Journal of Machine Learning Research 18 (2017) 1-40 Submitted 12/16; Revised 11/17; Published 12/17

# Robust and Scalable Bayes via a Median of Subset Posterior Measures

**Stanislav Minsker** MINSKER@USC.EDU
*Department of Mathematics*
*University of Southern California*
*Los Angeles, CA 90089, USA*

**Sanvesh Srivastava** SANVESH-SRIVASTAVA@UIOWA.EDU
*Department of Statistics and Actuarial Science*
*University of Iowa*
*Iowa City, IA 52242, USA*

**Lizhen Lin** LIZHEN.LIN@ND.EDU
*Department of Applied and Computational Mathematics and Statistics*
*The University of Notre Dame*
*Notre Dame, IN 46556, USA*

**David B. Dunson** DUNSON@DUKE.EDU
*Departments of Statistical Science, Mathematics, and ECE*
*Duke University*
*Durham, NC 27708, USA*

**Editor:** David Blei

## Abstract

We propose a novel approach to Bayesian analysis that is provably robust to outliers in the data and often has computational advantages over standard methods. Our technique is based on splitting the data into non-overlapping subgroups, evaluating the posterior distribution given each independent subgroup, and then combining the resulting measures. The main novelty of our approach is the proposed aggregation step, which is based on the evaluation of a median in the space of probability measures equipped with a suitable collection of distances that can be quickly and efficiently evaluated in practice. We present both theoretical and numerical evidence illustrating the improvements achieved by our method.

**Keywords:** Big data, geometric median, distributed computing, parallel MCMC, Wasserstein distance

## 1. Introduction

Contemporary data analysis problems pose several general challenges. One is resource limitations: massive data require computer clusters for storage and processing. Another problem occurs when data are severely contaminated by "outliers" that are not easily iden-

©2017 Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin and David B. Dunson.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v18/16-655.html.

tified and removed. Following Box and Tiao (1968), an outlier can be defined as *"being an observation which is suspected to be partially or wholly irrelevant because it is not generated by the stochastic model assumed."* While the topic of robust estimation has occupied an important place in the statistical literature for several decades and significant progress has been made in the theory of point estimation, robust Bayesian methods are not sufficiently well-understood.

Our main goal is to take a step towards solving these problems, proposing a general Bayesian approach that is

(i) provably robust to the presence of outliers in the data without any specific assumptions on their distribution or reliance on preprocessing;

(ii) scalable to big data sets through allowing computational algorithms to be implemented in parallel for different data subsets prior to an efficient aggregation step.

The proposed approach consists in splitting the sample into disjoint parts, implementing Markov chain Monte Carlo (MCMC) or another posterior sampling method to obtain draws from each "subset posterior" in parallel, and then using these draws to obtain weighted samples from the *median posterior (or M-Posterior)*, a new probability measure which is a (properly defined) median of a collection of subset posterior distributions. We show that, despite the loss of "interactions" among the data in different groups, the final result still admits strong guarantees; moreover, splitting the data gives certain advantages in terms of robustness to outliers.

In particular, we demonstrate that the M-Posterior is a probability measure centered at the "robust" estimator of the unknown parameter, the associated credible sets are often of the same "width" as the credible sets obtained from the usual posterior distribution and admit strong "frequentist" coverage guarantees (see Section 3.2 for exact statements).

The paper is organized as follows: Section 1.1 contains an overview of the existing literature and explains the goals that we aim to achieve in this work. Section 2 introduces the mathematical background and key facts used throughout the paper. Section 3 describes the main theoretical results for the median posterior. Section 4 presents details of algorithms, implementation, and numerical performance of the median posterior for several models. The simulation study and analysis of data examples convincingly show the robustness properties of the median posterior. In particular, we have used M-Posterior for scalable nonparametric Bayesian modeling of joint dependence in the multivariate categorical responses collected by the General Social Survey (`gss.norc.org`). Proofs that are omitted in the main text are contained in the supplementary material.

### 1.1 Discussion of Related Work

A. Dasgupta remarks that (see the discussion following the work by Berger, 1994): "Exactly what constitutes a study of Bayesian robustness is of course impossible to define." The popular definition which also indicates the main directions of research in this area is due to Berger (1994): "Robust Bayesian analysis is the study of the sensitivity of Bayesian answers

to uncertain inputs. These uncertain inputs are typically the model, prior distribution, or utility function, or some combination thereof." Notable works on robustness of Bayesian procedures to model misspecification include Doksum and Lo (1990) and Hoff (2007) that investigate methods based on conditioning on partial information, as well as a more recent paper by Miller and Dunson (2015) who introduced the notion of the "coarsened" posterior; however, its behavior in the presence of outliers is not explicitly addressed. Outliers are typically accommodated by either employing heavy-tailed likelihoods (e.g., Svensen and Bishop, 2005) or by attempting to identify and remove them as a first step (as in Box and Tiao, 1968 or Bayarri and Berger, 1994). The common assumption in the Bayesian literature is that the distribution of the outliers can be modeled (e.g., using a $t$-distribution, contamination by a larger variance parametric distribution, etc). In this paper, we instead bypass the need to place a model on the outliers and do not require their removal prior to analysis; similar approach has previously been advocated by Hooker and Vidyashankar (2014). We base inference on the *median posterior*, whose robustness can be formally and precisely quantified in terms of concentration properties around the true delta measure under the potential influence of outliers and contaminations of arbitrary nature.

Also relevant is the recent progress in scalable Bayesian algorithms. Most methods 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 and 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 that Bayesian approaches typically require an approximation to the full posterior distribution instead of just a point estimate of parameters.

Several major approaches exist for scalable Bayesian learning in a distributed setting. The first approach independently evaluates the likelihood for each data subset across multiple machines and returns 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 (see Smola and Narayanamurthy, 2010; Agarwal and Duchi, 2012). This approach is problem-specific and involves extensive communication among machines. The second approach uses a so-called stochastic approximation (SA) and successively learns "noisy" approximations to the full posterior distribution using data in small mini-batches. A group of methods based on this approach uses sampling-based techniques to explore the posterior distribution through modified Hamiltonian or Langevin dynamics (e.g., Welling and Teh, 2011; Ahn et al., 2012; Korattikara et al., 2013). Unfortunately, these methods fail to accommodate discrete-valued parameters and multimodality. Another subgroup of methods uses deterministic variational approximations and learns the variational parameters of the approximated posterior through an optimization-based approach (see Wang et al., 2011; Hoffman et al., 2013; Broderick et al., 2013). Although these techniques often have excellent predictive performance, it is well known (Bishop, 2006) that variational methods tend to substantially underestimate posterior uncertainty and provide a poor characterization of posterior dependence, while lacking theoretical guarantees.

Our approach instead falls in a 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 these methods simply average draws from each subset (Scott et al., 2013). Other alternatives use an approximation to the full posterior distribution based on kernel density estimates (Neiswanger et al., 2013) or the so-called Weierstrass transform (Wang and Dunson, 2013). These methods have limitations related to the dimension of the parameter, moreover, their applicability and theoretical justification are restricted to parametric models. Unlike the method proposed below, none of the aforementioned algorithms are provably robust. Another closely related method is the so-called WASP (Srivastava et al., 2015b). It is scalable but still lacks robustness guarantees.

Our work was inspired by recent multivariate median-based techniques for robust estimation developed in Minsker (2015) (see also Hsu and Sabato, 2013; Alon et al., 1996; Lerasle and Oliveira, 2011; Nemirovskiĭ and David, 1983 where similar ideas were applied in different frameworks).

## 2. Preliminaries

We proceed by recalling key definitions and facts which will be used throughout the paper, followed by the definition of the M-Posterior distribution in Section 2.4.

### 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 $B_1, \dots, B_N$ of radius $\varepsilon$ contained in $\mathbb{Y}$, i.e., $\bigcup_{j=1}^{N} B_j \subseteq \mathbb{Y}$.

Let $\{p_\theta,\ \theta \in \Theta\}$ be a family of probability density functions on $\mathbb{R}^p$. Let $l, u : \mathbb{R}^p \mapsto \mathbb{R}_+$ be two functions such that $l(x) \le u(x)$ for every $x \in \mathbb{R}^p$ and $d^2(l,u) := \int_{\mathbb{R}^p} (\sqrt{u} - \sqrt{l})^2(x)dx < \infty$. A bracket $[l, u]$ consists of all functions $g : \mathbb{R}^p \mapsto \mathbb{R}$ such that $l(x) \le g(x) \le u(x)$ for all $x \in \mathbb{R}^p$. For $A \subseteq \Theta$, the bracketing number $N_{[]}(\varepsilon, A, d)$ is defined as the smallest number $N$ such that there exist $N$ brackets $[l_i, u_i],\ i = 1, \dots, N$ satisfying $\{p_\theta,\ \theta \in A\} \subseteq \bigcup_{i=1}^{N} [l_i, u_i]$ and $d(l_i, u_i) \le \varepsilon$ for all $1 \le i \le N$.

For $y \in \mathbb{Y}$, $\delta_y$ 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.

We will say that $k : \mathbb{Y} \times \mathbb{Y} \mapsto \mathbb{R}$ is a *kernel* if it is a symmetric, positive definite function. Assume that $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$ is a reproducing kernel Hilbert space (RKHS) of functions $f : \mathbb{Y} \mapsto$

$\mathbb{R}$. Then $k$ is a *reproducing kernel* for $\mathbb{H}$ if for any $f \in \mathbb{H}$ and $y \in \mathbb{Y}$, $\langle f, k(\cdot, y)\rangle_{\mathbb{H}} = f(y)$ (see Aronszajn, 1950 for details).

For a square-integrable function $f \in L_2(\mathbb{R}^p)$, $\hat{f}$ stands for its Fourier transform. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer not greater than $x$.

Finally, given two nonnegative sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for some $C > 0$ and all $n$. Other objects and definitions are introduced in the course of exposition when necessity arises.

### 2.2 Generalizations of the Univariate Median

Let $\mathbb{Y}$ be a normed space with norm $\|\cdot\|$, and let $\mu$ be a probability measure on $(\mathbb{Y}, \|\cdot\|)$ equipped with Borel $\sigma$-algebra. Define the *geometric median* of $\mu$ by

$$x_* = \operatorname*{argmin}_{y \in \mathbb{Y}} \int_{\mathbb{Y}} (\|y - x\| - \|x\|)\, \mu(dx).$$

In this paper, we focus on the special case when $\mu$ is a uniform distribution on a finite collection of atoms $x_1, \ldots, x_m \in \mathbb{Y}$, so that

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

The geometric median exists under rather general assumptions; for example, if $\mathbb{Y}$ is a Hilbert space (this case will be our main focus; more general conditions were obtained by Kemperman, 1987). Moreover, it is well-known that in this situation $x_* \in \mathrm{co}(x_1, \ldots, x_m)$—the convex hull of $x_1, \ldots, x_m$ (meaning that there exist nonnegative $\alpha_j$, $j = 1 \ldots m$, $\sum_{j=1}^{m} \alpha_j = 1$ such that $x_* = \sum_{j=1}^{m} \alpha_j x_j$).

Another useful generalization of the univariate median is defined as follows. Let $(\mathbb{Y}, d)$ be a metric space with metric $d$, and $x_1, \ldots, x_k \in \mathbb{Y}$. Define $B_*$ to be the $d$-ball of minimal radius such that it is centered at one of $\{x_1, \ldots, x_m\}$ and contains at least half of these points. Then the median $\mathrm{med}_0(x_1, \ldots, x_m)$ of $x_1, \ldots, x_m$ is the center of $B_*$. In other words, let

$$\begin{aligned} \varepsilon_* := \inf \Big\{ \varepsilon > 0 :\ & \exists j = j(\varepsilon) \in \{1, \ldots, m\} \text{ and } I(j) \subset \{1, \ldots, m\} \text{ such that} \\ & |I(j)| > \frac{m}{2} \text{ and } \forall i \in I(j),\ d(x_i, x_j) \leq 2\varepsilon \Big\}, \end{aligned} \tag{2}$$

$j_* := j(\varepsilon_*)$, where ties are broken arbitrarily, and set

$$x_* = \mathrm{med}_0(x_1, \ldots, x_m) := x_{j_*}. \tag{3}$$

We will say that $x_*$ is the *metric median* of $x_1, \ldots, x_m$. Note that $x_*$ always belongs to $\{x_1, \ldots, x_m\}$ by definition. Advantages of this definition are its generality (only metric space structure is assumed) and simplicity of numerical evaluation since only the pairwise

distances $d(x_i, x_j),\ i, j = 1, \dots, m$ are required to compute the median. This construction was previously employed by Nemirovskiĭ and David (1983) in the context of stochastic optimization and is further studied by Hsu and Sabato (2013). A closely related notion of the median was used by Lopuhaa and Rousseeuw (1991) under the name of the "minimal volume ellipsoid" estimator.

Finally, we recall an important property of the median (shared both by $\mathrm{med}_g$ and $\mathrm{med}_0$) which states that it transforms a collection of independent, "weakly concentrated" estimators into a single estimator with significantly stronger concentration properties. Given $q, \alpha$ such that $0 < q < \alpha < 1/2$, define a nonnegative function $\psi(\alpha, q)$ via

$$\psi(\alpha, q) := (1-\alpha)\log\frac{1-\alpha}{1-q} + \alpha\log\frac{\alpha}{q}. \tag{4}$$

The following result is an adaptation of Theorem 3.1 in (Minsker, 2015):

**Theorem 1**

**a** *Assume that $(\mathbb{H}, \|\cdot\|)$ is a Hilbert space and $\theta_0 \in \mathbb{H}$. Let $\hat\theta_1, \dots, \hat\theta_m \in \mathbb{H}$ be a collection of independent random variables. Let $\kappa$ be a constant satisfying $0 \le \kappa < \frac{1}{3}$. Suppose $\varepsilon > 0$ is such that for all $j,\ 1 \le j \le \lfloor (1-\kappa)m \rfloor + 1$,*

$$\Pr\Big(\|\hat\theta_j - \theta_0\| > \varepsilon\Big) \le \frac{1}{7}. \tag{5}$$

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

$$\Pr\Big(\|\hat\theta_* - \theta_0\| > 1.52\varepsilon\Big) \le \left[e^{(1-\kappa)\psi\left(\frac{3/7-\kappa}{1-\kappa}, 1/7\right)}\right]^{-m}.$$

**b** *Assume that $(\mathbb{Y}, d)$ is a metric space and $\theta_0 \in \mathbb{Y}$. Let $\hat\theta_1, \dots, \hat\theta_m \in \mathbb{Y}$ be a collection of independent random variables. Let $\kappa$ be a constant satisfying $0 \le \kappa < \frac{1}{3}$. Suppose $\varepsilon > 0$ is such that for all $j,\ 1 \le j \le \lfloor (1-\kappa)m \rfloor + 1$,*

$$\Pr\Big(d(\hat\theta_j, \theta_0) > \varepsilon\Big) \le \frac{1}{4}. \tag{6}$$

*Let $\hat\theta_* = \mathrm{med}_0(\hat\theta_1, \dots, \hat\theta_m)$. Then*

$$\Pr\Big(d(\hat\theta_*, \theta_0) > 3\varepsilon\Big) \le \left[e^{(1-\kappa)\psi\left(\frac{1/2-\kappa}{1-\kappa}, 1/4\right)}\right]^{-m}.$$

**Proof** See Section A.1 in the appendix. ■

**Remark 2** *While we require $\kappa < 1/3$ above for clarity and to keep the constants small, we prove a slightly more general result that holds for any $\kappa < 1/2$.*

Theorem 1 implies that the concentration of the geometric median of independent estimators around the "true" parameter value improves geometrically fast with respect to the number of such estimators, while the estimation rate is preserved, up to a constant. In our case, the role of $\hat{\theta}_j$'s will be played by posterior distributions based on disjoint subsets of observations, viewed as elements of the space of signed measures equipped with a suitable distance.

