# Scalable and Robust Bayesian Inference via the Median Posterior

**Stanislav Minsker**[1] SMINSKER@MATH.DUKE.EDU
**Sanvesh Srivastava**[2,3] SS602@STAT.DUKE.EDU
**Lizhen Lin**[2] LIZHEN@STAT.DUKE.EDU
**David B. Dunson**[2] DUNSON@STAT.DUKE.EDU
Departments of Mathematics[1] and Statistical Science[2], Duke University, Durham, NC 27708
Statistical and Applied Mathematical Sciences Institute[3], 19 T.W. Alexander Dr, Research Triangle Park, NC 27709

## Abstract

Many Bayesian learning methods for massive data benefit from working with small subsets of observations. In particular, significant progress has been made in scalable Bayesian learning via stochastic approximation. However, Bayesian learning methods in distributed computing environments are often problem- or distribution-specific and use ad hoc techniques. We propose a novel general approach to Bayesian inference that is scalable and robust to corruption in the data. Our technique is based on the idea of splitting the data into several non-overlapping subgroups, evaluating the posterior distribution given each independent subgroup, and then combining the results. Our main contribution is the proposed aggregation step which is based on finding the geometric median of subset posterior distributions. Presented theoretical and numerical results confirm the advantages of our approach.

## 1. Introduction

Massive data often require computer clusters for storage and processing. In such cases, each machine in the cluster can only access a subset of data at a given point. Most learning algorithms designed for distributed computing share a common feature: they efficiently use the data subset available to a single machine and combine the "local" results for "global" learning, while minimizing communication among cluster machines (Smola & Narayanamurthy, 2010). A wide variety of optimization-based approaches are available for distributed learning (Boyd et al., 2011); however, the number of similar Bayesian methods is limited. One of the reasons for this limitation is related to Markov chain Monte Carlo (MCMC) - one of the key techniques for approximating the posterior distribution of parameters in Bayesian models. While there are many efficient MCMC techniques for sampling from posterior distributions based on small subsets of the data (called subset posteriors in the sequel), there is no widely accepted and theoretically justified approach for combining the subset posteriors into a single distribution for improved performance. To this end, we propose a new general solution for this problem based on evaluation of *the geometric median* of a collection of subset posterior distributions. The resulting measure is called the *M-posterior* ("median posterior").

Modern approaches to scalable Bayesian learning in a distributed setting fall into three major categories. Methods in the first category independently evaluate the likelihood for each data subset across multiple machines and return the likelihoods to a "master" machine, where they are appropriately combined with the prior using conditional independence assumptions of the probabilistic model. These two steps are repeated at every MCMC iteration (Smola & Narayanamurthy, 2010; Agarwal & Duchi, 2012). This approach is problem-specific and involves extensive communication among machines. Methods from the second category use stochastic approximation (SA) and successively learn "noisy" approximations to the full posterior distribution using data in small mini-batches. The accuracy of SA increases as it uses more observations. One subgroup of this category uses sampling-based methods to explore the posterior distribution through a modified Hamiltonian or Langevin Dynamics (Welling & Teh, 2011; Ahn et al., 2012; Korattikara et al., 2013). Unfortunately, these methods fail to accommodate discrete-valued parameters and multimodality. The other subgroup uses deterministic variational approximations and learns the parameters of the approximated posterior through an optimization-based method (Hoffman et al., 2013; Broderick et al., 2013). Although these approaches often have excellent predictive performance, it is well known that they tend to substan-

*Proceedings of the $31^{st}$ International Conference on Machine Learning*, Beijing, China, 2014. JMLR: W&CP volume 32. 

tially underestimate posterior uncertainty and lack theoretical guarantees.

Our approach instead falls in a third class of methods, which avoid extensive communication among machines by running independent MCMC chains for each data subset and obtaining draws from subset posteriors. These subset posteriors can be combined in a variety of ways. Some of the methods simply average draws from each subset (Scott et al., 2013), other use an approximation to the full posterior distribution based on kernel density estimators (Neiswanger et al., 2013), or the Weierstrass transform (Wang & Dunson, 2013). Unlike the method proposed in this work, none of the aforementioned algorithms are provably robust to the presence of outliers, moreover, they have limitations related to the dimension of the parameter.

We propose a different approximation to the full data posterior for each subset, with MCMC used to generate samples from these "noisy" subset posteriors in parallel. As a "de-noising step" that also induces robustness to outliers, we then calculate the *geometric median* of the subset posteriors, referred to as the M-posterior. By embedding the subset posteriors in a Reproducing Kernel Hilbert Space (RKHS), we facilitate computation of distances, allowing Weiszfeld's algorithm to be used to approximate the geometric median (Beck & Sabach, 2013). The M-posterior admits strong theoretical guarantees, is provably resistant to the presence outliers, efficiently uses all of the available observations, and is well-suited for distributed Bayesian learning. Our work was inspired by multivariate median-based techniques for robust point estimation developed by Minsker (2013) and Hsu & Sabato (2013) (see also Alon et al., 1996; Lerasle & Oliveira, 2011; Nemirovski & Yudin, 1983, where similar ideas were applied in different frameworks).

## 2. Preliminaries

We first describe the notion of geometric median and the method for calculating distance between probability distributions via embedding them in a Hilbert space. This will be followed by the formal description of our method and corresponding theoretical guarantees.

### 2.1. Notation

In what follows, $\|\cdot\|_2$ denotes the standard Euclidean distance in $\mathbb{R}^p$ and $\langle\cdot,\cdot\rangle_{\mathbb{R}^p}$ - the associated dot product. Given a totally bounded metric space $(\mathbb{Y}, d)$, the packing number $M(\varepsilon, \mathbb{Y}, d)$ is the maximal number $N$ such that there exist $N$ disjoint $d$-balls $B_1, \ldots, B_N$ of radius $\varepsilon$ contained in $\mathbb{Y}$, i.e., $\bigcup_{j=1}^{N} B_j \subseteq \mathbb{Y}$. Given a metric space $(\mathbb{Y}, d)$ and $y \in \mathbb{Y}$, $\delta_y$ denotes the Dirac measure concentrated at $y$. In other words, for any Borel-measurable $B$, $\delta_y(B) = I\{y \in B\}$, where $I\{\cdot\}$ is the indicator function. Other objects and definitions are introduced in the course of exposition when such necessity arises.

