LATENT DIRICHLET REALLOCATION FOR TERM SWAPPING

Creighton Heaukulani[†⋆]*, Jonathan Le Roux*[†] *and John R. Hershey*[†]

[†]Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA, USA

[⋆]University of Cambridge
Department of Engineering
Cambridge, UK

ABSTRACT

This paper is an extended abstract of a work in progress, which proposes *latent Dirichlet reallocation* (LDR), a probabilistic model for text data from different dialects over a shared vocabulary. LDR first uses a topic model to allocate word probabilities to vocabulary terms; it then uses a "subtopic" model to allow for a possible "reallocation" of probability between a few potentially swappable terms between dialects. An MCMC inference procedure is derived, combining Gibbs sampling with Hamiltonian Monte-Carlo. Finally, we demonstate the ability of LDR to correctly switch the probabilities for swappable terms under the subtopics using a toy example.

***Index Terms*—** topic models, dialect, information retrieval, Bayesian methods, machine learning

1. INTRODUCTION

Latent Dirichlet allocation (LDA) [1] and its extensions are popular dimension reduction techniques, commonly applied to the modeling of large text corpora. We present an extension of LDA to handle the case of modeling documents from different dialects. In this work, we consider dialects to be separate corpora, composed from the same vocabulary. We call a vocabulary term "universal" if its usage is equivalent across dialects, or "swappable" if it is dialect-specific. In practice, a swappable term may have a different meaning in another dialect and/or different dialects could have their own unique, equivalent term.

Consider for example, guides for different programming languages, product user manuals for different brands, or course catalogues from different universities. In these examples, sections of text in the different dialects may refer to the same topics, however, they may use different key terms. The problem we wish to address is: given query terms in one dialect, relevant sections in different dialects should be reliably returned. While models such as LDA can attempt to cluster text based on the topical similarity of common words across dialects, they will be insufficient as dialects increase in variation and the need for a more robust method arises. Such an application will have practical value in information retrieval, where searching for useful information in an unfamiliar domain can be a difficult task with differing key terminology.

Toward this goal, we will develop an extension of LDA based on a technique we call *latent Dirichlet reallocation* (LDR), which uses a topic model, similarly to LDA, to allocate a distribution over words to each document. Then, using a Bayesian method, LDR is able to reallocate word probabilities between a few vocabulary terms which are potentially swappable between dialects.

1.1. Related works

Related works on word-sense disambiguation using topic models [2] attempt to learn a polysemantic word's hidden sense according to a predefined labeled hierarchy of words. Other models for multilingual corpora require aligned or syntactically similar documents [3]. Others such as [4] work on unaligned text, however, they model corresponding topics in different vocabularies. In comparison, our method is completely unsupervised and models dialects within a shared vocabulary.

Highly related to our work in these respects is the *dialectical topic model* (diaTM) [5], which associates different documents in a corpus with different draws from both a mixture of dialects and a mixture of topics. That is, each word in every document has its own topic assignment as well as dialect assignment. Motivated by the three problems described above, we make different modeling assumptions. Firstly, LDR assumes that a corpus is associated with just one dialect and just a universal set of topics are shared. For the applications we are interested in, this is a natural assumption. For instance, you would train each user manual or catalogue as a different corpus and the sections comprising it as different documents. This information is available at training time and does not need to be automatically inferred.

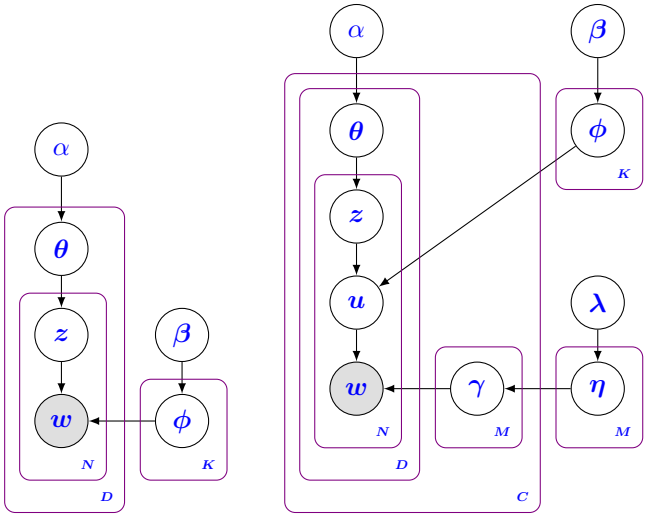
Further related works are the *topic-adapted latent Dirichlet allocation model* (τLDA) of [6], which models a technicality hierarchy in parallel with the topic hierarchy and the *hierarchical latent Dirichlet allocation* (hLDA) model of [7], which models a tree structured hierarchy for the learned topics using the nested Chinese restaurant process. These models are best suited to address documents of differing levels of specificity (called "technicality" in [6]), which is not the same as our dialect modeling problem. Again, in our applications, we do not assume different words in a document to be associated with different levels of technicality (or different dialects).