Parameter $\kappa$ allows taking corrupted observations into account: if the initial sample contains not more than $\lfloor \kappa m \rfloor$ outliers (of arbitrary nature), then at most $\lfloor \kappa m \rfloor$ estimators amongst $\{\theta_1, \dots, \theta_m\}$ can be affected but their median remains stable, still being close to the unknown $\theta_0$ with high probability. To clarify the notion of "robustness" that such a statement provides, assume that $\hat{\theta}_1, \dots, \hat{\theta}_m$ are consistent estimators of $\theta_0$ based on disjoint samples of size $n/m$ each. If $\frac{n}{m} \to \infty$, then $\frac{\kappa m}{n} \to 0$, hence the breakdown point of the estimator $\hat{\theta}_*$ is 0 is general. However, it is able to handle a number of outliers that grows like $o(n)$ while preserving consistency, which is the best one can hope for without imposing any additional assumptions on the underlying distribution, parameter of interest or nature of the outliers.

Let us also mention that the the geometric median of a collection of points in a Hilbert space belongs to the convex hull of these points. Thus, one can think about "downweighing" some observations (potential outliers) and increasing the weight of others, and geometric median gives a way to formalize this approach. The median $\mathrm{med}_0$ defined in (3) corresponds to the extreme case when all but one weight are equal to 0. Its potential advantage lies in the fact that its evaluation requires only the knowledge of pairwise distances $d(\hat{\theta}_i, \hat{\theta}_j),\ i, j = 1, \dots, m$, see (2).

### 2.3 Distances Between Probability Measures

Next, we discuss the special family of distances between probability measures that will be used throughout the paper. These distances provide the necessary structure to define and evaluate medians in the space of measures, as discussed above. Since one of our goals was to develop computationally efficient techniques, we focus on 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

$$\|P - Q\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \int_{\mathbb{X}} f(x) d(P - Q)(x) \right|. \tag{7}$$

Important special cases include the situation when

$$\mathcal{F} = \mathcal{F}_L := \{f : \Theta \mapsto \mathbb{R} \text{ s.t. } \|f\|_L \le 1\}, \tag{8}$$

where $\|f\|_L := \sup\limits_{x_1 \neq x_2} \frac{|f(x_1) - f(x_2)|}{\rho(x_1, x_2)}$ is the Lipschitz constant of $f$.

It is well-known (Dudley, 2002, Theorem 11.8.2) that in this case $\|P - Q\|_{\mathcal{F}_L}$ is equal to the Wasserstein distance (also known as the Kantorovich-Rubinstein distance)

$$d_{W_{1,\rho}}(P, Q) = \inf \Big\{ \mathbb{E}\rho(\boldsymbol{X}, \boldsymbol{Y}) : \mathcal{L}(\boldsymbol{X}) = P,\ \mathcal{L}(\boldsymbol{Y}) = Q \Big\}, \tag{9}$$

where $\mathcal{L}(\boldsymbol{Z})$ denotes the law of a random variable $\boldsymbol{Z}$ and the infimum on the right is taken over the set of all joint distributions of $(\boldsymbol{X}, \boldsymbol{Y})$ with marginals $P$ and $Q$.

Another fruitful structure emerges when $\mathcal{F}$ is a unit ball in a Reproducing Kernel Hilbert Space $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$ with a reproducing kernel $k : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$. That is,

$$\mathcal{F} = \mathcal{F}_k := \{f : \mathbb{X} \mapsto \mathbb{R},\ \|f\|_{\mathbb{H}} := \sqrt{\langle f, f\rangle_{\mathbb{H}}} \leq 1\}. \tag{10}$$

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

$$\|P - Q\|_{\mathcal{F}_k} = \left\| \int_{\mathbb{X}} k(x,\cdot)d(P-Q)(x) \right\|_{\mathbb{H}}. \tag{11}$$

It follows that $P \mapsto \int_{\mathbb{X}} k(x,\cdot)dP(x)$ is 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 one to use fast numerical methods to approximate the geometric median, see Section 4 below.

**Remark 3** *Note that when $P$ and $Q$ are discrete measures (e.g., $P = \sum_{j=1}^{N_1} \beta_j \delta_{z_j}$ and $Q = \sum_{j=1}^{N_2} \gamma_j \delta_{y_j}$), then*

$$\|P - Q\|^2_{\mathcal{F}_k} = \sum_{i,j=1}^{N_1} \beta_i\beta_j k(z_i, z_j) + \sum_{i,j=1}^{N_2} \gamma_i\gamma_j k(y_i, y_j) - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} \beta_i\gamma_j k(z_i, y_j). \tag{12}$$

In this paper, 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 its *strict positive definiteness*: we say that $k$ is *strictly positive definite* if it is bounded, measurable, and such that for all non-zero signed Borel measures $\nu$

$$\iint_{\mathbb{X}\times\mathbb{X}} k(x,y)d\nu(x)d\nu(y) > 0.$$

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 4** *Let $\mathbb{X} = \mathbb{R}^p$, $p \geq 1$. 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 5** *It is important to mention that in practical applications, we often deal with empirical measures based on a collection of MCMC samples from the posterior distribution. A natural question is the following: if $P$ and $Q$ are probability measures on $\mathbb{R}^D$ and $P_m$, $Q_n$ are their empirical versions, what is the size of the error*

$$e_{m,n} := \left| \|P - Q\|_{\mathcal{F}_k} - \|P_m - Q_n\|_{\mathcal{F}_k} \right| ?$$

*For i.i.d samples, a useful and favorable fact is that $e_{m,n}$ often does not depend on $D$: under weak assumptions on kernel $k$, $e_{m,n}$ has an upper bound of order $m^{-1/2} + n^{-1/2}$ (that is, $\lim_{m,n\to\infty} \Pr\left(e_{m,n} \geq C(m^{-1/2} + n^{-1/2})\right)$ can be made arbitrarily small by choosing $C$ big enough, see Corollary 12 in Sriperumbudur et al., 2009). On the other hand, the bound for the (stronger) Wasserstein distance is not dimension-free and is of order $m^{-1/(D+1)} + n^{-1/(D+1)}$. Similar error rates hold for empirical measures based on samples from Markov chains used to approximate invariant distributions, including MCMC samples (see Boissard and Le Gouic, 2014 and Fournier and Guillin, 2015).*

If $\mathbb{X}$ is a separable Hilbert space with dot product $\langle \cdot, \cdot \rangle_{\mathbb{X}}$ and $P_1, P_2$ are probability measures with

$$\int_{\mathbb{X}} \|x\|_{\mathbb{X}} dP_i(x) < \infty, \ i = 1, 2,$$

it will be useful to assume that the class $\mathcal{F}$ is chosen such that the distance between the measures is lower bounded by the distance between their means, namely

$$\left\| \int_{\mathbb{X}} x dP_1(x) - \int_{\mathbb{X}} x dP_2(x) \right\|_{\mathbb{X}} \leq C \|P_1 - P_2\|_{\mathcal{F}} \tag{13}$$

for some absolute constant $C > 0$. Clearly, this holds if $\mathcal{F}$ contains the set of continuous linear functionals $\mathbb{L} = \{x \mapsto \langle u, x \rangle_{\mathbb{X}}, \ u \in \mathbb{X}, \ \|u\|_{\mathbb{X}} \leq 1/C\}$, since

$$\left\| \int_{\mathbb{X}} x dP_i(x) \right\|_{\mathbb{X}} = \sup_{\|u\|_{\mathbb{X}} \leq 1} \int_{\mathbb{X}} \langle x, u \rangle_{\mathbb{X}} \, dP_i(x), \ i = 1, 2.$$

In particular, this is true for the Wasserstein distance $d_{W_{1,\rho}}(\cdot, \cdot)$ defined with respect to the metric $\rho$ such that $\rho(x, y) \geq c_1 \|x - y\|_{\mathbb{X}}$. Next, we will state a simple sufficient condition on the kernel $k(\cdot, \cdot)$ for (13) to hold for the unit ball $\mathcal{F}_k$.

**Proposition 6** *Let $\mathbb{X}$ be a separable Hilbert space, $k_0 : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$ - a characteristic kernel, and define*

$$k(x, y) := k_0(x, y) + \langle x, y \rangle_{\mathbb{X}} .$$

*Then $k$ is characteristic and satisfies (13) with $C = 1$.*

**Proof** Let $\mathbb{H}_1$ and $\mathbb{H}_2$ be two reproducing kernel Hilbert spaces with kernels $k_1$ and $k_2$ respectively. It is well-known (e.g., Aronszajn, 1950) that the space corresponding to kernel $k = k_1 + k_2$ is

$$\mathbb{H} = \{f = f_1 + f_2, \ f_1 \in \mathbb{H}_1, \ f_2 \in \mathbb{H}_2\}$$

with the norm $\|f\|_{\mathbb{H}}^2 = \inf\{\|f_1\|_{\mathbb{H}}^2 + \|f_2\|_{\mathbb{H}}^2,\ f_1 + f_2 = f\}$. Hence, the unit ball of $\mathbb{H}$ contains the unit balls of $\mathbb{H}_1$ and $\mathbb{H}_2$, so that for any probability measures $P, Q$

$$\|P - Q\|_{\mathcal{F}_k} \geq \max\left(\|P - Q\|_{\mathcal{F}_{k_1}}, \|P - Q\|_{\mathcal{F}_{k_2}}\right),$$

which easily implies the result. ■

The kernels of the form $k(x, y) = k_0(x, y) + \langle x, y\rangle_{\mathbb{X}}$ will prove especially useful in the situation when the parameter of interest is finite-dimensional (see Section 3.2 for details).

Finally, we recall the definition of the well-known Hellinger and total variation distances. 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 the Hellinger distance between $P$ and $Q$ is given by

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

The total variation distance between two probability measures defined on a $\sigma$-algebra $\mathfrak{B}$ is

$$\|P - Q\|_{TV} = \sup_{B\in\mathfrak{B}} |P(B) - Q(B)|.$$

We are ready to introduce the *median posterior* (or M-Posterior) distribution.

### 2.4 Construction of the M-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 $\theta \in \Theta$, $P_\theta$ is absolutely continuous with respect to Lebesgue measure $dx$ on $\mathbb{R}^D$ with $dP_\theta(\cdot) = p_\theta(\cdot)dx$. In what follows, we equip $\Theta$ with a metric (that we will refer to as the "Hellinger metric")

$$\rho(\theta_1, \theta_2) := h(P_{\theta_1}, P_{\theta_2}), \tag{14}$$

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

Let $k$ be a characteristic kernel defined on $\Theta \times \Theta$. Kernel $k$ defines a metric on $\Theta$ via

$$\rho_k(\theta_1, \theta_2) := \|k(\cdot, \theta_1) - k(\cdot, \theta_2)\|_{\mathbb{H}} = \left(k(\theta_1, \theta_1) + k(\theta_2, \theta_2) - 2k(\theta_1, \theta_2)\right)^{1/2}, \tag{15}$$

where $\mathbb{H}$ is the RKHS associated to $k$. We will assume that $(\Theta, \rho_k)$ is separable. Note that the Hellinger metric $\rho(\theta_1, \theta_2)$ is a particular case corresponding to the kernel

$$k_H(\theta_1, \theta_2) := \left\langle\sqrt{p_{\theta_1}}, \sqrt{p_{\theta_2}}\right\rangle_{L_2(dx)}.$$

All subsequent results apply to this special case. While this is a natural metric for the problem, the disadvantage of $k_H(\cdot, \cdot)$ is that it is often difficult to evaluate numerically. Instead, we will consider metrics $\rho_k$ that are dominated by $\rho$ (this is formalized in assumption 1).

Let $X_1, \ldots, X_n$ be i.i.d. $\mathbb{R}^D$-valued random vectors defined on a probability space $(\Omega, \mathcal{B}, P)$ with unknown distribution $P_0 := P_{\theta_0}$ for some $\theta_0 \in \Theta$. Bayesian inference of $P_0$ requires specifying a prior distribution $\Pi$ over $\Theta$ (equipped with the Borel $\sigma$-algebra induced by $\rho$). The posterior distribution given the observations $\mathcal{X}_n := \{X_1, \ldots, X_n\}$ is a random probability measure on $\Theta$ defined by

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

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

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

almost surely or in probability as $n \to \infty$ for a suitable sequence $\varepsilon_n \to 0$.

One of the questions 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? Even if there is only one outlier, the usual posterior distribution might concentrate most of its mass far from the true value $\theta_0$.

We proceed with a general description of our proposed algorithm for constructing a robust version of the posterior distribution. Let $1 \leq m \leq n/2$ be an integer. Divide the sample $\mathcal{X}_n$ into $m$ disjoint groups $G_1, \ldots, G_m$ of size $|G_j| \geq \lfloor n/m \rfloor$ each:

$$\{X_1, \ldots, X_n\} = \bigcup_{j=1}^m G_j, \ G_i \cap G_l = \emptyset \text{ for } i \neq j, \ |G_j| \geq \lfloor n/m \rfloor, \ j = 1 \ldots m.$$

A good choice of $m$ efficiently exploits the available computational resource while ensuring that the groups $G_j$s are sufficiently large.

Let $\Pi$ be a prior distribution over $\Theta$, and let

$$\left\{ \Pi^{(j)}(\cdot) := \Pi_{|G_j|}(\cdot|G_j), \ j = 1, \ldots, m \right\}$$

be the family of subset posterior distributions depending on disjoint subgroups $G_j$, $j = 1, \ldots, m$:

$$\Pi_{|G_j|}(B|G_j) := \frac{\int\limits_B \prod_{i \in G_j} p_\theta(X_i) d\Pi(\theta)}{\int\limits_\Theta \prod_{i \in G_j} p_\theta(X_i) d\Pi(\theta)}.$$

Define the M-Posterior as

$$\hat{\Pi}_{n,g} := \text{med}_g(\Pi^{(1)}, \ldots, \Pi^{(m)}), \tag{16}$$

or

$$\hat{\Pi}_{n,0} := \text{med}_0(\Pi^{(1)}, \ldots, \Pi^{(m)}), \tag{17}$$

where the medians $\mathrm{med}_g(\cdot)$ and $\mathrm{med}_0(\cdot)$ are evaluated with respect to $\|\cdot\|_{\mathcal{F}_L}$ or $\|\cdot\|_{\mathcal{F}_k}$ introduced in Section 2.2 above. Note that $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$ are always probability measures: indeed, due to the aforementioned properties of a geometric median, there exists $\alpha_1 \geq 0, \ldots, \alpha_m \geq 0$, $\sum\limits_{j=1}^{m} \alpha_j = 1$ such that $\hat{\Pi}_{n,g} = \sum\limits_{j=1}^{m} \alpha_j \Pi^{(j)}$, and $\hat{\Pi}_{n,0} \in \{\Pi^{(1)}(\cdot), \ldots, \Pi^{(m)}(\cdot)\}$ by definition.

