclidean space. In contrast, the natural gradient gives the direction of steepest descent by incorporating knowledge of geometry given by preconditioning matrix $G(\theta)$. Rest of the terms gives curvature of the manifold defined by $G(\theta)$ for small changes in θ. Note that the update equation depends on the choice of $G(\theta)$ which in turn depends on the choice of parametrization.

2.3 Dirichlet-multinomial Regression (DMR) for topic model with metadata

Text data sometimes contain meta information such as authors, publications, dates, etc. However, the standard LDA model does not incorporate any of these. Several extensions of the basic model have been proposed to take this meta data into account. But most of these extension form a fully generative story. In a recent work Minmo and McCallum ([1]), a regression model has been proposed in which they regress on the metadata. The DMR model is quite accommodative in terms of type of metadata, i.e. it can be continuous (e.g. rating) as well discrete (e.g. year, journal) keeping the inference simple. Below is the graphical model.

The graphical notation is exactly the same as shown in figure (2) however, $\theta_k \sim Dirichlet(\beta)$ for each topic k as in the standard LDA model, but in our model we propose a different data generation scheme (described in section 3).

3 Novel Contribution

The work by Mimno and McCallum [1] on DMR topic model use stochastic EM sampling scheme to train their model. However, since we would use SGRLD to perform inference, we have redefined the generative story of the model.

Notations

- x_d is the feature vector of the metadata
- N_d denotes the number of words in each document
- η_d is the topic proportion vector
- λ_k is the feature vector of each topic
- π_k is word proportion vector

Data Generation

1. For each topic $k = 1, \ldots, K$
   - Draw $\lambda_k \sim \mathcal{N}(0, \sigma^2 I)$
   - For each word w in the vocabulary
     - Draw $\theta_{kw} \sim Gamma(\beta, 1)$
     - Set $\pi_{kw} = \frac{\theta_{kw}}{\sum_w \theta_{kw}}$
2. For each document d
   - For each topic k, set $\alpha_{dk} = \exp(x_d^T \lambda_k)$

- Draw $\eta_d \sim Dirichlet(\alpha_d)$
- For each word $n = 1, \ldots, N_d$ in document d
  - Choose the topic for the word $z_{dn} \sim multinoulli(\eta_d)$
  - Generate word from the chosen topic $w_{dn} \sim multinoulli(\pi_{z_{dn}})$

The graphical notation of our model would be as shown below

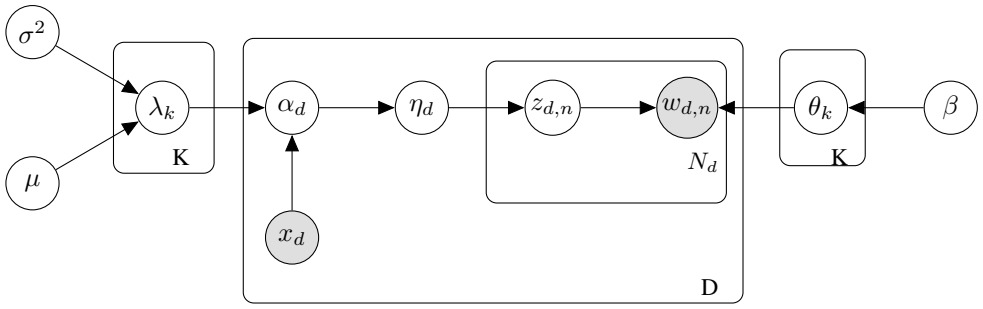


Figure 2: Graphical Model Notation

Note that due to the *nice* structure of our model we collapse η_d. We have three hyper-parameters namely μ, σ^2 and β. This leaves us with λ , θ and z as unknown parameters, whose updates are derived using SGRLD.

3.1 Parameter Updates

Note that the standard LDA model does not have any meta-information to model and thus has same α associated with every document. However, in our model we have α_d for each document. We derive the SGRLD updates as follows:

$$\theta^*_{kw} = \left| \theta_{kw} + \frac{\epsilon}{2}\left(\beta - \theta_{kw} + \frac{|D|}{|D_t|} \sum_{d \in D_t} \mathbb{E}_{z_d | w_d, \theta, \alpha}[n_{dkw} - \pi_{kw} n_{dk.}] \right) + (\theta_{kw})^{\frac{1}{2}} \zeta_{kw} \right| \tag{6}$$

, where $\zeta_{kw} \sim \mathcal{N}(0, \epsilon)$.