### 2.2. Geometric median

The goal of this section is to introduce the *geometric median*, a generalization of the univariate median to higher dimensions. Let $\mu$ be a Borel probability measure on a normed space $(\mathbb{Y}, \|\cdot\|)$. The *geometric median* $x_*$ of $\mu$ is defined as $x_* := \operatorname{argmin}_{y\in\mathbb{Y}} \int_{\mathbb{Y}} (\|y - x\| - \|x\|)\, \mu(dx)$. We use a special case of this definition and assume that $\mu$ is a uniform distribution on a collection of $m$ atoms $x_1, \ldots, x_m \in \mathbb{Y}$ (which will later correspond to $m$ subset posteriors identified with points in a certain space), so that

$$x_* = \operatorname{med}_g(x_1, \ldots, x_m) := \operatorname*{argmin}_{y\in\mathbb{Y}} \sum_{j=1}^{m} \|y - x_j\|. \quad (1)$$

Geometric median exists under rather general conditions, in particular, when $\mathbb{Y}$ is a Hilbert space. Moreover, in this case $x_* \in \operatorname{co}(x_1, \ldots, x_m)$ – the convex hull of $x_1, \ldots, x_m$ (in other words, there exist nonnegative $\alpha_j$, $j = 1 \ldots m$, $\sum_j \alpha_j = 1$ such that $x_* = \sum_{j=1}^m \alpha_j x_j$).

An important property of the geometric median states that it transforms a collection of independent and "weakly concentrated" estimators into a single estimator with significantly stronger concentration properties. Given $q, \alpha$ such that $0 < q < \alpha < 1/2$, define

$$\psi(\alpha, q) := (1 - \alpha) \log \frac{1 - \alpha}{1 - q} + \alpha \log \frac{\alpha}{q}. \quad (2)$$

The following result follows from Theorem 3.1 of Minsker (2013).

**Theorem 2.1.** *Assume that $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ is a Hilbert space and $\theta_0 \in \mathbb{H}$. Let $\hat\theta_1, \ldots, \hat\theta_m \in \mathbb{H}$ be a collection of independent random $\mathbb{H}$ - valued elements. Let the constants $\alpha, q, \nu$ be such that $0 < q < \alpha < 1/2$, and $0 \le \nu < \frac{\alpha - q}{1 - q}$. Suppose $\varepsilon > 0$ is such that for all $j$, $1 \le j \le \lfloor (1 - \nu)m \rfloor + 1$,*

$$\Pr\left(\|\hat\theta_j - \theta_0\| > \varepsilon\right) \le q. \quad (3)$$

*Let $\hat\theta_* = \operatorname{med}_g(\hat\theta_1, \ldots, \hat\theta_m)$ be the geometric median of $\{\hat\theta_1, \ldots, \hat\theta_m\}$. Then*

$$\Pr\left(\|\hat\theta_* - \theta_0\| > C_\alpha \varepsilon\right) \le \left[e^{(1-\nu)\psi\left(\frac{\alpha-\nu}{1-\nu}, q\right)}\right]^{-m},$$

*where $C_\alpha = (1 - \alpha)\sqrt{\frac{1}{1 - 2\alpha}}$.*

Theorem 2.1 implies that the concentration of the geometric median around the "true" parameter value improves geometrically fast with respect to the number $m$ of independent estimators, while the estimation rate is preserved. Parameter $\nu$ allows to take the corrupted observations into account: if the data contain not more than $\lfloor \nu m \rfloor$ outliers of arbitrary nature, then at most $\lfloor \nu m \rfloor$ estimators amongst $\{\theta_1, \dots, \theta_m\}$ can be affected. Parameter $\alpha$ should be viewed as a fixed quantity and can be set to $\alpha = 1/3$ for the rest of the paper.

### 2.3. RKHS and distances between probability measures

The goal of this section is to introduce a special family of distances between probability measures which provide a structure necessary to evaluate the geometric median in the space of measures. Since our goal is to develop computationally efficient techniques, we consider distances that admit accurate numerical approximation.

Assume that $(\mathbb{X}, \rho)$ is a separable metric space, and let $\mathcal{F} = \{f : \mathbb{X} \mapsto \mathbb{R}\}$ be a collection of real-valued functions. Given two Borel probability measures $P, Q$ on $\mathbb{X}$, define

$$\|P - Q\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \int_{\mathbb{X}} f(x) d(P - Q)(x) \right|. \quad (4)$$

An important special case arises when $\mathcal{F}$ is a unit ball in a RKHS $(\mathbb{H}, \langle \cdot, \cdot \rangle)$ with a reproducing kernel $k : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}^P$ so that [1]

$$\mathcal{F} = \mathcal{F}_k := \{f : \mathbb{X} \mapsto \mathbb{R},\ f \in \mathbb{H},\ \|f\|_{\mathbb{H}} := \sqrt{\langle f, f \rangle} \le 1\}. \quad (5)$$

Let $\mathcal{P}_k := \{P \text{ is a prob. measure}, \int_{\mathbb{X}} \sqrt{k(x,x)} dP(x) < \infty\}$, and assume that $P, Q \in \mathcal{P}_k$. If follows from Theorem 1 in (Sriperumbudur et al., 2010) that the corresponding distance between measures $P$ and $Q$ takes the form

$$\|P - Q\|_{\mathcal{F}_k} = \left\| \int_{\mathbb{X}} k(x, \cdot) d(P - Q)(x) \right\|_{\mathbb{H}}. \quad (6)$$

Note that when $P$ and $Q$ are discrete measures (say, $P = \sum_{j=1}^{N_1} \beta_j \delta_{z_j}$ and $Q = \sum_{j=1}^{N_2} \gamma_j \delta_{y_j}$), then

$$\|P - Q\|^2_{\mathcal{F}_k} = \sum_{i,j=1}^{N_1} \beta_i \beta_j k(z_i, z_j) + \sum_{i,j=1}^{N_2} \gamma_i \gamma_j k(y_i, y_j) - 2 \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \beta_i \gamma_j k(z_i, y_j). \quad (7)$$

The mapping $P \mapsto \int_{\mathbb{X}} k(x, \cdot) dP(x)$ is thus an embedding of $\mathcal{P}_k$ into the Hilbert space $\mathbb{H}$ which can be seen as an application of the "kernel trick" in our setting. The Hilbert space structure allows to use fast numerical methods to approximate the geometric median.