While $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$ possess several nice properties (such as robustness to outliers), in practice they often overestimate the uncertainty about $\theta_0$, especially when the number of groups $m$ is large: indeed, if for example $\theta \in \mathbb{R}$ and Bernstein-von Mises theorem holds, then each $\Pi_{|G_j|}(\cdot|G_j)$ is "approximately normal" with covariance $\frac{m}{n} I^{-1}(\theta_0)$ (here, $I(\theta_0)$ is the Fisher information). However, the asymptotic covariance of the posterior distribution based on the whole sample is $\frac{1}{n} I^{-1}(\theta_0)$.

To overcome this difficulty, we propose a modification of our approach where the random measures $\Pi_n^{(j)}$ are replaced by the *stochastic approximations* $\Pi_{|G_j|,m}(\cdot|G_j),\ j = 1, \ldots, m$ of the full posterior distribution. To this end, define the "stochastic approximation" based on the subsample $G_j$ as

$$\Pi_{|G_j|,m}(B|G_j) := \frac{\int\limits_B \Big(\prod_{i \in G_j} p_\theta(X_i)\Big)^m d\Pi(\theta)}{\int\limits_\Theta \Big(\prod_{i \in G_j} p_\theta(X_i)\Big)^m d\Pi(\theta)}, \tag{18}$$

where we assume that $p_\theta^m(\cdot)$ is an integrable function for all $\theta$. In other words, $\Pi_{|G_j|,m}(\cdot|G_j)$ is obtained as a posterior distribution given that each data point from $G_j$ is observed $m$ times. While each of $\Pi_{|G_j|,k}(\cdot|G_j)$ might underestimate uncertainly, the median $\hat{\Pi}^{\mathrm{st}}_{n,g}$ (or $\hat{\Pi}^{\mathrm{st}}_{n,0}$) of these random measures yields credible sets with much better coverage. This approach shows good performance in numerical experiments. One of our main results (see Section 3.2) provides a justification for this observation, albeit, under rather strong assumptions and for the parametric case.

## 3. Theoretical Analysis

We proceed with a discussion of the theoretical guarantees for the M-Posterior, starting with the contraction rates and robustness properties.

### 3.1 Convergence of Posterior Distribution and Robust Bayesian Inference

Our first result establishes the "weak concentration" property of the posterior distribution around the true parameter. Let $\delta_0 := \delta_{\theta_0}$ be the Dirac measure supported on $\theta_0 \in \Theta$. Recall the following version of Theorem 2.1 in Ghosal et al. (2000) (we state the result for the Wasserstein distance $d_{W_1,\rho}(\Pi_n(\cdot|\mathcal{X}_l), \delta_0)$ rather than the closely related contraction rate of the posterior distribution). Here, the Wasserstein distance is evaluated with respect to the "Hellinger metric" $\rho(\cdot,\cdot)$ defined in (14).

**Theorem 7** *Let $\mathcal{X}_l = \{X_1, \dots, X_l\}$ be an i.i.d. sample from $P_0$. Assume that $\varepsilon_l > 0$ and $\Theta_l \subset \Theta$ are such that for some constant $C > 0$*

(1) *the packing number satisfies* $\log M(\varepsilon_l, \Theta_l, \rho) \le l\varepsilon_l^2$,

(2) $\Pi(\Theta \setminus \Theta_l) \le \exp(-l\varepsilon_l^2(C+4))$,

(3) $\Pi\left(\theta : \ -P_0\left(\log \frac{p_\theta}{p_0}\right) \le \varepsilon_l^2, \ P_0\left(\log \frac{p_\theta}{p_0}\right)^2 \le \varepsilon_l^2\right) \ge \exp(-Cl\varepsilon_l^2)$.

*Then there exists $R = R(C)$ and a universal constant $\tilde{K}$ such that*

$$\Pr\left(d_{W_{1,\rho}}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)) \ge R\varepsilon_l + e^{-\tilde{K}l\varepsilon_l^2}\right) \le \frac{1}{l\varepsilon_l^2} + 4e^{-\tilde{K}l\varepsilon_l^2}. \tag{19}$$

**Proof** The proof closely mimics the argument behind Theorem 2.1 in (Ghosal et al., 2000). For reader's convenience, details are outlined in Section A.2 of the appendix. ■

Conditions of Theorem 7 are standard assumptions guaranteeing that the resulting posterior distribution contracts to the true parameter $\theta_0$ at the rate $\varepsilon_n$. Note that the bounds for the distance $d_{W_{1,\rho}}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)$ slightly differ from the contraction rate itself: indeed, we have

$$d_{W_{1,\rho}}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)) \le \varepsilon_l + \int\limits_{h(P_\theta, P_0) \ge \varepsilon_l} d\Pi_l(\cdot|\mathcal{X}_l),$$

hence to obtain the inequality $d_{W_{1,\rho}}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)) \lesssim \varepsilon_l$, we usually require $\int\limits_{h(P_\theta, P_0) \ge \varepsilon_l} d\Pi_l(\cdot|\mathcal{X}_l) \lesssim \varepsilon_l$, which adds an extra logarithmic factor in the parametric case.

Combination of Theorems 7 and 1 immediately yields the corollary for $\hat{\Pi}_{n,0}$. Let $\mathbb{H}$ be the reproducing kernel Hilbert space with the reproducing kernel

$$k_H(\theta_1, \theta_2) = \frac{1}{2}\left\langle \sqrt{p_{\theta_1}}, \sqrt{p_{\theta_2}} \right\rangle_{L_2(dx)}.$$

Let $f \in \mathbb{H}$ and note that, due to the reproducing property and Cauchy-Schwarz inequality, we have

$$\begin{aligned} f(\theta_1) - f(\theta_2) &= \langle f, k_H(\cdot, \theta_1) - k_H(\cdot, \theta_2)\rangle_{\mathbb{H}} \\ &\le \|f\|_{\mathbb{H}} \left\|k_H(\cdot, \theta_1) - k_H(\cdot, \theta_2)\right\|_{\mathbb{H}} = \|f\|_{\mathbb{H}} \ \rho(\theta_1, \theta_2). \end{aligned} \tag{20}$$

Therefore, $\mathcal{F}_k \subseteq \mathcal{F}_L$ and $\|P - Q\|_{\mathcal{F}_k} \le \|P - Q\|_{\mathcal{F}_L}$, where $\mathcal{F}_k$ and $\mathcal{F}_L$ were defined in (10) and (8) respectively, and the underlying metric structure is given by $\rho$. In particular, convergence with respect to $\|\cdot\|_{\mathcal{F}_L}$ implies convergence with respect to $\|\cdot\|_{\mathcal{F}_k}$.

**Corollary 8** *Let $X_1, \dots, X_n$ be an i.i.d. sample from $P_0$, and assume that $\hat{\Pi}_{n,g}$ is defined with respect to the norm $\|\cdot\|_{\mathcal{F}_L}$ as in (17) above. Set $l := \lfloor n/m \rfloor$, assume that conditions of Theorem 7 hold, and, moreover, that $\varepsilon_l$ satisfies*

$$\frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2} < \frac{1}{7}.$$

*Then*

$$\Pr\Big( \Big\| \delta_0 - \hat{\Pi}_{n,g} \Big\|_{\mathcal{F}_{k_H}} \geq 1.52 \Big( R\varepsilon_l + e^{-\tilde{K} l \varepsilon_l^2} \Big) \Big) \leq \Big[ e^{\psi(3/7,1/7)} \Big]^{-m} < 1.27^{-m}.$$

**Proof** It is enough to apply part (a) of Theorem 1 with $\kappa = 0$ to the independent random measures $\Pi_n(\cdot|G_j),\ j = 1, \dots, m$. Note that the "weak concentration" assumption (29) is implied by (19). ■

Once again, note the exponential improvement of concentration as compared to Theorem 7. It is easy to see that a similar statement holds for the median $\hat{\Pi}_{n,0}(\cdot)$ defined in (17) (even for the stronger Wasserstein distance $d_{W_{1,\rho}}(\delta_0, \hat{\Pi}_{n,0})$), modulo changes in constants.

**Remark 9** *The case when the sample $\mathcal{X}_n = \{X_1, \dots, X_n\}$ contains $\lfloor \kappa m \rfloor$ outliers (which can be completely arbitrary vectors in $\mathbb{R}^D$) for some $\kappa < 1/3$ can be handled similarly. In most examples throughout the paper, we state the results for the case $\kappa = 0$ for simplicity, keeping in mind that the generalization is a trivial corollary of Theorem 1. For example, if we allow $\lfloor \kappa m \rfloor$ outliers in the setup of Corollary 8, the resulting bounds becomes*

$$\Pr\Big( \Big\| \delta_0 - \hat{\Pi}_{n,g} \Big\|_{\mathcal{F}_{k_H}} \geq 1.52 \Big( R\varepsilon_l + e^{-\tilde{K} l \varepsilon_l^2} \Big) \Big) \leq \Big[ e^{(1-\kappa)\psi\left(\frac{3/7-\kappa}{1-\kappa}, 1/7\right)} \Big]^{-m}.$$

While the result of the previous statement is promising, numerical approximation and sampling from the "robust posterior" $\hat{\Pi}_{n,g}$ is often problematic due to the underlying geometry defined by the Hellinger metric, and the associated distance $\|\cdot\|_{\mathcal{F}_{k_H}}$ is hard to estimate in practice. Our next goal is to derive similar guarantees for the M-Posterior evaluated with respect to the computationally tractable family of distances discussed in Section 2.3 above.

To transfer the conclusions of Theorem 7 and Corollary 8 to the case of other kernels $k(\cdot,\cdot)$ and associated metrics $\rho_k(\cdot,\cdot)$, we need to guarantee the existence of tests versus the complements of the balls in these distances. Such tests can be obtained from comparison inequalities between distances.

**Assumption 1** *There exists $\gamma > 0$, $r(\theta_0) > 0$ and $\tilde{C}(\theta_0) > 0$ satisfying*

$$d(\theta, \theta_0) \geq \tilde{C}(\theta_0) \rho_k^{\gamma}(\theta, \theta_0) \ \text{ whenever } d(\theta, \theta_0) \leq r(\theta_0),$$

*where $d$ is the Hellinger distance or the Euclidean distance (in the parametric case).*

**Remark 10** *When $d$ is the Euclidean distance, we will impose an additional mild assumption guaranteeing existence of test versus the complements of the balls (for the Hellinger distance, this is always true, see Ghosal et al., 2000). Namely, we will assume that for every $n$ and every pair $\theta_1, \theta_2 \in \Theta$, there exists a test $\phi_n := \phi_n(X_1, \dots, X_n)$ such that for some $\gamma > 0$ and a universal constant $K > 0$*

$$\begin{aligned} \mathbb{E}_{P_{\theta_1}} \phi_n &\leq e^{-Knd^2(\theta_1,\theta_2)}, \\ \sup_{d(\theta,\theta_2) < d(\theta_1,\theta_2)/2} \mathbb{E}_{P_\theta} (1 - \phi_n) &\leq e^{-Knd^2(\theta_1,\theta_2)}. \end{aligned} \tag{21}$$

Below, we provide examples of kernels satisfying the stated assumption.

**Example 1 (Exponential families)** *Let* $\{P_\theta,\ \theta \in \Theta \subseteq \mathbb{R}^p\}$ *be of the form*

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

*where* $\langle\cdot,\cdot\rangle_{\mathbb{R}^p}$ *is the standard Euclidean dot product. Then the Hellinger distance can be expressed as (Nielsen and Garcia, 2011)*

$$h^2(P_{\theta_1}, P_{\theta_2}) = 1 - \exp\Big( -\frac{1}{2}\Big(G(\theta_1) + G(\theta_2) - 2G\Big(\frac{\theta_1+\theta_2}{2}\Big)\Big)\Big).$$

*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*

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

*hence assumption 1 holds with* $d$ *being the Hellinger distance,* $\gamma = 1$*,* $\tilde{C} = \frac{1}{\sqrt{2}}$ *and* $r(\theta_0) \equiv 1$ *for*

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

For finite-dimensional models, we will be especially interested in kernels $k(\cdot,\cdot)$ such that the associated metric $\rho_k(\cdot,\cdot)$ is bounded by the Euclidean distance. The following proposition gives a sufficient condition for this to hold.

Assume that kernel $k(\cdot,\cdot)$ satisfies conditions of Proposition 4 (in particular, $k$ is characteristic). Recall that by Bochner's theorem, there exists a finite nonnegative Borel measure $\nu$ such that $k(\theta) = \int\limits_{\mathbb{R}^p} e^{i\langle x,\theta\rangle} d\nu(x)$.

**Proposition 11** *Assume that* $\int\limits_{\mathbb{R}^p} \|x\|_2^2 d\nu(x) < \infty$*. Then there exists* $D_k > 0$ *depending only on* $k$ *such that for all* $\theta_1, \theta_2$*,*

$$\rho_k(\theta_1, \theta_2) \leq D_k \left\|\theta_1 - \theta_2\right\|_2.$$

**Proof** For all $z \in \mathbb{R}$, $|e^{iz} - 1 - iz| \leq \frac{|z|^2}{2}$, implying that

$$\begin{aligned}
\rho_k^2(\theta_1,\theta_2) &= \|k(\cdot,\theta_1) - k(\cdot,\theta_2)\|_{\mathbb{H}}^2 \\
&= 2k(0) - 2k(\theta_1-\theta_2) = 2\int\limits_{\mathbb{R}^p} (1 - e^{i\langle x, \theta_1-\theta_2\rangle}) d\nu(x) \\
&\leq \int\limits_{\mathbb{R}^p} \langle x, \theta_1-\theta_2\rangle_{\mathbb{R}^p}^2 \, d\nu(x) \leq \|\theta_1-\theta_2\|_2^2 \int\limits_{\mathbb{R}^p} \|x\|_2^2 d\nu(x).
\end{aligned}$$

■

Moreover, the result of the previous proposition clearly remains valid for kernels of the form

$$\tilde{k}(\theta_1, \theta_2) = k(\theta_1 - \theta_2) + c \langle \theta_1, \theta_2 \rangle_{\mathbb{R}^p}, \tag{22}$$

where $c > 0$ and $k$ satisfies the assumptions of proposition 11. For such a kernel, we have the obvious lower bound $\rho_{\tilde{k}}(\theta_1, \theta_2) \geq \sqrt{c} \|\theta_1 - \theta_2\|_2$, hence $\rho_{\tilde{k}}$ is equivalent (in the strong sense) to the Euclidean distance.

We are ready to state our main result for convergence with respect to the RKHS-induced distance $\|\cdot\|_{\mathcal{F}_k}$.

