WASP: Scalable Bayes via barycenters of subset posteriors

Sanvesh Srivastava[1,2] ss602@stat.duke.edu
Volkan Cevher[3] volkan.cevher@epfl.ch
Quoc Tran-Dinh[3] quoc.trandinh@epfl.ch
David B. Dunson[1] dunson@duke.edu
Department of Statistical Science, Duke University, Durham, North Carolina, USA[1]
Statistical and Applied Mathematical Sciences Institute (SAMSI), Durham, North Carolina, USA[2]
Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne, Switzerland[3]

Abstract

The promise of Bayesian methods for big data sets has not fully been realized due to the lack of scalable computational algorithms. For massive data, it is necessary to store and process subsets on different machines in a distributed manner. We propose a simple, general, and highly efficient approach, which first runs a posterior sampling algorithm in parallel on different machines for subsets of a large data set. To combine these subset posteriors, we calculate the Wasserstein barycenter via a highly efficient linear program. The resulting estimate for the Wasserstein posterior (WASP) has an atomic form, facilitating straightforward estimation of posterior summaries of functionals of interest. The WASP approach allows posterior sampling algorithms for smaller data sets to be trivially scaled to huge data. We provide theoretical justification in terms of posterior consistency and algorithm efficiency. Examples are provided in complex settings including Gaussian process regression and nonparametric Bayes mixture models.

1 Introduction

Efficient computation for massive data commonly relies on using small subsets of the data in parallel, combining results of local computations to obtain global results. The usual focus is on obtaining parameter estimates, which minimize a loss function based on the complete data, by dividing computation into subset-specific optimization problems. One widely used and well understood framework is ADMM [4; 6]. In Bayesian statistics, one is faced with the more challenging problem of approximating a posterior measure for the unknown parameters instead of just obtaining a single point estimate of these parameters. Posterior measures have the major advantage of providing a characterization of uncertainty in parameter learning and predictions; such uncertainty quantification is lacking for many optimization approaches. However, a fundamental disadvantage is the lack of scalable computational algorithms for accurately approximating posterior measures in general Bayesian models.

The main workhorse of Bayesian posterior computation is Markov chain Monte Carlo (MCMC) sampling, with variants such as sequential Monte Carlo (SMC) and adaptive Monte Carlo being also popular. Such Monte Carlo algorithms obtain samples from the posterior measure, which are used to estimate summaries of the posterior and predictive distributions of interest. These algorithms for posterior sampling are applicable to essentially any Bayesian model, but face computational bottlenecks in scaling up to large data sets. Such bottlenecks can potentially be addressed by using data subsets to define *subset posterior* measures, which provide a noisy approximation to the posterior measure based on the full data. After applying Monte Carlo algorithms to obtain samples from each subset posterior in parallel, the goal is then to efficiently combine these samples to obtain samples from an approximation to the posterior measure for the full data.

This article focuses on fundamentally improving the combining step, bypassing the need to introduce a kernel or tuning parameters, while massively speeding up computations. In particular, the proposed approach combines samples from subset posterior measures by calculating the barycenter with respect to a Wasserstein distance between measures. The resulting Wasserstein posterior (WASP) can be estimated efficiently via a linear program, and has an atomic form,

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

making calculation of posterior and predictive summaries trivial. The WASP framework is tuned for distributed computations. Posterior samples from local computations are combined in the final step to yield a posterior distribution that is close to the true posterior distribution with respect to a Wasserstein distance.

Current scalable Bayes methods belong to three major groups. The first group estimates an approximate posterior distribution that is closest to the true posterior while restricting the search to a parametric family. While these methods have restrictive distributional assumptions, they can be made computationally efficient [24; 3; 5; 15; 14; 19]. The second group exploits the analytic form of posterior and uses computer architecture to improve the sampling time and convergence [20; 22; 1]. This approach is ideal for large-scale applications that use simple parametric models. The third group obtains subset posteriors using some sampling algorithm and combines them by using kernel density estimation [18], Weierstrass transform [23], or minimizing a loss defined on the reproducing kernel Hilbert space (RKHS) embedding of the subset posteriors [16]. These methods are flexible in that they are not restricted to a parametric class of models; however, the results can vary significantly depending on the choice of kernels without a principled approach for kernel choice.

The WASP framework relies on two general assumptions that are frequently satisfied in Bayesian applications. These assumptions relate the parameter space and the associated space of probability measures:

(a) The parameter space is a metric space.

(b) The atomic approximations of the subset posteriors (empirical measures) can be obtained efficiently using a posterior sampler.

Assumption (a) is frequently satisfied in practice using the Euclidean distance. Assumption (b) can be typically satisfied if the subset size and the number of unknown parameters are not too large. Due to the geometric properties of the Wasserstein metric [21], the Wasserstein barycenter (WB) of the subset posteriors has highly appealing statistical and computational properties. We formulate a linear program (LP) to estimate the (atomic) WB of atomic approximations of subset posteriors. Exploiting the sparsity of the LP, we efficiently estimate the WB using standard software, such as Gurobi [13].

The WASP framework is inspired from recent developments in optimal transport problems [8; 9] and scalable Bayes methods [16]. Minsker et al. [16] proposed to use the geometric median of subset posteriors, calculated using a RKHS embedding that required choice of a kernel and corresponding bandwidth. The resulting median posterior provides a robust alternative to the true posterior. Inspired by this idea, we focus on approximating the true posterior instead of robustifying it. This removes the need for the RKHS embedding and vastly speeds up the computation time. Extending the Sinkhorn algorithm of Cuturi [8], Cuturi and Doucet [9] estimate the WB of empirical measures using entropy-smoothed subgradient methods. We instead estimate the WASP by solving a LP and efficiently obtain its solution by exploiting the sparsity of LP constraints.

2 Preliminaries

We first describe the Wasserstein space of probability distributions, with the Wasserstein distance as a natural metric. We then recall the relation between Wasserstein distance and the objective of optimal transportation problems. Based on these ideas, we highlight the role of the WB as a summary of a collection of posterior distributions for scalable Bayesian inference.

