Provable Nonparametric Bayesian Inference

Bo Dai, Niao He, Hanjun Dai, Le Song
Georgia Institute of Technology
{bodai, nhe6, hdai8}@gatech.edu, lsong@cc.gatech.edu

Abstract

Bayesian methods are appealing in their flexibility in modeling complex data and ability in capturing uncertainty in parameters. However, when Bayes' rule does not result in tractable closed-form, most approximate inference algorithms lack either scalability or rigorous guarantees. To tackle this challenge, we propose a simple yet provable algorithm, *Particle Mirror Descent* (PMD), to iteratively approximate the posterior density. PMD is inspired by stochastic functional mirror descent where one descends in the density space using a small batch of data points at each iteration, and by particle filtering where one uses samples to approximate a function. We prove result of the first kind that, with m particles, PMD provides a posterior density estimator that converges in terms of KL-divergence to the true posterior in rate $O(1/\sqrt{m})$. We demonstrate competitive empirical performances of PMD compared to several approximate inference algorithms in various models.

1 Introduction

Bayesian methods are attractive because of their ability in modeling complex data and capturing uncertainty in parameters. The crux of Bayesian inference is to compute the posterior distribution $p(\theta|X) = \frac{p(\theta)\prod_{n=1}^{N} p(x_n|\theta)}{\int p(\theta)\prod_{n=1}^{N} p(x_n|\theta)d\theta}$. It can be challenging to compactly represent, tractably compute or efficiently sample from the solution when the prior is not conjugate to the likelihood. Besides the intractability, large-scale datasets also pose additional challenges for Bayesian inference. To tackle these challenges, the optimization perspective of Bayasian inference provides us a chance to leverage recent advances from convex optimization algorithms. Zellner [1] showed that Bayes' rule can be obtained by solving the optimization problem

$$\min_{q(\theta)\in\mathcal{P}} \quad L(q) := KL(q(\theta)\,||\,p(\theta)) - \sum_{n=1}^{N}\left[\int q(\theta)\log p(x_n|\theta)\,d\theta\right], \tag{1}$$

where $\mathcal{P}$ is the space of density. We present a simple, flexible and provable algorithm, *Particle Mirror Descent* (PMD), to iteratively approximate the posterior density by solving optimization (1). The algorithm connects stochastic optimization, functional analysis, kernel density estimation and Monte Carlo approximation to Bayesian inference.

2 Error Tolerant Stochastic Mirror Descent in Density Space

We will resort to stochastic optimization to avoid scanning through the entire data in each gradient evaluation. In particular, the stochastic mirror descent framework [4] expands the usual stochastic gradient descent scheme to problems with non-Euclidean geometries. We now introduce the stochastic mirror descent algorithm in the context of minimizing the objective $L(q)$ in density space.

At t-th iteration, given a data point x_t drawn randomly from the data set X, the stochastic functional gradient of $L(q)$ with respect to $q(\theta) \in L_2$ is $g_t(\theta) = \log(q(\theta)) - \log(p(\theta)) - N\log p(x_t|\theta)$. The stochastic mirror descent updates the density by the prox-mapping

$$q_{t+1} = \mathbf{P}_{q_t}(\gamma_t g_t) := \mathrm{argmin}_{\widehat{q}(\theta)\in\mathcal{P}}\left\{\langle\widehat{q}(\theta), \gamma_t g(\theta)\rangle_{L_2} + KL(\widehat{q}(\theta)\|q(\theta))\right\}$$

where $\gamma_t > 0$. Since the domain is density space, the KL-divergence is a natural choice for the prox-function. For any $q \in \mathcal{P}$ and $g \in L_2$, the prox-mapping therefore leads to the update

$$q_{t+1}(\theta) \;=\; q_t(\theta)\exp(-\gamma_t g_t(\theta))/Z \;=\; q_t(\theta)^{1-\gamma_t}p(\theta)^{\gamma_t}p(x_t|\theta)^{N\gamma_t}/Z, \tag{2}$$

where the normalization $Z := \int q_t(\theta) \exp(-\gamma_t g_t(\theta))\, d\theta$. This update resembles to the Bayes' rule. However, an important difference here is that the posterior is updated using the *fractional power* of the previous solution, the prior and the likelihood. The stochastic mirror descent allows us to go through the dataset several passes to refine the solution.