**Theorem 12** *Assume that conditions of Theorem 7 hold with $\rho$ being the Hellinger or the Euclidean distance, and that assumption 1 is satisfied. In addition, let prior $\Pi$ be such that*

$$D_k^W := \int_\Theta \rho_k^W(\theta, \theta_0) d\Pi(\theta) < \infty$$

*for a sufficiently large $W$ (it follows from the proof that $W = \frac{4}{3} + \frac{4+2C}{3\tilde{K}}$ is sufficient, with $C$ and $\tilde{K}$ being the constants from the statement of Theorem 7). Then there exists a sufficiently large $R = R(\theta_0, \gamma) > 0$ and an absolute constant $\tilde{K}$ such that*

$$\Pr\left( \|\delta_0 - \Pi_l(\cdot|\mathcal{X}_l)\|_{\mathcal{F}_k} \geq R\varepsilon_l^{1/\gamma} + D_k e^{-\tilde{K} l \varepsilon_l^2/2} \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-\tilde{K} l \varepsilon_l^2}. \tag{23}$$

**Proof** The result essentially follows from the combination of Theorem 7 and assumption 1, see Section A.3 for details. ■

Theorem 12 yields the "weak" estimate that is needed to obtain the stronger bound for the M-Posterior distribution $\hat{\Pi}_{n,g}$. This is summarized in the following corollary:

**Corollary 13** *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 distance $\|\cdot\|_{\mathcal{F}_k}$ as in (17) above. Let $l := \lfloor n/m \rfloor$. Assume that conditions of Theorem 12 hold, and, moreover, $\varepsilon_l$ is such that*

$$\frac{1}{l\varepsilon_l^2} + 4e^{-\tilde{K} l \varepsilon_l^2} < \frac{1}{7}.$$

*Then*

$$\Pr\left( \left\|\delta_0 - \hat{\Pi}_{n,g}\right\|_{\mathcal{F}_k} \geq 1.52 \left( R\varepsilon_l^{1/\gamma} + D_k e^{-\tilde{K} l \varepsilon_l^2/2} \right) \right) \leq 1.27^{-m}. \tag{24}$$

**Proof** It is enough to apply parts (a) and (b) of Theorem 1 with $\kappa = 0$ to the independent random measures $\Pi_{|G_j|}(\cdot|G_j)$, $j = 1, \ldots, m$. Note that the "weak concentration" assumption (28) is implied by (23). ■

Note that if $\Theta \subseteq \mathbb{R}^p$ and kernel $k(\cdot, \cdot)$ is of the form (22), the previous corollary together with proposition 6 imply that

$$\Pr\left( \left\|\theta_* - \theta_0\right\|_2 \geq \frac{1.52R}{c} \varepsilon_l \right) \leq 1.27^{-m},$$

where $\theta_* = \int_\Theta \theta d\hat{\Pi}_{n,g}(\theta)$ is the mean of $\hat{\Pi}_{n,g}$. In other words, this shows that the M-Posterior mean is the robust estimator of $\theta_0$.

### 3.2 Bayesian Inference Based on Stochastic Approximation of the Posterior Distribution

As we have already mentioned in Section 2.4, when the number of disjoint subgroups $m$ is large, the resulting M-Posterior distribution is "too flat", which results in large credible sets and overestimation of uncertainty. Clearly, the source of the problem is the fact that each individual random measure $\Pi_{|G_j|}(\cdot|G_j),\ j=1,\dots,m$ is based on sample of size $l \simeq \frac{n}{m}$ which can be much smaller than $n$.

One way to reduce the variance of each subset posterior distribution is to repeat each observation in $G_j$ $m$ times (although other alternatives, such as bootstrap, are possible; for instance, bootstrap-related techniques have been investigated by Kleiner et al. (2014) in the frequentist setting), $\tilde{G}_j = \underbrace{\{G_j, \dots, G_j\}}_{m \text{ times}}$. Formal application of the Bayes rule in this situation yields a collection of new measures on the parameter space:

$$\Pi_{|G_j|,m}(B|G_j) := \frac{\int\limits_B \left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}{\int\limits_\Theta \left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)},$$

where we have assumed that $p_\theta(\cdot)$ is integrable. Here, $\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m$ can be viewed as an approximation of the full data likelihood. We call the random measure $\Pi_{|G_j|,m}(\cdot|G_j)$ the *$j$-th stochastic approximation* to the full posterior distribution.

Of course, such a "correction" negatively affects coverage properties of the credible sets associated with each measure $\Pi_{|G_j|}(\cdot|G_j)$. However, taking the median of stochastic approximations yields improved coverage of the resulting M-Posterior distribution. The main goal of this section is to establish an asymptotic statement in spirit of a Bernstein-von Mises theorem for the M-Posterior based on stochastic approximations $\Pi_{|G_j|,m}(B|G_j),\ j=1,\dots,m$.

We will start by showing that under certain assumptions the upper bounds for the convergence rates of $\Pi_{|G_j|,m}(\cdot|G_j)$ towards $\delta_0$ are the same as for $\Pi_{|G_j|}(\cdot|G_j)$, the "standard" posterior distribution given $G_j$.

For $A \subseteq \Theta$, let $N_{[]}(u, A, d)$ be the bracketing number of $\{p_\theta,\ \theta \in A\}$ with respect to the distance $d(l,u) := \int\limits_{\mathbb{R}^D} \left(\sqrt{l(x)} - \sqrt{u(x)}\right)^2 dx$, and let

$$H_{[]}(u; A) := \log N_{[]}(u, A, d)$$

be the *bracketing entropy*. In what follows, $B(\theta_0, r) := \{\theta \in \Theta :\ h(P_\theta, P_{\theta_0}) \le r\}$ denotes the "Hellinger ball" of radius $r$ centered at $\theta_0$.

**Theorem 14 (Wong et al., 1995, Theorem 1)** *There exist constants $c_j,\ j=1,\dots,4$ and $\zeta > 0$ such that if*

$$\int\limits_{\zeta^2/2^8}^{\sqrt{2}\zeta} H_{[]}^{1/2}\left(u/c_3; B(\theta_0, \zeta\sqrt{2})\right) du \le c_4\sqrt{l}\zeta^2,$$

*then*

$$P\left(\sup_{\theta:h(P_\theta,P_0)\geq\zeta}\prod_{j=1}^{l}\frac{p_\theta}{p_0}(X_j)\geq e^{-c_1 l\zeta^2}\right)\leq 4e^{-c_2 l\zeta^2}.$$

*In particular, one can choose* $c_1 = 1/24,\ c_2 = (4/27)(1/1926),\ c_3 = 10$ *and* $c_4 = (2/3)^{5/2}/512$.

In "typical" parametric problems ($\Theta\subseteq\mathbb{R}^p$), the bracketing entropy can be bounded as $H_{[]}(u;B(\theta_0,r))\leq C_1\log(C_2 r/u)$, whence the minimal $\zeta$ that satisfies conditions of Theorem 14 is of order $\zeta\simeq\sqrt{\frac{1}{l}}$. In particular, it is easy to check (e.g., using Theorem 2.7.11 in van der Vaart and Wellner, 1996) that this is the case when

(a) there exists $r_0>0$ such that

$$h\left(P_\theta,P_{\theta_0}\right)\geq K_1\|\theta-\theta_0\|_2$$

whenever $h\left(P_\theta,P_{\theta_0}\right)\leq r_0$, and

(b) there exists $\alpha>0$ such that for $\theta_1,\theta_2\in B(\theta_0,r_0)$,

$$|p_{\theta_1}(x)-p_{\theta_2}(x)|\leq F(x)\left\|\theta_1-\theta_2\right\|_2^\alpha$$

with $\int_{\mathbb{R}^D}F(x)dx<\infty$.

Application of theorem 14 to the analysis of "stochastic approximations" yields the following result.

**Theorem 15**
*Let* $\varepsilon_l>0$ *be such that conditions of Theorem 14 hold with* $\zeta:=\varepsilon_l$, *and*

*(a) for some* $C>0$

$$\Pi\left(\theta:\ -P_0\left(\log\frac{p_\theta}{p_0}\right)\leq\varepsilon_l^2,\ P_0\left(\log\frac{p_\theta}{p_0}\right)^2\leq\varepsilon_l^2\right)\geq\exp(-Cl\varepsilon_l^2),$$

*(b) k is a positive-definite kernel that satisfies assumption 1 for the Hellinger distance for some* $\tilde{C}(\theta_0)$ *and* $\gamma>0$.

*Then there exists* $\tilde{R}=\tilde{R}(C,\tilde{C},\gamma)>0$ *such that*

$$\Pr\left(\left\|\delta_0-\Pi_{l,m}(\cdot|\mathcal{X}_l)\right\|_{\mathcal{F}_k}\geq\tilde{R}\varepsilon_l^{1/\gamma}+e^{-ml\varepsilon_l^2}\right)\leq\frac{1}{l\varepsilon_l^2}+4e^{-c_2\tilde{C}^2\tilde{R}^{2\gamma}l\varepsilon_l^2}.$$

**Proof** See Section A.4 in the appendix. ■

**Remark 16** *Note that for the kernel* $k(\cdot,\cdot)$ *of the form (22), assumption 1 reduces to the inequality between the Hellinger and Euclidean distances.*

As before, Theorem 1 combined with the "weak concentration" inequality of Theorem 15 gives stronger guarantees for the median $\hat{\Pi}^{\text{st}}_{n,g}$ (or its alternative $\hat{\Pi}^{\text{st}}_{n,0}$) of

$$\Pi_{|G_1|,m}(\cdot|G_1), \ldots, \Pi_{|G_m|,m}(\cdot|G_m).$$

Exact statement is very similar in spirit to Corollary 13.

Our next goal is to obtain the result describing the asymptotic behavior of the M-Posterior distribution $\hat{\Pi}^{\text{st}}_{n,0}$ in the parametric case. We start with a result that addresses each individual stochastic approximation $\Pi_{|G_j|,m}(\cdot|G_j),\ j = 1, \ldots, m$. Assume that $\Theta \subseteq \mathbb{R}^p$ has non-empty interior. For $\theta \in \Theta$, let

$$I(\theta) := \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log p_\theta(X)\left(\frac{\partial}{\partial\theta}\log p_\theta(X)\right)^T\right]$$

be the Fisher information matrix (we are assuming that it is well-defined). We will say that the family $\{P_\theta,\ \theta \in \Theta\}$ is *differentiable in quadratic mean* (see Chapter 7 in van der Vaart, 2000 for details) if there exists $\dot{\ell}_{\theta_0} : \mathbb{R}^D \mapsto \mathbb{R}^p$ such that

$$\int_{\mathbb{R}^D}\left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2}h^T\dot{\ell}_{\theta_0}\sqrt{p_{\theta_0}}\right)^2 = o(\|h\|_2^2)$$

as $h \to 0$; usually, $\dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta}\log p_\theta(x)$. Next, define

$$\Delta_{l,\theta_0} := \frac{1}{\sqrt{l}}\sum_{j=1}^{l} I^{-1}(\theta_0)\dot{\ell}_{\theta_0}(X_j).$$

We will first state a preliminary result for each individual "subset posterior" distribution:

**Proposition 17** *Let $X_1, \ldots, X_l$ be an i.i.d. sample from $P_{\theta_0}$ for some $\theta_0$ in the interior of $\Theta$. Assume that*

*(a) the family $\{P_\theta,\ \theta \in \Theta\}$ is differentiable in quadratic mean;*

*(b) the prior $\Pi$ has a density (with respect to the Lebesgue measure) that is continuous and positive in the neighborhood of $\theta_0$;*

*(c) conditions of Theorem 14 hold with $\zeta = \frac{C}{\sqrt{l}}$ for some $C > 0$ and $l$ large enough.*

*Then for any integer $m \geq 1$,*

$$\left\|\Pi_{l,m}(\cdot|X_1, \ldots, X_l) - N\left(\theta_0 + \frac{\Delta_{l,\theta_0}}{\sqrt{l}}, \frac{1}{l\cdot m}I^{-1}(\theta_0)\right)\right\|_{\text{TV}} \to 0$$

*in $P_{\theta_0}$-probability as $l \to \infty$.*

**Proof** The proof follows standard steps (e.g., Theorem 10.1 in van der Vaart, 2000), where the existence of tests is substituted by the inequality of Theorem 14. See Section A.5 in the appendix for details. ■

The implication of this result for the M-Posterior is the following: if $k$ is the kernel of

type (22), for sufficiently regular parametric families (differentiable in quadratic mean, with "well-behaved" bracketing numbers, satisfying assumption 1 for the Euclidean distance with $\gamma = 1$) and regular priors, then

(a) the M-Posterior is well approximated by a normal distribution centered at the "robust" estimator $\theta^*$ of unknown $\theta_0$;

(b) the estimator $\theta^*$ is a center of the confidence set of level $1.15^{-m}$ and diameter of order $\sqrt{\frac{m}{n}}$ (same as we would expect for this level for the usual posterior distribution - however, the bound for the M-Posterior holds for finite sample sizes).