2.1 Notations

The ordered pair $(\Theta,\ d)$ represents a metric space Θ with metric d. In this work, we are only concerned with complete separable metric space (Polish space). The space of Borel probability measures on Θ is represented by $\mathcal{P}(\Theta)$. The symbol $\delta_{\boldsymbol{\theta}_0}$ denotes the Dirac measure concentrated at $\boldsymbol{\theta}_0 \in \Theta$, i.e., $\delta_{\boldsymbol{\theta}_0}(A) = 1\{\boldsymbol{\theta}_0 \in A\}$ for any Borel measurable set A. If ψ is a Borel map $\Theta \to \Theta$ and ν is a measure on Θ, then the push-forward of ν through ψ is the measure $\psi\#\nu$ defined as $\int_\Theta f(y)d(\psi\#\nu)(y) = \int_\Theta f(\psi(x))d\nu(x)$ for every continuous bounded function f on Θ. The N-dimensional simplex is $\Delta_N = \{\mathbf{a} \in \mathbb{R}^N\,|\forall n \leq N, 0 \leq a_n \leq 1, \sum_{n=1}^N a_n = 1\}$ and $\Theta^{(n)}$ represents the n-dimensional Cartesian product $\Theta \times \ldots \times \Theta$. We represent the Frobenius inner product between two matrices $\mathbf{A}$ and $\mathbf{B}$ as $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}^T\,\mathbf{B})$ and $\|\cdot\|_2$ denotes the standard Euclidean distance.

2.2 Wasserstein space and distance

The Wasserstein space of order $p \in [1, \infty)$ of probability measures on $(\Theta,\ d)$ for an arbitrary $\boldsymbol{\theta}_0 \in \Theta$ is defined as

$$\mathcal{P}_p(\Theta) := \left\{\mu \in \mathcal{P}(\Theta) : \int_\Theta d(\boldsymbol{\theta}_0, \boldsymbol{\theta})^p d\mu(\boldsymbol{\theta}) < \infty\right\}, \tag{1}$$

and $\mathcal{P}_p(\Theta)$ is independent of the choice of $\boldsymbol{\theta}_0$. The Wasserstein distance of order p between $\mu, \nu \in \mathcal{P}(\Theta)$

is defined as

$$W_p(\mu,\nu) := \left\{ \inf_{\tau \in \mathcal{T}(\mu,\nu)} \int_{\Theta^{(2)}} d(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)^p d\tau(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) \right\}^{1/p}, \tag{2}$$

where $\mathcal{T}(\mu,\nu)$ is the set of all probability measures on $\Theta^{(2)}$ with marginals μ and ν, respectively. W_p is attractive in that it metrizes the weak convergence on Θ and preserves the geometry of the space. An important topological property of $\mathcal{P}_p(\Theta)$ that will used later in proving posterior consistency is as follows.

Theorem 2.1 ([21]) *If (Θ, d) is a Polish space and $p \in [1,\infty)$, then $(\mathcal{P}_p(\Theta), W_p)$ is a Polish space. More specifically,*

(a) Given $\epsilon > 0$, an arbitrary $\boldsymbol{\theta}_0$ in a dense subset of Θ, and $\mu \in \mathcal{P}_2(\Theta)$, there exists a compact set $\Theta_\epsilon \subset \Theta$ such that $\int_{\Theta \setminus \Theta_\epsilon} d(\boldsymbol{\theta}_0,\boldsymbol{\theta})^p d\mu(\boldsymbol{\theta}) < \epsilon^p$.

(b) $\exists M_\epsilon < \infty$ such that $\| \boldsymbol{\theta} \| < M_\epsilon \, \forall \, \boldsymbol{\theta} \in \Theta_\epsilon$.

We focus on $\Theta \subset \mathbb{R}^D$, $d = \| \cdot \|_2$, and metric space $(\mathcal{P}_2(\Theta), W_2)$.

2.3 Wasserstein distance as a solution of an optimal transport problem

If μ and ν are discrete probability measures, then $W_2(\mu,\nu)$ (2) is the minimum objective function value of a discrete optimal transport problem [8]. Let μ and ν be atomic measures supported on $\Theta_1 \in \mathbb{R}^{N_1 \times D}$ and $\Theta_2 \in \mathbb{R}^{N_2 \times D}$ so that $\mu = \sum_{n=1}^{N_1} a_n \delta_{\boldsymbol{\theta}_{1_n}^T}$ and that $\nu = \sum_{n=1}^{N_2} b_n \delta_{\boldsymbol{\theta}_{2_n}^T}$, where $\mathbf{a} \in \Delta_{N_1}$, $\mathbf{b} \in \Delta_{N_2}$, and $\boldsymbol{\theta}_{i_n}^T$ is the n-th row of Θ_i. The discrete optimal transport problem is formulated in terms of (a) the matrix $\mathbf{M}_{12} \in \mathbb{R}_+^{N_1 \times N_2}$ of pairwise squared distances between the collection of atoms in Θ_1 and Θ_2; and (b) the optimal transport polytope that is the set of all feasible solutions called transport plans. The (i,j)-th entry of $\mathbf{M}_{12}$ is

$$[\mathbf{M}_{12}]_{ij} = \| \boldsymbol{\theta}_{1_i} - \boldsymbol{\theta}_{2_j} \|_2^2. \tag{3}$$

The optimal transport polytope is defined as

$$\mathcal{T}(\mathbf{a},\mathbf{b}) = \{ \mathbf{T} \in \mathbb{R}_+^{N_1 \times N_2} : \mathbf{T}\,\mathbf{1}_{N_2} = \mathbf{a}, \mathbf{T}^T\,\mathbf{1}_{N_1} = \mathbf{b} \}; \tag{4}$$

therefore, transport plan $\mathbf{T} \in \mathcal{T}(\mathbf{a},\mathbf{b})$ is a $N_1 \times N_2$ doubly stochastic matrix such that its row sums equal $\mathbf{a}$ and its column sums equal $\mathbf{b}$. Based on (3) and (4), the objective of discrete optimal transport problem is

$$d_{\mathbf{M}_{12}}(\mathbf{a},\mathbf{b}) := \min_{\mathbf{T} \in \mathcal{T}(\mathbf{a},\mathbf{b})} \langle \mathbf{T}, \mathbf{M}_{12} \rangle = W_2^2(\mu,\nu), \tag{5}$$

where the last equality is implied by the definition of W_2 (2); see [8] for details. The worst case complexity of solving (5) scales as $\mathcal{O}((N_1 N_2)^3 \log(N_1 N_2))$. Motivated by this limitation, Cuturi [8] smoothed the objective in (5) using entropy and derived an efficient algorithm for calculating optimal $\mathbf{T}$ based on Sinkhorn algorithm. This approach, however, has limited applications since the parameter that controls the amount of entropy-smoothing needs to be specified *apriori*, and the results can be sensitive to its choice.