Still $q_{t+1}(\theta)$ may not be tractable due to the normalization Z. In fact, we can show that stochastic mirror descent can tolerate additional error during each prox-mapping step, which will give us rooms to design more flexible and provable approximation algorithms. Given $\epsilon \geqslant 0$ and $g \in L_2$, we define the ϵ-prox-mapping of q as the set

$$\mathbf{P}_q^{\epsilon}(g) := \{\widehat{q} \in \mathcal{P} \,:\, KL(\widehat{q}||q) + \langle g, \widehat{q}\rangle_{L_2} \leqslant \min_{\widehat{q}\in\mathcal{P}}\{KL(\widehat{q}||q) + \langle g, \widehat{q}\rangle_{L_2}\} + \epsilon\},$$

and consider the update $\tilde{q}_{t+1}(\theta) \in \mathbf{P}_{\tilde{q}_t}^{\epsilon_t}(\gamma_t g_t)$. When $\epsilon_t = 0, \forall t$, this reduces to the usual stochastic mirror descent algorithm. The classical results regarding the convergence rate can also be extended.

Essentially, this implies that we can *approximate* the intermediate density $q_t(\theta)$ by some *tractable representation*. As long as the approximation error is not too large, the algorithm will still converge; and if the approximation does not involve costly computation, the overall algorithm will be efficient.

3 Particle Mirror Descent Algorithm

We will introduce two efficient strategies to approximate the intermediate density, one based on weighted particles and the other based on weighted kernel density estimator. Interestingly, these two methods resemble particle reweighting and rejuvenation in sequential Monte Carlo yet with notable differences.

3.1 Posterior Approximation Using Weighted Particle

Assume the support of prior $p(\theta)$ is the same as the true posterior $q^*(\theta) = p(\theta|X)$, such that $\mathcal{F}_p = \{q(\theta) = \alpha(\theta)p(\theta), \int \alpha(\theta)p(\theta)d\theta = 1, 0 \leqslant \alpha(\theta) \leqslant C\}$ and $q^*(\theta) \in \mathcal{F}_p$. We will simply draw a set of samples (or particles) from $p(\theta)$, and approximate the intermediate posterior using these particles. More specifically, we sample a set of locations, $\{\theta_i\}_{i=1}^m$, *i.i.d.* from $p(\theta)$ and fix them across iterations. Then given $\alpha_t(\theta_i)$ from previous iteration, we will approximate $q_{t+1}(\theta)$ as a set of weighted particles

$$\tilde{q}_{t+1}(\theta) = \sum\nolimits_{i=1}^m \alpha_{t+1}(\theta_i)\, \delta(\theta_i), \text{ where } \alpha_{t+1}(\theta_i) := \frac{\alpha_t(\theta_i) \exp(-\gamma_t g_t(\theta_i))}{\sum_{i=1}^m \alpha_t(\theta_i) \exp(-\gamma_t g_t(\theta_i))}.$$

The update is derived from the closed-form solution to the *exact* prox-mapping step (2), *i.e.*, $\alpha_{t+1}(\theta) = \alpha_t(\theta) \exp(-\gamma_t g_t(\theta))/Z$. Since Z is constant common to all $\alpha_t(\theta_i)$, and α_i is a ratio, Z can be ignored. One can simply update an unnormalized version of $\alpha_t(\theta)$, and then use them to compute α_i. In summary, we can simply update the set of working variable α_i as

$$\alpha_i \leftarrow \alpha_i^{1-\gamma_t} p(x_t|\theta_i)^{N\gamma_t}, \forall i \text{ and then } \alpha_i \leftarrow \frac{\alpha_i}{\sum_{i=1}^m \alpha_i}. \tag{3}$$

We can show that the convergence rate of such approximation is *independent* of the dimension. [1]