This is formalized below:

**Theorem 18**

*(a) Let $k$ be the kernel of type (22), and suppose that the assumptions of Proposition 17 hold. Moreover, let the prior $\Pi$ be such that $\int_{\mathbb{R}^p} \|\theta\|_2^2 d\Pi(\theta) < \infty$. Then for any fixed $m \geq 1$,*

$$\left\| \hat{\Pi}_{n,0}^{\mathrm{st}} - N\left(\theta^*, \frac{1}{n} I^{-1}(\theta_0)\right) \right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty,$$

*in $P_{\theta_0}$-probability when $n \to \infty$, where $\theta^*$ is the mean of $\hat{\Pi}_{n,0}^{\mathrm{st}}$.*

*(b) Assume that conditions (a), (b) of Theorem 15 hold with*

$$\varepsilon_l \gtrsim \frac{1}{\sqrt{l}} \simeq \sqrt{\frac{m}{n}}$$

*and $\gamma = 1$. Then for all $n \geq n_0$ and $\bar{R}$ large enough,*

$$\Pr\left( \|\theta^* - \theta_0\|_2 \geq \bar{R}\left(\varepsilon_l + e^{-ml\varepsilon_l^2}\right)\right) \leq 1.15^{-m}.$$

**Proof** (a) It is easy to see that convergence in total variation norm, together with an assumption that the prior distribution satisfies

$$\int_{\mathbb{R}^p} \|\theta\|_2^2 d\Pi(\theta) < \infty,$$

implies that the expectations converge in $P_{\theta_0}$-probability as well:

$$\left\| \int_\Theta \theta \left( d\Pi_{l,m}(\theta|\mathcal{X}_l) - dN\left(\theta_0 + \frac{\Delta_{l,\theta_0}}{\sqrt{l}}, \frac{1}{l \cdot m} I^{-1}(\theta_0)\right)(\theta)\right)\right\|_2 \to 0 \text{ as } l \to \infty.$$

Together with an observation that the total variation distance between $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ is bounded by the multiple of $\|\mu_1 - \mu_2\|_2$, it implies that we can replace $\theta_0 + \frac{\Delta_{l,\theta_0}}{\sqrt{l}}$ by the mean

$$\bar{\theta}_{l,m}(X_1, \ldots, X_l) := \int_\Theta \theta d\Pi_{l,m}(\theta|\mathcal{X}_l),$$

in other words, the conclusion of Proposition 17 can be stated as

$$\left\| \Pi_{l,m}(\cdot|\mathcal{X}_l) - N\left(\bar{\theta}_{l,m}, \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } l \to \infty,$$

in $P_{\theta_0}$-probability. Now assume that $m = \lfloor \frac{n}{l} \rfloor$ is fixed, and let $n, l \to \infty$. As before, let $G_1, \ldots, G_m$ be disjoint groups of i.i.d. observations from $P_{\theta_0}$ of cardinality $l$ each. Recall that, by the definition (3) of $\mathrm{med}_0(\cdot)$, $\Pi^{\mathrm{st}}_{n,0} = \Pi_{l,m}(\cdot|\mathcal{X}_{l_*})$ for some $l_* \le m$, and $\theta^* := \theta_{l_*,m}$ is the mean of $\Pi^{\mathrm{st}}_{n,0}$. Clearly, we have

$$\begin{aligned}
&\left\| \Pi^{\mathrm{st}}_{n,0} - N\left(\theta^*, \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)\right\|_{\mathrm{TV}} \le \\
&\max_{j=1,\ldots,m} \left\| \Pi_{l,m}(\cdot|G_j) - N\left(\bar{\theta}_{l,m}(G_j), \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)\right\|_{\mathrm{TV}} \to 0 \text{ as } n \to \infty. \qquad (25)
\end{aligned}$$

(b) Let $\varepsilon_l \ge C\sqrt{\frac{1}{l}}$ where $C$ large enough so that

$$\frac{1}{l\varepsilon_l^2} + 4e^{-c_2 \tilde{C}^2 \tilde{R}^2 l \varepsilon_l^2} \le \frac{1}{4},$$

where $c_2$, $\tilde{R}$ are the same as in Theorem 15. Applying Theorem 15, we get

$$\Pr\left( \|\delta_{\theta_0} - \Pi_{l,m}(\cdot|\mathcal{X}_l)\|_{\mathcal{F}_k} \ge \tilde{R}\varepsilon_l + e^{-ml\varepsilon_l^2}\right) \le \frac{1}{4}.$$

By part (b) of Theorem 1,

$$\left\| \hat{\Pi}^{\mathrm{st}}_{n,0} - \delta_{\theta_0} \right\|_{\mathcal{F}_{\tilde{k}}} \le 3\left(\tilde{R}\varepsilon_l + e^{-ml\varepsilon_l^2}\right)$$

with probability $\ge 1 - 1.15^{-m}$. Since kernel $\tilde{k}$ is of the type (22), proposition (6) implies that

$$\|\theta^* - \theta_0\|_2 \le \left\| \hat{\Pi}^{\mathrm{st}}_{n,0} - \delta_{\theta_0} \right\|_{\mathcal{F}_{\tilde{k}}}, \qquad (26)$$

and the result follows.

■

In particular, for $m = A\log(n)$ and $\varepsilon_l \simeq \sqrt{\frac{m}{n}}$, we obtain the bound

$$\Pr\left( \|\theta^* - \theta_0\|_2 \ge \bar{R}\sqrt{\frac{A\log n}{n}}\right) \le n^{-A}.$$

for some constant $\bar{R}$ *independent* of $m$. Note that $\theta^*$ itself depends on $m$, hence this bound is *not uniform*, and holds only for a given confidence level $1 - n^{-A}$.

It is convenient to interpret this (informally) in terms of the credible sets: to obtain the credible set with "frequentist" coverage level $\ge 1 - n^{-A}$, pick $m = A\log n$ and use the $(1 - n^{-A})$ - credible set of the M-Posterior $\hat{\Pi}^{\mathrm{st}}_{n,0}$.

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

**Input:**

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

**Initialize:**

1. Set $w_j^{(0)} := \frac{1}{m}$, $j = 1 \dots m$;
2. Set $Q_*^{(0)} := \frac{1}{m} \sum\limits_{j=1}^{m} Q_j$;

**repeat**

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

1. Update $w_j^{(t+1)} = \frac{\|Q_*^{(t)} - Q_j\|_{\mathcal{F}_k}^{-1}}{\sum\limits_{i=1}^{m} \|Q_*^{(t)} - Q_i\|_{\mathcal{F}_k}^{-1}}$; (apply (12) to evaluate $\|Q_*^{(t)} - Q_i\|_{\mathcal{F}_k}$);
2. Update $Q_*^{(t+1)} = \sum\limits_{j=1}^{m} w_j^{(t+1)} Q_j$;

**until** $\|Q_*^{(t+1)} - Q_*^{(t)}\|_{\mathcal{F}_k} \le \varepsilon$;
**Return:** $w_* := (w_1^{(t+1)}, \dots, w_m^{(t+1)})$.

## 4. Numerical Algorithms and Examples

In this section, we consider examples and applications in which comparisons are made for the inference based on the usual posterior distribution and on the M-Posterior. One of the well-known and computationally efficient ways to find the geometric median in Hilbert spaces is the famous *Weiszfeld's algorithm*, introduced in Weiszfeld (1936). Details of implementation are described in Algorithms 1 and 2. Algorithm 1 is a particular case of Weiszfeld's algorithm applied to subset posterior distributions and distance $\|\cdot\|_{\mathcal{F}_k}$, while Algorithm 2 shows how to obtain an approximation to M-Posterior given the samples from $\Pi_{n,m}(\cdot|G_j)$, $j = 1 \dots m$. Note that the subset posteriors $\Pi_{n,m}(\cdot|G_j)$ whose "weights" $w_{*,j}$ in the expression of the M-Posterior are small (in our case, smaller than $1/(2m)$) are excluded from the analysis. Our extensive simulations show the empirical evidence in favor of this additional thresholding step.

Detailed discussion of convergence rates and acceleration techniques for Weiszfeld's method from the viewpoint of modern optimization can be found in Beck and Sabach (2013). For alternative approaches and extensions of Weiszfeld's algorithm, see Bose et al. (2003), Ostresh (1978), Overton (1983), Chandrasekaran and Tamir (1990), Cardot et al. (2012), Cardot et al. (2013), among other works.

In all numerical simulations below, we use stochastic approximations and the corresponding M-Posterior $\hat{\Pi}_{n,g}^{\mathrm{st}}$, unless noted otherwise. The posterior $\hat{\Pi}_{n,0}^{\mathrm{st}}$ based on the metric median is mainly of theoretical interest, while numerical results obtained for it are typically inferior compared to the geometric median, so we do not discuss them further.

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

**Input:**

1. Samples $\{Z_{j,i}\}_{i=1}^{S_j} \sim \Pi_{n,m}(\cdot|G_j),\ j = 1 \dots m$ (see equation (18));

**Do:**

1. $Q_j := \frac{1}{S_j} \sum_{i=1}^{S_j} \delta_{Z_{j,i}}$, $j = 1 \dots m$ - empirical approximations of $\Pi_{n,m}(\cdot|G_j)$.
2. Apply Algorithm 1 to $Q_1, \dots, Q_m$; return $w_* = (w_{*,1} \dots w_{*,m})$;
3. For $j = 1, \dots, m$, 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:** $\hat{\Pi}_{n,g}^{\text{st}} := \sum_{i=1}^m \hat{w}_i^* Q_i$.

Before presenting the results of numerical analysis, let us remark on two important computational aspects.

**Remark 19**

*(a) The number of subsets $m$ appears as a "free parameter" entering the theoretical guarantees for M-Posterior. One interpretation of $m$ (in terms of the credible sets) is given in the end of Section 3.2. Our results also imply that partitioning the data into $m = 2k + 1$ subsets guarantees robustness to the presence of $k$ outliers of arbitrary nature.*

*In many applications, $m$ is dictated by the sample size and computational resources (e.g., the number of available machines). In Section A.6, we describe a heuristic approach to selection of $m$ that shows good practical performance. As a rule of a thumb, we recommend choosing $m \lesssim \sqrt{n}$ as larger values of $m$ lead to an M-Posterior that overestimates uncertainty. This heuristic is supported by the numerical results presented below.*

*(b) It is easy to get a general idea regarding the potential improvement in computational time complexity achieved by the M-Posterior. Given the data set $\mathcal{X}_n = \{X_1, \dots, X_n\}$ of size $n$, let $t(n)$ be the running time of the algorithm (e.g., MCMC) that outputs a single observation from the posterior distribution $\Pi_n(\cdot|\mathcal{X}_n)$. If the goal is to obtain $S$ samples from the posterior, then the total running time is $O(S \cdot t(n))$. Let us compare this time with the running time needed to obtain $S$ samples from the $M$-posterior given that the algorithm is running on $m$ machines in parallel. In this case, we need to generate $O(S)$ samples from each of $m$ subset posteriors, which is done in time $O\left(S \cdot t\left(\frac{n}{m}\right)\right)$, where $S$ is typically large and $m \ll n$. According to Theorem 7.1 in Beck and 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 $O(S^2)$ (which follows from (12)), so that the total running time is*

$$O\left(S \cdot t\left(\frac{n}{m}\right) + \frac{S^2}{\varepsilon}\right). \tag{27}$$

*The term $\frac{S^2}{\varepsilon}$ can be refined in several ways via application of more advanced optimization techniques (see the aforementioned references). If, for example, $t(n) \simeq n^r$ for some $r \geq 1$, then $\frac{S}{m} \cdot t\left(\frac{n}{m}\right) \simeq \frac{1}{m^{1+r}} S n^r$ which should be compared to $S \cdot n^r$ required by the standard approach.*

*To give a specific example, consider an application of* (27) *in the context of Gaussian process (GP) regression. If* $n$ *is the number of training samples, then GP regression has* $O(n^3) + O(Sn^2)$ *asymptotic time complexity to obtain* $S$ *samples from the posterior distribution of GP (Rasmussen and Williams, 2006, Algorithm 2.1). Assuming we have access to* $m$ *machines, the time complexity to obtain* $S$ *samples from M-Posterior in GP regression is* $O\left(\left(\frac{n}{m}\right)^3 + S\left(\frac{n}{m}\right)^2 + \frac{S^2}{\varepsilon}\right)$. *If for example* $S = cn$ *for some* $c > 0$ *and* $m^2 < n\varepsilon$, *we get* $O(m^2)$ *improvement in running time.*

*(c) In many cases, replacing the "subset posterior" by the stochastic approximation does not result in increased sampling complexity: indeed, the log-likelihood in the sampling algorithm for the subset posterior is simply multiplied by* $m$ *to obtain the sampler for the stochastic approximation. We have included the description of a modified Dirichlet mixture model in the supplementary material as an illustration.*

### 4.1 Numerical Analysis: Simulated Data

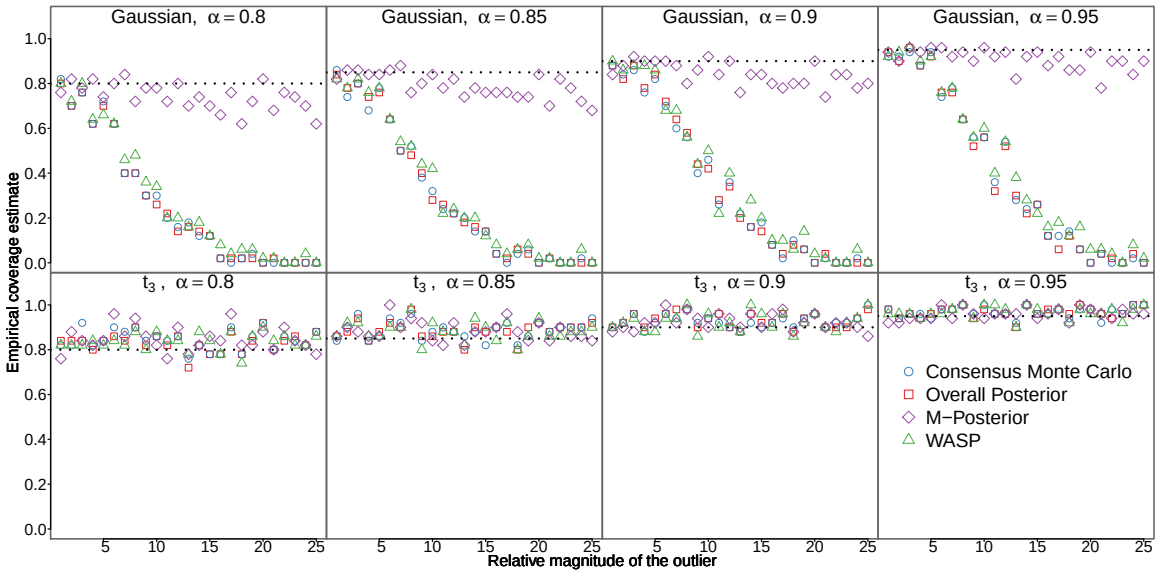



Figure 1: Effect of the outlier on empirical coverage of (1-$\alpha$)100% credible intervals (CIs). The x-axis represents the outlier magnitude. The y-axis represents the empirical coverage computed from 50 simulation replications. The dotted horizontal lines show the theoretical frequentist coverage.

This section demonstrates the effect of magnitude of an outlier on the posterior distribution of the mean parameter $\mu$. We empirically show that M-Posterior of $\mu$ is a robust alternative to the overall posterior. To this end, we used the simplest univariate Gaussian model $\{P_\mu = \mathcal{N}(\mu, 1),\ \mu \in \mathbb{R}\}$.

We simulated 25 data sets containing 200 observations each. Each data set $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,200})$ contained 199 independent observations from the standard Gaussian distribution ($x_{i,j} \sim$

$\mathcal{N}(0,1)$ for $i = 1,\ldots,25$ and $j = 1,\ldots,199$). The last entry in each data set, $x_{i,200}$, was an outlier, and its value increased linearly for $i = 1,\ldots,25$: $x_{i,200} = i\max(|x_{i,1}|,\ldots,|x_{i,199}|)$. The index of outlier was unknown to every sampling algorithm. The true variance of observations was fixed at 1 and was assumed to be known. The algorithms for sampling from the subset and overall posterior distributions of $\mu$ were implemented using Stan language (Carpenter et al., 2016) based on Stan's default Gaussian prior on $\mu$. We used two likelihoods for the data, one was the standard Gaussian likelihood and the other was a (more robust) Student's $t$-distribution likelihood with 3 degrees of freedom ($t_3$). We generated 1000 samples from each posterior distribution $\Pi_{200}(\cdot|\,\mathbf{x}_i)$ for $i = 1,\ldots,25$. Setting $m = 10$ in Algorithm 1, we generated 1000 samples from every subset posterior $\Pi_{200,10}(\cdot|G_{j,i})$, $j = 1,\ldots,10$ to form the empirical measures $Q_{j,i}$; here, $\cup_{j=1}^{10} G_{j,i} = \mathbf{x}_i$. Using these $Q_{j,i}$s, Algorithm 2 generated 10000 samples from the M-Posterior $\hat{\Pi}^{\text{st}}_{2000,g}(\cdot|\,\mathbf{x}_i)$ for each $i = 1,\ldots,25$. This process was replicated 50 times.