2.4 Wasserstein barycenter

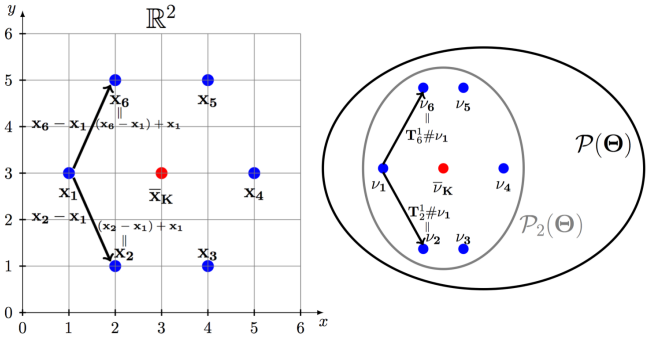

Figure 1: Barycenter in $\mathbb{R}^2$ and $\mathcal{P}_2(\Theta)$. The Euclidean and Wasserstein barycenters are represented as $\overline{\mathbf{x}}_K$ and $\overline{\nu}_K$ (in red) when $K = 6$ and $\lambda_k = 1/6$ in (6) and (7). The arrows represent constant speed geodesics in $\mathbb{R}^2$ and $\mathcal{P}_2(\Theta)$, respectively. The geodesic in $\mathbb{R}^2$ is the straight line joining two points, where as the geodesic in $\mathcal{P}_2(\Theta)$ corresponds to the measure preserving map $\mathbf{T}_j^1$ such that $\nu_j = \mathbf{T}_j^1 \# \nu_1$ for $j = 2, 6$.

WB generalizes the Euclidean barycenter (EB) to $\mathcal{P}_2(\Theta)$. If $\mathbf{x}_1, \ldots, \mathbf{x}_K \equiv \mathbf{x}_{1:K} \in \mathbb{R}^D$, then their EB $\overline{\mathbf{x}}_{K,\boldsymbol{\lambda}} = \sum_{k=1}^K \lambda_k \mathbf{x}_k$ for $\boldsymbol{\lambda} \in \Delta_K$ is such that

$$\sum_{k=1}^K \lambda_k \| \mathbf{x}_k - \overline{\mathbf{x}}_{K,\boldsymbol{\lambda}} \|_2^2 = \inf_{\mathbf{y} \in \mathbb{R}^D} \sum_{k=1}^K \lambda_k \| \mathbf{x}_k - \mathbf{y} \|_2^2; \tag{6}$$

see Figure 1. Generalizing (6) to $\mathcal{P}_2(\Theta)$, Agueh and Carlier [2] showed that if $\nu_1, \ldots, \nu_K \equiv \nu_{1:K} \in \mathcal{P}_2(\Theta)$, then their WB $\overline{\nu}_{K,\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda} \in \Delta_K$ is such that

$$\sum_{k=1}^K \lambda_k W_2^2(\nu_k, \overline{\nu}_{K,\boldsymbol{\lambda}}) = \inf_{\nu \in \mathcal{P}_2(\Theta)} \sum_{k=1}^K \lambda_k W_2^2(\nu_k, \nu); \tag{7}$$

see Figure 1. Agueh and Carlier [2] also showed that $\overline{\nu}_{K,\boldsymbol{\lambda}}$ in (7) can be obtained as a solution to a LP problem posed as a multimarginal optimal transportation problem [11]. We only present their main result that relates $\overline{\nu}_{K,\boldsymbol{\lambda}}$ and $\nu_{1:K}$. Recall that if σ is a Borel map $\mathbb{R}^D \to \mathbb{R}^D$, then the push-forward of μ through σ is the measure $\sigma \# \mu$; see Section 2.1. If $\mathbf{T}_k^1$ represents

the measure preserving map from ν_1 to ν_k such that $\nu_k = \mathbf{T}_k^1 \# \nu_1$ for $k = 1, \ldots, K$, then

$$\overline{\nu}_{K,\boldsymbol{\lambda}} := \left(\sum_{k=1}^{K} \lambda_k \, \mathbf{T}_k^1 \right) \# \nu_1 \tag{8}$$

generalizes the EB (6) to WB (7) in $\mathcal{P}_2(\Theta)$ [2]. We use this result later in proving Theorem 3.3; see Theorem 4.1 and Proposition 4.2 of [2] for greater details. We also note that there are many formulations of (7) in literature but have been solved using different tools or appear under different names [9]. Extending the Sinkhorn algorithm [8], Cuturi and Doucet [9] propose two fast algorithms for calculating entropy-smoothed versions of $\overline{\nu}_{K,\boldsymbol{\lambda}}$ using gradient based methods. The WASP framework reformulates (7) as a sparse LP problem that is computationally efficient without requiring any entropy-smoothing.

3 Contributions and main results

This section proposes to combine a collection of subset posteriors using their WB called WASP. We prove that WASP is strongly consistent at the true value $\boldsymbol{\theta}_0 \in \Theta$. Specifically, WASP converges weakly to $\delta_{\boldsymbol{\theta}_0}$ in $(\mathcal{P}_2(\Theta), W_2)$. We also modify the subset posteriors appropriately so that the uncertainty quantification of the WASP is well-calibrated. Finally, we reformulate (7) as a sparse LP for fast estimation of WASP.

3.1 Wasserstein Barycenter for scalable Bayesian inference

We first highlight important topological properties of the metric space $(\Theta, \|\cdot\|_2)$ that will be used for proving theoretical properties of the WASP. Let $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ be a family of probability distributions parameterized by $\Theta \subset \mathbb{R}^D$. The norm topology of $(\mathbb{R}^D, \|\cdot\|_2)$ when restricted to Θ implies that $(\Theta, \|\cdot\|_2)$ is also a Polish space. Using Theorem 2.1, W_2 metrizes the topology of weak convergence in $\mathcal{P}_2(\Theta)$. For all $\boldsymbol{\theta} \in \Theta$, $P_{\boldsymbol{\theta}}$ is assumed to be absolutely continuous with respect to the Lebesgue measure dx on $\mathbb{R}^D$ so that $dP(\cdot\,|\,\boldsymbol{\theta}) = p(\cdot\,|\,\boldsymbol{\theta})dx$. All statements regarding the convergence of measures in the context of WASP are in the metric space $(\mathcal{P}_2(\Theta), W_2)$.