In this work, we will only consider *characteristic* kernels, which means that $\|P - Q\|_{\mathcal{F}_k} = 0$ if and only if $P = Q$. It follows from Theorem 7 in (Sriperumbudur et al., 2010) that a sufficient condition for $k$ to be characteristic is *strict positive definiteness*: we say that $k$ is *strictly positive definite* if it is measurable, bounded, and for all non-zero signed Borel measures $\mu$, $\iint_{\mathbb{X} \times \mathbb{X}} k(x,y) d\mu(x) d\mu(y) > 0$. When $\mathbb{X} = \mathbb{R}^p$, a simple sufficient criterion for the kernel $k$ to be characteristic follows from Theorem 9 in (Sriperumbudur et al., 2010):

**Proposition 2.2.** *Let $\mathbb{X} = \mathbb{R}^p$, $p \ge 1$. Assume that $k(x,y) = \phi(x - y)$ for some bounded, continuous, integrable, positive-definite function $\phi : \mathbb{R}^p \mapsto \mathbb{R}$.*

1. *Let $\widehat{\phi}$ be the Fourier transform of $\phi$. If $|\widehat{\phi}(x)| > 0$ for all $x \in \mathbb{R}^p$, then $k$ is characteristic;*
2. *If $\phi$ is compactly supported, then $k$ is characteristic.*

**Remark 2.3.**

*(a) It is important to mention that in practical applications, we (almost) always deal with empirical measures based on a collection of independent samples from the posterior. A natural question is the following: if $P$ and $Q$ are probability distributions on $\mathbb{R}^p$ and $P_n$, $Q_m$ are their empirical versions, what is the size of the error $e_{m,n} := \left| \|P - Q\|_{\mathcal{F}_k} - \|P_m - Q_n\|_{\mathcal{F}_k} \right|$? A useful fact is that $e_{m,n}$ often does not depend on $p$: under weak assumptions on $k$, $e_{n,m}$ has an upper bound of order $m^{-1/2} + n^{-1/2}$ (see corollary 12 in Sriperumbudur et al., 2009).*

*(b) Choice of the kernel determines the "richness" of the space $\mathbb{H}$ and, hence, the relative strength of induced norm $\|\cdot\|_{\mathcal{F}_k}$. For example, the well-known family of Matérn kernels leads to Sobolev spaces (Rieger & Zwicknagl, 2009). Gaussian kernels often yields good results in applications.*

Finally, we recall the definition of the well-known Hellinger distance. Assume that $P$ and $Q$ are probability measures on $\mathbb{R}^D$ which are absolutely continuous with respect to Lebesgue measure with densities $p$ and $q$ respectively. Then

$$h(P, Q) := \sqrt{\frac{1}{2} \int_{\mathbb{R}^D} \left( \sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx}$$

is the Hellinger distance between $P$ and $Q$.

## 3. Contributions and main results

### 3.1. Construction of "robust posterior distribution"

Let $\{P_\theta,\ \theta \in \Theta\}$ be a family of probability distributions over $\mathbb{R}^D$ indexed by $\Theta$. Suppose that for all $\theta \in \Theta$, $P_\theta$ has

[1] We will say that $k$ is a kernel if it is a symmetric, positive definite function; it is a reproducing kernel for $\mathbb{H}$ and such that for any $f \in \mathbb{H}$ and $x \in \mathbb{X}$, $\langle f, k(\cdot, x) \rangle_{\mathbb{H}} = f(x)$ (see Aronszajn, 1950, for details).

a Radon-Nikodym derivative $p_\theta(\cdot) = \frac{dP_\theta}{dx}(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^D$. In what follows, we equip $\Theta$ with a "Hellinger metric"

$$\rho(\theta_1, \theta_2) := h(P_{\theta_1}, P_{\theta_2}), \tag{8}$$

and assume that the metric space $(\Theta, \rho)$ is separable.

Let $X_1, \dots, X_n$ be i.i.d. $\mathbb{R}^D$-valued random vectors defined on some probability space $(\Omega, \mathcal{B}, \Pr)$ with unknown distribution $P_0 := P_{\theta_0}$ for $\theta_0 \in \Theta$. A usual way to "estimate" $P_0$ in Bayesian statistics consists in defining a *prior* distribution $\Pi$ over $\Theta$ (equipped with the Borel $\sigma$-algebra induced by $\rho$), so that $\Pi(\Theta) = 1$. The *posterior* distribution given the observations $\mathcal{X}_n := \{X_1, \dots, X_n\}$ is a random probability measure on $\Theta$ defined by

$$\Pi_n(B|\mathcal{X}_n) := \int_B \frac{\prod_{i=1}^n p_\theta(X_i)}{\int_\Theta \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta)} d\Pi(\theta)$$

for all Borel measurable sets $B \subseteq \Theta$. It is known (Ghosal et al., 2000) that under rather general assumptions the posterior distribution $\Pi_n$ "contracts" towards $\theta_0$, meaning that

$$\Pi_n(\theta \in \Theta : \rho(\theta, \theta_0) \geq \varepsilon_n | \mathcal{X}_n) \to 0$$

in probability as $n \to \infty$ for a suitable sequence $\varepsilon_n \to 0$. One of the question that we address can be formulated as follows: what happens if some observations in $\mathcal{X}_n$ are corrupted, e.g., if $\mathcal{X}_n$ contains outliers of arbitrary nature and magnitude? In this case, the usual posterior distribution might concentrate "far" from the true value $\theta_0$, depending on the amount of corruption in the sample. We show that it is possible to modify existing inference procedures via a simple and computationally efficient scheme that improves robustness of the underlying method.

We proceed with the general description of the proposed algorithm. Let $1 \leq m \leq n/2$ be an integer, and divide the sample $\mathcal{X}_n$ into $m$ disjoint groups $G_j,\ j = 1 \dots m$ of size $|G_j| \geq \lfloor n/m \rfloor$ each: $\mathcal{X}_n = \bigcup_{j=1}^m G_j$, $G_i \cap G_l = \emptyset$ for $i \neq j$. A typically choice of $m$ is $m \simeq \log n$, so that the groups $G_j$ are sufficiently large (however, other choices are possible as well depending on concrete practical scenario).