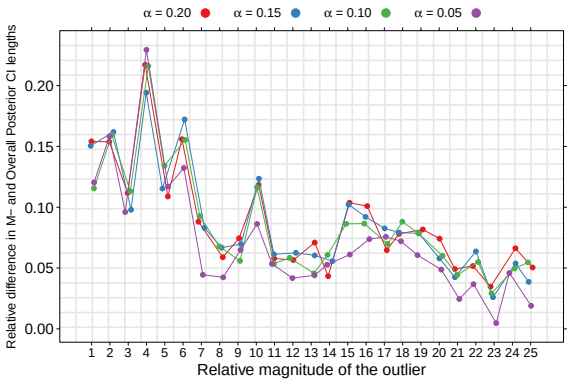



Figure 2: Calibration of uncertainty quantification of M-Posterior. The x-axis represents the outlier magnitude that increases from 1 to 25. The y-axis represents the relative difference between M-Posterior and overall posterior CI lengths. A value close to 0 represents that the M-Posterior CIs are well-calibrated.

We used Consensus MCMC (Scott et al., 2013) and WASP (Srivastava et al., 2015a) as representative methods for scalable MCMC approach, and compared their performance with M-Posterior. Across 50 simulation replications with the Gaussian and $t_3$ likelihood, we estimated the empirical coverage of (1-$\alpha$)100% credible intervals (CIs) for the "consensus posterior", WASP, the overall posterior, and the M-Posterior for $\alpha = 0.2, 0.15, 0.10$, and 0.05. The empirical coverages of M-Posterior's CIs showed robustness to magnitude of the outlier in the Gaussian likelihood model. On the contrary, performance of the WASP, consensus posterior and overall posterior deteriorated fairly quickly across all $\alpha$'s leading to 0% empirical coverage as magnitude of the outlier increased from $i = 1$ to $i = 25$ (Figure 1). When the more robust $t_3$ likelihood was used, empirical coverages of CIs computed using the four methods were close to their theoretical values for all four $\alpha$'s. We compared uncertainty quantification of the M-Posterior with that of the overall posterior using relative lengths of their CIs, with zero value corresponding to identical lengths and a positive value to wider CIs of the M-Posterior. We found that widths of CIs for both posteriors were

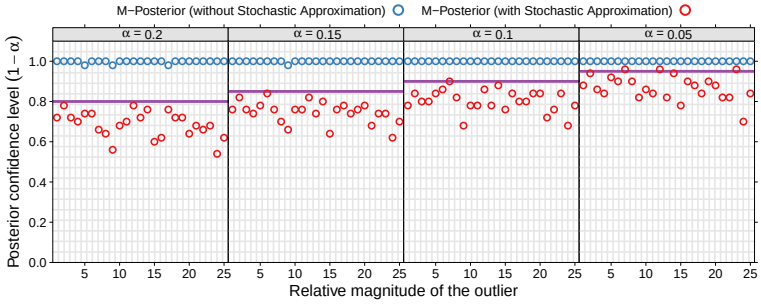



Figure 3: Effect of stochastic approximation on empirical coverage of (1-$\alpha$)100% CIs. The x-axis represents the outlier magnitude that increases from 1 to 25. The y-axis represents the fraction of times the CIs of M-Posteriors with and without stochastic approximation include the true mean over 50 replications. The horizontal lines (in violet) show the theoretical frequentist coverage.

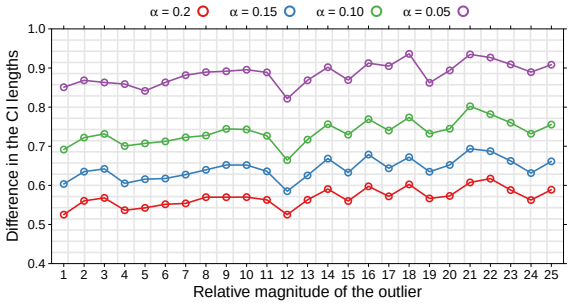
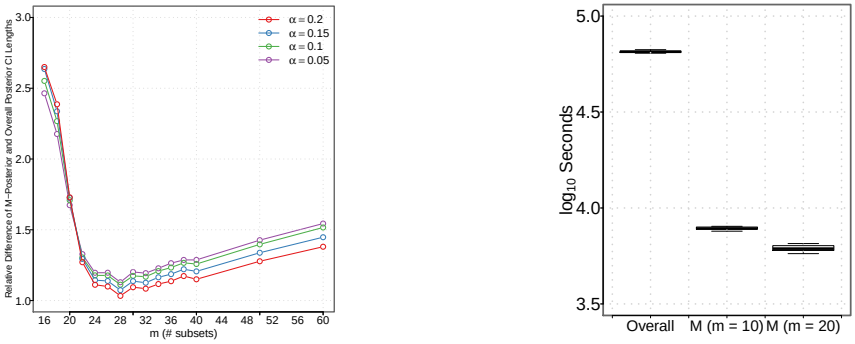



Figure 4: Effect of stochastic approximation on the length of (1-$\alpha$)100% CIs. The x-axis represents the outlier magnitude that increases from 1 to 25. The y-axis represents the differences in the lengths of the CIs of M-Posteriors without and with stochastic approximation.

fairly similar for $i = 1, \ldots, 25$, with M-Posterior's CIs being slightly wider in absence of large outliers (Figure 2).

Stochastic approximation was important for proper calibration of uncertainty quantification. The empirical coverages of (1-$\alpha$)100% CIs of the M-Posterior without stochastic approximation overcompensated for uncertainty at all levels of $\alpha$ (Figure 3). Similarly, lengths of the CIs of M-Posterior without stochastic approximation are wider than those with stochastic approximation (Figure 4). Both these observations showed that stochastic approximation led to shorter CIs for M-Posterior that had empirical coverages close to their theoretical values.

The number of subsets $m$ had an effect on credible interval length of the M-Posterior. We modified the simulation above and generated 1000 observations $\mathbf{x}$, with the last 10 observations in $\mathbf{x}$ being outliers with value $x_j = 25 \max(|x_1|, \ldots, |x_{990}|)$, $j = 991, \ldots, 1000$. The simulation setup was replicated 50 times. M-Posteriors were obtained for $m = 16, 18, \ldots, 40, 50, 60$. Across all values of $m$, M-Posterior's CI was compared to the CI of the overall posterior




(a) Sensitivity to the choice of $m$

(b) Time

Figure 5: (a) Effect of $m$ on the length of (1-$\alpha$)100% CIs. The x-axis represents different choices of $m$. The y-axis represents the relative difference between M-Posterior and overall posterior CI lengths (median across all 25 outlier magnitudes and 50 replications). (b) Computation time to estimate overall posterior and M-Posterior (M) with $m = 10$ and $m = 20$ in real data analysis.

after removing the outliers; the relative difference of M-Posterior and the overall posterior CI lengths decreases for $m \geq 22 > 2k$, where $k = 10$ is the number of outliers, remains stable as $m$ increases to $m = 38$, and grows for larger values of $m$ (Figure 5a). This demonstrates that inference based on M-Posterior was not too sensitive to the choice of $m$ for a wide range of values.

### 4.2 Real Data Analysis: General Social Survey

The General Social Survey (GSS; `gss.norc.org`) has collected responses to questions about evolution of American society since 1972. We selected data for 9 questions from different social topics: "happy" (happy), "Bible is a word of God" (bible), "support capital punishment" (cap), "support legalization of marijuana" (grass), "support premarital sex" (pre), "approve bible prayer in public schools" (prayer), "expect US to be in world war in 10 years" (uswar), "approve homosexual sex relations" (homo), "support abortion" (abort). These questions were in the survey since 1988 and their answers were converted to two levels: *yes* or *no*. Missing data were imputed based on the average, resulting in a data set with approximately 28,000 respondents.

We use a Dirichlet process (DP) mixture of product multinomial distributions, probabilistic parafac (p-parafac), to model the multivariate dependence among the 9 questions. Let $c_k \in \{\text{yes}, \text{no}\}$ represent the response to $k$th question, $k = 1, \ldots, 9$, then $\boldsymbol{\pi}_{c_1,\ldots,c_9}$ is the joint probability of response $c = (c_1, \ldots, c_9)$ for the 9 questions. Using $\boldsymbol{\pi}_{c_1,\ldots,c_9}$, we estimated the joint probability of response to two questions $\boldsymbol{\pi}_{c_i,c_j}$ for every $i$ and $j$ in $\{1, \ldots, 9\}$; see Section 6 of the supplementary material for the p-parafac generative model and sampling algorithms. The GSS data were randomly split into 10 test and training data sets. Samples from the overall posteriors of $\boldsymbol{\pi}_{c_i,c_j}$s were obtained using the Gibbs sampling algorithm of Dunson and Xing (2009). We chose $m$ as 10 and 20 and estimated M-Posteriors for $\boldsymbol{\pi}_{c_i,c_j}$s in

four steps: training data were randomly split into $m$ subsets, samples from subset posteriors were obtained after modifying the original sampler using stochastic approximation, weights of subsets posteriors were estimated using Algorithm 2, and atoms with estimated weights below $\frac{1}{2m}$ were removed.

M-Posterior had similar uncertainty quantification as the overall posterior while being more efficient. M-Posterior was at least 10 ($m = 20$) and 8 times ($m = 10$) faster than the overall posterior and it used less than 25% of the memory resources required by the overall posterior (Figure 5b). The lengths of credible intervals for $\boldsymbol{\pi}_{c_i,c_j}$s obtained using the overall posterior and the two M-Posteriors were very similar, but the overall posterior's coverage of the maximum likelihood estimators for $\boldsymbol{\pi}_{c_i,c_j}$s obtained from the test data was worse than that of the two M-Posteriors (Table 1).

| Empirical Coverage of (1-$\alpha$)100% Credible Intervals | | | | |
|---|---|---|---|---|
| $\alpha$ | 0.05 | 0.10 | 0.15 | 0.20 |
| Overall Posterior | 0.56 (0.10) | 0.52 (0.01) | 0.49 (0.02) | 0.46 (0.02) |
| M-Posterior ($m = 10$) | 0.71 (0.02) | 0.65 (0.04) | 0.62 (0.04) | 0.60 (0.04) |
| M-Posterior ($m = 20$) | 0.74 (0.03) | 0.70 (0.03) | 0.66 (0.03) | 0.63 (0.03) |
| Length of (1-$\alpha$)100% Credible Intervals (in $10^{-2}$) | | | | |
| $\alpha$ | 0.05 | 0.10 | 0.15 | 0.20 |
| Overall Posterior | 1.20 (0.12) | 1.08 (0.11) | 1.00 (0.10) | 0.95 (0.10) |
| M-Posterior ($m = 10$) | 1.70 (0.13) | 1.54 (0.12) | 1.43 (0.11) | 1.35 (0.10) |
| M-Posterior ($m = 20$) | 1.88 (0.11) | 1.70 (0.10) | 1.59 (0.09) | 1.49 (0.09) |

Table 1: Empirical coverage and lengths of (1-$\alpha$)100% credible intervals. The results are averaged across all joint probabilities $\boldsymbol{\pi}_{c_i,c_j}$s and 10 folds of cross-validation. Monte Carlo errors are in parentheses.

## Acknowledgments

Authors were partially supported by grant R01-ES-017436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). Stanislav Minsker acknowledges support from NSF grants DMS-1712956, FODAVA CCF-0808847, DMS-0847388, ATD-1222567. Lizhen Lin was partially supported by NSF CAREER DMS-1654579 and IIS-1663870.

## Appendix A. Proofs.

### A.1 Proof of Theorem 1

We will prove a slightly more general result:

**Theorem 20**

**a** *Assume that* $(\mathbb{H}, \|\cdot\|)$ *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 variables. Let the constants* $\alpha, q, \gamma$ *be such that* $0 < q < \alpha < 1/2$, *and* $0 \leq \gamma < \frac{\alpha - q}{1-q}$. *Suppose* $\varepsilon > 0$ *is such that for all* $j,\ 1 \leq j \leq \lfloor (1-\gamma)m \rfloor + 1$,

$$\Pr\left(\|\hat{\theta}_j - \theta_0\| > \varepsilon\right) \leq q. \tag{28}$$

*Let* $\hat{\theta}_* = \mathrm{med}_g(\hat{\theta}_1, \ldots, \hat{\theta}_m)$ *be the geometric median of* $\{\hat{\theta}_1, \ldots, \hat{\theta}_m\}$. *Then*

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

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

**b** *Assume that* $(\mathbb{Y}, d)$ *is a metric space and* $\theta_0 \in \mathbb{Y}$. *Let* $\hat{\theta}_1, \ldots, \hat{\theta}_m \in \mathbb{Y}$ *be a collection of independent random variables. Let the constants* $q, \gamma$ *be such that* $0 < q < \frac{1}{2}$ *and* $0 \leq \gamma < \frac{1/2 - q}{1-q}$. *Suppose* $\varepsilon > 0$ *are such that for all* $j,\ 1 \leq j \leq \lfloor (1-\gamma)m \rfloor + 1$,

$$\Pr\left(d(\hat{\theta}_j, \theta_0) > \varepsilon\right) \leq q. \tag{29}$$

*Let* $\hat{\theta}_* = \mathrm{med}_0(\hat{\theta}_1, \ldots, \hat{\theta}_m)$. *Then*

$$\Pr\left(d(\hat{\theta}_*, \theta_0) > 3\varepsilon\right) \leq e^{-m(1-\gamma)\psi\left(\frac{1/2-\gamma}{1-\gamma}, q\right)}.$$

To get the bound stated in the paper, take $q = \frac{1}{7}$ and $\alpha = \frac{3}{7}$ in part **(a)** and $q = \frac{1}{4}$ in part **b**.

We start by proving part **a**. To this end, we will need the following lemma (see lemma 2.1 in Minsker, 2015):

**Lemma 21** *Let* $\mathbb{H}$ *be a Hilbert space,* $x_1, \ldots, x_m \in \mathbb{H}$ *and let* $x_*$ *be their geometric median. Fix* $\alpha \in \left(0, \frac{1}{2}\right)$ *and assume that* $z \in \mathbb{H}$ *is such that* $\|x_* - z\| > C_\alpha r$, *where*

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

*and* $r > 0$. *Then there exists a subset* $J \subseteq \{1, \ldots, m\}$ *of cardinality* $|J| > \alpha m$ *such that for all* $j \in J$, $\|x_j - z\| > r$.