We now recall some basic concepts from nonparametric Bayes theory. Most of these concepts and definitions are based on fundamental results of [12]. Let $\mathcal{C}$ be the Borel σ-field on Θ and Π_n be a (prior) probability measure on $(\Theta, \mathcal{C})$. Suppose that we observe random variables $(X_1, \ldots, X_n) \equiv X^{(n)}$ that are independent and identically distributed as $P_{\boldsymbol{\theta}_0}$ for some unknown $\boldsymbol{\theta}_0 \in \Theta$. Assume that random variables $X^{(n)}$ are defined on the fixed measurable space $(\Omega, \mathcal{A})$ and that $P_{\boldsymbol{\theta}_0}^{(n)} = P_{\boldsymbol{\theta}_0}^{\infty}(X^{(n)})^{-1}$ for all n and $\boldsymbol{\theta}_0 \in \Theta$ is unknown. Given a Bayesian model, we obtain the (random) posterior distribution $\Pi_n(\cdot|X^{(n)})$. A version of Bayes theorem implies that for all Borel measurable $\mathcal{U} \subset \Theta$

$$\Pi_n(\mathcal{U}\,|X^{(n)}) = \frac{\int_{\mathcal{U}} \prod_{i=1}^{n} p(X_i|\,\boldsymbol{\theta}) d\,\Pi_n(\boldsymbol{\theta})}{\int_{\Theta} \prod_{i=1}^{n} p(X_i|\,\boldsymbol{\theta}) d\,\Pi_n(\boldsymbol{\theta})}. \tag{9}$$

The following is a useful notion of consistency that characterizes the behavior of random posterior distribution $\Pi_n(\mathcal{U}\,|X^{(n)})$ as $n \to \infty$.

Definition 3.1 *(Strong Consistency) A posterior distribution $\Pi_n(\cdot|X^{(n)})$ is said to be strongly consistent at $\boldsymbol{\theta}_0$ if $\Pi_n(\cdot|X^{(n)}) \to_w \delta_{\boldsymbol{\theta}_0}$ as $n \to \infty$ a.s. $[P_{\boldsymbol{\theta}_0}^{\infty}]$.*

This is a stronger notion of consistency in that strong consistency of $\Pi_n(\cdot|X^{(n)})$ implies that there exists a consistent estimator of $\boldsymbol{\theta}_0$. It is well-known that under fairly general conditions $\Pi_n(\cdot|X^{(n)})$ is strongly consistent at $\boldsymbol{\theta}_0$ [12]. A necessary condition for posterior consistency to hold states that the prior must assign positive probabilities to every Kullback-Leibler (KL) neighborhood of $\boldsymbol{\theta}_0$.

Definition 3.2 *(KL property) Let $\boldsymbol{\theta} \sim \Pi$. Then Π has KL property at $\boldsymbol{\theta}_0 \in \Theta$, if $\Pi(\mathcal{K}_\epsilon(\boldsymbol{\theta}_0)) > 0 \ \forall \epsilon > 0$, where $\mathcal{K}_\epsilon(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} : KL(p_{\boldsymbol{\theta}_0}||q_{\boldsymbol{\theta}}) < \epsilon\}$, for densities $p_{\boldsymbol{\theta}_0}$ and q_θ with respect to the reference measure ν, and $KL(f||g) = \int f \log \frac{f}{g} d\nu$. This property is represented as $\boldsymbol{\theta}_0 \in KL(\Pi)$.*

Intuitively, the KL property requires the prior to assign positive probability to any small neighborhood of $\boldsymbol{\theta}_0$. If this property is not satisfied, then it is well-known that the posterior $\Pi_n(\cdot|X^{(n)})$ might fail to concentrate around $\boldsymbol{\theta}_0$ even when $n \to \infty$.

Posterior sampling for massive data is frequently intractable in Bayesian applications. A typical divide-and-conquer strategy randomly partitions $X^{(n)}$ into K subsets $X_{[k]}$ and obtains subset posteriors $\Pi_{k_n}(\cdot|X_{[k]})$ for $k = 1, \ldots, K$. Without loss of generality, we assume that each subset is of size m so that $n = Km$. The following lemma is used to prove strong consistency of subset posteriors. It is similar in spirit to Theorem 2.1 of Ghosal et al. [12]. All proofs are included in the supplementary materials.

Lemma 3.1 *Let Θ be a compact subset of $\mathbb{R}^D$, let $\mathcal{P}_2(\Theta)$ be the Wasserstein space of probability measures parametrized by $\boldsymbol{\theta} \in \Theta$, and let the true measure $\delta_{\boldsymbol{\theta}_0} \in \mathcal{P}_2(\Theta)$. Given $\epsilon > 0$, Θ_ϵ is a compact subset of Θ that satisfies (a) in Theorem 2.1, M_ϵ is a large number that satisfies (b) in Theorem 2.1, and Π_n satisfies the KL property $\forall n$ such that $\liminf_{n\to\infty} \Pi_n(\mathcal{K}_\epsilon(\boldsymbol{\theta}_0)) > 0$ $\forall \epsilon > 0$. Then, $\Pi_n(\cdot|X^{(n)})$ is strongly consistent at $\boldsymbol{\theta}_0$.*

Intuitively, Lemma 3.1 states that the posterior measure $\Pi_n(\cdot|X^{(n)})$ assigns probability 1 to any neighborhood of $\boldsymbol{\theta}_0$ as $n \to \infty$ if conditions (a) and (b) of Theorem 2.1 and the KL property hold. Furthermore, if $m \to \infty$, then the following proposition proves that $\Pi_{k_n}(\cdot|X_{[k]})$ are strongly consistent at $\boldsymbol{\theta}_0$ for $k = 1, \ldots, K$ as a straightforward application of Lemma 3.1.

