Provable Bayesian Inference via Particle Mirror Descent

Bo Dai
Georgia Tech
bodai@gatech.edu

Niao He
UIUC
niaohe@illinois.edu

Hanjun Dai
Georgia Tech
hanjundai@gatech.edu

Le Song
Georgia Tech
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 mixture models, logistic regression, sparse Gaussian processes and latent Dirichlet allocation on large scale datasets.

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, $p(\theta|X) \propto p(\theta) \prod_{n=1}^{N} p(x_n|\theta)$, of a parameter $\theta \in \mathbb{R}^d$ given a set of N data points $X = \{x_n\}_{n=1}^{N}$ from $\mathbb{R}^D$, with a prior distribution $p(\theta)$ and a model of data likelihood $p(x|\theta)$. For many nontrivial models from real-world applications, the prior might not be conjugate to the likelihood or might contain hierarchical structure. Therefore, computing the posterior often results in intractable integration and poses computational challenges. Typically, one resorts to approximate inference such as sampling, *e.g.*, MCMC (Neal, 1993) and SMC (Doucet et al., 2001), or variational inference (Jordan et al., 1998; Wainwright and Jordan, 2008).

Two longstanding challenges in approximate Bayesian inference are i) *provable convergence* and ii) *data-intensive computation* at each iteration. MCMC is a general algorithm known to generate samples from distribution that converges to the true posterior. However, in order to generate a single sample at every iteration, it requires a complete scan of the dataset and evaluation of the likelihood at each data point, which is computationally expensive. To address this issue, approximate sampling algorithms have been proposed which use only a small batch of data points at each iteration (*e.g.* Chopin, 2002; Balakrishnan and Madigan, 2006; Welling and Teh, 2011; Maclaurin and Adams, 2014). Chopin (2002); Balakrishnan and Madigan (2006) extend the sequential Monte Carlo (SMC) to Bayesian inference on static models. However, these algorithms rely on Gaussian distribution or kernel density estimator as transition kernel for efficiency, which breaks down the convergence guarantee of SMC. On the other hand, the stochastic Langevin dynamics algorithm (SGLD) (Welling and Teh, 2011) and its derivatives (Ahn et al., 2012; Chen et al., 2014; Ding et al., 2014) combine ideas from stochastic optimization and Hamiltonian Monte Carlo, and are proven to converge in terms of integral approximation, as recently shown in (Teh et al., 2014; Vollmer et al, 2015). Still, it is unclear whether the *dependent* samples generated reflects convergence to the true posterior. FireflyMC (Maclaurin and Adams, 2014), introduces auxiliary variables to switch on and off data points to save computation for likelihood evaluations, but this algorithm requires the knowledge of lower bounds of likelihood that is model-specific and may be hard to calculate.

In another line of research, the variational inference algorithms (Jordan et al., 1998; Wainwright and Jordan, 2008; Minka, 2001) attempt to approximate the entire posterior density by optimizing information divergence (Minka, 2005). The recent derivatives (Hoffman et al., 2013) avoid examination of all the data in each update. However, the

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51. Copyright 2016 by the authors.

major issue for these algorithms is the absence of theoretical guarantees. This is due largely to the fact that variational inference algorithms typically choose a parametric family to approximate the posterior density, which can be far from the true posterior, and require to solve a highly non-convex optimization problem. In most cases, these algorithms optimize over simple exponential family for tractability. More flexible variational families have been explored but largely restricted to mixture models (Jaakkola and Jordon, 1999; Gershman et al., 2012). In these cases, it is often difficult to quantify the approximation and optimization error at each iteration, and analyze how the error accumulates across the iterations. Therefore, a provably convergent variational inference algorithm is still needed.

In this paper, we present such a simple and provable nonparametric inference algorithm, *Particle Mirror Descent* (PMD), to iteratively approximate the posterior density. PMD relies on the connection that Bayes' rule can be expressed as the solution to a convex optimization problem over the density space (Williams, 1980; Zellner, 1988; Zhu et al., 2014). However, directly solving the optimization will lead to both computational and representational issues: one scan over the entire dataset at each iteration is needed, and the exact function update has no closed-form. To address these issues, we draw inspiration from two sources: (**i**) stochastic mirror descent, where one can instead descend in the density space using a small batch of data points at each iteration; and (**ii**) particle filtering and kernel density estimation, where one can maintain a tractable approximate representation of the density using samples. In summary, PMD possesses a number of desiderata:

Simplicity. PMD applies to many probabilistic models, even with *non-conjugate priors*. The algorithm is summarized in just a few lines of codes, and only requires the value of likelihood and prior, unlike other approximate inference techniques (Welling and Teh, 2011; Gershman et al., 2012; Paisley et al., 2012; Hoffman et al., 2013, *e.g.*), which typically require their first and/or second-order derivatives.

Flexibility. Different from other variational inference algorithms, which sacrifice the model flexibility for tractability, our method approximates the posterior by particles or kernel density estimator. The flexibility of nonparametric model enables PMD to capture multi-modal in posterior.

Stochasticity. At iteration t, PMD only visits a mini-batch of data to compute the stochastic functional gradient, and samples $O(t)$ points from the solution. Hence, it avoids scanning over the whole dataset in each update.

Theoretical guarantees. We show the density estimator provided by PMD converges in terms of both integral approximation and KL-divergence to the true posterior density in rate $O(1/\sqrt{m})$ with m particles. To our best knowledge, these results are the first of the kind in Bayesian inference for estimating posterior.