Assume that event $\mathcal{E} := \Big\{\|\hat\theta_* - \theta_0\| > C_\alpha \varepsilon\Big\}$ occurs. Lemma 21 implies that there exists a subset $J \subseteq \{1,\dots,m\}$ of cardinality $|J| \geq \alpha k$ such that $\|\hat\theta_j - \theta_0\| > \varepsilon$ for all $j \in J$, hence

$$\begin{aligned}\Pr(\mathcal{E}) \leq &\Pr\left(\sum_{j=1}^m I\left\{\|\hat\theta_j - \theta_0\| > \varepsilon\right\} > \alpha m\right) \leq \\ &\Pr\left(\sum_{j=1}^{\lfloor(1-\gamma)m\rfloor+1} I\left\{\|\hat\theta_j - \theta_0\| > \varepsilon\right\} > (\alpha-\gamma)m\frac{\lfloor(1-\gamma)m\rfloor+1}{\lfloor(1-\gamma)m\rfloor+1}\right) \leq \\ &\Pr\left(\sum_{j=1}^{\lfloor(1-\gamma)m\rfloor+1} I\left\{\|\hat\theta_j - \theta_0\| > \varepsilon\right\} > \frac{\alpha-\gamma}{1-\gamma}\big(\lfloor(1-\gamma)m\rfloor+1\big)\right).\end{aligned}$$

If $W$ has Binomial distribution $W \sim B(\lfloor(1-\gamma)m\rfloor+1, q)$, then

$$\begin{aligned}\Pr\Bigg(\sum_{j=1}^{\lfloor(1-\gamma)m\rfloor+1} I\left\{\|\hat\theta_j - \theta_0\| > \varepsilon\right\} > \frac{\alpha-\gamma}{1-\gamma}\big(\lfloor(1-\gamma)m\rfloor+1\big)\Bigg) \leq \\ \Pr\left(W > \frac{\alpha-\gamma}{1-\gamma}\big(\lfloor(1-\gamma)m\rfloor+1\big)\right)\end{aligned}$$

(see Lemma 23 in Lerasle and Oliveira, 2011 for a rigorous proof of this fact). Chernoff bound (e.g., Proposition A.6.1 in van der Vaart and Wellner, 1996), together with an obvious bound $\lfloor(1-\gamma)m\rfloor + 1 > (1-\gamma)m$, implies that

$$\Pr\left(W > \frac{\alpha-\gamma}{1-\gamma}\big(\lfloor(1-\gamma)m\rfloor+1\big)\right) \leq \exp\left(-m(1-\gamma)\psi\left(\frac{\alpha-\gamma}{1-\gamma}, q\right)\right).$$

To establish part **b**, we proceed as follows: let $\mathcal{E}_1$ be the event

$$\mathcal{E}_1 = \{\text{more than a half of events } d(\hat\theta_j, \theta_0) \leq \varepsilon,\ j = 1\dots m \text{ occur}\}.$$

Assume that $\mathcal{E}_1$ occurs. Then we clearly have $\varepsilon_* \leq \varepsilon$, where $\varepsilon_*$ is defined in equation (2.2) of the paper: indeed, for any $\theta_{j_1},\ \theta_{j_2}$ such that $d(\hat\theta_{j_i}, \theta_0) \leq \varepsilon,\ i = 1, 2$, triangle inequality gives $d(\theta_{j_1}, \theta_{j_2}) \leq 2\varepsilon$. By the definition of $\hat\theta_*$, inequality $d(\hat\theta_*, \hat\theta_j) \leq 2\varepsilon_* \leq 2\varepsilon$ holds for at least a half of $\{\hat\theta_1, \dots, \hat\theta_m\}$, hence, it holds for some $\hat\theta_{\tilde\jmath}$ with $d(\hat\theta_{\tilde\jmath}, \theta_0) \leq \varepsilon$. In turn, this implies (by triangle inequality) $d(\hat\theta_*, \theta_0) \leq 3\varepsilon$. We conclude that

$$\Pr\Big(d(\hat\theta_*, \theta_0) > 3\varepsilon\Big) \leq \Pr(\mathcal{E}_1).$$

The rest of the proof repeats the argument of part **a** since

$$\Pr(\mathcal{E}_1^c) = \Pr\left(\sum_{j=1}^m I\left\{d(\hat\theta_j, \theta_0) > \varepsilon\right\} \geq \frac{m}{2}\right),$$

where $\mathcal{E}_1^c$ is the complement of $\mathcal{E}_1$.

### A.2 Proof of Theorem 7

By the definition of Wasserstein distance $d_{W_1}$,

$$d_{W_1,\rho}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)) = \int_\Theta \rho(\theta, \theta_0) d\Pi_l(\theta|X_1, \ldots, X_l). \tag{30}$$

(recall that $\rho$ is the Hellinger distance). Let $R$ be a large enough constant to be determined later. Note that the Hellinger distance is uniformly bounded by 1. Using (30), it is easy to see that

$$d_{W_1,\rho}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)) \leq R\varepsilon_l + \int\limits_{\rho(\theta,\theta_0)\geq R\varepsilon_l} d\Pi_l(\cdot|\mathcal{X}_l). \tag{31}$$

To this end, it remains to estimate the second term in the sum above. We will follow the proof of Theorem 2.1 in Ghosal et al. (2000). Bayes formula implies that

$$\Pi_l(\theta : \rho(\theta, \theta_0) \geq R\varepsilon_l|\mathcal{X}_l) = \int\limits_{\rho(\theta,\theta_0)\geq R\varepsilon_l} \frac{\prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)}{\int\limits_\Theta \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)}.$$

Let

$$A_l = \left\{\theta : -P_0\left(\log\frac{p_\theta}{p_0}\right) \leq \varepsilon_l^2, P_0\left(\log\frac{p_\theta}{p_0}\right)^2 \leq \varepsilon_l^2\right\}.$$

For any $C_1 > 0$, Lemma 8.1 Ghosal et al. (2000) yields

$$\Pr\left\{\int\limits_\Theta \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) dQ(\theta) \leq \exp\big(-(1+C_1)l\varepsilon_l^2\big)\right\} \leq \frac{1}{C_1^2 l\varepsilon_l^2}.$$

for every probability measure $Q$ on the set $A_l$. Moreover, by the assumption on the prior $\Pi$,

$$\Pi(A_l) \geq \exp\left(-Cl\varepsilon_l^2\right).$$

Consequently, with probability at least $1 - \frac{1}{C_1^2 l\varepsilon_l^2}$,

$$\int\limits_\Theta \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) \geq \exp\Big(-(1+C_1)l\varepsilon_l^2\Big)\Pi(A_l) \geq \exp(-(1+C_1+C)l\varepsilon_l^2).$$

Define the event $B_l = \left\{\int\limits_\Theta \prod\limits_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) \leq \exp\Big(-(1+C_1+C)l\varepsilon_l^2\Big)\right\}$.

Let $\Theta_l$ be the set satisfying conditions of Theorem 3.1. Then by Theorem 7.1 in Ghosal et al. (2000), there exist test functions $\phi_l := \phi_l(X_1, \ldots, X_l)$ and a universal constant $K$ such that

$$\begin{aligned} \mathbb{E}_{P_0}\phi_l &\leq 2e^{-Kl\varepsilon_l^2}, \\ \sup_{\theta\in\Theta_l, h(P_\theta,P_0)\geq R\varepsilon_l} \mathbb{E}_{P_\theta}(1-\phi_l) &\leq e^{-KR^2\cdot l\varepsilon_l^2}, \end{aligned} \tag{32}$$

where $KR^2 - 1 > K$.

Note that

$$\Pi_l(\theta : \rho(\theta, \theta_0) \geq R\varepsilon_l | X_1, \dots, X_l) = \Pi_l(\theta : \rho(\theta, \theta_0) \geq R\varepsilon_l | X_1, \dots, X_l)(\phi_l + 1 - \phi_l).$$

For the first term,

$$\mathbb{E}_{P_0}\Big[\Pi_l(\theta : \rho(\theta, \theta_0) \geq R\varepsilon_l | X_1, \dots, X_l) \cdot \phi_l\Big] \leq \mathbb{E}_{P_0}\phi_l \leq 2e^{-Kl\varepsilon_l^2}. \tag{33}$$

Next, by the definition of $B_l$, we have

$$\begin{aligned}
&\Pi_l(\theta : \rho(\theta, \theta_0) \geq R\varepsilon_l | \mathcal{X}_l)(1 - \phi_l) = \\
&\frac{\int\limits_{\rho(\theta,\theta_0)\geq R\varepsilon_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)(1 - \phi_l)}{\int\limits_{\Theta} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)} \left(I\{B_l\} + I\{B_l^c\}\right) \\
&\leq I\{B_l\} + e^{(1+C_1+C)l\varepsilon_l^2} \int\limits_{\rho(\theta,\theta_0)\geq R\varepsilon_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)(1 - \phi_l).
\end{aligned} \tag{34}$$

To estimate the second term of last equation, note that

$$\begin{aligned}
&\mathbb{E}_{P_0} \int\limits_{\rho(\theta,\theta_0)\geq R\varepsilon_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)(1 - \phi_l) \leq \\
&\mathbb{E}_{P_0}\left( \int\limits_{\theta\in\Theta\setminus\Theta_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)(1 - \phi_l) + \int\limits_{\{\Theta_l \cap \rho(\theta,\theta_0)\geq R\varepsilon_l\}} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)(1 - \phi_l) \right) \leq \\
&\Pi(\Theta\setminus\Theta_l) + \int\limits_{\{\Theta_l \cap \rho(\theta,\theta_0)\})\geq R\varepsilon_l} \mathbb{E}_{P_0}\left( \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)(1 - \phi_l) \right) \leq \\
&e^{-l\varepsilon_l^2(C+4)} + e^{-KR^2 \cdot l\varepsilon_l^2} \leq 2e^{-l\varepsilon_l^2(C+4)}
\end{aligned} \tag{35}$$

for $R \geq \sqrt{(C+4)/K}$. Set $C_1 = 1$ and note that $I\{B_l\} = 1$ with probability $P(B_l) \leq 1/l\varepsilon_l^2$. It follows from (33), (34) and (35) and Chebyshev's inequality that for any $t > 0$

$$\begin{aligned}
\Pr\Big(\Pi_l(\theta : \rho(\theta, \theta_0) \geq R\varepsilon_l | \mathcal{X}_l) \geq t\Big) &\leq \Pr(B_l) + \frac{2e^{-Kl\varepsilon_l^2}}{t} + \frac{2\exp\left(-2l\varepsilon_l^2\right)}{t} \\
&\leq \frac{1}{l\varepsilon_l^2} + \frac{2e^{-Kl\varepsilon_l^2}}{t} + \frac{2\exp\left(-2l\varepsilon_l^2\right)}{t}.
\end{aligned}$$

Finally, for a constant $\tilde{K} = \min(K/2, 1)$ and $t = e^{-\tilde{K}l\varepsilon_l^2}$, we obtain

$$\begin{aligned}
\Pr\Big(\Pi_l(\theta : \rho(\theta, \theta_0) \geq R\varepsilon_l | \mathcal{X}_l) \geq t\Big) &\leq \frac{1}{l\varepsilon_l^2} + 2e^{-Kl\varepsilon_l^2/2} + 2\exp\left(-l\varepsilon_l^2\right) \\
&\leq \frac{1}{l\varepsilon_l^2} + 4e^{-\tilde{K}l\varepsilon_l^2},
\end{aligned}$$

which yields the result.

### A.3 Proof of Theorem 12

From equation (20) in the paper and proceeding as in proof of Theorem 7, we get that

$$\|\delta_0 - \Pi_l(\cdot|\mathcal{X}_l)\|_{\mathcal{F}_k} \leq R\varepsilon_l^{1/\gamma} + \int\limits_{\rho_k(\theta,\theta_0)\geq R\varepsilon_l^{1/\gamma}} \rho_k(\theta,\theta_0)d\Pi_l(\cdot|\mathcal{X}_l).$$

By Hölder's inequality,

$$\int\limits_{\rho_k(\theta,\theta_0)\geq R\varepsilon_l^{1/\gamma}} \rho_k(\theta,\theta_0)d\Pi_l(\cdot|\mathcal{X}_l) \leq$$

$$\left[\int\limits_{\Theta} \rho_k^w(\theta,\theta_0)d\Pi_l(\cdot|\mathcal{X}_l)\right]^{1/w} \left[\int\limits_{\rho_k(\theta,\theta_0)\geq R\varepsilon_l^{1/\gamma}} d\Pi_l(\cdot|\mathcal{X}_l)\right]^{1/q}$$

with $w > 1$ and $q = \frac{w}{w-1}$.

Define the event

$$B_l = \left\{\int\limits_{\Theta} \prod_{i=1}^{l} \frac{p_\theta}{p_0}(X_i)d\Pi(\theta) \geq \exp\Big(-(1+C_1+C)l\varepsilon_l^2\Big)\right\}.$$

Following in the proof of Theorem 3.1, we note that $\Pr(B_l) \geq 1 - \frac{1}{C_1^2 l\varepsilon_l^2}$ (where constants $C, C_1$ are the same as in the proof of Theorem 7). Also, note that

$$\mathbb{E}_{P_0}\left[\int_\Theta \rho_k^w(\theta,\theta_0)\prod_{i=1}^{l}\frac{p_\theta}{p_0}(X_i)d\Pi(\theta)\right] = \int_\Theta \rho_k^w(\theta,\theta_0)d\Pi(\theta)$$

Hence, with probability $\geq 1 - e^{-\tilde{K}l\varepsilon_l^2}$,

$$\int_\Theta \rho_k^w(\theta,\theta_0)\prod_{i=1}^{l}\frac{p_\theta}{p_0}(X_i)d\Pi(\theta) \leq e^{\tilde{K}l\varepsilon_l^2}\int_\Theta \rho_k^w(\theta,\theta_0)d\Pi(\theta),$$

where $\tilde{K} = \min(K/2, 1)$ is the same universal constant as in the proof of Theorem 7. Writing

$$\left[\int\limits_{\Theta} \rho_k^w(\theta,\theta_0)d\Pi_l(\cdot|\mathcal{X}_l)\right]^{1/w} = \left[\frac{\int_\Theta \rho_k^w(\theta,\theta_0)\prod_{i=1}^{l}\frac{p_\theta}{p_0}(X_i)d\Pi(\theta)}{\int_\Theta \prod_{i=1}^{l}\frac{p_\theta}{p_0}(X_i)d\Pi(\theta)}\right]^{1/w},$$

we deduce that with probability $\geq 1 - e^{-\tilde{K}l\varepsilon_l^2} - \frac{1}{l\varepsilon_l^2}$ (where we set $C_1 := 1$),

$$\left[\int\limits_{\Theta} \rho_k^w(\theta,\theta_0)d\Pi_l(\cdot|\mathcal{X}_l)\right]^{1/w} \leq e^{\frac{2+C+\tilde{K}}{w}l\varepsilon_l^2}\left[\int_\Theta \rho_k^w(\theta,\theta_0)d\Pi(\theta)\right]^{1/w}.$$

By Theorem 7.1 in Ghosal et al. (2000), there exist test functions $\phi_l := \phi_l(X_1, \dots, X_l)$ and a universal constant $K$ such that

$$\begin{aligned} \mathbb{E}_{P_0}\phi_l &\le 2e^{-Kl\varepsilon_l^2}, \\ \sup_{\theta\in\Theta_l, d(\theta,\theta_0)\ge\tilde{R}\varepsilon_l} \mathbb{E}_{P_\theta}(1-\phi_l) &\le e^{-K\tilde{R}^2\cdot l\varepsilon_l^2}, \end{aligned}$$

where $K\tilde{R}^2 - 1 > K$ and $d(\cdot,\cdot)$ is the Hellinger or Euclidean distance. It immediately follows from Assumption 3.4 that

$$\left\{\theta : \rho(\theta,\theta_0) \ge \tilde{C}_0(\theta_0)R^\gamma\varepsilon_l\right\} \supseteq \left\{\theta : \rho_k(\theta,\theta_0) \ge R\varepsilon_l^{1/\gamma}\right\},$$

hence the test functions $\phi_l$ (for $\tilde{R} := \tilde{C}_0(\theta_0)R^\gamma$) satisfy

$$\begin{aligned} \mathbb{E}_{P_0}\phi_l &\le 2e^{-Kl\varepsilon_l^2}, \\ \sup_{\theta\in\Theta_l, \rho_k(\theta,\theta_0)\ge R\varepsilon_l^{1/\gamma}} \mathbb{E}_{P_\theta}(1-\phi_l) &\le e^{-KR^2\cdot l\varepsilon_l^2}. \end{aligned} \tag{36}$$

Repeating the steps of the proof of Theorem 7, we see that if $R$ is chosen large enough, then

$$\int\limits_{\rho_k(\theta,\theta_0)\ge R\varepsilon_l^{1/\gamma}} d\Pi_l(\cdot|\mathcal{X}_l) \le e^{-\tilde{K}l\varepsilon_l^2}$$

with probability $\ge 1 - \frac{1}{l\varepsilon_l^2} - 4e^{-\tilde{K}l\varepsilon_l^2}$. Combining the bounds, we obtain that, for $w := \frac{4}{3} + \frac{4+2C}{3\tilde{K}}$, with probability $\ge 1 - \frac{1}{l\varepsilon_l^2} - 5e^{-\tilde{K}l\varepsilon_l^2}$,

$$\int\limits_{\rho(\theta,\theta_0)\ge R\varepsilon_l^{1/\gamma}} \rho_k(\theta,\theta_0)d\Pi_l(\cdot|\mathcal{X}_l) \le e^{-\tilde{K}l\varepsilon_l^2/2}\left[\int_\Theta \rho_k^w(\theta,\theta_0)d\Pi(\theta)\right]^{1/w},$$

hence the result follows.