Proposition 3.1 *Under the conditions of Lemma 3.1, if $m \to \infty$, then subset posteriors $\Pi_{k_n}(\cdot|X_{[k]})$ for $k = 1, \ldots, K$ are strongly consistent at $\boldsymbol{\theta}_0$.*

It is clear that the subset posteriors use only $\frac{1}{K}$ fraction of the whole data, so the credible intervals obtained from Π_{k_n}s will be wider than the credible interval obtained from a posterior distribution that uses the whole data; that is, subset posteriors over-estimate the uncertainty in the unknown parameters. Minsker et al. [16] addressed this issue by using a "stochastic approximation (SA) trick." This idea compensates for the data lost due to partitioning by adding $(K-1)$ extra copies of $X_{[k]}$ in each subset for $k = 1, \ldots, K$. The resulting subset posteriors are noisy estimates of the full data posterior, and do not vary systematically from the overall posterior in mean, variance, or shape.

WASP also uses the SA trick for each of the subset posteriors, obtaining subset posteriors that are distributed randomly around the overall posterior. The subset posteriors using SA trick are defined as

$$\Pi^{\mathrm{SA}}_{k_n}(\mathcal{U} \,|\, \underbrace{X_{[k]}, \ldots, X_{[k]}}_{K}) = \frac{\int_{\mathcal{U}} \left(\prod_{i=1}^{m} p(X_{[k]_i} \,|\, \boldsymbol{\theta})\right)^K d\,\Pi_n(\boldsymbol{\theta})}{\int_{\Theta} \left(\prod_{i=1}^{m} p(X_{[k]_i} \,|\, \boldsymbol{\theta})\right)^K d\,\Pi_n(\boldsymbol{\theta})}. \tag{10}$$

for $k = 1, \ldots, K$ and Borel measurable $\mathcal{U} \subset \Theta$. The next proposition states that SA-corrected subset posteriors are also strongly consistent at $\boldsymbol{\theta}_0$.

Proposition 3.2 *Under the conditions of Lemma 3.1, the subset posteriors with stochastic approximation $\Pi^{\mathrm{SA}}_{k_n}(\cdot|X_{[k]})$ for $k = 1, \ldots, K$ are strongly consistent at $\boldsymbol{\theta}_0$.*

This lemma is follows from Lemma 3.1 by noticing that $m \to \infty \implies Km \to \infty$.

The WASP $\overline{\Pi}_n$ is the WB $\overline{\Pi}^{\mathrm{SA}}_{K,\boldsymbol{\lambda}}$ for a given $\boldsymbol{\lambda} \in \Delta_K$ and SA-corrected subset posteriors $\Pi^{\mathrm{SA}}_{1_n:K_n}$ (7). Following (8), $\overline{\Pi}_n$ has the following analytic form

$$\overline{\Pi}_n(\cdot|X^{(n)}) := \overline{\Pi}^{\mathrm{SA}}_{K,\boldsymbol{\lambda}} = \left(\sum_{k=1}^{K} \lambda_k \, \mathbf{T}^1_k\right) \# \, \Pi^{\mathrm{SA}}_{1_n}, \tag{11}$$

where $\Pi^{\mathrm{SA}}_{k_n} = \mathbf{T}^1_k \# \Pi^{\mathrm{SA}}_{1_n}$ is the push-forward of SA-corrected subset posterior $\Pi^{\mathrm{SA}}_{1_n}$ through the measure preserving map $\mathbf{T}^1_k$ for $k = 1, \ldots, K$. The following theorem states the main results of this subsection that $\overline{\Pi}_n$ (11) is strongly consistent at $\boldsymbol{\theta}_0$

Theorem 3.3 *If all the conditions of Lemma 3.1 hold, then $\overline{\Pi}_n$ (11) is strongly consistent at $\boldsymbol{\theta}_0$.*

Intuitively, this theorem states that the mean of noisy approximations of the true posterior in $\mathcal{P}_2(\Theta)$ is also a good approximation of the true posterior.

Theorem 3.3 can be improved substantially to yield rate of contraction using the construction of sieves introduced by [12]. Given a decreasing sequence $(\epsilon_n)_{n\in\mathbb{N}}$, prior sequence $(\Pi_n)_{n\in\mathbb{N}} \in \mathcal{P}_2(\Theta)$, and universal constants $B, b > 0$, there exists sequence of increasing compact parameter spaces $\Theta_n \subset \Theta$ and *polynomially* increasing sequence $(M_n)_{n\in\mathbb{N}}$ such that

(A1) $\Pi_n(\Theta_n^c) \leq B\exp(-bn)$ ("*tight prior*");

(A2) $M_n^2 \exp(-nb) \to 0$ as $n \to \infty$ ("*polynomially increasing M_ns*");

(A3) $\Theta_n = \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta}\|_2 \leq M_n\}$ ("*polynomially bounded parameter space*");

(A4) the packing number satisfies $\log N(\epsilon_n, \Theta_n, \|.\|_2) \leq n\epsilon_n^2$.

Then, the following theorem states that under assumptions (A1) - (A4), $\Pi_n(\cdot|X_{1:n})$ is strongly consistent at $\boldsymbol{\theta}_0$ if the prior satisfies the KL property.

Theorem 3.4 *Let Θ be compact subset of $\mathbb{R}^D$, let $\mathcal{P}_2(\Theta)$ be the Wasserstein space of probability measures parametrized by $\boldsymbol{\theta} \in \Theta$, and let the true measure $\delta_{\boldsymbol{\theta}_0} \in \mathcal{P}_2(\Theta)$. Assume that the sequences $(\Theta_n)_{n\in\mathbb{N}} \subset \Theta$ and $(M_n)_{n\in\mathbb{N}}$ satisfy Assumptions (A1) – (A3) for universal constants B, b and sequences $(\epsilon_n)_{n\in\mathbb{N}}$, $(\Pi_n)_{n\in\mathbb{N}} \in \mathcal{P}_2(\Theta)$ and that Π_n satisfies the KL property $\forall n$ such that $\liminf_{n\to\infty} \Pi_n(\mathcal{K}_\epsilon(\boldsymbol{\theta}_0)) > 0 \ \forall \epsilon > 0$. Then, $\Pi_n(\cdot|X_{1:n})$ is strongly consistent at $\boldsymbol{\theta}_0$.*

Further, it follows from Theorem 3.4 of [16] and [12, Section 5] that if we choose $\epsilon_n \simeq \sqrt{\frac{K \log n}{n}}$ and we take $K = \mathcal{O}(\log n)$ in Theorem 3.4, then we differ from the optimal rate of $n^{-1/2}$ by a factor of only $\log n$.