Theorem 1 *Assume $p(\theta)$ has the same support as the true posterior $q^*(\theta)$,* i.e.*, $0 \leqslant q^*(\theta)/p(\theta) \leqslant C$. Assume further model $\|p(x|\theta)^N\|_\infty \leqslant \rho, \forall x$. Then $\forall f(\theta)$ bounded and integrable, the t-step PMD algorithm with stepsize $\gamma_t = \frac{\eta}{t}$ returns m weighted particles such that*

$$\mathbb{E}\left[|\langle \tilde{q}_t - q^*, f\rangle|\right] \leqslant \frac{2\sqrt{\max\{C, \rho e^M)\}}\|f\|_\infty}{\sqrt{m}} + \max\left\{\sqrt{KL(q^*||p)}, \frac{\eta M}{\sqrt{2\eta - 1}}\right\}\frac{\|f\|_\infty}{\sqrt{T}}$$

where $\langle \tilde{q}_t - q^*, f\rangle := \int (\tilde{q}_t(\theta) - q^*(\theta))f(\theta)d\theta$ and $M := \max_{t=1,\ldots,T} \|g_t\|_\infty$.

3.2 Posterior Approximation Using Weighted Kernel Density Estimator

In general, a $\pi(\theta)$ far way from the true posterior will leads to particle depletion and inaccurate estimation of the posterior. To deal with this issue, we developed an algorithm based on weighted kernel density estimator. The algorithm leverages the error tolerate stochastic mirror descent, and alternate between sampling from weighted density estimators, and updating the kernel density estimator.

More specifically, given $\tilde{q}_t(\theta)$ from previous iteration which is supposed to be easy to sample from, we will approximate $q_{t+1}(\theta)$ as a weighted kernel density estimator

$$\tilde{q}_{t+1}(\theta) = \sum\nolimits_{i=1}^m \alpha_i\, K_h(\theta - \theta_i), \text{ where } \alpha_i := \frac{\exp(-\gamma_t g_t(\theta_i))}{\sum_{i=1}^m \exp(-\gamma_t g_t(\theta_i))},\ \{\theta_i\}_{i=1}^m \overset{i.i.d.}{\sim} \tilde{q}_{t-1}(\theta), \tag{4}$$

[1] Due to space limitation, we ignore the proofs. For the details of the proof of theorems 1 and 2, please refer to our full version in http://arxiv.org/abs/1506.03101.

where $h > 0$ is the bandwidth parameter and $K_h(\theta) := \frac{1}{h^d} K(\theta/h)$ is a smoothing kernel. The update is again derived based on the closed-form solution to the *exact* prox-mapping step (2). However, the particle location in this case is sampled from the previous solution $\tilde{q}_t(\theta)$. The idea here is that $\tilde{q}_t^+(\theta) = \tilde{q}_t(\theta)\exp(-\gamma_t g_t(\theta))/Z$ can be viewed as an importance weighted version of $\tilde{q}_t(\theta)$ with weights equal to $\exp(-\gamma_t g_t(\theta))/Z$. If we want to approximate $\tilde{q}_t^+(\theta)$, we can sample m locations from $\tilde{q}_t(\theta)$ and associate each location the normalized weight α_i. To obtain a density for re-sampling in the next iteration, we place a kernel function $K_h(\theta)$ on each sampled location. Since α_i is a ratio, we can avoid evaluating the normalization factor Z when computing α_i.

In summary, we can simply update the set of working variable α_i as

$$\alpha_i \leftarrow \tilde{q}_{t-1}(\theta_i)^{-\gamma_t} p(\theta_i)^{\gamma_t} p(x_t|\theta_i)^{N\gamma_t}, \forall i,$$

with $\alpha_i \leftarrow \frac{\alpha_i}{\sum_{i=1}^m \alpha_i}$. This weighted kernel density estimation step is equivalent to solving the prox-mapping step $\mathbf{P}_{\tilde{q}_t}^{\epsilon_t}(\gamma_t g_t)$ as we discussed in Section 1. We can formally provide the rate of convergence of weighted KDE approximation in terms of KL-divergence, with assumptions

A. Kernel $K(\cdot)$ is a β-valid density kernel with a compact support and there exists $\mu, \nu, \delta > 0$ such that $\int K(z)^2\, dz \leqslant \mu^2$, $\int \|z\|^\beta |K(z)| dz \leqslant \nu$ and $K(z) \geqslant \delta^{-1}$ almost surely.