### A.4 Proof of Theorem 15

The proof strategy is similar to Theorem 7. Note that

$$d_{W_{1,\rho}}(\delta_0, \Pi_{n,m}(\cdot|\mathcal{X}_l)) \le R\varepsilon_l + \int\limits_{h(P_\theta,P_0)\ge R\varepsilon_l} d\Pi_l(\cdot|\mathcal{X}_l), \tag{37}$$

where $h(\cdot,\cdot)$ is the Hellinger distance.

Let $\mathcal{E}_l := \{\theta : h(P_\theta, P_0) \ge R\varepsilon_l\}$. By the definition of $\Pi_{n,m}$, we have

$$\Pi_{n,m}(\mathcal{E}_l|\mathcal{X}_l) = \frac{\int\limits_{\mathcal{E}_l}\left(\prod_{j=1}^l \frac{p_\theta}{p_0}(X_j)\right)^m d\Pi(\theta)}{\int\limits_{\Theta}\left(\prod_{j=1}^l \frac{p_\theta}{p_0}(X_j)\right)^m d\Pi(\theta)}. \tag{38}$$

To bound the denominator from below, we proceed as before. Let

$$\Theta_l = \left\{ \theta : -P_0 \left( \log \frac{p_\theta}{p_0} \right) \le \varepsilon_l^2, P_0 \left( \log \frac{p_\theta}{p_0} \right)^2 \le \varepsilon_l^2 \right\}.$$

Let $B_l$ be the event defined by

$$B_l := \left\{ \int\limits_{\Theta_l} \left( \prod_{i=1}^{l} \frac{p_\theta}{p_0}(X_i) \right)^m dQ(\theta) \le \exp(-2ml\varepsilon_l^2) \right\},$$

where $Q$ is a probability measure supported on $\Theta_l$. Lemma 8.1 in Ghosal et al. (2000) yields that $\Pr(B_l) \le \frac{1}{l\varepsilon_l^2}$ for any $Q$, in particular, for the conditional distribution $\Pi(\cdot|\Theta_l)$. We conclude that

$$\int\limits_{\Theta} \left( \prod_{j=1}^{l} \frac{p_\theta}{p_0}(X_j) \right)^m d\Pi(\theta) \ge \Pi(\Theta_l) \exp(-2ml\varepsilon_l^2) \ge \exp(-(2m + C)l\varepsilon_l^2).$$

To estimate the numerator in (38), note that if Theorem 3.10 holds for $\gamma = \varepsilon_l$, then it also holds for $\gamma = L\varepsilon_l$ for any $L \ge 1$. This observation implies that

$$\sup_{\theta \in \mathcal{E}_l} \left( \prod_{j=1}^{l} \frac{p_\theta}{p_0}(X_j) \right)^m \le e^{-c_1 R^2 ml\varepsilon_l^2}$$

with probability $\ge 1 - 4e^{-c_2 R^2 l\varepsilon_l^2}$, hence

$$\int\limits_{\mathcal{E}_l} \left( \prod_{j=1}^{l} \frac{p_\theta}{p_0}(X_j) \right)^m d\Pi(\theta) \le e^{-c_1 R^2 ml\varepsilon_l^2}$$

with the same probability. Choose $R = R(C)$ large enough so that $c_1 m R^2 \ge 3m + C$. Putting the bounds for the numerator and denominator of (38) together, we get that with probability $\ge 1 - \frac{1}{l\varepsilon_l^2} - 4e^{-c_2 R^2 l\varepsilon_l^2}$,

$$\Pi_{n,m}(\mathcal{E}_l | \mathcal{X}_l) \le e^{-ml\varepsilon_l^2}.$$

The result now follows from (37).

### A.5 Proof of Proposition 17

The proof follows a standard pattern: on the first step, we show that it is enough to consider the posterior obtained from the prior restricted to a large compact set, and then proving the theorem for the prior with compact support. The second part mimics the classical argument exactly (e.g., see van der Vaart, 2000).

To show that one can restrict the prior to the compact set, it is enough to establish that for $R$ large enough and $\mathcal{E}_l := \{\theta : \|\theta - \theta_0\| \geq \frac{R}{\sqrt{l}}\}$,

$$\Pi_{n,m}(\mathcal{E}_l|\mathcal{X}_l) = \frac{\int\limits_{\mathcal{E}_l} \left(\prod_{j=1}^{l} \frac{p_\theta}{p_0}(X_j)\right)^m d\Pi(\theta)}{\int\limits_{\Theta} \left(\prod_{j=1}^{l} \frac{p_\theta}{p_0}(X_j)\right)^m d\Pi(\theta)} \tag{39}$$

can be made arbitrarily small. This follows from the inclusion

$$\mathcal{E}_l \subseteq \left\{\theta : h(P_\theta, P_{\theta_0}) \geq \tilde{C}\frac{R}{\sqrt{l}}\right\}$$

(due to the assumed inequality between Hellinger and Euclidean distances) and the bounds for the numerator and the denominator of (39) established in the proof of Theorem 15.

### A.6 Selection of the optimal number of subsets $m$

The following heuristic approach picks the median among the candidate $M$-posteriors. Namely, start by evaluating the M-Posterior for each $m$ in the range of candidate values $[m_1, m_2]$:

$$\begin{aligned}
\hat{\Pi}^g_{n,m_1} &:= \mathrm{med}_g(\Pi_n^{(1)}, \ldots, \Pi_n^{(m_1)}),\\
\hat{\Pi}^g_{n,m_1+1} &:= \mathrm{med}_g(\Pi_n^{(1)}, \ldots, \Pi_n^{(m_1+1)}),\\
&\vdots\\
\hat{\Pi}^g_{n,m_2} &:= \mathrm{med}_g(\Pi_n^{(1)}, \ldots, \Pi_n^{(m_2)})
\end{aligned}$$

and choose $m_* \in [m_1, m_2]$ such that

$$\hat{\Pi}^g_{n,m_*} = \mathrm{med}_0\left(\hat{\Pi}^g_{n,m_1}, \hat{\Pi}^g_{n,m_1+1}, \ldots, \hat{\Pi}^g_{n,m_2}\right), \tag{40}$$

where $\mathrm{med}_0$ is the *metric median* defined in (3) in Section 2.2.

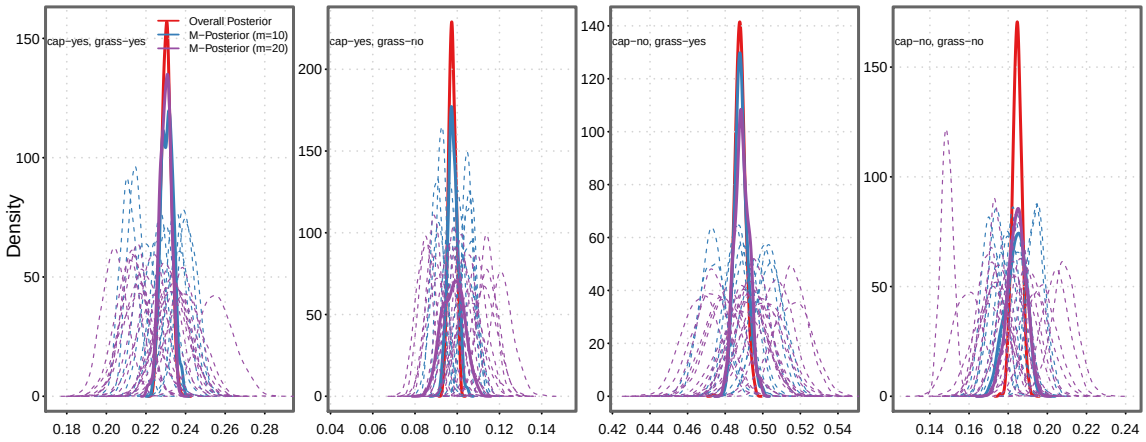
## References


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

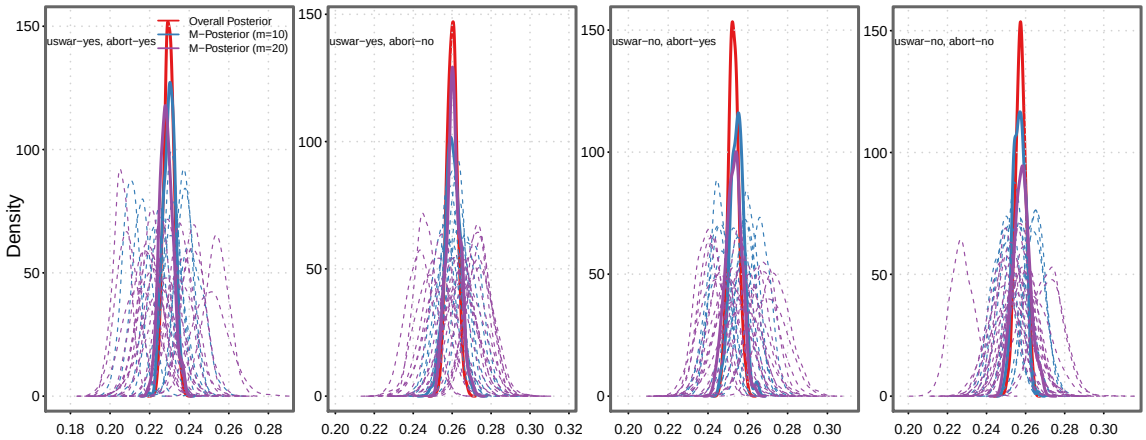
Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.

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




(a) Support capital punishment (cap) and legalization of marijuana (grass)




(b) Expect US to be in world war in 10 years (uswar) and support abortion (abort)

Figure 6: Kernel density estimators of subset posteriors (dashed), M-Posterior with $m = 10$ (blue) and $m = 20$ (violet), and the overall posterior (red) for the four responses. In particular, the model "recognizes" that "expect US to be in world war in 10 years" (uswar) and "support abortion" (abort) should be independent.

M. J. Bayarri and James O. Berger. Robust Bayesian bounds for outlier detection. *Recent Advances in Statistics and Probability*, pages 175–190, 1994.

Amir Beck and Shoham Sabach. Weiszfeld's method: old and new results. *Preprint*, 2013. Available at `https://iew3.technion.ac.il/Home/Users/becka/Weiszfeld_review-v3.pdf`.

James O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. *Ann. Inst. H. Poincaré Probab. Statist.*, 50(2):539–563, 2014.

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

George EP Box and George C Tiao. A Bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129, 1968.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2013.

Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Recursive estimation of the conditional geometric median in Hilbert spaces. *Electronic Journal of Statistics*, 6:2535–2562, 2012.

Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43, 2013.

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37, 2016.

Ramaswamy Chandrasekaran and Arie Tamir. Algebraic optimization: the Fermat-Weber location problem. *Mathematical Programming*, 46(1-3):219–224, 1990.

Kjell A Doksum and Albert Y Lo. Consistent and robust Bayes procedures for location based on partial information. *The Annals of Statistics*, pages 443–453, 1990.

Richard M Dudley. *Real Analysis and Probability*, volume 74. Cambridge University Press, 2002.

David B Dunson and Chuanhua Xing. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

Peter D Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007.

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

Giles Hooker and Anand N Vidyashankar. Bayesian model robustness via disparities. *Test*, 23(3):556–584, 2014.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *arXiv preprint arXiv:1307.1827*, 2013.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., second edition, 2009.

J. H. B. Kemperman. The median of a finite measure on a Banach space. *Statistical Data Analysis Based on the $L_1$-norm and Related Methods, North-Holland, Amesterdam*, pages 217–230, 1987.

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: cutting the Metropolis-Hastings budget. *arXiv preprint arXiv:1304.5299*, 2013.

Matthieu Lerasle and Roberto I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

Hendrik P Lopuhaa and Peter J Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, pages 229–248, 1991.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *arXiv preprint arXiv:1506.06101*, 2015.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*, 2013.

Arkadi S. Nemirovskiĭ and B. Yudin David. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

Frank Nielsen and Vincent Garcia. Statistical exponential families: a digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2011.

Lawrence M. Ostresh. On the convergence of a class of iterative methods for solving the Weber location problem. *Operations Research*, 26(4):597–609, 1978.

Michael L. Overton. A quadratically convergent method for minimizing a sum of Euclidean norms. *Mathematical Programming*, 27(1):34–63, 1983.

Carl Edward Rasmussen and Christopher KI Williams. Gaussian processes for machine learning. *the MIT Press*, 2, 2006.

Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: the consensus Monte Carlo algorithm, 2013.

Alexander Smola and Shravan Narayanamurthy. An Architecture for Parallel Topic Models. In *Very Large Databases (VLDB)*, 2010.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics, $\phi$-divergences and binary classification. *arXiv:0901.2698*, 2009.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 99:1517–1561, 2010.

Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 912–920, 2015a.

Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable Bayes via barycenter in Wasserstein space. *arXiv preprint arXiv:1508.05880*, 2015b.

Markus Svensen and Christopher M Bishop. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005.

Aad van der Vaart. *Asymptotic Statistics*. Cambridge university press, 2000.

Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

Chong Wang, John W Paisley, and David M Blei. Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

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

Endre Weiszfeld. Sur un problème de minimum dans l'espace. *Tohoku Mathematical Journal*, 1936.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, 2011.

Wing Hung Wong, Xiaotong Shen, et al. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.