Let $\Pi$ be a prior distribution over $\Theta$, and let $\left\{\Pi_n^{(j)}(\cdot) := \Pi_n(\cdot|G_j),\ j = 1 \dots m\right\}$ be the family of posterior distributions depending on disjoint subgroups $G_j,\ j = 1 \dots m$:

$$\Pi_n(B|G_j) := \int_B \frac{\prod_{i \in G_j} p_\theta(X_i)}{\int_\Theta \prod_{i \in G_j} p_\theta(X_i) d\Pi(\theta)} d\Pi(\theta).$$

Define the *"median posterior" (or M-posterior)* $\hat{\Pi}_{n,g}$ as

$$\hat{\Pi}_{n,g} := \mathrm{med}_g(\Pi_n^{(1)}, \dots, \Pi_n^{(m)}), \tag{9}$$

where the median $\mathrm{med}_g(\cdot)$ is evaluated with respect to $\|\cdot\|_{\mathcal{F}_k}$ introduced in (1) and (5). Note that $\hat{\Pi}_{n,g}$ is always a probability measure: due to the aforementioned properties of a geometric median, there exists $\alpha_1 \geq 0, \dots, \alpha_m \geq 0$, $\sum_{j=1}^m \alpha_j = 1$ such that $\hat{\Pi}_{n,g} = \sum_{j=1}^m \alpha_j \Pi_n^{(j)}$. In practice, small weights $\alpha_j$ are set to 0 for improved performance; see Algorithm 2 for details of implementation.

While $\hat{\Pi}_{n,g}$ possesses several nice properties (such as robustness to outliers), in practice it often overestimates the uncertainty about $\theta_0$, especially when the number of groups $m$ is large. To overcome this difficulty, we suggest a modification of our approach where the random measures $\Pi_n^{(j)}$ (subset posteriors) are replaced by the *stochastic approximations* $\Pi_{n,m}(\cdot|G_j),\ j = 1 \dots m$ of the full posterior distribution. To this end, define the stochastic approximation to the full posterior based on the subsample $G_j$ as

$$\Pi_{n,m}(B|G_j) := \int_B \frac{\left(\prod_{i \in G_j} p_\theta(X_i)\right)^{\lfloor n/|G_j| \rfloor} d\Pi(\theta)}{\int_\Theta \left(\prod_{i \in G_j} p_\theta(X_i)\right)^{\lfloor n/|G_j| \rfloor} d\Pi(\theta)}. \tag{10}$$

In other words, $\Pi_{n,k}(\cdot|G_j)$ is obtained as a posterior distribution given that each data point from $G_j$ is observed $\lfloor n/|G_j| \rfloor$ times. Similarly to $\hat{\Pi}_{n,g}$, we set

$$\hat{\Pi}_{n,g}^{\mathrm{st}} := \mathrm{med}_g(\Pi_{n,m}(\cdot|G_1), \dots, \Pi_{n,m}(\cdot|G_m)). \tag{11}$$

While each of $\Pi_{n,k}(\cdot|G_j)$ might be "unstable", the geometric median $\hat{\Pi}_{n,g}^{\mathrm{st}}$ of these random measures improves stability and yields smaller credible sets with good coverage properties. Practical performance of $\hat{\Pi}_{n,g}^{\mathrm{st}}$ is often superior as compared to $\hat{\Pi}_{n,g}$ in our experiments. In all numerical simulations below, we evaluate $\hat{\Pi}_{n,g}^{\mathrm{st}}$ unless noted otherwise.

### 3.2. Convergence of posterior distribution and applications to robust Bayesian inference

Let $k$ be a characteristic kernel defined on $\Theta \times \Theta$; $k$ defines a metric on $\Theta$

$$\begin{aligned} \rho_k(\theta_1, \theta_2) &:= \|k(\cdot, \theta_1) - k(\cdot, \theta_2)\|_{\mathbb{H}} \\ &= \Big(k(\theta_1, \theta_1) + k(\theta_2, \theta_2) - 2k(\theta_1, \theta_2)\Big)^{1/2}, \end{aligned} \tag{12}$$

where $\mathbb{H}$ is the RKHS associated to kernel $k$. We will assume that $(\Theta, \rho_k)$ is separable.

**Assumption 3.1.** *Let $h(P_{\theta_1}, P_{\theta_2})$ be the Hellinger distance between $P_{\theta_1}$ and $P_{\theta_2}$. Assume there exist positive constants $\gamma$ and $\tilde{C}$ such that for all $\theta_1, \theta_2 \in \Theta$,*

$$h(P_{\theta_1}, P_{\theta_2}) \geq \tilde{C} \rho_k^\gamma(\theta_1, \theta_2).$$

**Example 3.2.** *Let $\{P_\theta,\ \theta \in \Theta \subseteq \mathbb{R}^p\}$ be the exponential family*

$$\frac{dP_\theta}{dx}(x) := p_\theta(x) = \exp\Big(\langle T(x), \theta\rangle_{\mathbb{R}^p} - G(\theta) + q(x)\Big),$$

*where $\langle\cdot,\cdot\rangle_{\mathbb{R}^p}$ is the standard Euclidean dot product. Then the Hellinger distance can be expressed as $h^2(P_{\theta_1}, P_{\theta_2}) = 1 - \exp\Big(-\frac{1}{2}\Big(G(\theta_1)+G(\theta_2)-2G\Big(\frac{\theta_1+\theta_2}{2}\Big)\Big)\Big)$ (Nielsen & Garcia, 2011). If $G(\theta)$ is convex and its Hessian $D^2G(\theta)$ satisfies $D^2G(\theta) \succeq A$ uniformly for all $\theta \in \Theta$ and some symmetric positive definite operator $A : \mathbb{R}^p \mapsto \mathbb{R}^p$ , then*

$$h^2(P_{\theta_1}, P_{\theta_2}) \geq 1 - \exp\Big(-\frac{1}{8}(\theta_1-\theta_2)^T A(\theta_1-\theta_2)\Big),$$

*hence assumption 3.1 holds with $C_k = \frac{1}{\sqrt{2}}$ and $\gamma = 1$ for*

$$k(\theta_1,\theta_2) := \exp\left(-\frac{1}{8}(\theta_1-\theta_2)^T A(\theta_1-\theta_2)\right).$$

*In particular, it implies that for the family $\{P_\theta = N(\theta,\Sigma),\ \theta \in \mathbb{R}^D\}$ with $\Sigma \succ 0$ and the kernel*

$$k(\theta_1,\theta_2) := \exp\left(-\frac{1}{8}(\theta_1-\theta_2)^T\Sigma^{-1}(\theta_1-\theta_2)\right),$$

*assumption 3.1 holds with $C_k = \frac{1}{\sqrt{2}}$ and $\gamma = 1$ (moreover, it holds with equality rather than inequality).*

Assume that $\Theta \subset \mathbb{R}^p$ is compact, and let $k(\cdot,\cdot)$ be a kernel defined on $\mathbb{R}^p \times \mathbb{R}^p$. Suppose that $k$ satisfies conditions of proposition 2.2 (in particular, $k$ is characteristic). Recall that by Bochner's theorem, there exists a finite nonnegative Borel measure $v$ such that $k(\theta) = \int\limits_{\mathbb{R}^p} e^{i\langle x,\theta\rangle} dv(x)$.