B. The logarithmic of the prior and likelihood belong to $(\beta; \mathcal{L})$-Hölder class.

Theorem 2 *Based on the above assumptions, when setting* $\gamma_t = \min\{\frac{2}{t+1}, \frac{\delta}{M m_t^{\beta/(d+2\beta)}}\}$,

$$\mathbb{E}[KL(q^*||\tilde{q}_T)] \leqslant \frac{2\max\left\{KL(q^*||\tilde{q}_1), M^2\right\}}{T} + \mathcal{C}_1 \frac{\sum_{t=1}^T t^2 m_t^{-2\beta/d+2\beta}}{T^2}$$

where $M := \max_{t=1,\ldots,T} \|g_t\|_\infty$, $\mathcal{C}_1 := O(1)(\mu + \nu\mathcal{L})^2\delta$ *with* $O(1)$ *being a constant.*

Algorithm 1 Particle Mirror Descent Algorithm

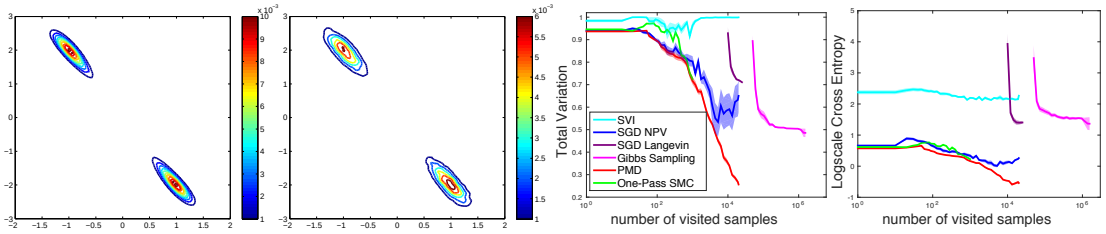
Input: Data set $X = \{x_n\}_{n=1}^N$, prior $p(\theta)$
Output: posterior density estimator $\tilde{q}_T(\theta)$
Initialize $\tilde{q}_1(\theta) = p(\theta)$
for $t = 1, 2, \ldots, T-1$ **do**
  $x_t \overset{unif.}{\sim} X$
  if Good $p(\theta)$ is provided **then**
    $\{\theta_i\}_{i=1}^{m_t} \overset{i.i.d.}{\sim} \pi(\theta)$ when $t = 1$
    $\alpha_i \leftarrow \alpha_i^{1-\gamma_t} p(x_t|\theta_i)^{N\gamma_t}, \forall i$
    $\alpha_i \leftarrow \frac{\alpha_i}{\sum_{i=1}^{m_t}\alpha_i}, \forall i$
    $\tilde{q}_{t+1}(\theta) = \sum_{i=1}^{m_t} \alpha_i\, \delta(\theta_i)$
  else
    $\{\theta_i\}_{i=1}^{m_t} \overset{i.i.d.}{\sim} \tilde{q}_t(\theta)$
    $\alpha_i \leftarrow \tilde{q}_t(\theta_i)^{-\gamma_t} p(\theta_i)^{\gamma_t} p(x_t|\theta_i)^{N\gamma_t}, \forall i$
    $\alpha_i \leftarrow \frac{\alpha_i}{\sum_{i=1}^{m_t}\alpha_i}, \forall i$
    $\tilde{q}_{t+1}(\theta) = \sum_{i=1}^{m_t} \alpha_i K_{h_t}(\theta - \theta_i)$
  end if
end for