An important, distinguishing, and novel objective of our model is that LDR attempts to directly identify the subcluster of key terms which are swappable across dialects.

2. LATENT DIRICHLET REALLOCATION

We will first state the model for LDR, then, in section 3, we will explain the motivation and intuition behind the modeling.

LDR considers documents from one dialect to constitute a corpus $c = 1, \dots, C$. Here, a topic $z \in \{1, \dots, K\}$ is a distribution over "subtopics" $u \in \{1, \dots, M\}$, which are distributions over vocabulary terms indexed by $\{1, \dots, V\}$. We associate each document $d = 1, \dots, D_c$ with a distribution over topics $\boldsymbol{\theta}_{c,d}$, drawn from a symmetric Dirichlet Distribution shared across all corpora. For each word $n = 1, \dots, N_{c,d}$, a topic is drawn according to $\boldsymbol{\theta}_{c,d}$, then a subtopic is drawn from a topic dependent multinomial $\boldsymbol{\phi}_k$ (each an



(a) LDA (b) LDR

Fig. 1. Graphical model representation for (a) latent Dirichlet allocation and (b) latent Dirichlet reallocation

M-vector), depending on the topic. A vocabulary term is then selected from a multinomial $\boldsymbol{\gamma}_{c,m}$ which depends on both the corpus and subtopic assignment. We will also be placing symmetric Dirichlet priors over the $\boldsymbol{\phi}_k$. A key feature of our model will be subtopic-dependent Dirichlet priors $\boldsymbol{\eta}_m$ we place on the $\boldsymbol{\gamma}_{c,m}$, the motivation for which will be detailed in section 3.2.

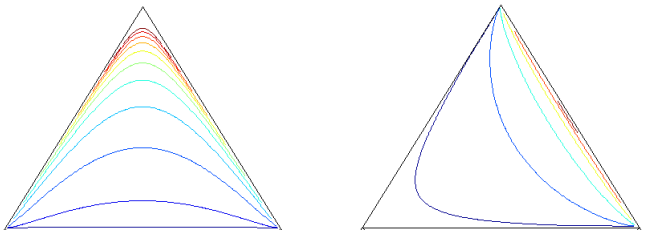
In detail, we specify a priori a number of topics K, subtopics M and the vocabulary size V, where we assume $K \ll M < V$. The model has two scalar parameters α and β for symmetric Dirichlet distributions, and a scalar λ parameterizing an exponential distribution. The generative model is

$$
\begin{aligned}
&1.\quad \boldsymbol{\eta}_m|\lambda \sim \exp(\lambda);\ m = 1,\ldots,M;\\
&2.\quad \boldsymbol{\phi}_k|\beta \sim \mathrm{Dir}(\beta);\ k = 1,\ldots,K;\\
&3.\quad \text{For } c = 1,\ldots,C:\\
&\quad a)\quad \boldsymbol{\gamma}_{c,m}|\boldsymbol{\eta}_m \sim \mathrm{Dir}(\boldsymbol{\eta}_m),\ m = 1,\ldots,M,\\
&\quad b)\quad \theta_{c,d}|\alpha \sim \mathrm{Dir}(\alpha),\ d = 1,\ldots,D_c;\\
&4.\quad \text{For } c = 1,\ldots,C,\ d = 1,\ldots,D_c,\ n = 1,\ldots,N_{c,d}:\\
&\quad a)\quad z_{c,d,n}|\boldsymbol{\theta}_{c,d} \sim \mathrm{Mult}(\boldsymbol{\theta}_{c,d}),\\
&\quad b)\quad u_{c,d,n}|z_{c,d,n};\boldsymbol{\phi}_{1:K} \sim \mathrm{Mult}(\boldsymbol{\phi}_{z_{c,d,n}}),\\
&\quad c)\quad w_{c,d,n}|u_{c,d,n};\boldsymbol{\gamma}_{c,1:M} \sim \mathrm{Mult}(\boldsymbol{\gamma}_{c,u_{c,d,n}}).
\end{aligned}
\tag{1}
$$

A graphical model for LDR can be seen in Fig. 1(b). For readers familiar with LDA, its graphical model is given in figure 1(a) for comparison (see [1] for details).

