Importance Weighted Consensus Monte Carlo for Distributed Bayesian Inference

Qiang Liu
Computer Science
Dartmouth College
qiang.liu@dartmouth.edu

Abstract

The recent explosion in big data has created a significant challenge for efficient and scalable Bayesian inference. In this paper, we consider a divide-and-conquer setting in which the data is partitioned into different subsets with communication constraints, and a proper combination strategy is used to aggregate the Monte Carlo samples drawn from the local posteriors based on the dataset subsets. We propose a new importance weighted consensus Monte Carlo method for efficient Bayesian inference in this setting. Our method outperforms the previous one-shot combination strategies in terms of accuracy, and is more computation- and communication-efficient than the previous iterative combination methods that require iterative re-sampling and communication steps. We provide two practical versions of our approach, and illustrate their properties both theoretically and empirically.

1 INTRODUCTION

Bayesian inference provides a powerful paradigm for reasoning with uncertain data by reducing inference problems into computational problems that can be routinely solved using efficient methods like Monte Carlo (MC) or Markov chain Monte Carlo (MCMC) methods. However, the recent explosion in big data has created a significant challenge for efficient and scalable Bayesian inference due to the difficulty for evaluating the likelihood across all the data points; traditional methods like Gibbs sampling and Metropolis-Hastings are extremely slow when the size of datasets is large.

We consider a divide-and-conquer approach for Bayesian computation under big data, in which case we partition the data into multiple subsets, and draw posterior samples on each subset separately, and then combine the results properly. Typical combination methods mostly rely on first approximating the subset posteriors using certain density estimation method and then combine the corresponding estimated densities. For example, the *consensus Monte Carlo* by Scott et al. (2013) fits the subset samples using normal distributions in which case the density combination reduces to a simple weighted linear averaging on the samples; Neiswanger et al. (2013) instead approximates each subset posterior using kernel density estimator (KDE), and uses another MCMC to draw sample from the product of KDEs. These methods are "one-shot" in that they do not require any further communication beyond passing the posterior samples; however, these methods critically rely on the qualities of the density estimators and often do not perform well in practice. Other *iterative* methods (e.g., Wang & Dunson, 2013; Xu et al., 2014) propose to iteratively resample from the local posteriors with adjusted local priors to enforce a consistency between the subset posteriors. Although being able to improve the performance iteratively, these methods require to re-draw the samples repeatedly, resulting higher computation and communication costs.

We propose a new importance weighted consensus Monte Carlo method for efficient distributed Bayesian inference. The key ingredient of our method is an importance weighted consensus strategy that efficiently combines the subset samples by leveraging their likelihood information. Our method performs significantly better than the pervious one-shot methods based on density estimations that solely rely on the subset samples and ignore the likelihood information. In addition, we show that our method can perform as efficient as the iterative combination methods, but with much less communication and computational costs.

Related Work The divided-and-conquer approach for scalable Bayesian computation has been studied in a series of recent works (Huang & Gelman, 2005; Scott et al., 2013; Neiswanger et al., 2013; Wang & Dunson, 2013; Xu et al., 2014; Minsker et al., 2014; Rabinovich et al., 2015). Another major approach for scalable Bayesian inference is based on efficient subsampling; see e.g., Korattikara et al. (2014); Welling & Teh (2011); Maclaurin & Adams (2014);

Bardenet et al. (2014). Despite being a relatively new topic, there are already a rich set of comprehensive reviews (Bardenet et al., 2015; Green et al., 2015; Zhu et al., 2014; Baker et al., 2015; Angelino et al., 2015).

Outline The rest of this paper is organized as follows. Section 2 introduces backgrounds and review the existing methods. Section 3 introduces our main method, where we propose two practical versions of our method and study their properties. Section 4 presents empirical results on both simulated and real-world datasets. The conclusion is made in Section 5.

2 BACKGROUND AND EXISTING METHODS

Consider a probabilistic model $p(D|x)$ where x is a random parameter with prior $p(x)$ and D is the observed data. Bayesian computation involves inferring the posterior distribution $f(x) \propto p(D|x)p(x)$, often in terms of calculating posterior moments of form $\mathbb{E}_f[h(x)] = \int h(x)f(x)dx$, where $h(x)$ is a test function, including the mean, variance or credible intervals. Typical Monte Carlo methods work by drawing samples from the posterior $\{x_i\}_{i=1}^n \sim f(x)$, and approximating the posterior moments by $\sum_{i=1}^n h(x_i)/n$; this gives a consistent estimator with mean squared error $\mathrm{var}_f[h(x)]/n$ with i.i.d. samples. Unfortunately, directly sampling from $p(D|x)$ requires to repeatedly evaluate the posterior probability and can be prohibitively slow when the number of data instances in D is very large.

We consider a divided-and-conquer approach in which case the data D is partitioned into m independent, nonoverlapping subsets $D^1, \ldots, D^m$, so that we have

$$f(x) \overset{def}{=} p(x|D) \propto p(x) \prod_{k=1}^{m} p(D^k|x).$$

This allows us to decompose the global posterior $f(x)$ into a product of "local posteriors" $f_k(x)$:

$$f(x) \propto \prod_k f_k(x), \qquad f_k(x) = p(D^k|x)p(x)^{1/m},$$

where each local posterior $f_k(x)$ receives $1/m$ of the original prior. Note that we do not assume D^m to have the same size nor follow a same probabilistic model.

Since each subset contains less data points, it is easier to sample from each of the local posteriors independently, which can be done in a parallel fashion. A critical problem, however, is how to inference about the global posterior $f(x)$, or estimate $\mathbb{E}_f[h(x)]$, using the samples from the local posteriors $\{x_i^k\}_{i=1}^n \sim f_k(x)$, $k = 1, \ldots, m$.

Existing Methods Useful perspectives can be obtained by considering the special case when $f_k(x)$ are assumed to be normal distributions, e.g., $f_k(x) = \mathcal{N}(x; \mu_k, \Sigma_k)$. This is justified by Bernstein-Von Mises Theorem, which says that the posterior $f_k(x)$ is close to normal when the number of data points in D^k is large. An important property of Gaussian distributions is that the product $f \propto \prod_k f_k$ of the densities f_k is equivalent to the density function of a weighted averaging $\bar{x} = \sum_k w_k x^k$, where $w_k = (\sum_k \Sigma_k^{-1})^{-1}\Sigma_k^{-1}$ and $x^k \sim f_k$. This motivates the consensus Monte Carlo (CMC) method (Scott et al., 2013) that combines the subset posterior samples by

$$\bar{x}_i = \sum_k \hat{w}_k x_i^k, \quad \hat{w}_k = (\sum_k \hat{\Sigma}_k^{-1})^{-1}\hat{\Sigma}_k^{-1},$$

where the exact covariance matrix Σ_k is replaced by the empirical covariance matrix $\hat{\Sigma}_k$ of $\{x_i^k\}_{i=1}^n$.

Unfortunately, CMC does not provide guarantees for non-Gaussian cases. Neiswanger et al. (2013) proposed a more general approach which approximates each subset posterior $f_k(x)$ with a Gaussian kernel density estimator (KDE), and then sample from the product of the KDEs using MCMC. This methods, however, does not scale well in high dimensions due to the use of non-parametric density estimation; in particular, the MSE of this method is $O(n^{-2/(2+d)})$, where d is the dimension of x; when $d > 2$, this is worse than the typical parametric rate $O(n^{-1/2})$ that we would get from the global posterior sampling.