**Proposition 3.3.** *Assume that $\int\limits_{\mathbb{R}^p} \|x\|_2^2 dv(x) < \infty$ and for all $\theta_1, \theta_2 \in \Theta$ and some $\gamma > 0$,*

$$h(P_{\theta_1}, P_{\theta_2}) \geq c(\Theta)\|\theta_1 - \theta_2\|_2^\gamma. \tag{13}$$

*Then assumption 3.1 holds with $\gamma$ as above and $\tilde{C} = \tilde{C}(k, c(\Theta), \gamma)$.*

Let $\delta_0 := \delta_{\theta_0}$ be the Dirac measure concentrated at $\theta_0 \in \Theta$ corresponding to the "true" distribution $P_0$.

**Theorem 3.4.** *Let $\mathcal{X}_l = \{X_1, \ldots, X_l\}$ be an i.i.d. sample from $P_0$. Assume that $\varepsilon_l > 0$ and $\Theta_l \subset \Theta$ are such that for a universal constant $K > 0$ and some constant $C > 0$*

1) $\log M(\varepsilon_l, \Theta_l, \rho) \leq l\varepsilon_l^2$,
2) $\Pi(\Theta \setminus \Theta_l) \leq \exp(-l\varepsilon_l^2(C+4))$,
3) $\Pi\left(\theta :\ -P_0\left(\log\frac{p_\theta}{p_0}\right) \leq \varepsilon_l^2,\ P_0\left(\log\frac{p_\theta}{p_0}\right)^2 \leq \varepsilon_l^2\right) \geq \exp(-l\varepsilon_l^2 C)$,
4) $e^{-Kl\varepsilon_l^2/2} \leq \varepsilon_l$.

*Moreover, let assumption 3.1 be satisfied. Then there exists a sufficiently large $M = M(C, K, \tilde{C}) > 0$*

$$\Pr\left(\|\delta_0 - \Pi_l(\cdot|\mathcal{X}_l)\|_{\mathcal{F}_k} \geq M\varepsilon_l^{1/\gamma}\right) \leq \frac{1}{Cl\varepsilon_l^2} + e^{-Kl\varepsilon_l^2/2}. \tag{14}$$

Note that the right-hand side in (14) may decay very slowly with $l$. This is where the properties of the geometric median become useful. Combination of Theorems 3.4 and 2.1 yields the following inequality for $\hat{\Pi}_{n,g}$ which is our main theoretical result.

**Corollary 3.5.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample from $P_0$, and assume that $\hat{\Pi}_{n,g}$ is defined with respect to the $\|\cdot\|_{\mathcal{F}_k}$ as in (9) above. Let $l := \lfloor n/m \rfloor$. Assume that conditions of Theorem 3.4 hold, and, moreover, $\varepsilon_l$ is such that $q := \frac{1}{Cl\varepsilon_l^2} + 4e^{-Kl\varepsilon_l^2/2} < \frac{1}{2}$. Let $\alpha$ be such that $q < \alpha < 1/2$. Then*

$$\Pr\left(\left\|\delta_0 - \hat{\Pi}_{n,g}\right\|_{\mathcal{F}_k} \geq C_\alpha M\varepsilon_l^{1/\gamma}\right) \leq \left[e^{\psi(\alpha,q)}\right]^{-m}, \tag{15}$$

*where $C_\alpha = (1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$ and $M$ is as in Theorem 3.4.*

The case when the sample $\mathcal{X}_n$ contains $\lfloor \nu m \rfloor$ outliers of arbitrary nature can be handled similarly. This more general bound is readily implied by Theorem 2.1.

For many parametric models (see Section 5 in Ghosal et al. (2000)), (15) holds with $\varepsilon_l \simeq \tau\sqrt{\frac{m\log(n/m)}{n}}$ for $\tau$ small enough. If $m \simeq \log n$ (which is a typical scenario), then $\varepsilon_l \simeq \sqrt{\frac{\log^2 n}{n}}$. At the same time, if we use the "full posterior" distribution $\Pi_n(\cdot|\mathcal{X}_n)$ (which corresponds to $m = 1$), conclusion of Theorem 3.4 is that $\Pr\left(\|\delta_0 - \Pi_n(\cdot|\mathcal{X}_n)\|_{\mathcal{F}_k} \geq M\left(\frac{\log n}{n}\right)^{1/(2\gamma)}\right) \lesssim \log^{-1} n$, while Corollary 3.5 yields a much stronger bound for $\hat{\Pi}_{n,g}$: $\Pr\left(\left\|\delta_0 - \hat{\Pi}_{n,g}\right\|_{\mathcal{F}_k} \geq C_\alpha M\left(\frac{\log^2 n}{n}\right)^{1/(2\gamma)}\right) \leq r_n^{-\log n}$ for some $1 > r_n \to 0$.

Theoretical guarantees for $\hat{\Pi}^{\mathrm{St}}_{n,g}$, the median of "stochastic approximations" defined in (11), are very similar to results of Corollary 3.5, but we omit exact formulations here due to the space constraints.

## 4. Numerical Experiments

In this section, we describe a method based on Weiszfeld's algorithm for implementing the M-posterior, and compare the performance of M-posterior with the usual posterior in the tasks involving simulated and real data sets.

We start with a short remark discussing the improvement in computational time complexity achieved by M-posterior.