3. MODELING REALLOCATIONS BETWEEN TERMS

We now present a thorough motivation for LDR, focusing on intuition. The number of subtopics M characterizes LDR as a wide spectrum of models. Our motivation begins by considering large M, where we can interpret subtopics as "meanings" an observation should express; the exact vocabulary term used is corpus dependent. However, in practice, we scale M down, where we will adopt a topic modeling view.



(a) $(0.7, 0.7, 1.2)$ (b) $(0.7, 1.2, 1.2)$

Fig. 2. Examples of sparse Dirichlet distributions

3.1. Subtopics as meanings

The M subtopics can be viewed as intended "meanings" for a word w_i. It then makes sense to draw the subtopic u_i from a topic-dependent distribution, independent of the dialect, and for the word probabilities to depend on both the intended meaning and the dialect. Ideally, every universal term will correspond to its own subtopic. For the swappable terms, the model should group those which are equivalent in meaning and associate one subtopic with all terms in the group. Now consider selecting a subtopic associated with a universal term: in this case, the word has already been determined and an appropriately structured subtopic-dependent multinomial $\boldsymbol{\gamma}_{c_i,u_i}$ should not reallocate probability to another term. However, when a subtopic corresponding to a swappable group is selected, $\boldsymbol{\gamma}_{c_i,u_i}$ needs to correctly reallocate highest probability to the term appropriate for the dialect. The next section describes our technique for automatically learning these appropriately-structured multinomials.

3.2. Sparse Dirichlet priors for reallocation

We take a Bayesian approach to automatically learn the appropriate $\boldsymbol{\gamma}_{c,m}$ by giving it a subtopic-dependent, asymmetric Dirichlet prior parameterized by $\boldsymbol{\eta}_m$. This Dirichlet distribution is defined over the probability simplex in $V-1$ dimensions, which is a polytope with each vertex corresponding to a vocabulary term. Furthermore, we want it to be sparse, i.e., for it to place weight on either a $(p-1)$-face of the polytope (corresponding to p swappable terms under subtopic m) or on a single vertex (a universal term under m). An example of a sparse Dirichlet distribution favoring a vertex and another hugging a polytope face are shown in Fig. 2 (Note that the Dirichlet distribution itself is not sparse, but instead the draws from it should favor sparsity).

3.3. Relaxing combinatorial search to Bayesian reallocation

Consider, more fundamentally, the problem of learning word-equivalence across dialects. We need to form the multinomial probabilities over terms for each dialect (here $\boldsymbol{\gamma}_c$) to best explain the data. This entails finding an optimal sparse selection of terms to represent swappable terms in the dialect. This combinatorial search is a computationally intractable task. By taking a Bayesian approach and using a subtopic-dependent Dirichlet prior shared across dialects, we are in effect relaxing this combinatorial search to a continuous optimization which is automatically performed during inference.

3.4. From meanings to subtopics

As mentioned at the beginning of this section, we are motivated by a value of M close to V, with every vocabulary term (or group of

swapped terms) to have a corresponding word-multinomial per corpus, dictating exactly when the terms should be used. However, this not only entails fitting $CMV \approx CV^2$ word-multinomial parameters, which is a daunting computational task. We thus decrease the value of M, moving our viewpoint from one of selecting words to represent meanings, to one of topic modeling, performing two reductions onto lower dimensional subspaces.

4. INFERENCE BY MCMC

For observation w_i, we sample the topic z_i and subtopic u_i assignments using Gibbs sampling [8], [9], which samples from the conditional distribution given all other variables in the model. Unfortunately, the exponential and Dirichlet distributions are not conjugate, so implementing Variational Bayes [1] or Gibbs sampling for the subtopic priors $\boldsymbol{\eta}$ will not be straightforward. We use *Hamiltonian Monte-Carlo* (also known as *hybrid Monte-Carlo*) [10], which is based on Hamiltonian dynamics and avoids the random walk behavior of other Gibbs sampling methods.