In fact, we argue that inferring the global posterior $f \propto \prod_k f_k$ using only the subset samples $x_i^k \sim f_k(x)$ is fundamentally difficult, since it involves evaluating non-linear functionals of form $\int h(x) \prod_k f_k(x)dx$ for which certain non-parametric density estimates of f_k are unavoidable, and subjects to non-parametric minimax lower bounds that are generally worse than $O(n^{-1/2})$; see, for example, Birge & Massart (1995); Krishnamurthy et al. (2015, 2014) for discussions related to estimating the simpler form $\int f_1(x)f_2(x)dx$.

Therefore, acquiring further information is critical for improving the performance. Several authors (Wang & Dunson, 2013; Xu et al., 2014) have proposed to iteratively adjust and resample from the local posteriors to improve the results. Specifically, they set the local posteriors to be $f_k(x) \propto p(D^k|x)p_k(x)$, where $p_k(x)$ is a local "prior" that is adjusted iteratively. In particular, Xu et al. (2014) takes $p_k(x) = \prod_{k' \neq k} \hat{g}_{k'}(x)$ where $\hat{g}_k$ is an approximation of $p(D|x)/p(D^k|x)$ based on the current subset samples. In this way, we have $f_k \approx f$, and hence the subset samples can be treated as drawn from the global posterior f approximately.

In Wang & Dunson (2013), $p_k(x)$ are instead chosen to enforce the local samples $\{x_i^k\}_{i=1}^n$ to be consistent with each other; in particular, it is based on the observation that

$$\prod_k f_k(x) = \int \prod_k f_k(x^k) \exp\Big[-\frac{(x^k - x)^2}{2h^2}\Big] dx^k + O(h^2)$$

for small h, and evaluates the above integral using a Gibbs sampler that alternatively sample $\{x^k\}$ and x. This methods, however, critically depends on the value of h, since large h gives poor approximation (we need $h = O(n^{-1/4})$ to obtain an $O(n^{-1/2})$ approximation error), while small h makes Gibbs sampler difficult to converge. Wang & Dunson (2013) proposed to gradually decrease the value of h, making it essentially an annealed MCMC (Gibbs sampling) algorithm over the augmented distribution of $\{x^k\}$ and x; it is therefore difficult to formally guarantee the convergence of this algorithm. In addition, each iteration of the Gibbs sampling has a relatively expensive computation and communication cost in that it requires a fully convergent sampling from the local posteriors as well as communicating the subset posterior samples between the local subsets and the fusion center.

The above methods use only the information in the local posterior samples and do not make use of the values of their posterior probabilities which can carry important information. In this work, we propose a new combination method that avoids the density estimation using an important sampling strategy that assigns importance weights to the subset samples based on their likelihood values. We show that our method can significantly improve over the one-shot combination methods based on density estimations, while avoiding the expensive resampling steps in the iterative methods. We provide two versions of our method: our *Method I* is a valid importance sampling estimator and hence provides a consistent estimator with a typical parametric $O(n^{-1/2})$ estimator for generic non-Gaussian cases; our *Method II* provides a heuristic that works exceptionally well when f_k are nearly Gaussian, although without a formal consistency guarantee in generic cases.

3 IMPORTANCE WEIGHTED CONSENSUS MONTE CARLO

Assume we want to combine the local samples $\boldsymbol{x}_i = [x_i^1, \ldots, x_i^m]$ via a generic consensus function $\bar{x}_i = \phi(x_i^1, \ldots, x_i^m) = \phi(\boldsymbol{x})$; this includes, but does not limit to, the weighted averaging function $\bar{x} = \sum_{k=1}^m w_k x^k$. The key component of our approach is an auxiliary distribution over $[x^1, \ldots, x^m]$ under which the consensus $\bar{x} = \phi(\boldsymbol{x})$ is distributed according to the global posterior $f = \prod_k f_k$.

Proposition 3.1. *Let $g(x^1, \ldots, x^m)$ be an arbitrary density function, and $g(\bar{x})$ is the corresponding density function of $\bar{x} = \phi(\boldsymbol{x})$, that is,*

$$g(\bar{x}) = \int_{S_{\bar{x}}} g(x^1, \ldots, x^m) dS_{\bar{x}},$$

where the integral is over on the surface $S_{\bar{x}} = \{\boldsymbol{x} \colon \phi(\boldsymbol{x}) = \bar{x}\}$. We define an auxiliary distribution

$$\begin{aligned} p(x^1, \ldots, x^m) &= f(\bar{x}) g(x^1, \ldots, x^m \mid \bar{x}) \\ &= f(\bar{x}) g(x^1, \ldots, x^m) / g(\bar{x}), \end{aligned}$$

then the distribution $p(\bar{x})$ of $\bar{x}$ under $p(x^1, \ldots, x^m)$ equals $f(\cdot)$, that is, $p(\bar{x}) = f(\bar{x})$, and hence

$$\mathbb{E}_f[h(x)] = \mathbb{E}_p[h(\bar{x})].$$

for any function $h(x)$.

Proof. Simply note that

$$p(\bar{x}) = \int_{S_{\bar{x}}} f(\bar{x}) g(x^1, \ldots, x^m) / g(\bar{x}) dS_{\bar{x}} = f(\bar{x}). \quad \square$$

This result allows us to transform the estimation problem of $f(x)$ to that of a higher dimensional distribution $p(x^1, \ldots, x^m)$. Now given the local posterior samples $\{x_i^k\}_{i=1}^n \sim f_k$, we can treat $\boldsymbol{x}_i = [x_i^1, \ldots, x_i^m]$ as drawn from the product distribution $\prod_k f_k(x_k) \overset{def}{=} q(\boldsymbol{x})$. Using $q(\boldsymbol{x})$ as a proposal distribution allows us to construct a convenient importance sampling estimator:

$$\mathbb{E}_f[h(x)] = \mathbb{E}_p[h(\bar{x})] \approx \frac{\sum_{i=1}^n w(\boldsymbol{x}_i) h(\bar{x}_i)}{\sum_{i=1}^n w(\boldsymbol{x}_i)} \overset{def}{=} \hat{z}_h, \quad (1)$$

where $\bar{x}_i = \phi(\boldsymbol{x}_i)$ and the estimator $\hat{z}_h$ is a self normalized importance sampling estimator with importance weights

$$w(\boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} = \frac{g(x_i^1, \ldots, x_i^m) \prod_k f_k(\bar{x}_i)}{g(\bar{x}_i) \prod_k f_k(x_i^k)}. \quad (2)$$

Note that we do not need to know the normalization constants in $f_k(x)$ to calculate $\hat{z}_h$; calculating the normalization constants is often a critically difficult task.

Since $\hat{z}_h$ is a standard importance sampling estimator, it forms a consistent estimator for $z_h = \mathbb{E}_f[h(x)]$ in that $\Pr(\lim_{n\to\infty} \hat{z}_h = z_h) = 1$ if $q(\boldsymbol{x}) > 0$ whenever $p(\boldsymbol{x}) > 0$ (see e.g., Theorem 9.2 in Owen (2013)). In addition, the asymptotic MSE $\mathbb{E}[(\hat{z}_h - z_h)^2]$ can be calculated using the Delta method:

$$\mathbb{E}[(\hat{z}_h - z_h)^2] \asymp \frac{1}{n} \mathbb{E}_q[(h(\bar{x}) - z_h)^2 w(\boldsymbol{x})^2]. \quad (3)$$

Therefore, $\hat{z}_h$ approximates z_h with a typical parametric $O(n^{-1/2})$ error rate.