In the remainder, we will introduce the optimization view of Bayes' rule before presenting our algorithm, and then we provide both theoretical and empirical supports of PMD.

Throughout this paper, we denote KL as the Kullback-Leibler divergence, function $q(\theta)$ as q, a random sequence as $\theta_{[t]} := [\theta_1, \ldots, \theta_t]$, integral $f(\cdot)$ w.r.t. some measure $\mu(\theta)$ over support Ω as $\int f(\theta)\mu(d\theta)$, or $\int f(\theta)d\theta$ without ambiguity, $\langle \cdot, \cdot \rangle_{L_2}$ as the L_2 inner product, and $\|\cdot\|_p$ as the L_p norm for $1 \leqslant p \leqslant \infty$.

2 Optimization View of Bayesian Inference

Our algorithm stems from the connection between Bayes' rule and optimization. Williams (1980); Zellner (1988); Zhu et al. (2014) showed that Bayes' rule

$$p(\theta|X) = \frac{p(\theta)\prod_{n=1}^{N} p(x_n|\theta)}{p(X)}$$

where $p(X) = \int p(\theta)\prod_{n=1}^{N} p(x_n|\theta)d\theta$, can be obtained by solving the optimization problem

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

where $\mathcal{P}$ is the valid density space. The objective, $L(q)$, is continuously differentiable with respect to $q \in \mathcal{P}$ and one can further show that

Lemma 1 *Objective function $L(q)$ defined on $q(\theta) \in \mathcal{P}$ is 1-strongly convex w.r.t. KL-divergence.*

Despite of the closed-form representation of the optimal solution, it can be challenging to compactly represent, tractably compute, or efficiently sample from the solution. The normalization, $p(X) = \int p(\theta)\prod_{n=1}^{N} p(x_n|\theta)d\theta$, involves high dimensional integral and typically does not admit tractable closed-form computation. Meanwhile, the product in the numerator could be arbitrarily complicated, making it difficult to represent and sample from. However, this optimization perspective provides us a way to tackle these challenges by leveraging recent advances from optimization algorithms.

2.1 Stochastic Mirror Descent in Density Space

We will resort to stochastic optimization to avoid scanning the entire dataset for each gradient evaluation. The stochastic mirror descent (Nemirovski et al., 2009) expands the usual stochastic gradient descent scheme to problems with non-Euclidean geometries, by applying unbiased stochastic subgradients and Bregman distances as prox-map functions. We now explain in details, the stochastic mirror descent algorithm in the context of Bayesian inference.

At t-th iteration, given a data point x_t drawn randomly from the dataset, 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 iterates over the prox-mapping step $q_{t+1} = \mathbf{P}_{q_t}(\gamma_t g_t)$, where $\gamma_t > 0$ is the stepsize and

$$\mathbf{P}_q(g) := \mathrm{argmin}_{\widehat{q}(\theta)\in\mathcal{P}} \{\langle \widehat{q}, g\rangle_{L_2} + KL(\widehat{q}||q)\}.$$

Since the domain is density space, KL-divergence is a natural choice for the prox-function. The prox-mapping therefore admits the closed-form

$$\begin{aligned} q_{t+1}(\theta) &= q_t(\theta)\exp(-\gamma_t g_t(\theta))/Z \qquad (2)\\ &= q_t(\theta)^{1-\gamma_t} p(\theta)^{\gamma_t} p(x_t|\theta)^{N\gamma_t}/Z, \end{aligned}$$

where $Z := \int q_t(\theta)\exp(-\gamma_t g_t(\theta))\, d\theta$ is the normalization. This update is similar 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. Still computing $q_{t+1}(\theta)$ can be intractable due to the normalization Z.

2.2 Error Tolerant Stochastic Mirror Descent

To handle the intractable integral normalization at each prox-mapping step, we will consider a modified version of the stochastic mirror descent algorithm which can tolerate additional error in the prox-mapping step. Given $\epsilon \geqslant 0$ and $g \in L_2$, we define the ϵ-prox-mapping of q as the set

$$\begin{aligned} \mathbf{P}^{\epsilon}_q(g) := \{\widehat{q} \in \mathcal{P} \,:\, & KL(\widehat{q}||q) + \langle g, \widehat{q}\rangle_{L_2} \qquad (3)\\ & \leqslant \min_{\widehat{q}\in\mathcal{P}}\{KL(\widehat{q}||q) + \langle g, \widehat{q}\rangle_{L_2}\} + \epsilon\}, \end{aligned}$$

and consider the update $\tilde{q}_{t+1}(\theta) \in \mathbf{P}^{\epsilon_t}_{\tilde{q}_t}(\gamma_t g_t)$. 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 as below

Theorem 2 *Let $q^* = \mathrm{argmin}_{q\in\mathcal{P}} L(q)$, stochastic mirror descent with inexact prox-mapping after T steps gives the recurrence:*

$$\mathbb{E}[KL(q^*||\tilde{q}_{t+1})] \leqslant \epsilon_t + (1-\gamma_t)\mathbb{E}[KL(q^*||\tilde{q}_t)] + \frac{\gamma_t^2}{2}\mathbb{E}\|g_t\|^2_\infty$$

Remark 1: As shown in the classical analysis of stochastic mirror descent, we could also provide a non-asymptotic convergence results in terms of objective error at average solutions, *e.g.*, simple average $\bar{q}_T = \sum_{t=1}^T \gamma_t \tilde{q}_t / \sum_{t=1}^T \gamma_t$ in Appendix B.