4.1. Sampling the topic and subtopic assignments

We sample the topic and subtopic assignments as a block. For Gibbs sampling, we follow standard derivations, as in [8], to obtain the required conditional distribution for observation w_i

$$\begin{aligned} p(z_i = k,\ u_i = m | \boldsymbol{W}, \boldsymbol{Z}_{-i}, \boldsymbol{U}_{-i}, \boldsymbol{\eta}; \alpha, \beta, \lambda) \\ \propto \frac{n^{(c_i)}_{-i,m,w_i} + \eta_{m,w_i}}{\sum_j \left(n^{(c_i)}_{-i,m,j} + \eta_{m,j} \right)} \ \frac{n_{-i,k,m} + \beta}{\sum_m \left(n_{-i,k,m} \right) + M\beta} \\ \times \frac{n^{(c_i,d_i)}_{-i,k} + \alpha}{\sum_k \left(n^{(c_i,d_i)}_{-i,k} \right) + K\alpha}. \end{aligned} \tag{2}$$

where the subscript $-i$ indicates removal of the current observation in the following counts: $n^{(c_i)}_{-i,m,w_i}$, the number of times word w_i is assigned to subtopic m in corpus c_i; $n_{-i,k,m}$, the number of observations assigned to both subtopic m and topic k; and $n^{(c_i,d_i)}_{-i,k}$, the number of observations assigned to topic k in document d_i (in corpus c_i). Once each observation is sampled, we update the count matrices, which can be stored efficiently in sparse matrix format.

4.2. Sampling the sparse Dirichlet priors

Given an assignment for each observation of z and u, we sample the sparse Dirichlet priors $\boldsymbol{\eta}_m, m = 1, \ldots, M$ using Hamiltonian Monte-Carlo (HMC). Originating in the statistical physics literature, HMC is motivated by Hamiltonian dynamics in order to avoid the random walk nature of other MCMC procedures.

We do not provide the details of the HMC algorithm here, only the required expressions. However, one may refer to [10] and [11] for the details. Firstly, the conditional distribution for $\boldsymbol{\eta}_m$ gives the proportionality

$$\begin{aligned} p(\boldsymbol{\eta}_m | \boldsymbol{\eta}_{-m}, \boldsymbol{W}; \boldsymbol{U}, \lambda) \\ \propto \left[\frac{\Gamma\left(\sum_j \eta_{m,j} \right)}{\prod_j \Gamma(\eta_{m,j})} \right]^C \left[\prod_c \frac{\prod_j^V \Gamma(n^c_{jm} + \eta_{m,j})}{\Gamma\left(\sum_j n^c_{jm} + \eta_{m,j} \right)} \right] \\ \times \exp\left\{ -\lambda \sum_j \eta_{m,j} \right\}. \end{aligned} \tag{3}$$

The HMC algorithm requires the state variables to be unbounded, thus, we transform $\eta_{m,j} \in (0, \infty)$ to $x_j \in (-\infty, \infty)$ using $x_j = \log(\eta_{m,j})$. Using the change of variables theorem, the prior distribution for the sparse Dirichlet parameters $\boldsymbol{\eta}_m$ expressed in terms of $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x}|\lambda) \propto \exp\left\{ \sum_{j=1}^V x_j - \lambda \sum_{j=1}^V \exp(x_j) \right\} \tag{4}$$

and, thus, the negative of the j-th gradient component of the log probability $\mathcal{L}_{\boldsymbol{x}}$ is given by

$$\begin{aligned} -\frac{\partial \mathcal{L}_{\boldsymbol{x}}}{\partial x_j} = -1 + \exp(x_j) \Bigg\{ -\sum_c \Psi(n^c_{jm} + \exp(x_j)) \\ + \left[\sum_c \Psi\left(\sum_j n^c_{jm} + \exp(x_j) \right) \right] \\ - C\Psi\left(\sum_j \exp(x_j) \right) + C\Psi(\exp(x_j)) + \lambda \Bigg\}. \end{aligned} \tag{5}$$