3.3 Overall Algorithm

We present the overall algorithm, Particle Mirror Descent (PMD), in Algorithm 1, incorporating the two strategies from section 3.1 and 3.2. PMD takes as input N samples $X = \{x_n\}_{n=1}^N$, a prior $p(\theta)$ over the model parameter and the likelihood $p(x|\theta)$, and outputs the posterior density estimator $\tilde{q}_T(\theta)$ after T iterations. At each iteration, PMD will maintain an approximate $\tilde{q}_t(\theta)$ of the posterior $p(\theta|X)$. We note that our algorithm can also take a mini-batch of points in each iteration, and the guarantees still hold. The proposed two strategies have their own merits: the computation cost of the first strategy in each step is linear to the number of particles, m; while the cost is $O(m^2)$ in the second strategy because of the KDE. However, the KDE in second strategy could adapt the support and relax the requirement of $\pi(\theta)$ in first strategy. In practice, we could combine the proposed two strategies to balance cost and performance. In the beginning stage, we use the second strategy to locate the support. Since the number of samples is small, the computational cost is tolerable. After several iterations, we will have an estimate of the support of the posterior, and we could start the first strategy based on such estimator.

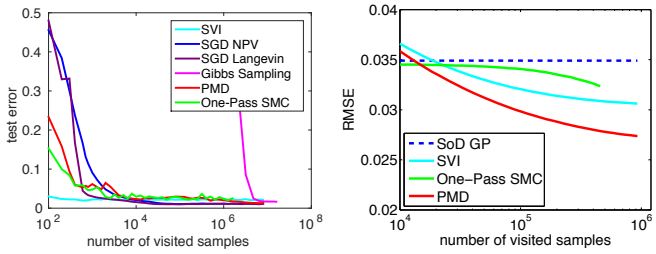
4 Experiments

We conduct experiments on mixture models, Bayesian logistic regression, and sparse Gaussian processes [5, 6] to demonstrate the advantages of the proposed algorithm in capturing multiple modes, dealing with non-conjugate models, and handling real-world applications, respectively. We combine two proposed strategies to balance the computational cost and performance. For the mixture model and logistic regression, we compare our algorithm three sampling algorithms, *i.e.*, one-pass sequential Monte Carlo (one-pass SMC) [7], stochastic gradient Langevin dynamics (SGD



(1) True Posterior (2) Estimated Posterior (3) Total Variation (4) Cross Entropy

Figure 1: Experimental results for mixture model on synthetic dataset.



(1) Logistic regression on MNIST (2) Sparse GP on music data

Figure 2: Experimental results on several different models for real-world datasets.

Langevin) [8] and Gibbs sampling, and two variational inference methods, *i.e.*, stochastic variational inference (SVI) [9] and stochastic variant of nonparametric variational inference (SGD NPV) [10]. For sparse GP model, we compared with several existing inference algorithms designed specifically for the model. For the derivation details of PMD for different models, details of experiments and additional experiments results, please refer to our full version in http://arxiv.org/abs/1506.03101.

Mixture Models. We conduct comparison on a simple yet interesting mixture model [8], the observations $x_i \sim p\mathcal{N}(\theta_1, \sigma_x^2) + (1-p)\mathcal{N}(\theta_1+\theta_2, \sigma_x^2)$ and $\theta_1 \sim \mathcal{N}(0, \sigma_1^2)$, $\theta_2 \sim \mathcal{N}(0, \sigma_2^2)$, where $(\sigma_1, \sigma_2) = (1, 1)$, $\sigma_x = 2.5$ and $p = 0.5$. We repeat the experiments 10 times and report the average results. The actual recovered posterior distribution of our method are illustrated in Figure 1 (1)(2) as a concrete example. PMD fits both modes well and recovers nicely the posterior while other algorithms either miss a mode or fail to fit the multimodal density. Figure 1 (3)(4) compares the algorithms in terms of total variation and cross entropy. PMD achieves the best performance with fewer observations demonstrating that our algorithm is able to take the advantages of nonparametric model to capture the modes and to adapt to the shape of posterior automatically.