Remark 2: For simplicity, we present the algorithm with stochastic gradient estimated by a single data point. The mini-batch trick is also applicable to reduce the variance of stochastic gradient, and convergence remains the same order but with an improved constant.

Allowing error in each step gives us room to design more flexible algorithms. Essentially, this implies that we can *approximate* the intermediate density 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 still be efficient.

3 Particle Mirror Descent Algorithm

We introduce two efficient strategies to approximate prox-mappings, one based on weighted particles and the other based on weighted kernel density estimator. The first strategy is designed for the situation when the prior is a "good" guess of the true posterior, while the second strategy works for general situations. Interestingly, these two methods resemble particle reweighting and rejuvenation respectively in sequential Monte Carlo yet with notable differences.

3.1 Posterior Approximation Using Weighted Particle

We first consider the situation when we are given a "good" prior, such that $p(\theta)$ has the same support as the true posterior $q^*(\theta)$, *i.e.*, $0 \leqslant q^*(\theta)/p(\theta) \leqslant C$. We will simply maintain a set of samples (or particles) from $p(\theta)$, and utlize them to estimate the intermediate prox-mappings. Let $\{\theta_i\}_{i=1}^m \sim p(\theta)$ be a set of fixed *i.i.d.* samples. We approximate $q_{t+1}(\theta)$ as a set of weighted particles

$$\begin{aligned} \tilde{q}_{t+1}(\theta) &= \textstyle\sum_{i=1}^m \alpha_i^{t+1}\,\delta(\theta_i), \qquad (4)\\ \alpha_i^{t+1} &:= \frac{\alpha_i^t \exp(-\gamma_t g_t(\theta_i))}{\sum_{i=1}^m \alpha_i^t \exp(-\gamma_t g_t(\theta_i))}, \forall t \geqslant 1. \end{aligned}$$

The update is derived from the closed-form solution to the *exact* prox-mapping step (2). Since the normalization is a constant common to all components, one can simply update the set of working variable α_i as

$$\begin{aligned} \alpha_i &\leftarrow \alpha_i^{1-\gamma_t} p(x_t|\theta_i)^{N\gamma_t}, \forall i \qquad (5)\\ \alpha_i &\leftarrow \frac{\alpha_i}{\sum_{i=1}^m \alpha_i}. \end{aligned}$$

We show that the one step approximation (4) incurs a dimension-*independent* error when estimating the integration of a function.

Theorem 3 *For any bounded and integrable function f,* $\mathbb{E}\left[\left|\int \tilde{q}_t(\theta)f(\theta)d\theta - \int q_t(\theta)f(\theta)d\theta\right|\right] \leqslant \frac{2C\|f\|_\infty}{\sqrt{m}}$.

Remark. Please refer to the Appendix C for details. When the model has several latent variables $\theta = (\xi, \zeta)$ and some parts of the variables have closed-form update in (2). *e.g.*, sparse GPs and LDA (refer to Appendix F), we could incorporate such structure information into algorithm by decomposing the posterior $q(\theta) = q(\xi)q(\zeta|\xi)$. When $p(\xi)$ satisfies the condition, we could sample $\{\xi_i\}_{i=1}^m \sim p(\xi)$ and approximate the posterior with summation of several functions, i.e., in the form of $q(\theta) \approx \sum \alpha_i q(\zeta|\xi_i)$.

3.2 Posterior Approximation Using Weighted Kernel Density Estimator

In general, sampling from prior $p(\theta)$ that are not so "good" will lead to particle depletion and inaccurate estimation of the posterior. To alleviate particle degeneracy, we propose to estimate the prox-mappings via weighted kernel density estimator (KDE). The weighted KDE prevents particles from dying out, in a similar fashion as kernel smoothing variant SMC (Doucet et al., 2001) and one-pass SMC (Balakrishnan and Madigan, 2006), but with guarantees.

More specifically, we approximate $q_{t+1}(\theta)$ via a weighted kernel density estimator

$$\tilde{q}_{t+1}(\theta) = \sum\nolimits_{i=1}^{m} \alpha_i \, K_h(\theta - \theta_i), \qquad (6)$$
$$\alpha_i := \frac{\exp(-\gamma_t g_t(\theta_i))}{\sum_{i=1}^{m} \exp(-\gamma_t g_t(\theta_i))}, \quad \{\theta_i\}_{i=1}^{m} \overset{i.i.d.}{\sim} \tilde{q}_t(\theta),$$

where $h > 0$ is the bandwidth parameter and $K_h(\theta) := \frac{1}{h^d} K(\theta/h)$ is a smoothing kernel. The update serves as an ϵ-prox-mapping (3) based on the closed-form solution to the *exact* prox-mapping step (2). Unlike the first strategy, 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

$$\alpha_i \leftarrow \tilde{q}_t(\theta_i)^{-\gamma_t} p(\theta_i)^{\gamma_t} p(x_t|\theta_i)^{N\gamma_t}, \forall i \qquad (7)$$
$$\alpha_i \leftarrow \frac{\alpha_i}{\sum_{i=1}^{m} \alpha_i}.$$

Intuitively, the sampling procedure gradually adjusts the support of the intermediate distribution towards that of the true posterior, which is similar to "rejuvenation" step. The reweighting procedure gradually adjusts the shape of the intermediate distribution on the support. Same as the mechanism in Doucet et al. (2001); Balakrishnan and Madigan (2006), the weighted KDE could avoid particle depletion.

We demonstrate that the estimator in (6) in one step possesses similar estimation properties as standard KDE for densities (for details, refer to the Appendix D).