The MSE (3) depends on the test function $h(\cdot)$; a more generic measure of efficiency that is independent of $h(x)$ is the variance of the importance weights,

$$\mathrm{var}_q(w(\boldsymbol{x})) = \int q(\boldsymbol{x}) \left[\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} - 1\right]^2 d\boldsymbol{x}, \quad (4)$$

or equivalently, the effective sample size (ESS) $n/(\mathrm{var}_q(w(\boldsymbol{x})) + 1)$. We would like to have the importance weights $w(\boldsymbol{x})$ to be as uniform as possible, having a small variance or a large effective sample size [1] (ideally $w(\boldsymbol{x}) = 1,\ \forall \boldsymbol{x}$, in which case $q(\boldsymbol{x}) = p(\boldsymbol{x})$).

3.1 OPTIMAL CHOICE OF $\phi(\cdot)$ AND $g(\cdot)$

The estimator $\hat{z}_h$ depends on both the consensus function $\phi(\cdot)$ and the auxiliary distribution $g(\cdot)$. In this section, we discuss the optimal choice of $\phi(\cdot)$ and $g(\cdot)$ in terms of minimizing the variance $\mathrm{var}_q(w(\boldsymbol{x}))$ of the importance weights.

Proposition 3.2. *(i). The optimal $g(x^1, \ldots, x^m)$ that minimizes the variance $\mathrm{var}_q(w(\boldsymbol{x}))$ is*

$$g^*(x^1, \ldots, x^m) = \prod_{k=1}^{m} f_k(x^k),$$

in which case $g^*(\bar{x}) = \int_{S_{\bar{x}}} \prod_{k=1}^m f_k(x^k) dS_{\bar{x}}$ with $S_{\bar{x}} = \{\boldsymbol{x} \colon \bar{x} = \phi(\boldsymbol{x})\}$ and $w(\boldsymbol{x}) = \prod_{k=1}^m f_k(\bar{x})/g^*(\bar{x})$.

(ii). With $g = g^*$, the optimal consensus function $\bar{x} = \phi(\boldsymbol{x})$ should be chosen such that $g^*(\bar{x}) \propto \prod_k f_k(\bar{x})$. In the special case when $f_k(x)$ are Gaussian, that is, $f_k(x) = \mathcal{N}(x; \mu_k, \Sigma_k)$, the optimal $\phi(\cdot)$ is the weighted averaging $\phi(x) = \sum_{k=1}^m w_k x^k$ with $w_k = (\sum_k \Sigma_k^{-1})^{-1}\Sigma_k^{-1}$, and in this case, we have $w(\boldsymbol{x}) = 1,\ \forall \boldsymbol{x}$ and $\mathrm{var}_q(w(x)) = 0$. But there is no closed form for such an optimal $\phi(\cdot)$ in general cases.

Proof. (i). Since $\mathbb{E}_q[w(\boldsymbol{x})] = 1$, minimizing the variance $\mathrm{var}_q(w(\boldsymbol{x}))$ is equivalent to minimizing $\mathbb{E}_q[w(\boldsymbol{x})^2]$,

$$\mathbb{E}_q[w(\boldsymbol{x})^2] = \int \frac{p(\boldsymbol{x})^2}{q(\boldsymbol{x})} d\boldsymbol{x} = \int \frac{\prod_k f_k(\bar{x})^2}{g(\bar{x})^2} \Phi_g(\bar{x}) d\bar{x},$$

where $\Phi_g(\bar{x}) = \int_{S_{\bar{x}}} \frac{g(\boldsymbol{x})^2}{\prod_k f_k(x^k)} dS_{\bar{x}}$; one can show that g^* minimizes $\Phi_g(\bar{x})$ for any fixed $\bar{x}$, and hence minimizes $\mathbb{E}_q[w(\boldsymbol{x})^2]$.

(ii). With $g = g^*$, we have $\Phi_{g^*}(\bar{x}) = g^*(\bar{x})$ and hence $\mathbb{E}_q[w(\boldsymbol{x})^2] = \int \frac{\prod_k f_k(\bar{x})^2}{g^*(\bar{x})} d\bar{x}$, which is minimized when $g^*(\bar{x}) \propto \prod_k f_k(\bar{x})$. □

Remark With $g^*(\boldsymbol{x}) = \prod_k f_k(x^k)$, our estimator $\hat{z}_h$ can be treated as simply an importance sampler on $f(\bar{x})$ with proposal $g^*(\bar{x})$. The difficulty, however, is that $g^*(\bar{x})$ is usually intractable to calculate, making it essential to find suboptimal $g(\boldsymbol{x})$ that is more computationally tractable.

Algorithm 1 Importance Weighted Consensus Monte Carlo

Input: Samples from the local posteriors $\{x_i^k\}_{i=1}^n \sim f_k, \forall k = 1, \ldots, m$. Test function $h(x)$.

Output: Estimate $\mathbb{E}(h(x))$ under the global posterior $f(x) \propto \prod_k f_k(x)$.

Consensus: Let $\hat{\Sigma}_k$ be the empirical covariance matrix of subsample $\{x_i^k\}_{i=1}^n$. Calculate

$$\bar{x}_i = \sum_i w_k x_i^k, \text{ where } w_k = (\sum_k \hat{\Sigma}_k^{-1})^{-1}\hat{\Sigma}_k^{-1}.$$

Reweighting: Calculate the importance weights $w(\boldsymbol{x}_i)$ by *Method I* as defined in (5) or *Method II* in (6).

Estimating: $\mathbb{E}(h) \approx \sum_i w_i h(\bar{x}_i)/\sum_i w_i$

3.2 PRACTICAL IMPLEMENTATION

Although the optimal choices in Proposition 3.2 are intractable in general cases, we can leverage Bernstein-von Mises theorem to obtain near optimal choices. In particular, Proposition 3.2 justified the use of the weighted averaging $\phi(\boldsymbol{x}) = \sum_{k=1}^m w_k x^k$, with weights decided by the empirical variance $w_k = (\sum_k \hat{\Sigma}_k^{-1})^{-1}\hat{\Sigma}_k^{-1}$, which is the same as the consensus MC in Scott et al. (2013). Our method also requires to set a good auxiliary distribution $g(\boldsymbol{x})$; we explore two simple choices in this paper:

1. *Method I.* Motivated by Bernstein-von Mises theorem, we approximate each $f_k(x)$ by a Gaussian $\hat{f}_k(x) = \mathcal{N}(x;\ \hat{\mu}_k, \hat{\Sigma}_k)$, where $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are the empirical mean and covariance matrices of the k-th local sample $\{x_i^k \colon i \in [n]\}$, respectively. We then construct the auxiliary distribution $g(\boldsymbol{x})$ to be $g(\boldsymbol{x}) = \prod_k \hat{f}_k(x_k)$, under which we have $g(\bar{x}) = \mathcal{N}(\bar{x};\ \bar{\mu}, \bar{\Sigma})$, with $\bar{\mu} = \sum_k w_k \hat{\mu}_k$, and $\bar{\Sigma}^{-1} = (\sum_k w_k \hat{\Sigma}_k^{-1})$. In this case, the importance weight in (2) reduces to

$$w(\boldsymbol{x}_i) = \frac{\prod_k f_k(\bar{x}_i)}{\mathcal{N}(\bar{x}_i; \bar{\mu}, \bar{\Sigma})} \cdot \frac{\prod_k \mathcal{N}(x_i^k; \hat{\mu}_k, \hat{\Sigma}_k)}{\prod_k f_k(x_i^k)}. \quad (5)$$