**Algorithm 1** Evaluating the geometric median of probability distributions via Weiszfeld's algorithm

**Input:**
1. Discrete measures $Q_1, \ldots, Q_m$;
2. The kernel $k(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$;
3. Threshold $\varepsilon > 0$;

**Initialize:**
1. Set $w_j^{(0)} := \frac{1}{m}, j = 1 \ldots m$;
2. Set $Q_*^{(0)} := \frac{1}{m} \sum_{j=1}^{m} Q_j$;

**repeat**

Starting from $t = 0$, for each $j = 1, \ldots, m$:

1. Update $w_j^{(t+1)} = \frac{\|Q_*^{(t)} - Q_j\|_{\mathcal{F}_k}^{-1}}{\sum_{i=1}^{m} \|Q_*^{(t)} - Q_i\|_{\mathcal{F}_k}^{-1}}$; (apply (7) to evaluate $\|Q_*^{(t)} - Q_i\|_{\mathcal{F}_k}$);
2. Update $Q_*^{(t+1)} = \sum_{j=1}^{m} w_j^{(t+1)} Q_j$;

**until** $\|Q_*^{(t+1)} - Q_*^{(t)}\|_{\mathcal{F}_k} \leq \varepsilon$

**Return:** $w_* := (w_1^{(t+1)}, \ldots, w_m^{(t+1)})$.

**Algorithm 2** Approximating the M-posterior distribution

**Input:**
1. Samples $\{Z_{j,i}\}_{i=1}^{N_j} \sim$ i.i.d. from $\Pi_{n,m}(\cdot|G_j)$, $j = 1 \ldots m$ (see equation (10));

**Do:**
1. $Q_j := \frac{1}{N_j} \sum_{i=1}^{N_j} \delta_{Z_{j,i}}, j = 1 \ldots m$ - empirical approximations of $\Pi_{n,m}(\cdot|G_j)$.
2. Apply Algorithm 1 to $Q_1, \ldots, Q_m$; return $w_* = (w_{*,1} \ldots w_{*,m})$;
3. For $j = 1, \ldots, m$, set $\bar{w}_j := w_{*,j} I\{w_{*,j} \geq \frac{1}{2m}\}$; define $\hat{w}_j^* := \bar{w}_j / \sum_{i=1}^{m} \bar{w}_i$.

**Return:** $\hat{\Pi}_{n,g}^{\text{st}} := \sum_{i=1}^{m} \hat{w}_i^* Q_i$.

Given the data set $\mathcal{X}_n$ of size $n$, let $t(n)$ be the running time of the subroutine (e.g., MCMC) that outputs a single observation from the posterior distribution $\Pi_n(\cdot|\mathcal{X}_n)$. Assuming that our goal is to obtain a sample of size $N$ from the (usual) posterior, the total computational complexity is $O\left(N \cdot t(n)\right)$. We compare this with the running time needed to obtain a sample of same size $N$ from the $M$-posterior given that the algorithm is running on $m$ machines ($m \ll n$) in parallel. In this case, we need to generate $O\left(N/m\right)$ samples from each of $m$ subset posteriors, which is done in time $O\left(\frac{N}{m} \cdot t\left(\frac{n}{m}\right)\right)$. According to Theorem 7.1 in (Beck & Sabach, 2013), Weiszfeld's algorithm approximates the M-posterior to degree of accuracy $\varepsilon$ in at most $O(1/\varepsilon)$ steps, and each of these steps has complexity $O(N^2)$ (which follows from (7)), so that the total running time is $O\left(\frac{N}{m} \cdot t\left(\frac{N}{m}\right) + \frac{N^2}{\varepsilon}\right)$. If, for example, $t(n) \simeq n^r$ for some $r \geq 1$, then $\frac{N}{m} \cdot t\left(\frac{n}{m}\right) \simeq \frac{1}{m^{1+r}} N n^r$ which should be compared to $N \cdot n^r$ required by the standard approach.

### 4.1. Simulated data

The first example uses data from a Gaussian distribution with known variance and unknown mean, and demonstrates the effect of the magnitude of an outlier on the posterior distribution of the mean parameter. The second example demostrates the robustness and scalability of nonparametric regression using M-posterior in presence of outliers.

#### 4.1.1. Univariate Gaussian data

The goal of this example is to demonstrate the effect of a large outlier on the posterior distribution of the mean parameter $\mu$. We generated 25 sets containing 100 observations each. Every sample $\{\mathbf{x}_i\}_{i=1}^{25}$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,100})$, contains 99 independent observations from the standard Gaussian distribution ($x_{i,j} \sim \mathcal{N}(0,1)$ for $i = 1, \ldots, 25$ and $j = 1, \ldots, 99$), while the last entry in each sample, $x_{i,100}$, is an outlier, and its value linearly increases for $i = 1, \ldots, 25$, namely, $x_{i,100} = i \max(|x_{i,1}|, \ldots, |x_{i,99}|)$. Index of an outlier is unknown to the algorithm, while the variance of observations is known. We use a flat (Jeffreys) prior on the mean $\mu$ and obtain its posterior distribution, which is also Gaussian with mean $\frac{\sum_{j=1}^{100} x_{ij}}{100}$ and variance $\frac{1}{100}$. We generate 1000 samples from each posterior distribution $\Pi_{100}(\cdot|\mathbf{x}_i)$ for $i = 1, \ldots, 25$. Algorithm 2 generates 1000 samples from the M-posterior $\hat{\Pi}_{100,g}^{\text{st}}(\cdot|\mathbf{x}_i)$ for each $i = 1, \ldots, 25$: to this end, we set $m = 10$ and generate 100 samples from every $\Pi_{100,10}(\cdot|G_{j,i})$, $j = 1, \ldots, 10$ to form the empirical measures $Q_{j,i}$; here, $\cup_{j=1}^{10} G_{j,i} = \mathbf{x}_i$. Consensus MCMC (Scott et al., 2013) as a representative for scalable MCMC methods, and compared its performance with M-posterior when the number of data subsets is fixed.

Figure 1 compares the performance of the "consensus posterior", the overall posterior and the M-posterior using the empirical coverage of (1-$\alpha$)100% credible intervals (CIs) calculated across 50 replications for $\alpha = 0.2, 0.15, 0.10,$ and 0.05. The empirical coverages of M-posterior's CIs show robustness to the size of an outlier. On the contrary, performance of the consensus and overall posteriors deteriorate fairly quickly across all $\alpha$'s leading to 0% empirical coverage as the outlier strength increases from $i = 1$ to $i = 25$.

#### 4.1.2. Gaussian process regression

We use function $f_0(x) = 1 + 3\sin(2\pi x - \pi)$ and simulate 90 (case 1) and 980 (case 2) values of $f_0$ at equidistant $x$'s in $[0, 1]$ (hereafter $x_{1:90}$ and $x_{1:980}$) corrupted by Gaussian noise with mean 0 and variance 1. To demonstrate the robustness of M-posterior in nonparametric regression, we added 10 (case 1) and 20 (case 2) outliers (sampled on the uniform grids of corresponding sizes) to

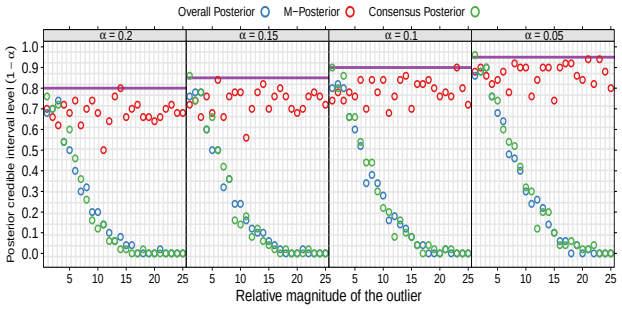


*Figure 1.* Effect of outlier on the empirical coverage of (1-$\alpha$)100% credible intervals (CIs). The $x$-axis represents the outlier magnitude. The $y$-axis represents the fraction of times the CIs include the true mean over 50 replications. The panels show the coverage results when $\alpha = 0.2, 0.15, 0.10$, and 0.05. The horizontal lines (in violet) show the theoretical coverage.