Theorem 4 *Let q_t be a $(\beta;\mathcal{L})$-Hölder density function, and K be a β-valid density kernel, and the kernel bandwidth chosen as $h = O(m^{-\frac{1}{d+2\beta}})$. Then, under some mild conditions, $\mathbb{E}\left\|\tilde{q}_t(\theta) - q_t(\theta)\right\|_1 = O(m^{-\frac{\beta}{d+2\beta}})$.*

A kernel function $K(\cdot)$ is called β-valid, if $\int z^s K(z)dz = 0$ holds true for any $s = (s_1, \ldots, s_d) \in \mathbb{N}^d$ with $|s| \leqslant \lfloor \beta \rfloor$. Notice that all spherically symmetric and product kernels satisfy the condition. For instance, the Gaussian kernel $K(\theta) = (2\pi)^{-d/2}\exp(-\left\|\theta\right\|^2/2)$ satisfies the condition with $\beta = 1$, and it is used throughout our experiments. Theorem 4 implies that the weighted KDE achieves the *minmax* rate for density estimation in $(\beta;\mathcal{L})$-Hölder function class (Delyon and Juditsky, 1996), where β stands for the smoothness parameter and $\mathcal{L}$ is the corresponding Lipschitz constant. With further assumption on the smoothness of the density, the weighted KDE can achieve even better rate. For instance, if β scales linearly with dimension, the error of weighted KDE can achieve a rate independent of the dimension.