3.2 Wasserstein barycenter of empirical probability measures based on LP

The analytic form of $\overline{\Pi}_n$ (11) is tractable but for most practical problems the maps $\mathbf{T}^1_k$s are analytically intractable. One solution is to estimate $\overline{\Pi}_n$ from posterior samples of $\Pi^{\mathrm{SA}}_{k_n}$s. Several simulation-based approaches, such as MCMC, SMC, and importance sampling, can be used to generate samples from $\Pi^{\mathrm{SA}}_{k_n}$s for

a large class of models. In this work, we focus on the special case when the reference measure is a uniform distribution on a finite collection of atoms in Θ.

We assume that the subset posteriors are empirical measures and their atoms are simulated from the subset posteriors using a sampler. Let $\widetilde{\Theta} \in \mathbb{R}^{N\times D}$ be a collection of N such posterior samples $\in \Theta$. If $\widetilde{\boldsymbol{\theta}}_n^T$ represents the n-th row of $\widetilde{\Theta}$, then the empirical probability measure corresponding to $\widetilde{\Theta}$ is defined as

$$\pi_N = \sum_{n=1}^{N} \frac{1}{N} \delta_{\widetilde{\boldsymbol{\theta}}_n^T}. \tag{12}$$

Empirical measures are routinely used to approximate posterior measures; however, for the approximation of the joint measure to be accurate, the number of atoms N must be very large. The WASP combines $\left(\Pi_{k_n}^{\text{SA}}\right)_{k=1}^{K}$, which are assumed to be empirical probability measures, by estimating their barycenter $\overline{\Pi}_n$. The estimation procedure is such that $\overline{\Pi}_n$ is estimated as an empirical probability measure.

We first set up the problem of WASP estimation in form of (5). Following (12), assume that posterior samples from the k-th subset posterior $\Pi_{k_n}^{\text{SA}}$ are summarized as the matrix $\widetilde{\Theta}_k \in \mathbb{R}^{N_k\times D}$, where N_k is the number of posterior samples and D is the dimension of the parameter space. The empirical measure corresponding to subset posterior $\Pi_{k_n}^{\text{SA}}$ is defined as

$$\Pi_{k_n}^{\text{SA}} = \sum_{i=1}^{N_k} \frac{1}{N_k} \delta_{\widetilde{\boldsymbol{\theta}}_{k_{n_i}}^T} \equiv \sum_{i=1}^{N_k} b_{k_i} \delta_{\widetilde{\boldsymbol{\theta}}_{k_{n_i}}^T}, \; \mathbf{b}_k = \frac{\mathbf{1}_{N_k}}{N_k}. \tag{13}$$

The empirical measures for $\left(\Pi_{k_n}^{\text{SA}}\right)_{k=1}^{K}$ are defined similarly using $\widetilde{\Theta}_k \in \mathbb{R}^{N_k\times D}$ and $b_k = \frac{\mathbf{1}_{N_k}}{N_k}$ for $k = 1, \ldots, K$. Given $\widetilde{\Theta}_{1:K}$, define the "overall" sample matrix $\widetilde{\Theta}$ by stacking $\widetilde{\Theta}_{1:K}$ along the rows such that $\widetilde{\Theta}^T = \left[\widetilde{\Theta}_1^T \ldots \widetilde{\Theta}_k^T \ldots \widetilde{\Theta}_K^T\right]$. Using $\widetilde{\Theta}$, define WASP as the empirical probability measure

$$\overline{\Pi}_n = \sum_{n=1}^{N} a_n \delta_{\boldsymbol{\theta}_n^T}, \text{ where } \mathbf{a} \in \Delta_N \tag{14}$$

is unknown and $N := \sum_{k=1}^{K} N_k$. The idea here is that the problem of combining subset posteriors to yield a valid probability measure is equivalent to estimating $\mathbf{a}$ in (14) for all the atoms across all subset posteriors. If $\mathbf{a}$ is known and $\mathbf{M}_k := \mathbf{M}_{\widetilde{\Theta}\widetilde{\Theta}_k} \in \mathbb{R}_+^{N\times N_k}$ is defined as

$$\mathbf{M}_k = \text{diag}(\widetilde{\Theta}\widetilde{\Theta}^T)\, \mathbf{1}_{N_k}^T + \mathbf{1}_N \,\text{diag}(\widetilde{\Theta}_k\widetilde{\Theta}_k^T)^T - 2\widetilde{\Theta}\widetilde{\Theta}_k^T$$

following (3), then (5) implies that

$$W_2^2(\overline{\Pi}_n, \Pi_{k_n}^{\text{SA}}) = \min_{\mathbf{T}_k \in \mathcal{T}(\mathbf{a},\mathbf{b}_k)} \langle \mathbf{T}_k, \mathbf{M}_k \rangle, \tag{15}$$

where $\mathcal{T}(\mathbf{a}, \mathbf{b}_k)$ is defined in (4). The optimal transport plan between $\mathbf{a}$ and $\mathbf{b}_k$, $\widehat{\mathbf{T}}_k(\mathbf{a})$, is obtained as

$$\widehat{\mathbf{T}}_k(\mathbf{a}) = \underset{\mathbf{T}_k \in \mathcal{T}(\mathbf{a},\mathbf{b}_k)}{\text{argmin}} \langle \mathbf{T}_k, \mathbf{M}_k \rangle \tag{16}$$

for $k = 1, \ldots, K$. To account for unknown $\mathbf{a}$, an extension of (16) based on (5) and (7) under the assumption that $\lambda_k = 1/K$ yields

$$\widehat{\mathbf{T}}_1, \ldots, \widehat{\mathbf{T}}_K, \widehat{\mathbf{a}} = \underset{\substack{\mathbf{T}_1,\ldots,\mathbf{T}_K,\mathbf{a},\\ \mathbf{a}\in\Delta_N,\\ \mathbf{T}_k\in\mathcal{T}(\mathbf{a},\mathbf{b}_k)\; k=1,\ldots,K}}{\text{argmin}} \sum_{k=1}^{K} \langle \mathbf{T}_k, \mathbf{M}_k \rangle. \tag{17}$$