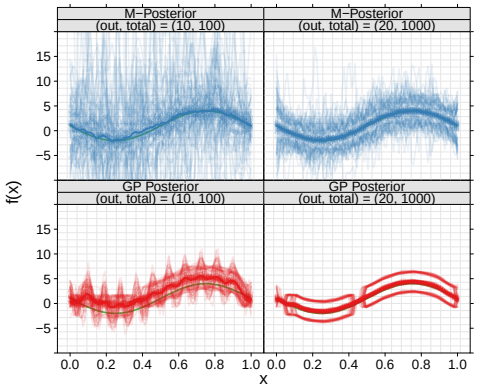


*Figure 2.* Performance of M-posterior in Gaussian process (GP) regression. The top and bottom row of panels show simulation results for M-posterior (in blue) and GP regression (in red). The size of data increases from column 1 to 2. The true noiseless curve $f_0(x)$ is in green. The shaded regions around the curves represent 95% confidence bands obtained over 30 replicated data sets.

the data sets such that $f_0(x_{91:100}) = 10 \max(f_0(x_{1:90}))$ and $f_0(x_{981:1000}) = 10 \max(f_0(x_{1:980}))$.

The `gausspr` function in `kernlab` R package (Karatzoglou et al., 2004) is used for GP regression. Based on the standard convention in GP regression, the noise variance (or "nugget effect") is fixed at 0.01. Using these settings for GP regression without the "standard" posterior, `gausspr` obtains an estimator $\widehat{f}_1$ and a 95% confidence band for the values of the regression function at 100 equally spaced grid points $y_{1:100}$ in $[0, 1]$ (note that these locations are different from the observed data). Algorithm 2 performs GP regression with M-posterior and obtains an estimator $\widehat{f}_2$ described below. The posterior draws across $y_{1:100}$ are obtained in cases 1 and 2 as follows. First, $\{(x_i, f_i)\}$ are split into $m = 10$ and 20 subsets (each living on its own uniform grid) respectively, and `gausspr` estimates the posterior mean $\mu_j$ and covariance $\mathbf{\Sigma}_j$ for each data subset, $j = 1, \ldots, m$. These estimates correspond to the Gaussian distributions $\Pi_j(\cdot|\mu_j, \mathbf{\Sigma}_j)$ that are used to generate 100 posterior draws at $y_{1:100}$ each. These draws are further employed to form the empirical versions of subset posteriors. Finally, Weiszfeld's algorithm is used to combine them. Next, we obtained 1000 samples from the M-posterior $\Pi_g(\cdot|\{(x_i, f_i)\})$. The median of these 1000 samples at each location on the grid $y_{1:100}$ represents the estimator $\hat{f}_2$. Its 95% confidence band corresponds to 2.5% and 97.5% quantiles of the 1000 posterior draws across $y_{1:100}$.

Figure 2 summarizes the results of GP regression with and without M-posterior across 30 replications. In case 1, GP regression without M-posterior is extremely sensitive to the outliers, resulting in $\widehat{f}_1$ that is shifted above the truth and distorted near the $x$'s that are adjacent to the outliers; in turn, this affects the coverage of 95% confidence bands and results in the "bumps" that correspond to the location of outliers. In contrast, GP regression using M-posterior produces $\widehat{f}_2$ which is close to the true curve in both cases; however, in case 1, when the number of data points is small, the 95% bands are unstable.

An attractive property of M-posterior based GP regression is that numerical instability due to matrix inversion can be avoided by working with multiple subsets. We investigated such cases when the number of data points $n$ was greater than $10^4$. Chalupka et al. (2012) compare several low rank matrix approximations techniques used to avoid matrix inversion in massive data GP computation. M-posterior–based GP computation does not use approximations to obtain subset posteriors. By increasing the number of subsets ($m$), M-posterior based GP regression is both computationally feasible and numerically stable for cases when $n = \mathcal{O}(10^6)$ and $m = \mathcal{O}(10^3)$. On the contrary, standard GP regression using the whole data set was intractable for data size greater than $10^4$ due of numerical instabilities in matrix inversion. In general, for $n$ data points and $m$ subsets, the computational complexity for GP with M-posterior is $\mathcal{O}(n(\frac{n}{m})^2)$; therefore, $m > 1$ is computationally better than working with the whole data set. By carefully choosing the $\frac{n}{m}$ ratio depending on the available computational resources and $n$, GP regression with M-posterior is a promising approach for GP regression for massive data without low rank approximations.

### 4.2. Real data: PdG hormone levels vs day of ovulation

North Carolina Early Pregnancy Study (NCEPS) measured urinary pregnanediol-3-glucuronide (PdG) levels, a progesterone metabolite, in 166 women from the day of ovulation across 41 time points (Baird et al., 1999). These data have two main features that need to be modeled. First, the data contains information about women in different stages

of conception and non-conception ovulation cycles, so the probabilistic model should be flexible and free of any restrictive distributional assumptions; therefore, we choose non-parametric regression of log PdG on day of ovulation using GP regression. Second, missing data and extreme observations are very common in these data due the nature of observations and diversity of subjects in the study; therefore, we use M-posterior based GP regression as a robust approach to automatically account for outliers and possible model misspecification.

NCEPS data have missing PdG levels for multiple women across multiple time points. We discarded subjects that did not have data for at least half the time points, which left us with 3810 PdG levels across 41 time points. The size of the data enables the use of `gausspr` function for GP regression, and its results are compared against M-posterior based GP regression (similar to Section 4.1.2). The data was divided into 10 subsets. On each stage, 9 of them were used to evaluate the M-posterior while the remaining was a test set; the process was repeated 10 times for different test subsets. Our goal is to obtain posterior predictive intervals for log PdG levels given the day of ovulation. Figure 3 shows the posterior predictive distribution obtained via GP regression with and without M-posterior. Across all folds, the uncertainty quantification based on M-posterior is much better than its counterpart without the M-posterior. The main reason for this poor performance of "vanilla" GP regression is that NCEPS data have many data points for each day relative to ovulation, but with many outliers and missing data. The vanilla GP regression does not account for the latter feature of NCEPS data, thus leading to over-optimistic uncertainty estimates across all folds. M-posterior automatically accounts for outliers and model misspecification; therefore, it leads to reliable posterior predictive uncertainty quantification across all folds.