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
        $\{\theta_i\}_{i=1}^{m_t} \overset{i.i.d.}{\sim} p(\theta)$ 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
```

Essentially, the weighted KDE step provides an ϵ-prox-mapping $\mathbf{P}_{\tilde{q}_t}^{\epsilon_t}(\gamma_t g_t)$ (3) in density space as we discussed in Section 2. The inexactness is therefore determined by the number of samples m_t and kernel bandwidth h_t used in the weighted KDE.

3.3 Overall Algorithm

We present the overall algorithm, Particle Mirror Descent (PMD), in Algorithm 1. The algorithm is based on stochastic mirror descent incorporated with two strategies from section 3.1 and 3.2 to compute prox-mapping. 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 takes the stochastic functional gradient information and computes an inexact prox-mapping $\tilde{q}_t(\theta)$ through either weighted particles or weighted kernel density estimator. Note that as discussed in Section 2, we can also take a batch of points at each iteration to compute the stochastic gradient in order to reduce variance.

In Section 4, we will show that, with proper setting of step-size γ, Algorithm 1 converges in rate $O(1/\sqrt{m})$ using m particles, in terms of either integral approximation or KL-divergence, to the true posterior.

In practice, we could combine the proposed two algorithms to reduce the computation cost. In the beginning stage, we adopt the second strategy. The computation cost is affordable for small number of particles. After we achieve a reasonably good estimator of the posterior, we could switch to the first strategy using large size particles to get better rate.

4 Theoretical Guarantees

In this section, we show that PMD algorithm (**i**) given good prior $p(\theta)$, achieves a dimension independent, sublinear rate of convergence in terms of integral approximation; and (**ii**) in general cases, achieves a dimension dependent, sublinear rate of convergence in terms of KL-divergence with proper choices of stepsizes.

4.1 Weak Convergence of PMD

The weighted particles approximation, $\tilde{q}_t(\theta) = \sum_{i=1}^m \alpha_i \,\delta(\theta_i)$, returned by Algorithm 1 can be used directly for Bayesian inference. That is, given a function f, $\int q^*(\theta) f(\theta) d\theta$ can be approximated as $\sum_{i=1}^m \alpha_i f(\theta_i)$. We will analyze its ability in approximating integral, which is commonly used in sequential Monte Carlo for dynamic models (Crisan and Doucet, 2002) and stochastic Langevin dynamics (Vollmer et al, 2015). For simplicity, we may write $\sum_{i=1}^m \alpha_i f(\theta_i)$ as $\int \tilde{q}_t(\theta) f(\theta) d\theta$, despite of the fact that $\tilde{q}_t(\theta)$ is not exactly a density here. We show a sublinear rate of convergence *independent* of the dimension exists.

Theorem 5 (Integral approximation) *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*

$$\begin{aligned}&\mathbb{E}\left[\left|\int \tilde{q}_T(\theta) f(\theta) d\theta - \int q^*(\theta) f(\theta) d\theta\right|\right] \\ \leqslant\ & \frac{2\sqrt{\max\{C, \rho e^M\}}\|f\|_\infty}{\sqrt{m}} \\ +\ & \max\left\{\sqrt{KL(q^*||p)}, \frac{\eta M}{\sqrt{2\eta - 1}}\right\} \frac{\|f\|_\infty}{\sqrt{T}}\end{aligned}$$

where $M := \max_{t=1,\ldots,T} \|g_t\|_\infty$.

Remark. The condition for the models, $\|p(x|\theta)^N\|_\infty \leqslant \rho, \forall x$, is mild, and there are plenty of models satisfying such requirement. For examples, in binary/multi-class logistic regression, probit regression, as well as latent Dirichlet analysis, $\rho \leqslant 1$. Please refer to details in Appendix C. The proof combines the results of the weighted particles for integration, and convergence analysis of mirror descent. One can see that the error consists of two terms, one from integration approximation and the other from optimization error. To achieve the best rate of convergence, we need to balance the two terms. That is, when the number particles, m, scales linearly with the number of iterations, we obtain an overall convergence rate of $O(\frac{1}{\sqrt{T}})$. In other words, if the number of particles is fixed to m, we could achieve the convergence rate $O(\frac{1}{\sqrt{m}})$ with $T = O(m)$ iterations.

4.2 Strong Convergence of PMD

In general, when the weighted kernel density approximation scheme is used, we show that PMD enjoys a much stronger convergence, *i.e.*, the KL-divergence between the generated density and the true posterior converges sublinearly. Throughout this section, we merely assume that

- The prior and likelihood belong to $(\beta; \mathcal{L})$-Hölder class.
- Kernel $K(\cdot)$ is a β-valid density kernel with a compact support and there exists $\mu, \nu, \delta > 0$ such that $\int K(z)^2 \, dz \leqslant \mu^2$, $\int \|z\|^\beta |K(z)| dz \leqslant \nu$.
- There exists a bounded support Ω such that $\tilde{q}_t^+$ almost surely bounded away from $\Delta^{-1} > 0$.

Note that the above assumptions are more of a brief characteristics of the commonly used kernels and inferences problems in practice rather than an exception. The first condition clearly holds true when the logarithmic of the prior and likelihood belongs to C_∞ with bounded derivatives of all orders, as assumed in several literature (Teh et al., 2014; Vollmer et al, 2015). The second condition is for regularizing kernels for density estimation. The third condition is for characterizing the estimator over its support. These assumptions automatically validate all the conditions required to apply Theorem 4 and the corresponding high probability bounds (stated in Corollary 17 in appendix). Let the kernel bandwidth $h_t = m_t^{-1/(d+2\beta)}$, we immediately have that with high probability,

$$\|\tilde{q}_{t+1} - \mathbf{P}_{\tilde{q}_t}(\gamma_t g_t)\|_1 \leqslant O(m_t^{-\beta/(d+2\beta)}).$$

Directly applying Theorem 2, and solving the recursion following (Nemirovski et al., 2009), we establish the convergence results in terms of KL-divergence.

Theorem 6 (KL-divergence) *Based on the above assumptions, when setting $\gamma_t = \min\{\frac{2}{t+1}, \frac{\Delta}{M m_t^{\beta/(d+2\beta)}}\}$,*

$$\begin{aligned}\mathbb{E}[KL(q^*||\tilde{q}_T)] \quad \leqslant\ & \frac{2\max\{D_1, M^2\}}{T} \\ +\ & \mathcal{C}_1 \frac{\sum_{t=1}^T t^2 m_t^{-\frac{2\beta}{d+2\beta}}}{T^2} + \mathcal{C}_2 m_T^{-\frac{\beta}{d+2\beta}}\end{aligned}$$