2. *Method II.* Instead of approximating each f_k, we explicitly set the auxiliary distribution $g(\boldsymbol{x})$ to be the optimal choice suggested in Proposition 3.2, that is, $g(\boldsymbol{x}) = \prod_k f_k(x^k)$, and then approximate the corresponding $g(\bar{x})$ with a Gaussian distribution, that is, we approximate $g(\bar{x})$ by $\hat{g}(\bar{x}) = \mathcal{N}(\bar{x}, \bar{\mu}, \bar{\Sigma})$, which gives an importance weight of form

$$w(\boldsymbol{x}_i) = \frac{\prod_k f_k(\bar{x}_i)}{\mathcal{N}(\bar{x}_i; \bar{\mu}, \bar{\Sigma})}. \quad (6)$$

This can be justified in two possible scenarios: a) when each f_k is close to Gaussian by Bernstein-von Mises theorem, the distribution $g(\bar{x})$ of their averaging $\bar{x} =$

[1] A simple connection between (4) and (3) is that $(\mathrm{var}_q(w(\boldsymbol{x})) + 1)/n$ can be treated as the expectation of the MSE $\mathbb{E}[(\hat{z}_h - z_h)^2]$ when the value of $h(x)$, $\forall x$ is drawn from standard normal distribution.

$\sum_{k=1}^{m} w_k x_i^k$ should also be close to Gaussian; b) when the number m of subsets is large, the averaging $\bar{x} = \sum_{k=1}^{m} w_k x_i^k$ is approximately Gaussian by the central limit theorem.

Comparing with (5), the importance weight in (6) simply drops the terms that involve x_i^k, and hence should have smaller variance, but with the risk of introducing additional biases. We note that although *Method I* is a valid importance sampling (IS) estimator, and gives consistent estimates in general cases, *Method II* is no longer a valid IS estimator, and hence is not consistent for general non-Gaussian distributions. Nevertheless, we find that *Method II* often performs surprisingly well in practice, and has attractive theoretical properties when f_k are indeed Gaussian.

3.3 GAUSSIAN CASES

It is illustrative to study the properties of *Method I* and *Method II* under the simple case when f_k are Gaussian. In particular, we show that, despite being inconsistent for non-Gaussian cases, *Method II* is guaranteed to outperform *Method I* in Gaussian cases, that is, it exploits the Gaussianity more aggressively.

Assume $f_k(x) = \mathcal{N}(x; \mu_k, \Sigma_k)$ and denote by $\hat{f}_k(x) = \mathcal{N}(x; \hat{\mu}_k, \hat{\Sigma}_k)$, where $\hat{\mu}_k, \hat{\Sigma}_k$ are the empirical mean and covariance of the local sample $\{x_i^k\}_{i=1}^n$ on the k-th subset. Let $\theta = \{\mu_k, \Sigma_k^{-1} \colon \forall k\}$ be the set of true parameters and $\hat{\theta} = \{\hat{\mu}_k, \hat{\Sigma}_k^{-1} \colon \forall k\}$ the empirical estimates; correspondingly we denote $q(\boldsymbol{x}|\theta) = \prod_k f_k(x^k)$ and $q(\boldsymbol{x}|\hat{\theta}) = \prod_k \hat{f}_k(x^k)$. Then $\hat{\theta}$ is the maximum likelihood estimator of θ based on data $\{\boldsymbol{x}_i\}_{i=1}^n$.

Denote by $\hat{z}_h^{I}$ and $\hat{z}_h^{II}$ the estimates given by *Method I* and *Method II*, respectively. Let $t(\boldsymbol{x}) = h(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$, then

$$\hat{z}_h^{I} = \frac{1}{n} \sum_{i=1}^{n} t(\boldsymbol{x}_i), \qquad \hat{z}_h^{II} = \frac{\sum_{i=1}^{n} \omega(\boldsymbol{x}_i) t(\boldsymbol{x}_i)}{\sum_{i=1}^{n} \omega(\boldsymbol{x}_i)}.$$

where $\omega(\boldsymbol{x}_i) = \frac{q(\boldsymbol{x}_i|\theta)}{q(\boldsymbol{x}_i|\hat{\theta})}$. Note that here the weights $\omega(\boldsymbol{x}_i)$ should be very close to one when the sample size n is large since $\hat{\theta}$ is a maximum likelihood estimator of θ (assuming all f_k are Gaussian). However, as observed in Henmi et al. (2007); Henmi & Eguchi (2004), the $\omega(\boldsymbol{x}_i)$ in fact can act as a control variate to cancel part of the variance in $t(\boldsymbol{x}_i)$. As a result, $\hat{z}_h^{II}$ is guaranteed to have lower variance than $\hat{z}_h^{I}$ when all f_k are Gaussian.

Lemma 3.3 (Henmi et al. (2007))**.** *Assume each f_k is Gaussian, e.g., $f_k(x) = \mathcal{N}(x; \mu_k, \Sigma_k)$, $\forall k \in [m]$. Denote by $\hat{s} = \sum_{i=1}^{n} \nabla_\theta \log q(\boldsymbol{x}_i|\theta)/n$, then $\mathbb{E}\hat{s} = 0$ and*

$$\hat{z}_h^{II} = \hat{z}_h^{I} - \mathbb{E}[\hat{z}_h^{I} \hat{s}^\top][\mathrm{var}(\hat{s})]^{-1} \hat{s} + O_p(1/n).$$

Proof. See the proof of Theorem 1 and Equation (10) in Henmi et al. (2007). □

Therefore, $\hat{z}_h^{II}$ is asymptotically equivalent to a variance reduced version of $\hat{z}_h^{I}$ by using the score function $\hat{s}$ as a control variate; see e.g., Owen (2013) for background on control variate.

Theorem 3.4. *Assume $f_k(x) = \mathcal{N}(x; \mu_k, \Sigma_k)$, $\forall k \in [m]$. Denote by $\mathrm{MSE}(\hat{z}_h) = \lim_{n \to +\infty} n\mathbb{E}[(\hat{z}_h - z_h)^2]$ the asymptotic mean square error of $\hat{z}_h$, then we have*

$$\mathrm{MSE}(\hat{z}_h^{I}) = \mathrm{var}_f(h(x)) \geq \mathrm{MSE}(\hat{z}_h^{II}), \tag{7}$$

where $\mathrm{var}_f(h(x))$ is the variance of $h(x)$ under the global posterior $f(x) \propto \prod_k f_k(x)$.