The algorithm chooses to explore areas of the state space using the gradient of the current state, say $\boldsymbol{x}^\tau$, and a corresponding vector of momentum variables $\boldsymbol{p}^\tau$. The evolution of the Hamiltonian dynamics in the system, here given by τ and interpreted as "time", can be approximated using various methods. We use "leapfrog" steps [11], which are popular in the machine learning literature. In order to sample from the marginal distribution for $\boldsymbol{x}$, samples are drawn from the phase space $\{\boldsymbol{x}, \boldsymbol{p}\}$ and the samples $\boldsymbol{p}$ are simply discarded. Proofs of the validity of this procedure can be found in [11]. In summary, inference will proceed as follows:

1. Initialize the values of $\boldsymbol{U}, \boldsymbol{Z}, \boldsymbol{\eta}$.
2. For each observation w_i, sample z_i and u_i from (2) by Gibbs sampling and update the count matrices.
3. Given the assignments $\boldsymbol{Z}$ and $\boldsymbol{U}$, sample each $\boldsymbol{\eta}_m$, $m = 1, \ldots, M$ from (3) using HMC.
4. Iterate between steps 2 and 3 until convergence.

In practice, we run this procedure for a suitable "burn-in" period, to allow the Markov Chain to approach the target distribution, and discard the burn-in samples.

Given any single sample for $\{\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{\eta}\}$, following [8], we can use the counts to estimate the values for $\boldsymbol{\gamma}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ as follows: for any c, d, k, and/or m, the components of $\boldsymbol{\theta}_{c,d}$, $\boldsymbol{\phi}_k$, and $\boldsymbol{\gamma}_{c,m}$ can be estimated, respectively, by

$$\begin{aligned} \hat{\theta}_{c,d,k} = \frac{n^{(c,d)}_k + \alpha}{\sum_{k'} (n_{k'}) + K\alpha}, \quad \hat{\phi}_{k,m} = \frac{n_{k,m} + \beta}{\sum_{m'} (n_{k,m}) + M\beta}, \\ \text{and} \quad \hat{\gamma}_{c,m,j} = \frac{n^c_{m,j} + \eta_{m,j}}{\sum_{j'} \left(n^c_{m,j'} + \eta_{m,j'} \right)}, \end{aligned} \tag{6}$$

where the counts are analogous to those used in equations (2) and (3), however, including all observations. These values are the predictive distributions over the new words j, topics k and subtopics m, given the $\boldsymbol{U}, \boldsymbol{Z}, \boldsymbol{W}$ and $\boldsymbol{\eta}$.

5. EXPERIMENTS

We experiment with LDR on a toy data example as a sanity check to ensure the model learns swappable terms correctly. We compose

a corpus out of four short documents taken from Wikipedia articles on computer science. We make a second corpus out of three of these documents, randomly chosen. The documents are stemmed and stop words are removed. We choose several key terms from the first corpus and replace every instance of those terms in the second corpus with a different, new term. For example, in the second corpus, one replacement we make is "system" $\rightarrow$ "systemC2", so that the documents in each corpus are identical, with exception of these swapped terms.

We set $M = V = 544$ and $\alpha = 1/K$, $\beta = 1/M$ and $\lambda = 12$. We run LDR on these corpora for a 1000-iteration burn-in and use a sample to compute the estimates for $\boldsymbol{\gamma}_{c=1,1:M}$ and $\boldsymbol{\gamma}_{c=2,1:M}$ according to (6). For illustration, consider the term "system" which was replaced with "systemC2" in corpus two. We find that the term "system" has the highest probability (highest entry in $\boldsymbol{\gamma}_{1,m}$) in corpus one under subtopic $m = m^\star = 384$. Since this subtopic should represent the meaning "system", we expect the same subtopic in corpus two to switch its probability onto the term "systemC2". The comparison of $\boldsymbol{\gamma}_{1,m^\star}$ and $\boldsymbol{\gamma}_{2,m^\star}$ is shown in Fig. 3. For visualization, we have only show the terms immediately before and after "system" (fourth bar) and "systemC2" (eigth bar), and instead of the true probabilities, we have plotted the proportional probabilities for each corpus, with $\boldsymbol{\gamma}_{1,m^\star}$ on the bottom and $\boldsymbol{\gamma}_{2,m^\star}$ on top. As expected, the term "system", represented by the fourth bar, has very high probability in corpus one and low probability in corpus two. Conversely, "systemC2", represented by the eigth bar, has high probability in corpus two but low probability in corpus one. The figure does not show the true probability masses for each of the other terms; in reality, the mass differs from term to term, but they were all significantly lower than that for the key term "system" or "systemC2" (depending on the corpus). Importantly, we can see that LDR correctly learned that the probability for non-swappable terms remain relatively the same across the two different corpora. We saw similar results for the subtopic maximizing the probability of "system" in corpus two and for the other terms we manually swapped.