where $M := \max_{t=1,\ldots,T} \|g_t\|_\infty$, $D_1 = KL(q^*||\tilde{q}_1)$, $\mathcal{C}_1 := O(1)(\mu + \nu\mathcal{L})^2 \mu^2 \Delta$, and $\mathcal{C}_2 := O(1) M (\mu + \nu\mathcal{L})$ with $O(1)$ being a constant.

Remark. Unlike Theorem 5, the convergence results are established in terms of the KL-divergence, which is a stronger criterion and can be used to derive the convergence under other divergences (Gibbs and Su, 2002). To our best knowledge, these results are the first of its kind for estimating posterior densities in literature. One can immediately see that the final accuracy is essentially determined by two sources of errors, one from noise in applying stochastic gradient, the other from applying weighted kernel density estimator. For the last iterate, an overall $O(\frac{1}{T})$ convergence rate can be achieved when $m_t = O(t^{2+d/\beta})$.

Table 1: Summary of the related inference methods

Methods	Provable	Convergence Criterion	Convergence Rate	Cost		Black Box
				Computation per Iteration	Memory	
SVI	No	–	–	$\Omega(d)$	$O(d)$	No
NPV	No	–	–	$\Omega(dm^2N + d^2N)$	$O(dm)$	No
Static SMC	No	–	–	$\Omega(dm)$	$O(dm)$	Yes
SGLD	Yes	$\lvert\langle q - q^*, f\rangle\rvert$	$O(m^{-\frac{1}{3}})$	$\Omega(d)$	$O(dm)$	Yes
PMD	Yes	$\lvert\langle q - q^*, f\rangle\rvert$	$O(m^{-\frac{1}{2}})$	$\Omega(dm)$	$O(dm)$	Yes
		$KL(q^*\|\|q)$	$O(m^{-\frac{1}{2}})$	$\Omega(dm^2)$	$O(dm)$	

There is an explicit trade-off between the overall rate and the total number of particles: the more particles we use at each iteration, the faster algorithm converges. One should also note that in our analysis, we explicitly characterize the effect of the smoothness of model controlled by β, which is assumed to be infinite in existing analysis of SGLD. When the smoothness parameter $\beta >> d$, the number of particles is no longer depend on the dimension. That means, with memory budget $O(dm)$, *i.e.*, the number of particles is set to be $O(m)$, we could achieve a $O(1/\sqrt{m})$ rate.

Open question. It is worth mentioning that in the above result, the $O(1/T)$ bound corresponding to the stochasticity is tight (see Nemirovski et al. (2009)), and the $O(m^{-\frac{\beta}{d+2\beta}})$ bound for KDE estimation is also tight by itself (see (Barron and Yang, 1995)). An interesting question here is whether the overall complexity provided here is indeed optimal? This is out of the scope of this paper, and we will leave it as an open question.

5 Related Work

PMD connects stochastic optimization, Monte Carlo approximation and functional analysis to Bayesian inference. Therefore, it is closely related to two different paradigms of inference algorithms derived based on either optimization or Monte Carlo approximation.

Relation to SVI. From the optimization point of view, the proposed algorithm shares some similarities to stochastic variational inference (SVI) (Hoffman et al., 2013)–both algorithms utilize stochastic gradients to update the solution. However, SVI optimizes a *surrogate of the objective*, the evidence lower bound (ELBO), with respect to a *restricted parametric* distribution[1]; while the PMD directly optimizes the objective over *all valid densities* in a nonparametric form. Our flexibility in density space eliminates the bias and leads to favorable convergence results.

Relation to SMC. From the sampling point of view, PMD and the particle filtering/sequential Monte Carlo (SMC) (Doucet et al., 2001) both rely on importance sampling. In the framework of SMC sampler (Del Moral et al., 2006), the static SMC variants proposed in (Chopin, 2002; Balakrishnan and Madigan, 2006) bares some resemblances to the proposed PMD. However, their updates come from completely different origins: the static SMC update is based on Monte Carlo approximation of Bayes' rule, while the PMD update based on inexact prox-mappings. On the algorithmic side, (**i**) the static SMC re-weights the particles with likelihood while the PMD re-weights based on functional gradient, which can be fractional power of the likelihood; and (**ii**) the static SMC only utilizes each datum once while the PMD allows multiple pass of the datasets. Most importantly, on the theoretical side, PMD is guaranteed with convergence in terms of both KL-divergence and integral approximation for *static model*, while SMC is only rigoriously justified for *dynamic models*. It is unclear whether the convergence still holds for these extensions.

Summary of the comparison. We summarize the comparison between PMD and static SMC, SGLD, SVI and NPV in Table 1. For the connections to other inference algorithms, including Annealed IS (Neal, 2001), general SMC sampler (Del Moral et al., 2006), stochastic gradient dynamics family, and nonparametric variational inference, please refer to Appendix G. Given dataset $\{x_i\}_{i=1}^N$, the model $p(x|\theta)$, $\theta \in \mathbb{R}^d$ and prior $p(\theta)$, whose value and gradient could be computed, we set PMD, static SMC, SGLD and NPV to keep m samples/components, so that they have the same memory cost and comparable convergence rate in terms of m. Therefore, SGLD runs $O(m)$ iterations. Meanwhile, by balancing the optimization error and approximation in PMD, we have PMD running $O(m)$ for integal approximation and $O(\sqrt{m})$ for KL-divergence. For static SMC, the number of iteration is $O(N)$. From Table 1, we can see that there exists a delicate trade-off between computation, memory cost and convergence rate for the approximate inference methods.

1. The static SMC uses simple normal distribution (Chopin, 2002) or kernel density estimation (Balakrishnan and Madigan, 2006) for rejuvenation. However, such moving kernel is purely heuristic and it is unclear whether the convergence rate of SMC for dynamic system (Crisan and Doucet, 2002; Gland and Oudjane, 2004) still holds for static models. To ensure the convergence of static SMC, MCMC is needed

[1] Even in (Gershman et al., 2012), "nonparametric variational inference" (NPV) uses the mixture of Gaussians as the variational family which is still parametric.

in the rejuvenation step. The MCMC step requires to browse all the previously visited data, leading to extra computation cost $\Omega(dmt)$ and memory cost $O(dt)$, and hence violating the memory budget requirement. We emphasize that even using MCMC in static SMC for rejuvenation, the conditions required for static SMC is more restricted. We discuss the conditions for convergence of SMC and PMD using particles approximation in Appendix C.