If we represent

$$\text{vec}(\mathbf{M}) = \text{vec}([\mathbf{M}_1 \ldots, \mathbf{M}_k, \ldots, \mathbf{M}_K]),$$
$$\text{vec}(\mathbf{T}) = \text{vec}([\mathbf{T}_1 \ldots, \mathbf{T}_k, \ldots, \mathbf{T}_K]),$$

and $\mathbf{b}^T = (\mathbf{b}_1^T, \ldots, \mathbf{b}_k^T, \ldots, \mathbf{b}_K^T)$, then (17) reduces to

$$\min_{\binom{\text{vec}(\mathbf{T})}{\mathbf{a}}} \text{vec}(\mathbf{M})^T \text{vec}(\mathbf{T}) + \mathbf{0}_{N\times 1}^T \mathbf{a}$$
$$\text{such that } \mathbf{A}\begin{pmatrix}\text{vec}(\mathbf{T})\\ \mathbf{a}\end{pmatrix} = \mathbf{c}, \; \begin{pmatrix}\text{vec}(\mathbf{T})\\ \mathbf{a}\end{pmatrix} \geq 0, \tag{18}$$

where

$$\mathbf{A} = \begin{pmatrix}\mathbf{0}_{1\times N^2} & \mathbf{1}_{1\times N}\\ \mathbf{F} & -\mathbf{G}\\ \mathbf{H} & \mathbf{0}_{N\times N}\end{pmatrix} \quad \mathbf{c}^T = [1 \quad \mathbf{0}_{1\times KN} \quad \mathbf{b}^T]$$
$$\mathbf{F} = \text{bdiag}\left(\mathbf{1}_{N_1}^T \otimes \mathbf{I}_N, \ldots, \mathbf{1}_{N_k}^T \otimes \mathbf{I}_N, \ldots, \mathbf{1}_{N_K}^T \otimes \mathbf{I}_N\right),$$
$$\mathbf{G} = \mathbf{1}_K \otimes \mathbf{I}_N,$$
$$\mathbf{H} = \text{bdiag}\left(\mathbf{I}_{N_1} \otimes \mathbf{1}_N^T, \ldots, \mathbf{I}_{N_k} \otimes \mathbf{1}_N^T, \ldots, \mathbf{I}_{N_K} \otimes \mathbf{1}_N^T\right), \tag{19}$$

where bdiag denotes a block-diagonal matrix. The optimum $[\text{vec}(\widehat{\mathbf{T}})^T \quad \widehat{\mathbf{a}}^T]^T$ of (18) corresponds to the optimum $\widehat{\mathbf{T}}_1, \ldots, \widehat{\mathbf{T}}_K, \widehat{\mathbf{a}}$ of (17). The constraints (19) of the LP (18) are sparse; Gurobi is used to obtain the solution efficiently.

4 Experiments

In this section we illustrate the computational gains and generality of the WASP framework using simulated and real data analyses.

4.1 Artificial data

We use the simulation example in the `GPML` MATLAB toolbox for demonstrating the performance of WASP for large scale GP regression. Using the function $f(x) = \sin(x) + \sqrt{x} + \epsilon$, we simulated two data sets

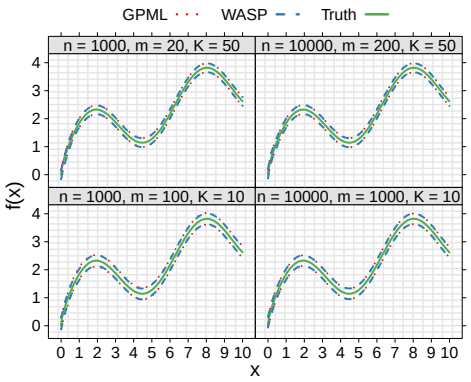

Figure 2: Comparison of GPML and WASP in Gaussian Process (GP) regression. Size of the data set increases across columns and the number of subsets increases from bottom to top. WASP results are in excellent agreement with GPML results while being substantitally faster.

of size 1000 (case 1) and 10000 (case 2) with Gaussian noise of mean 0 and variance 0.04. The `GPML` toolbox was used to obtain the estimate of $\widehat{f}(x)$ across a grid of 1000 xs in both these cases. GPML's performance in fitting GP regressions of the size 1000 was fairly reasonable; however, its performance for exact inference decayed exponentially for data sets of size $\mathcal{O}(10^4)$ and became impractically slow for data sets of size $\mathcal{O}(10^5)$ or larger.

We split the data sets in cases 1 and 2 into 10 and 50 subsets to demonstrate the performance of the WASP framework in massive GP computations. We used subset posteriors $\Pi_{k_n}^{\mathrm{SA}}(\cdot|(x_i, f_i)s)$ to obtain 1000 fs across 1000 xs as samples from these atoms with probabilities equal to their WASP weights. The 95% credible intervals for $\widehat{f}$ are calculated from the 2.5% and 97.5% quantiles of the 1000 posterior draws of fs across 1000 xs. The results of posterior uncertainty quantified by the 95% credible intervals of `GPML` and WASP show an excellent agreement with each other (Figure 2); however, WASP's computations were substantially faster than those of `GPML` because matrix inversions for data subsets of smaller size were stable and fast. On the contrary, `GPML` relied on inverting the matrix of dimensions of order 10^4.

The WASP framework offers an attractive approach for large scale GP regression. Exact inference for GP regression involves matrix inversion of size equal to the data set. This becomes infeasible when the size of the data set reaches $\mathcal{O}(10^4)$. Chalupka et al. [7] compared several low rank matrix approximations to avoid matrix inversion in massive data GP computation. Such approximations can be avoided by using WASP for combining GP regression on data subsets of smaller size for which matrix inversions are stable. Assume that data subset of size m (say < 500) are such that exact inference for GP is feasible due to matrix inversion, then K can be chosen such that $O(Km^3) < O(n^3)$. For all such choices of K and m, it is computationally appealing to use the WASP framework for GP regression over low rank or other approximations for GP regression.

4.2 Real data

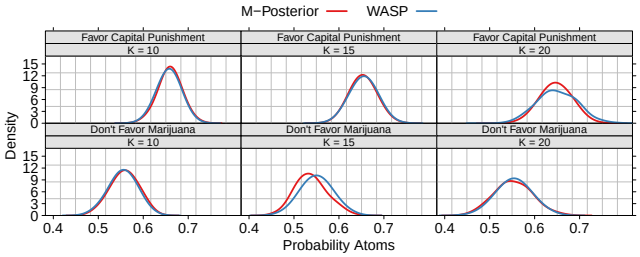

Figure 3: Comparison of M-Posterior and WASP for probabilistic parafac model for marginal probabilities of Mar and Cap responses in GSS data. The bottom row represents marginal probability of not favoring marijuana and the top row represents support for capital punishment. The subset size varies across columns.