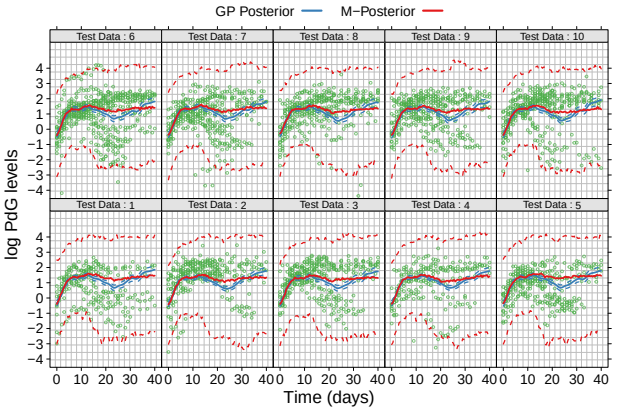


*Figure 3.* Comparison of 95% posterior predictive intervals for GP regression with and without M-posterior. The x and y axes represent day of ovulation and log PdG levels. Panels show the result of GP regression of log PdG on day of ovulation across 10-folds of NCEPS data. The mean dependence log PdG on day of ovulation is shown using solid lines. The dotted lines around the solid curves represent the 95% posterior predictive intervals. The intervals for GP regression without M-posterior severely underestimate the uncertainty in log PdG levels. On the other hand, posterior predictive intervals of M-posterior appear to be reasonable and the trend of dependence of log PdG on day of ovulation is stable.

## 5. Discussion

We presented a general approach to scalable and robust Bayesian inference based on the evaluation of the geometric median of subset posterior distributions (M-posterior). To the best of our knowledge, this is the first technique that is provably robust and computationally efficient. The key to making inference tractable is to embed the subset posterior distributions in a suitable RKHS, and pose the aggregation problem as convex optimization in this space, which in turn can be solved using Weiszfeld's algorithm, a simple gradient descent-based method. Unlike popular point estimators, the M-posterior distribution can be used for summarizing uncertainty in the parameters of interest. Another advantage of our approach is scalability, so that it can be used for distributed Bayesian learning: first, since it combines the subset posteriors using a gradient-based algorithm, it naturally scales to massive data. Second, since Weiszfeld's algorithm only uses the samples from subset posteriors, it can be implemented in a distributed setting via MapReduce/Hadoop.

Several important questions are not included in the present paper and will be addressed in subsequent work. These topic include applications to other types of models; alternative data partition methods and connections to the inference based on the usual posterior distribution; different choices of distances, notions of the median and related subset posterior aggregation methods. More efficient computational alternatives and extensions of Weiszfeld's algorithm (which is currently used due to its simplicity, stability, and ease of implementation) can be developed for estimating the M-posterior; see (Bose et al., 2003; Cardot et al., 2013; Vardi & Zhang, 2000), among other works. Applications of distributed optimization methods, such as ADMM (Boyd et al., 2011), is another potentially fruitful approach.

## Acknowledgments

Authors were supported by grant R01-ES-017436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). S. Srivastava was supported by NSF under Grant DMS-1127914 to SAMSI. S. Minsker acknowledges support from NSF grants FODAVA CCF-0808847, DMS-0847388, ATD-1222567.

## References


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

Ahn, S., Korattikara, A., and Welling, M. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of ICML*, 2012.

Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. In *Proceedings of the 28th ACM symposium on Theory of computing*, pp. 20–29. ACM, 1996.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Baird, D. D., Weinberg, C. R., Zhou, H., Kamel, F., McConnaughey, D. R., Kesner, J. S., and Wilcox, A. J. Preimplantation urinary hormone profiles and the probability of conception in healthy women. *Fertility and sterility*, 71(1):40–49, 1999.

Beck, A. and Sabach, S. Weiszfeld's method: old and new results. *Preprint*, 2013.

Bose, P., Maheshwari, A., and Morin, P. Fast approximations for sums of distances, clustering and the Fermat–Weber problem. *Computational Geometry*, 24(3):135–146, 2003.

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

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., and Jordan, M. Streaming variational Bayes. In *NIPS*, pp. 1727–1735, 2013.

Cardot, H., Cenac, P., and Zitt, P.-A. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19 (1):18–43, 2013.

Chalupka, K., Williams, C. K., and Murray, I. A framework for evaluating approximation methods for gaussian process regression. *arXiv:1205.6326*, 2012.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *JMLR*, 14:1303–1347, 2013.

Hsu, D. and Sabato, S. Loss minimization and parameter estimation with heavy tails. *arXiv:1307.1827*, 2013.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.

Korattikara, A., Chen, Y., and Welling, M. Austerity in MCMC land: cutting the Metropolis-Hastings budget. *arXiv:1304.5299*, 2013.

Lerasle, M. and Oliveira, R. I. Robust empirical mean estimators. *arXiv:1112.3914*, 2011.

Minsker, S. Geometric median and robust estimation in Banach spaces. *arXiv:1308.1334*, 2013.

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

Nemirovski, A. and Yudin, D. Problem complexity and method efficiency in optimization. 1983.

Nielsen, F. and Garcia, V. Statistical exponential families: a digest with flash cards. *arXiv:0911.4863*, 2011.

Rieger, C. and Zwicknagl, B. Deterministic error analysis of support vector regression and related regularized kernel methods. *JMLR*, 10:2115–2132, 2009.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. Bayes and Big Data: the consensus Monte Carlo algorithm. In *EFaB Bayes 250 workshop*, volume 16, 2013.

Smola, A. J. and Narayanamurthy, S. An architecture for parallel topic models. In *Very Large Databases (VLDB)*, 2010.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. On integral probability metrics, $\phi$-divergences and binary classification. *arXiv:0901.2698*, 2009.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. Hilbert space embeddings and metrics on probability measures. *JMLR*, 99: 1517–1561, 2010.

Vardi, Yehuda and Zhang, Cun-Hui. The multivariate $L_1$-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

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

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