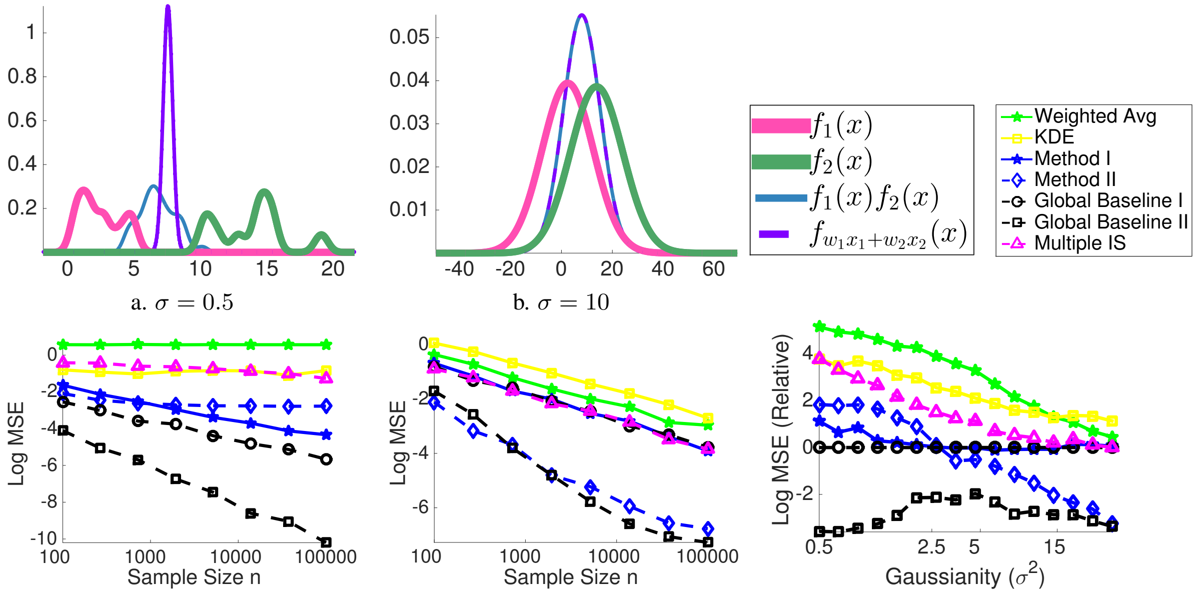
Proof. The fact that $\mathrm{MSE}(\hat{z}_h^{I}) \geq \mathrm{MSE}(\hat{z}_h^{II})$ is a result of Lemma 3.3 by the property of control variate (also see Henmi et al. (2007, Theorem 1)). The proof of $\mathrm{MSE}(\hat{z}_h^{I}) = \mathrm{var}_f(h(x))$ is shown in the Appendix. □

Therefore, despite being inconsistent in general non-Gaussian cases, *Method II* is guaranteed to outperform *Method I* in Gaussian cases, that is, it relies on a stronger assumption, and works well if the assumption is indeed satisfied. In contrast, *Method I* is more robust in that it is a consistent estimator for generic non-Gaussian f_k (but may also have large variances in bad cases). In practical cases when the size of each local dataset D^k is large, each f_k is close to Gaussian by Bernstein-von-Mises theorem, and we observe that *Method II* often performs better than *Method I* empirically.

3.4 FURTHER DISCUSSIONS

Our algorithm is summarized in Algorithm 1. We further discuss some issues here.

Communication Cost. Our methods outperform the previous *one-shot* combination methods such as Scott et al. (2013) and Neiswanger et al. (2013) in that *Method I* gives a consistent estimator for general f_k with a parametric $O(n^{-1/2})$ rate, while *Method II* provides exceptionally good estimates when f_k are (nearly) Gaussian. This is not surprising given that our methods leverage more information: it depends on both the local posterior samples x_i^k and their (unnormalized) likelihoods $f_k(x_i^k)$, as well as the (unnormalized) likelihoods $f_k(\bar{x}_i)$ of the combined sample $\bar{x}_i$. Therefore, compared with the *one-shot* methods, our methods require two additional rounds of communication between the subsets and the fusion center to evaluate and communicate the likelihood of the combined sample $\{\bar{x}_i\}$. The overall communication cost of our method is $O(mn(2d+1))$, where m is the number of machines, n is the Monte Carlo sample size and d is the dimension of the parameter, while that of Neiswanger et al. (2013) is $O(mnd)$, and that of Scott et al. (2013) is $O(d^2)$ which only needs to communicate the first two moments of the subset samples.



a. $\sigma = 0.5$

b. $\sigma = 10$

c. Result with fixed $\sigma = 0.5$

d. Result with fixed $\sigma = 10$

e. Varying σ from 0.5 to 10

Figure 1: Results on the toy Gaussian mixture model. (a)-(b) The simulated distributions $f_1(x)$ and $f_2(x)$ when $\sigma = 0.5$ (highly non-Gaussian) and $\sigma = 10$ (highly Gaussian), respectively. (c)-(d) The MSE for estimating the mean $\mathbb{E}_f[x]$ under the two cases shown in (a) and (b), respectively. (e) The relative MSE compared to *Global Baseline I* when we vary σ from 0.5 to 10, so that $f_1(x)$ and $f_2(x)$ change from being highly non-Gaussian to highly Gaussian.

Meanwhile, our method is still much more communication-efficient compared with the *iterative* combination methods that require more iterative rounds of communications, and resampling steps. The communication complexity of Wang & Dunson (2013) with T iterations is $O(mndT)$, and that of Xu et al. (2014) is $O(nd^2T)$ because it only passes the first two empirical moments instead of the samples. Our method has a significant advantage because the main practical bottleneck is often the number of communication rounds, regardless of the amount of information exchanged at each round. In our empirical results, we show that our methods work competitively with the iterative methods at their convergence.

Computational Cost. The total computational cost of our combination method is $O(nm(d^3 + L))$ where d^3 is due to the inverse of the covariance matrices and L denotes the cost for evaluating the local posterior probability $f_k(x)$; this is slightly worse than the linear averaging (Scott et al., 2013) which costs $O(nmd^3)$, but has advantage over Neiswanger et al. (2013) which requires a full MCMC procedure over the product of the KDEs for the combination. Further, the *iterative* combination methods have significantly higher computational cost, because they requires to re-draw subset samples iteratively, which is often much more expensive than the combination steps.

Random Permutation. The total size of all the local posterior samples is mn, while the size of the combined sample $\{\bar{x}_i\}$ is only n, that is, we lose a size of $(m-1)n$ when making the combination. To obtain more combined samples, we can randomly permute each subset sample $\{x_i^k : i \in [n]\}$ and combined the permuted subset samples, and repeat the process for multiple times. However, our empirical results do not suggest a significant improvement of performance by using multiple random permutations (e.g., we did not find significant improvement by averaging 10 random permutations).