We now compare WASP's performance with that of M-Posterior using the General Social Survey (GSS) data set from 2008 - 2010 for about 4100 responders that were used by Minsker et al. [17]. Following their approach, we use a Dirichlet Process mixture of product multinomial distributions, probabilistic parafac (p-parafac), to model multivariate dependence in these data; see [10] for details about the model. The details of the generative model and Gibbs sampler are found in the Appendix D of Minsker et al. [17].

Our interest lies in comparing the final marginals obtained using M-Posterior and WASP for different subset sizes. We varied the size of data subsets as $K = 10, 15$, and 20. For each of these subsets, we modified the original Gibbs sampler for p-parafac using the stochastic approximation trick and obtained 200 posterior draws. These samples were then combined separately using M-Posterior and WASP. In addition, application of M-posterior required specification of the radial basis function kernel for measuring distance between different subset posteriors. Similar to GP regression results, we observe that the M-Posterior and WASP marginals agree very closely with each other across all subset sizes.

The results of the p-parafac model also agrees with our intuition. Americans who do not favor capital punishment are more likely to vote in favor of legalization of marijuana. Both M-Posterior and WASP agree across all subsets and both categories. While Minsker

et al. [17] used M-Posterior arguing for need for robustness of Bayesian methods in surveys, our results show that even if WASP is not robust to outliers, it yields marginals that are close to the M-Posterior.

Discarding the robustness guarantees leads to several advantages of WASP over M-Posterior. The WASP framework does not require a kernel for measuring distances between the subset posteriors. M-Posterior obtains weights from the Weiszfeld algorithm. Since none of these weights are zero, one needs to rely on heuristics such as hard thresholding to truncate small atomic weights to zero for interpretable posterior approximation. WASP does not require such heuristics because the optimum is obtained at extreme points. Furthermore, WASP is obtained by solving a sparse LP that is computationally more efficient than the iterative Weiszfeld algorithm for estimating M-Posterior.

5 Discussion

We have presented the Wasserstein Posterior (WASP) framework as a general approach for scalable Bayesian computations. The assumptions of WASP framework are fairly general that ensure wide applicability. Specifically, it requires that computations with data subsets are feasible so that any existing sampler can be used to obtain atomic approximations of the subset posteriors. These atomic subset posteriors are then combined using the Wasserstein barycenter (WB). Being a natural generalization of the Euclidean barycenter to the space of probability measures, the WB is an ideal choice for combining subset posteriors that are noisy approximations of the true posterior. We exploited the structure of the problem to estimate the WB by efficiently solving a sparse LP.

The idea of solving LP for efficient Bayesian inference can be extended in many directions. We used off-the-shelf solver for our experiments and were able to solve LPs of the order 10^6 by exploiting sparsity of the WASP objective; however, solving LPs of higher dimensions becomes problematic. We plan to use recent developments in primal-dual methods to improve the computational efficiency of the LP solver. The optimal transport plan, which was not used in the current approach, could be used for designing samplers when the number of parameters is large and ordinary samplers fail to converge to their stationary distribution. While we have illustrated WASP's applications in the context of scalable Bayesian computations, its reliance on the Wasserstein space and metric could be used for obtaining barycenters in other spaces, such as shape spaces.

6 Acknowledgment

DBD and SS were partially supported by grant R01-ES-017436 from the National Institute of Environmental Health Sciences of the National Institutes of Health. SS was also supported by the National Science Foundation under Grant DMS-1127914 to SAMSI. VC and QTD were supported in part by the European Commission under grants MIRG-268398 and ERC Future Proof and by the Swiss Science Foundation under grants SNF 200021-132548, SNF 200021-146750, and SNF CRSII2-147633.

References

[1] Agarwal, A. and J. C. Duchi (2012). Distributed delayed stochastic optimization. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 5451–5452. IEEE.

[2] Agueh, M. and G. Carlier (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis 43*(2), 904–924.

[3] Ahn, S., A. Korattikara, and M. Welling (2012). Bayesian posterior sampling via stochastic gradient fisher scoring. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*.

[4] Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

[5] Broderick, T., N. Boyd, A. Wibisono, A. C. Wilson, and M. Jordan (2013). Streaming variational bayes. In *Advances in Neural Information Processing Systems*, pp. 1727–1735.

[6] Cevher, V., S. Becker, and M. Schmidt (2014). Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *Signal Processing Magazine, IEEE 31*(5), 32–43.

[7] Chalupka, K., C. K. Williams, and I. Murray (2012). A framework for evaluating approximation methods for gaussian process regression. *arXiv preprint arXiv:1205.6326*.

[8] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300.

[9] Cuturi, M. and A. Doucet (2014). Fast computation of Wasserstein barycenters. In *Proceedings*

of the 31st International Conference on Machine Learning, JMLR W&CP, Volume 32.

[10] Dunson, D. B. and C. Xing (2009). Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association 104*(487), 1042–1051.

[11] Gangbo, W. and A. Swiech (1998). Optimal maps for the multidimensional Monge-Kantorovich problem. *Communications on Pure and Applied Mathematics 51*(1), 23–45.

[12] Ghosal, S., J. K. Ghosh, and A. W. Van Der Vaart (2000). Convergence rates of posterior distributions. *Annals of Statistics 28*(2), 500–531.

[13] Gurobi Optimization Inc. (2014). *Gurobi Optimizer Reference Manual Version 6.0.0*.

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

[15] Korattikara, A., Y. Chen, and M. Welling (2013). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. *arXiv preprint arXiv:1304.5299*.

[16] Minsker, S., S. Srivastava, L. Lin, and D. Dunson (2014a). Scalable and robust bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1656–1664.

[17] Minsker, S., S. Srivastava, L. Lin, and D. B. Dunson (2014b). Robust and scalable bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.

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

[19] Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2013). Bayes and big data: the consensus Monte Carlo algorithm.

[20] Smola, A. J. and S. Narayanamurthy (2010). An Architecture for Parallel Topic Models. In *Very Large Databases (VLDB)*.

[21] Villani, C. (2008). *Optimal transport: old and new*. Springer.

[22] Wang, C., J. W. Paisley, and D. M. Blei (2011). Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, pp. 752–760.

[23] Wang, X. and D. B. Dunson (2013). Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.

[24] Welling, M. and Y. W. Teh (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688.