We note that when viewing the true word probabilities under the $\boldsymbol{\gamma}$, while the key term does have significantly higher probability than any other term under a subtopic, it does not hold the majority of the mass in the multinomial distribution over words. For example, under $\boldsymbol{\gamma}_{1,384}$, the key term "system" only has a mass of 0.18. In practice, we would like the key term to dominate the multinomial distribution. This is also seen in the sparse Dirichlet priors $\boldsymbol{\eta}$, where the mass in the high-probability term clusters do not contain the majority of the mass. This is likely due to the fact that the exponential prior on λ is too restrictive. To alleviate this, the authors are currently investigating the use of a Gamma prior on λ.

6. CONCLUSION AND FURTHER WORK

This extended abstract has described this work in progress and has demonstrated its ability to correctly model swapped terms on a toy example. Our next step is to conduct large experiments on real data to test LDR. These experiments should investigate two things: 1) attempt to verify whether LDR successfully clusters swappable terms under subtopics, and 2) compare its performance on modeling documents from different dialects to other models such as [1], [5], or [7]. In particular, experiments should be designed to evaluate the ability of different models to correctly associate sections of text from different dialects, given some query terms in one dialect. Such an experiment could be carried out with a labeled dataset, for example, and the authors are working on collecting a realistically sized dataset of one of the motivating examples given in the introduction.

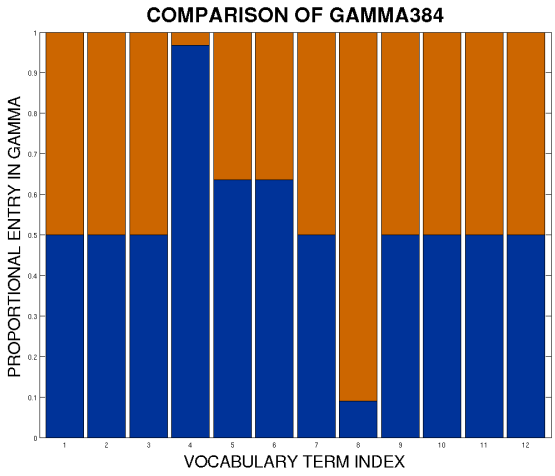


Fig. 3. A toy, two dialect example comparing the relative entries in $\boldsymbol{\gamma}_{1,384}$ (bottom) and $\boldsymbol{\gamma}_{2,384}$ (top). LDR correctly learns that terms four and eight are swappable across the two corpora under this subtopic. See section 5 for details.

7. REFERENCES

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[2] J. Boyd-Graber, D. Blei, and X. Zhu, "A topic model for word sense disambiguation," in *Proc. EMNLP*, 2007.

[3] D. Mimno, H.M. Wallach, J. Naradowsky, D.A. Smith, and A. McCallum, "Polylingual topic models," in *Proc. EMNLP*, Aug. 2009.

[4] J. Boyd-Graber and D. Blei, "Multilingual topic models for unaligned text," in *Proc. UAI*, June 2009.

[5] S.Crain, S. Yang, H. Zha, and Y. Jiao, "Dialect topic modeling for improved comsumer medical search," in *AMIA Annu. Symp. Proc.*, Nov. 2010, pp. 132–136.

[6] S.H. Yang, S.P. Crain, and H. Zha, "Bridging the language gap: Topic adaptation for documents with different technicality," in *Proc. AISTATS*, Apr. 2011.

[7] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, pp. 1–30, 2010.

[8] T.L. Griffiths and M. Steyvers, "Finding scientific topics," *PNAS*, , no. 1, pp. 5228–5235, Apr. 2004.

[9] W. Gilks, S. Richardson, and D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall, Suffolk, 1996.

[10] S. Duane, A.D. Kennedy, B.J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216 – 222, 1987.

[11] R.M. Neal, "Probabilistic inference using Markov Chain Monte Carlo methods," Tech. Rep. CRG-TR-93-1, Dept. of Computer Science, University of Toronto, Toronto, Canada, September 1993.