4 EXPERIMENTS

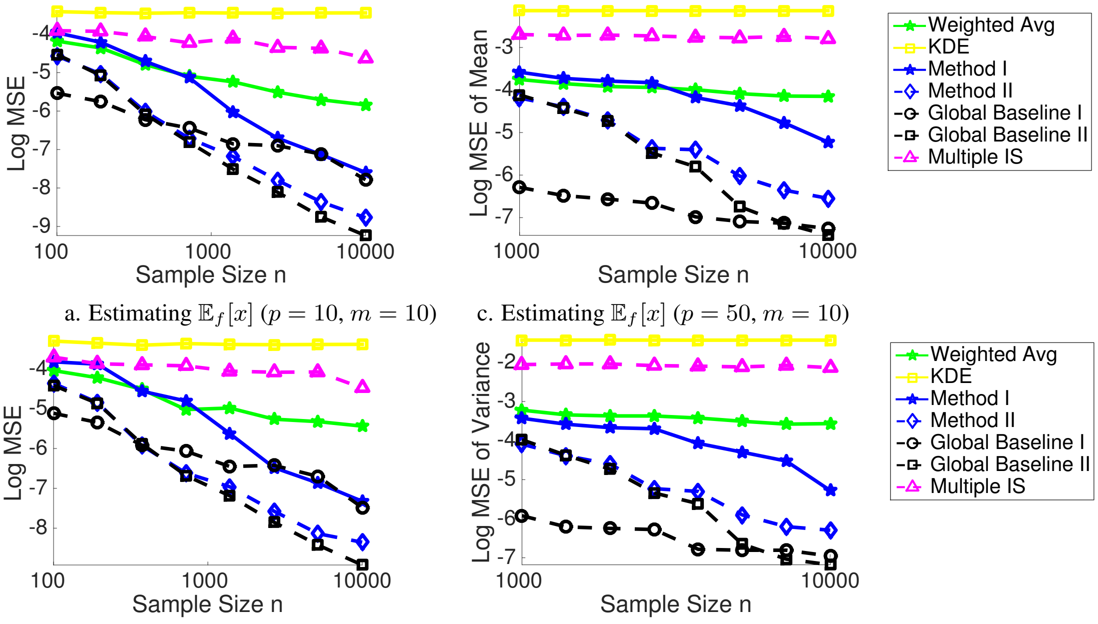
We report empirical results for our method using a toy example of mixture of Gaussians as well as a Bayesian probit model with both simulated and real world datasets. We compare with the following algorithms:

1. Our *Method I* and *Method II* as shown in Algorithm 1.

2. *Global Baseline I* and *Global Baseline II*, which draw sample $\{x_i^*\}$ from the global posterior $f(x) \propto \prod_k f_k(x)$, and estimate $\mathbb{E}_f(h(x))$ by

$$\hat{z}_h^{*I} = \frac{1}{n}\sum_i h(x_i^*), \qquad \hat{z}_h^{*II} = \frac{\sum_i \omega(x_i^*)h(x_i^*)}{\sum_i \omega(x_i^*)},$$

respectively, where $\omega(x_i^*) = f(x_i^*)/\hat{f}(x_i^*)$, and $\hat{f}(x) = \mathcal{N}(x;\ \hat{\mu}^*,\ \hat{\Sigma}^*)$ with $\hat{\mu}^*$ and $\hat{\Sigma}^*$ being the empirical mean and covariance matrix of $\{x_i^*\}$. Note that $\hat{z}_h^{*I}$ and $\hat{z}_h^{*II}$ can be treated as the global version of *Method I* and *Method II*, respectively; following Henmi et al. (2007), we can show



a. Estimating $\mathbb{E}_f[x]$ ($p = 10$, $m = 10$)

c. Estimating $\mathbb{E}_f[x]$ ($p = 50$, $m = 10$)

b. Estimating $\mathbb{E}_f[x^2]$ ($p = 10$, $m = 10$)

d. Estimating $\mathbb{E}_f[x^2]$ ($p = 50$, $m = 10$)

Figure 2: Results on probit regression model with simulated data D of size 12,000, partitioned into $m = 10$ subsets. (a)-(b) The MSE for estimating $\mathbb{E}_f[x]$ and $\mathbb{E}_f[x^2]$, respectively, when the dimension p of x is 10. (c)-(d) The results when the dimension p increases to 50. All the results are averaged over 500 random trials.

that $\hat{z}_h^{*II}$ always has smaller variance than $\hat{z}_h^{*I}$ when $f(x)$ is a Gaussian distribution.

3. *Weighted Avg*, which is the consensus Monte Carlo method by Scott et al. (2013).

4. The *KDE* method by Neiswanger et al. (2013).[2]

5. An naive *multiple importance weighted estimator (Multiple IS)* in which each subset sample $\{x_i^k\}_{i=1}^n \sim f_k$ is used to directly construct an importance sampling estimator for $\mathbb{E}_f(h(x))$:

$$\hat{z}_h^k = \frac{\sum_i h(x_i^k) w(x_i^k)}{\sum_i w(x_i^k)}, \quad \text{where} \quad w(x_i^k) = \frac{f(x_i^k)}{f_k(x_i^k)},$$

and the results from different subsets are combined by a weighted linear averaging:

$$\hat{z}_h^{MIS} = \frac{\sum_k v_k \hat{z}_h^k}{\sum_i v_k},$$

where v_k is chosen to be $v_k = 1/\hat{\text{var}}(\hat{z}_h^i)$.

6. The sampling via moment sharing (SMS) method by Xu et al. (2014),[3] which iteratively adjusts the local priors and draw local samples repeatedly.

4.1 TOY EXAMPLE

We first consider two Gaussian mixtures with 10 components,

$$f_k(x) = \frac{1}{10} \sum_{j=1}^{10} \mathcal{N}(x; \mu_{jk}, \sigma^2), \quad k = 1, 2,$$

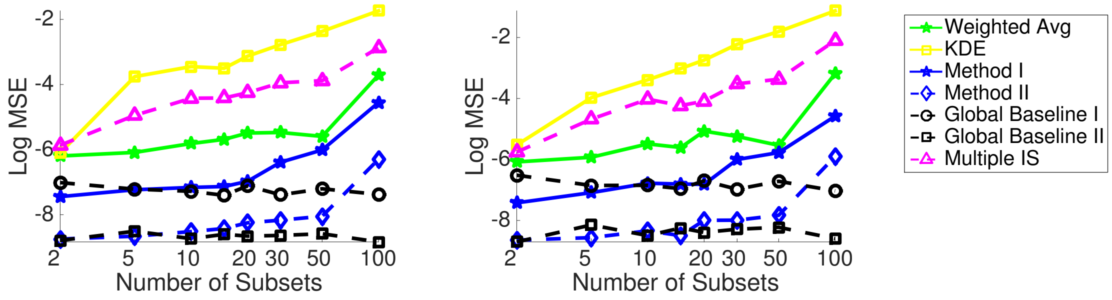
where μ_{jk} is randomly drawn from $\text{Uniform}([0, 10])$ for $f_1(x)$ and $\text{Uniform}([10, 20])$ for $f_2(x)$. The variance σ^2 is used to adjust the Gaussianity of $f_k(x)$. With a small σ (see Figure 1a), $f_1(x)$ and $f_2(x)$ are highly multi-modal, and are far away from each other; with a large σ (see Figure 1b), $f_1(x)$ and $f_2(x)$ become close to Gaussian and have a significant overlap with each other.

Figure 1(a) & (b) also shows the shapes of the corresponding product $f(x) \propto f_1(x) f_2(x)$ and the density function $f_{\bar{x}}(x)$ of the weighted averaging $\bar{x} = w_1 x_1 + w_2 x_2$ with $w_i \propto 1/\text{var}_{f_i}(x)$. We see that with a small σ, $f_{\bar{x}}(x)$ is very different from $f(x)$ but still covers a large part of $f(x)$, and hence can serve as a good importance sampling proposal (as approximately used in our methods). With a large σ, both $f(x)$ and $f_{\bar{x}}(x)$ are Gaussian like and are almost identical with each other.