Bayesian Logistic Regression. We test our algorithm on logistic regression with non-conjugate prior for handwritten digits classification on the MNIST8M 8 vs. 6 dataset. The dataset contains about 1.6M training samples and 1932 testing samples. We initialize all the inference algorithms with prior distribution and stop the stochastic algorithms after they pass through the whole dataset 5 times, while SMC only pass the dataset once. We repeat the experiments 10 times and the results are reported in Figure 2(1). Obviously, Gibbs sampling [11], which needs scan the whole set, is not suitable to large-scale problem. The SVI performs best at the beginning stage. This is expectable because searching in the Gaussian family is simpler comparing to nonparametric density family. However, it is should be noticed that our algorithm achieves comparable performance with the general nonparametric form when feeding enough data, 98.8%, to SVI which use carefully designed lower bound of the log-likelihood [12]. SGD NPV is flexible with mixture models family, however, its speed becomes the bottleneck. In SGD NPV, the gain from using stochastic gradient is dragged down by using L-BFGS to optimize the second-order approximation of the evidence lower bound.

Sparse Gaussian Processes. We conduct the comparison on sparse GPs for the task to predict the year of songs [13]. In this task, we compare with one-pass SMC and subset of data approximation (SoD) [5]. We also compared with the extended version of the SVI [6] to sparse GPs. The dataset contains about 0.5M songs, each of which represented by 90-dimension features. We stop the stochastic algorithms after they pass through the whole dataset 2 times, while SMC only pass the dataset once. We use 16 particles in both SMC and PMD. Gaussian RBF kernel is used in the model. The number of inducing inputs in sparse GP is set to be 2^{10}, and all the other hyperparameters of sparse GP are fixed for all the inference methods. We rerun experiments 10 times and the results are reported in Figure. 2(2). Our algorithm achieves the best RMSE 0.027, significant better than one-pass SMC, SVI and the baseline SoD.

References

[1] Arnold Zellner. Optimal Information Processing and Bayes's Theorem. *The American Statistician*, 42(4), November 1988.

[2] Tom Minka. Divergence measures and message passing. Report 173, Microsoft Research, 2005.

[3] T. Minka. *Expectation Propagation for approximative Bayesian inference*. PhD thesis, MIT Media Labs, Cambridge, USA, 2001.

[4] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. on Optimization*, 19(4):1574–1609, January 2009. ISSN 1052-6234.

[5] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[6] James Hensman, Nicoló Fusi, and Neil D. Lawrence. Gaussian processes for big data. *CoRR*, abs/1309.6835, 2013. URL http://arxiv.org/abs/1309.6835.

[7] Suhrid Balakrishnan and David Madigan. A one-pass sequential monte carlo method for bayesian analysis of massive datasets. *Bayesian Analysis*, 1(2):345–361, 06 2006.

[8] Max Welling and Yee-Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

[9] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

[10] Samuel Gershman, Matt Hoffman, and David M. Blei. Nonparametric variational inference. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 663–670, New York, NY, USA, 2012. ACM.

[11] C.C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1), 2006 145-168.

[12] T Jaakkola and Michael I Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

[13] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[14] Dan Crisan and Arnaud Doucet. A survey of convergence results on particle filtering methods for practitioners. *Signal Processing, IEEE Transactions on*, 50(3):736–746, 2002.

[15] Franois Le Gland and Nadia Oudjane. Stability and uniform approximation of nonlinear filters using the hilbert metric and application to particle filters. *The Annals of Applied Probability*, 14(1):pp. 144–187, 2004. ISSN 10505164. URL http://www.jstor.org/stable/4140493.

[16] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, London, 1995.

[17] Luc Devroye and Làszlò Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley and Sons, 1985.

[18] S. Patterson and Y. W. Teh. Stochastic gradient Riemannian Langevin dynamics on the probability simplex. In *Advances in Neural Information Processing Systems*, 2013.

[19] K. R. Canini, L. Shi, and T. L. Griff iths. Online inference of topics with latent dirichlet allocation. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

[20] A. Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan Narayanamurthy, and A. J. Smola. Scalable inference in latent variable models. In *Proceedings of The 5th ACM International Conference on Web Search and Data Mining (WSDM)*, 2012.

[21] David Mimno, Matt Hoffman, and David Blei. Sparse stochastic inference for latent dirichlet allocation. In *International Conference on Machine Learning*, 2012. URL http://arxiv.org/abs/1206.6425.