2. Comparing with SGLD, the cost of PMD at each iteration is higher. However, PMD converges in rate of $O(m^{-\frac{1}{2}})$, faster than SGLD, $O(m^{-\frac{1}{3}})$, in terms of integral approximation and KL-divergence which is more stringent if *all the orders* of derivatives of stochastic gradient is bounded. Moreover, even for the integral approximation, SGLD converges only when f having weak Taylor series expansion, while for PMD, f is only required to be bounded. The SGLD also requires the stochastic gradient satisfying several extra conditions to form a Lyapunov system, while such conditions are not needed in PMD.

6 Experiments

We conduct experiments on mixture models, logistic regression, sparse Gaussian processes and latent Dirichlet allocation to demonstrate the advantages of PMD in capturing multiple modes, dealing with non-conjugate models and incorporating special structures, respectively.

Competing algorithms. For the mixture model and logistic regression, we compare our algorithm with five general approximate Bayesian inference methods, including three sampling algorithms, *i.e.*, one-pass sequential Monte Carlo (one-pass SMC) (Balakrishnan and Madigan, 2006) which is an improved version of the SMC for Bayesian inference (Chopin, 2002), stochastic gradient Langevin dynamics (SGD Langevin) (Welling and Teh, 2011) and Gibbs sampling, and two variational inference methods, *i.e.*, stochastic variational inference (SVI) (Hoffman et al., 2013) and stochastic variant of nonparametric variational inference (SGD NPV) (Gershman et al., 2012). For sparse GP and LDA, we compare with the existing large-scale inference algorithms designed specifically for the models.

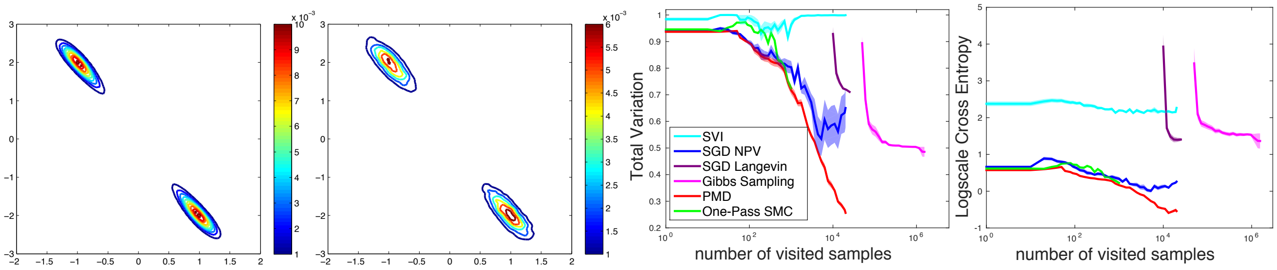
Evaluation criterion. For the synthetic data generated by mixture model, we could calculate the true posterior, Therefore, we evaluate the performance directly through total variation and KL-divergence (cross entropy). For the experiments on logistic regression, sparse GP and LDA on real-world datasets, we use indirect criteria which are widely used (Chen et al., 2014; Ding et al., 2014; Hensman et al., 2013; Patterson and Teh, 2013; Hoffman et al., 2013) because of the intractability of the posterior. We keep the same memory budget for Monte Carlo based algorithms if their computational cost is acceptable. To demonstrate the efficiency of each algorithm in utilizing data, we use the number of data visited cumulatively as x-axis.

For the details of the model specification, experimental setups, additional results and algorithm derivations for sparse GP and LDA, please refer to the Appendix H.

Mixture Models. We conduct comparison on a simple yet interesting mixture model (Welling and Teh, 2011), 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 $p = 0.5$. The means of two Gaussians are tied together making θ_1 and θ_2 correlated in the posterior. We generate 1000 data from the model with $(\theta_1, \theta_2) = (1, -2)$. This is one mode of the posterior, there is another equivalent mode at $(\theta_1, \theta_2) = (-1, 2)$. We initialize all algorithms with prior on (θ_1, θ_2). We repeat the experiments 10 times and report the average results. We keep the same memory for all except SVI. The true posterior and the one generated by our method is illustrated in Figure 1 (1)(2). PMD fits both modes well and recovers nicely the posterior while other algorithms either miss a mode or fail to fit the multimodal density. For the competitors' results, please refer to Appendix H. PMD achieves the best performance in terms of total variation and cross entropy as shown in Figure 1 (3)(4). This experiment clearly indicates our algorithm is able to take advantages of nonparametric model to capture multiple modes.

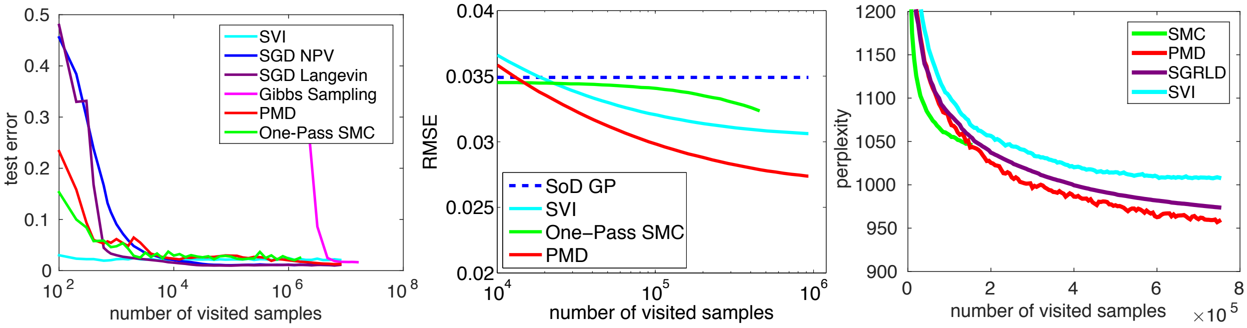
Bayesian Logistic Regression. We test our algorithm on logistic regression with non-conjugate prior for handwritten digits classification on the MNIST8M 8 vs. 6 dataset. The dataset contains about 1.6M training samples and 1932 testing samples. We initialize all algorithms with same prior and terminate the stochastic algorithms after 5 passes through the dataset. We keep 1000 samples for Monte Carlo based algorithms, except Gibbs sampling whose computation cost is unaffordable. We repeat the experiments 10 times and the results are reported in Figure 2(1). Obviously, Gibbs sampling (Holmes and Held, 145-168), which needs to scan the whole dataset, is not suitable for large-scale problem. In this experiment, SVI performs best at first, which is expectable because learning in the Gaussian family is simpler comparing to nonparametric density family. Our algorithm achieves comparable performance in nonparametric form after fed with enough data, 98.8%, to SVI which relies on carefully designed lower bound of the log-likelihood (Jaakkola and Jordan, 1997). SGD NPV is flexible with mixture models family, however, its speed becomes the bottleneck. For SGD NPV, the speed is dragged down for the use of L-BFGS to optimize the second-order approximation of ELBO.

Sparse Gaussian Processes. We use sparse GPs models to predict the year of songs (Bertin-Mahieux et al., 2011). In this task, we compare to the SVI for sparse GPs (Hoffman et al., 2013; Hensman et al., 2013) and one-pass SMC. We also included subset of data approxima-



(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 (3) LDA on wikipedia data

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

tion (SoD) (Quiñonero-Candela and Rasmussen, 2005) as baseline. The data contains about 0.5M songs, each represented by 90-dimension features. We terminate the stochastic algorithms after 2 passes of dataset. We use 16 particles in both SMC and PMD. The number of inducing inputs in sparse GP is set to be 2^{10}, and other hyperparameters of sparse GP are fixed for all methods. We run experiments 10 times and results are reported in Figure. 2(2). Our algorithm achieves the best RMSE 0.027, significantly better than one-pass SMC and SVI.