Figure 1(c) & (e) shows the MSE of different algorithms when estimating the posterior mean $\mathbb{E}_f(h(x))$ with $h(x) = x$. Figure 1(c) shows the results of different algorithms when $\sigma = 0.5$ (the highly non-Gaussian case), in which we find that *Method I* works better than *Method II* and

[2] We used the code available at `https://www.cs.cmu.edu/~wdn/research/embParMCMC/index.html`.

[3] We used the code available at `https://github.com/BigBayes/SMS`



a. Estimating $\mathbb{E}_f[x]$ ($p = 10, n = 5 \times 10^3$) b. Estimating $\mathbb{E}_f[x^2]$ ($p = 10, n = 5 \times 10^3$)

Figure 3: Results on probit regression model with simulated data D of size $12,000$, partitioned into m subsets, with m ranging from 2 to 100. (a)-(b) The MSE for estimating $\mathbb{E}_f[x]$ and $\mathbb{E}_f[x^2]$, respectively; the posterior sample size is fixed to be $n = 5 \times 10^3$ in both cases. All the results are averaged over 500 random trials.

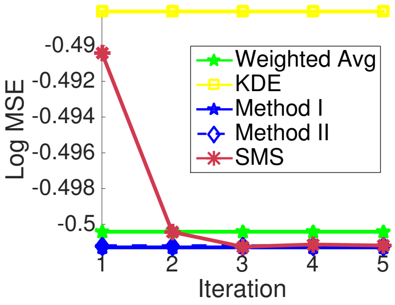


Figure 4: Comparing with SMS by Xu et al. (2014) on Bayesian probit regression. The setting is the same as that in Figure 3, except with fixed subset number $m = 10$.

Method II is clearly inconsistent in that its MSE stops decrease when the sample increases. Figure 1d shows the result of different algorithms when $\sigma = 10$ (the almost Gaussian case), in which we find that *Method II* works better than *Method I* as predicted by Theorem 3.4.

Figure 1(e) shows the results of different algorithms when we range σ from 0.5 to 10 (from highly non-Gaussian to highly Gaussian), and we can find that the performance of *Method I* converges to that of *Global baseline I* as predicted by Theorem 3.4, and that of *Method II* converges to *Global baseline II*. In all the cases, we find that both *Weighted Avg* and *KDE* perform much worse. *Multiple IS* tends to perform well when f_1 and f_2 are close to each other, but is worse when they are far apart from each other.

4.2 BAYESIAN PROBIT REGRESSION

We consider the Bayesian probit regression model for binary classification. Let $D = \{\chi_\ell, \zeta_\ell\}_{\ell=1}^N$ be a set of observed data with p-dimensional features $\chi_\ell \in \mathbb{R}^p$ and binary labels $\zeta_\ell \in \{0, 1\}$. The probit model is

$$p(D|x) = \prod_{\ell=1}^{N} \left[\zeta_\ell \Phi(x^\top \chi_\ell) + (1 - \zeta_\ell)(1 - \Phi(x^\top \chi_\ell))\right],$$

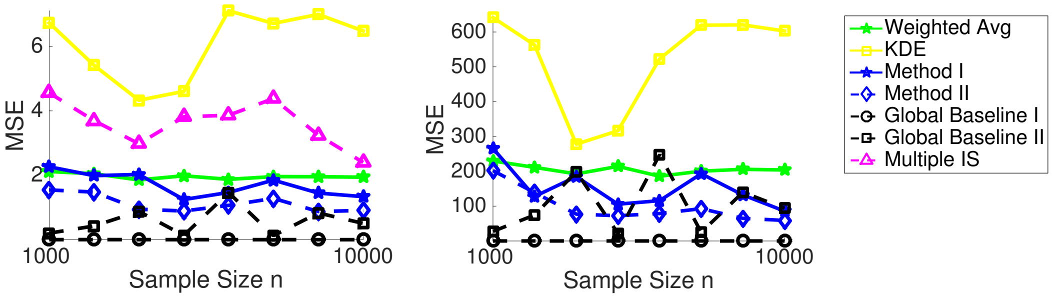
where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. We use an uninformative Gaussian prior $p(x) = \mathcal{N}(x; 0, 0.1)$ on x throughout our experiments.

We start with testing our methods on simulated datasets, where we first generate a true value of x with 50% zero elements and 50% elements drawn randomly from standard normal distribution, and then simulate a dataset $D = \{\chi_\ell, \zeta_\ell\}_{\ell=1}^N$ that is subsequently evenly partitioned into m subsets $\{D^k\}_{k=1}^m$, each of which includes N/m data points. We simulate $N = 12,000$ number of points throughout our experiments.

Figure 2(a) & (b) show the mean square error when estimating the posterior mean $\mathbb{E}_f[x]$ and the second order moment $\mathbb{E}_f[x^2]$, respectively, both when the dataset D is partitioned into $m = 10$ subsets (so that each subset D^k receives $1,200$ data points). We can see that as the posterior sample size n of the subset samples $\{x_i^k\}_{i=1}^n \sim p(x|D^k)$ increases, our *Method I* and *Method II* match closely with the *Global Baseline I* and *Global Baseline II*, respectively; this may suggest that the local posteriors f_k are close to Gaussian in this case. The other methods, including *Weighted Avg*, *KDE* and *Multiple IS*, work significantly worse than both of our methods.

Figure 2(c) & (d) shows the results under the same setting as Figure 2(a) & (b), except when the dimension p increases to 50, where we observe that our *Method I* and *Method II* match less closely with the corresponding global baselines, but still tend to significantly outperform all the other distributed algorithms.

Figure 3(a) & (b) show the results when we fix the dimension $p = 10$ and a posterior sample size of $n = 5 \times 10^3$ and partition the dataset D into different number m of subsets (so that each subset D^k receives $12,000/m$ data points), with m range from 2 to 100. We observed that our *Method I* and *Method II* again match closely with the *Global Baseline I* and *Global Baseline II*, except when the partition number m is very high (e.g., $m \geq 30$ for *Method I*, and



a. Estimating $\mathbb{E}_f[x]$ ($p = 54$, $m = 10$) b. Estimating $\mathbb{E}_f[x^2]$ ($p = 54$, $m = 10$)

Figure 5: Probit regression on the CoverType dataset. The dataset is partitioned evenly into 10 subsets. The result of *Multiple IS* in (b) is much worse than the other methods and is not shown in the figure.

$m \geq 100$ for *Method II*).

Figure 4 shows the result when we compare our methods with the iterative SMS method by Xu et al. (2014), where we find that our *Method I* and *Method II* tend to perform as well as SMS at its convergence, but has the advantage of requiring no iterative communication or re-sampling. We also tested the Weistrass sampler by Wang & Dunson (2013) (result does not report), but find it often performs worse (results similar to Figure 3 of Xu et al. (2014)).

In addition, we experimented with an iterative version of our method which introduces local priors $p_k(x)$ that satisfy $\prod_k p_k(x) = p(x)$ where $p(x)$ is the original global prior, and iteratively updates $p_k(x)$ to make the local posteriors $f_k(x) = p(D^k|x)p_k(x)$ match with each other. We observe that this iterative version does not improve the result significantly, likely because the non-iterative version of our method is already good enough.