Note that since $\theta_{kw} > 0$, we use the mirroring trick in the update above. Also, note that this update requires posterior over z_d. We shall implement Gibbs sampling to infer the same.

$$p(z_{di} = k \mid w_d, \alpha_d, \theta) = \frac{\left(\alpha_{dk} + n^{\setminus i}_{dk.}\right)\theta_{kw_{di}}}{\sum_k \left(\alpha_{dk} + n^{\setminus i}_{dk.}\right)\theta_{kw_{di}}}$$

where $\setminus i$ represents a count excluding the topic assignment variable we are updating

Note that since λ is an unconstrained parameter, we would update it using standard SGLD.

$$\frac{\partial \mathcal{L}}{\partial \lambda_{kt}} = \sum_d x_{dt} \exp\left(x_d^T \lambda_k\right) \times$$
$$\Big(\Psi\big(\sum_k \exp\left(x_d^T \lambda_k\right)\big) - \Psi\big(\sum_k \exp\left(x_d^T \lambda_k\right) + n_d\big) +$$
$$\Psi\big(\exp\left(x_d^T \lambda_k\right) + n_{k|d}\big) - \Psi\big(\exp\left(x_d^T \lambda_k\right)\big)\Big) - \frac{\lambda_{kt}}{\sigma^2}$$
$$= -\frac{\lambda_{kt}}{\sigma^2} + \sum_d f(x_d)$$

where $\mathcal{L}$ is the complete log-likelihood

$$\Delta \lambda_{kt} = \frac{\epsilon_t}{2}\Big(-\frac{\lambda_{kt}}{\sigma^2} + \frac{|S_d|}{D} \sum_{d \in S_d} f(x_d)\Big) + \eta_t$$

where $\eta_t \sim \mathcal{N}(0, \epsilon_t)$

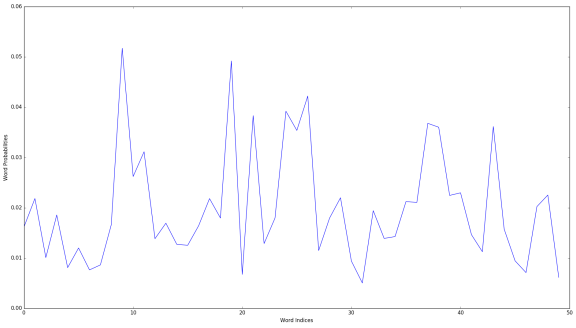
4 Experimental Results

We have implemented the DMR topic model using metadata. Mimno and McCallum[1] have implemented this model using stochastic EM. We have implemented this model using a combination of SGLD and SGRLD algorithms according to the update equations derived in the previous section. We have used a dummy dataset containing 100 documents, each document having 10 words. The dataset we used can be found at `https://github.com/mpkato/dmr`. Given the small size of the dataset, we have chosen number of topics to be 3. The implementation of our algorithm mainly relies on the collapsed Gibbs sampling of topic variable z_d's required for the parameter updates. This was the computationally challenging step, we shall discuss this in more detail by the end of this section.
Evaluation of the model is carried out using held-out validation. The train to test split proportions are set to $9 : 1$ as used by Patterson and Teh[3]. We use perplexity as the evaluation measure. The perplexity plot obtained for our model is given in figure 4.
Since we have implemented our model on a dummy dataset, we now demonstrate a few results that justify that model is giving desired results. We show evolution of word proportion vectors π_k for the first topic. Figure 3 shows the proportion π_0 at iterations $i = 0, 100, 200$. We can observe that initially no word has a large peek as all the π_k's are generated using $Dirichlet(\beta)$. However after a few iterations we observe that word proportions start concentrating on a few words for a given topic, as we would expect for a topic model. Moreover, Figure 5 shows the plot of all the π_k's and we observe that peaks are different for different topics, as we would expect.

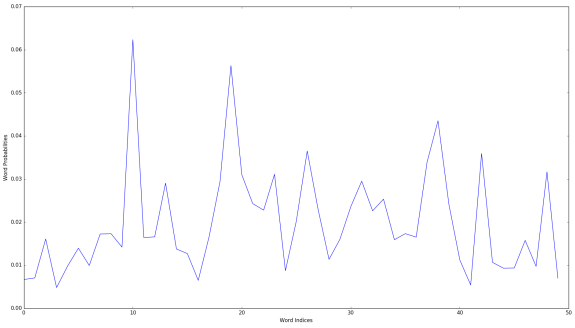
We have not implemented the model on a real dataset because of the scalability issue. Patterson and Teh[3] have stated in the section of results on Wikipedia corpus that they have not used Gibbs sampling since the dataset is very large. Once this issue has been addressed, only step remaining is to tune the hyper-parameters.