Latent Dirichlet Allocation. We compare to SVI (Hoffman et al., 2013), stochastic gradient Riemannian Langevin dynamic (SGRLD) (Patterson and Teh, 2013), and SMC specially designed for LDA (Canini et al., 2009) on Wikipedia dataset (Patterson and Teh, 2013). The dataset contains 0.15M documents, about 2M words and 8000 vocabulary. Since we evaluate their performances in terms of perplexity, which is integral over posterior, we do not need to recover the posterior, and therefore, we follow the same setting in (Ahmed et al., 2012; Mimno et al., 2012), where one particle is used in SMC and PMD to save the cost. We set topic number to 100 and fix other hyperparameters to be fair to all algorithms. We stop the stochastic algorithms after 5 passes of dataset. The results are reported in Figure 2(3). The top words from several topics found by our algorithm are illustrated in Appendix H. Our algorithm achieves the best perplexity, significantly better than SGRLD and SVI. In this experiment, SMC performs well at the beginning since it treats each documents equally and updates with full likelihood. However, SMC only uses each datum once, while the stochastic algorithms, *e.g.*, SGRLD, SVI and our PMD, could further refine the solution by running the dataset multiple times.

7 Conclusion

Our work contributes towards achieving better trade-off between *efficiency*, *flexibility* and *provability* in approximate Bayesian inference from *optimization perspective*. The proposed algorithm, *Particle Mirror Descent*, successfully combines stochastic mirror descent and nonparametric density approximation. Theoretically, the algorithm enjoys a rate $O(1/\sqrt{m})$ in terms of both integral approximation and KL-divergence, with $O(m)$ particles. Practically, the algorithm achieves competitive performance to existing state-of-the-art inference algorithms in mixture models, logistic regression, sparse Gaussian processes and latent Dirichlet analysis on several large-scale datasets.

Acknowledgements

We thank the anonymous referees for their valuable suggestions. This project was supported in part by NSF/NIH BIGDATA 1R01GM108341, ONR N00014-15-1-2340, NSF IIS-1218749, and NSF CAREER IIS-1350983.

References

A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. J. Smola. Scalable inference in latent variable models. In *The 5th ACM International Conference on Web Search and Data Mining*, 2012.

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *International Conference on Machine Learning*, 2012.

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

A. Barron and Y. Yang. Information theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27:1564–1599, 1995.

T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *International Conference on Music Information Retrieval*, 2011.

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

T. Chen, E. B. Fox, and C. Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, 2014.

N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–551, 2002.

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

P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

B. Delyon and A. Juditsky. On minimax wavelet estimators. *Applied and Computational Harmonic Analysis*, 3 (3):215 – 228, 1996.

L. Devroye and L. Györfi. *Nonparametric Density Estimation: The L_1 View*. John Wiley and Sons, 1985.

N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In *Advances in Neural Information Processing Systems 27*, 2014.

A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

S. Gershman, M. Hoffman, and D. M. Blei. Nonparametric variational inference. In *International Conference on Machine Learning*, 2012.

A. Gibbs and F. E.Su. On choosing and bounding probability metrics. *International statistical review*, 70: 419–435, 2002

F. L. Gland and N. 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.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

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

A. Ihler and D. McAllester. Particle belief propagation. In *the Twelfth International Conference on Artificial Intelligence and Statistics*, 2009.

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

T. S. Jaakkola and M. I. Jordon. Learning in graphical models. chapter Improving the Mean Field Approximation via the Use of Mixture Distributions, pages 163–173. MIT Press, Cambridge, MA, USA, 1999.

M. I. Jordan, Z. Gharamani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. Kluwer Academic, 1998.

T. Lienart, Y. W. Teh, and A. Doucet. Expectation particle belief propagation. In *Advances in Neural Information Processing Systems*, 2015.

D. Maclaurin and R. P. Adams. Firefly monte carlo: Exact MCMC with subsets of data. In *the Thirtieth Conference on Uncertainty in Artificial Intelligence*, 2014.

D. Mimno, M. Hoffman, and D. Blei. Sparse stochastic inference for latent dirichlet allocation. In *International Conference on Machine Learning*, 2012.

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

T. Minka. Divergence measures and message passing. Report 173, Microsoft Research, 2005.

R. M. Neal. Defining priors for distributions using dirichlet diffusion trees. Technical report, University of Toronto, 2001.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical report, University of Toronto, 1993.

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.

J. W. Paisley, D. M. Blei, and M. I. Jordan. Variational bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.

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

J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

S. J. Vollmer, K. C. Zygalakis and Y. W. Teh. (non-) asymptotic properties of stochastic gradient langevin dynamics. *submitted*, 2015.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *the Fourteenth Conference on Artificial Intelligence and Statistics*, 2011.

E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

Y. W. Teh, A. H. Thiéry, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *submitted*, 2014.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1 – 2):1–305, 2008.

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

M. Welling and Y.W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, 2011.

P. M. Williams. Bayesian conditionalisation and the principle of minimum information. *British Journal for the Philosophy of Science*, 31(2):131–144, 1980.

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

J. Zhu, N. Chen, and E. P. Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, 15(1): 1799–1847, January 2014.