Binary CoverType Dataset We then test our methods on the Forest Covertype dataset from the UCI machine learning repository (Bache & Lichman, 2013); it has 54 features, and is reprocessed to get binary labels following Collobert et al. (2002). For our experiment, we take the first 12,000 data points, and partition them into 10 subsets. The results of different algorithms are shown in Figure 5, in which we see that our *Method I* and *Method II* still perform significantly better than the other distributed algorithms.

5 CONCLUSION

We propose an importance weighted consensus Monte Carlo approach for distributed Bayesian inference. Two practical versions of our method are proposed, and their properties are studied both theoretically and empirically. Our methods have significant advantages over the previous one-shot methods based on density estimates in terms of accuracy, as well as the iterative methods in terms of computational and communication costs.

APPENDIX

Proof of Theorem 3.4. We only prove $\mathrm{MSE}(z_h^I) = \mathrm{var}_f(h(x))$ here; the fact that $\mathrm{MSE}(z_h^I) \geq \mathrm{MSE}(z_h^{II})$ can be found in Henmi et al. (2007, Theorem 1).

Let $\mathrm{MLE}_n(\hat{z}_h^I) = n\mathbb{E}[(\hat{z}_h^I - z_h)^2]$; using the Delta method we can show that $\mathrm{MLE}_n(z_h^I) \asymp \mathbb{E}_q[(h(\bar{x}) - z_h)^2 w_n(\boldsymbol{x})^2]$, with

$$w_n(\boldsymbol{x}) = \frac{p_n(\boldsymbol{x})}{q(\boldsymbol{x})} = \frac{\mathcal{N}(\bar{x}, \mu_0, \Sigma_0)}{\mathcal{N}(\bar{x}, \hat{\mu}_0, \hat{\Sigma}_0)} \prod_k \frac{\mathcal{N}(x^k, \mu_k, \Sigma_k)}{\mathcal{N}(x^k, \hat{\mu}_k, \hat{\Sigma}_k)},$$

where $q(\boldsymbol{x}) = \prod_k \mathcal{N}(x_k;\ \mu_k, \Sigma_k)$, and $p_n(\boldsymbol{x}) = \mathcal{N}(\bar{x};\ \mu_0, \Sigma_0) \prod_k \mathcal{N}(x_k;\ \hat{\mu}_k, \hat{\Sigma}_k)/\mathcal{N}(\bar{x};\ \hat{\mu}_0, \hat{\Sigma}_0)$; here $w_n(\boldsymbol{x})$ and $p_n(\boldsymbol{x})$ are indexed with sample size n since they dependent on the empirical means and variances. Since $\mathrm{var}_f(h(x)) = \mathbb{E}_{p_n}[(h(\bar{x}) - z_h)^2]$ by Proposition 3.1, we have

$$\begin{aligned}
&\mathrm{MSE}_n(z_h^I) - \mathrm{var}_f(h(x)) \\
&\asymp \mathbb{E}_q[(h(\bar{x}) - z_h)^2 w_n(\boldsymbol{x})^2] - \mathbb{E}_{p_n}[(h(\bar{x}) - z_h)^2] \\
&= \mathbb{E}_{p_n}[(h(\bar{x}) - z_h)^2 w_n(\boldsymbol{x})] - \mathbb{E}_{p_n}[(h(\bar{x}) - z_h)^2] \\
&= \mathbb{E}_{p_n}[(h(\bar{x}) - z_h)^2 (w_n(\boldsymbol{x}) - 1)]
\end{aligned}$$

Using Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\big(\mathrm{MSE}(z_h^I) - \mathrm{var}_f(h(x))\big)^2 \\
&\leq \mathbb{E}_{p_n}[(h(\bar{x}) - z_h)^4] \cdot \mathbb{E}_{p_n}[(w_n(\boldsymbol{x}) - 1)^2] \\
&= \mathbb{E}_f[(h(x) - z_h)^4] \cdot \mathbb{E}_q[w_n(\boldsymbol{x})(w_n(\boldsymbol{x}) - 1)^2].
\end{aligned}$$

Therefore, we just need to show that $\mathbb{E}_q[w_n(\boldsymbol{x})(w_n(\boldsymbol{x}) - 1)^2] \to 0$. This can be done using dominant convergence theorem: Let $\psi_n(\boldsymbol{x}) = q(\boldsymbol{x}) w_n(\boldsymbol{x})(w_n(\boldsymbol{x}) - 1)^2$, then we have $\psi_n(x) \to 0$, $\forall \boldsymbol{x}$ since $\hat{\mu}_k \to \mu_k$, $\hat{\Sigma}_k \to \Sigma_k$ and hence $w_n(\boldsymbol{x}) \to 1$ for $\forall \boldsymbol{x}$; in addition, we can show that $|\psi_n(\boldsymbol{x})| \leq q(\boldsymbol{x})^{1/2}$ for large enough n and $q(\boldsymbol{x})^{1/2}$ is an integrable function.

□

References

Angelino, E., Johnson, M. J., and Adams, R. P. Patterns of scalable Bayesian inference. *http://arxiv.org/abs/1602.05221*, 2015.

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Baker, J., Fearnhead, P., and Fox, E. Computational statistics for big data. 2015.

Bardenet, R., Doucet, A., and Holmes, C. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *ICML*, pp. 405–413, 2014.

Bardenet, R., Doucet, A., and Holmes, C. On Markov chain Monte Carlo methods for tall data. *arXiv preprint arXiv:1505.02827*, 2015.

Birge, L. and Massart, P. Estimation of integral functionals of a density. *The Annals of Statistics*, pp. 11–29, 1995.

Collobert, R., Bengio, S., and Bengio, Y. A parallel mixture of SVMs for very large scale problems. *Neural computation*, 14(5):1105–1114, 2002.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862, 2015.

Henmi, M. and Eguchi, S. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91(4):929–941, 2004.

Henmi, M., Yoshida, R., and Eguchi, S. Importance sampling via the estimated sampler. *Biometrika*, 94(4):985–991, 2007.

Huang, Z. and Gelman, A. Sampling for Bayesian computation with large datasets. *Available at SSRN 1010107*, 2005.

Korattikara, A., Chen, Y., and Welling, M. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *ICML*, 2014.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. Nonparametric estimation of Renyi divergence and friends. In *ICML*, 2014.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. On estimating L_2^2 divergence. In *AISTATS*, 2015.

Maclaurin, D. and Adams, R. P. Firefly Monte Carlo: Exact MCMC with subsets of data. In *UAI*, 2014.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. B. Robust and scalable Bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*, 2014.

Neiswanger, W., Wang, C., and Xing, E. Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*, 2013.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Rabinovich, M., Angelino, E., and Jordan, M. I. Variational consensus Monte Carlo. *arXiv preprint arXiv:1506.03074*, 2015.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E., and McCulloch, R. Bayes and big data: The consensus Monte Carlo algorithm. In *EFaBBayes 250 conference*, volume 16, 2013.

Wang, X. and Dunson, D. B. Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, pp. 681–688, 2011.

Xu, M., Lakshminarayanan, B., Teh, Y. W., Zhu, J., and Zhang, B. Distributed Bayesian posterior sampling via moment sharing. In *NIPS*, pp. 3356–3364, 2014.

Zhu, J., Chen, J., and Hu, W. Big learning with Bayesian methods. *arXiv preprint arXiv:1411.6370*, 2014.