5 What we learned from the project

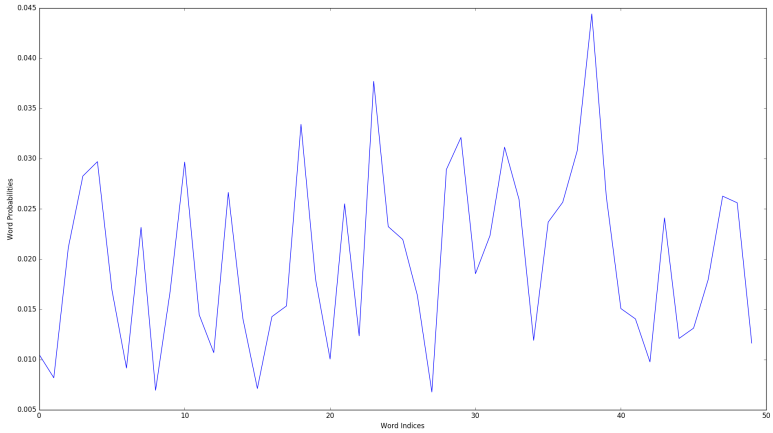
We learned a new approach to scale machine learning models which basically combines two fairly old ideas of stochastic optimization and Langevin dynamics. We also learned of new topic model called DMR which incorporates meta-information of each document. We modified the generative story of DMR and inferred the model parameters using SGRLD.



(a) Initial Word Proportion for Topic 0



(b) Word Proportion for Topic 0 after 100 iterations



(c) Word Proportion for Topic 0 after 200 iterations

Figure 3

6 Future Possibilities

We can train and test our model on a real dataset and compare it with the existing approaches. We expect the model to give better results compared to standard LDA model due to incorporation of meta data. Also, this model is an extension of DMR topic model because our model takes a fully Bayesian approach. The applications of DMR topic model in terms of discovering patters such as author profile and evolution of popular topics over time can similarly be explored for our model. SGRLD can also be applied to Poisson Matrix factorization. Infact, it can be implemented on various machine learning models that have constrained parameters and has scalability issues and one can compare its performance with the existing state-of-the-art methods.

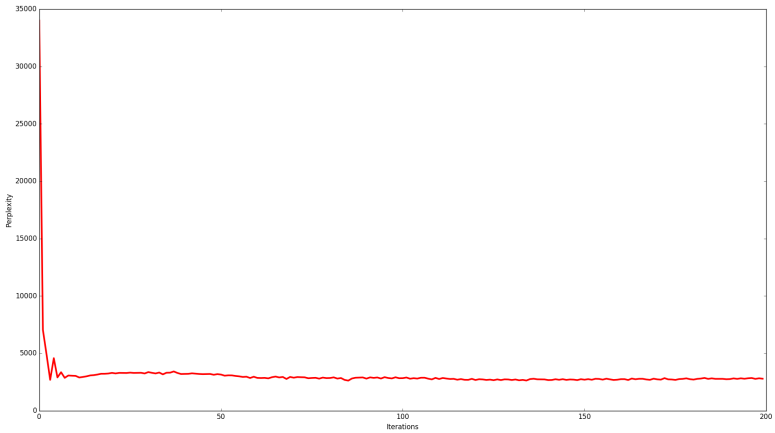


Figure 4: Test set perplexities

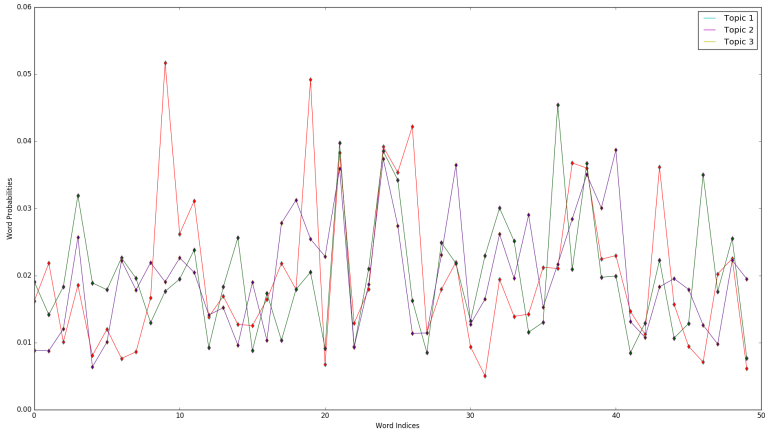


Figure 5: Evolution of π_k with iterations

Acknowledgments

We would like to thank Prof. Piyush Rai for suggesting an interesting topic and backing us up with constant help and support.

References

[1] David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

[2] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

[3] 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.

[4] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.