Distributed Posteriors

Bobby He Deborah Sulem Hector McKimm James Thornton

May 29, 2019

Abstract

In practice, some datasets may be too large to be efficiently processed on one machine. This work presents some methods for distributed Bayesian inference on big data, in the case of both parametric and non-parametric models.

1 Introduction

In many applications statistical methods struggle with the size of large datasets. In the ubiquitous Metropolis-Hastings (MH) algorithm, for instance, the full dataset must be processed at each iteration. In practice, this can become computationally infeasible when the number of observations is large. Techniques designed to deal with 'big data' can largely be categorised into 'Sub-sampling' approaches (Bardenet et al., 2017) or 'Divide and Conquer' (D&C). The first approach uses a random subset of the data for each iteration in an algorithm. The second approach splits the data into batches, where each batch is analysed separately and the results then somehow recombined, similar to map-reduce (Dean and Ghemawat, 2008). We will also refer to this general method as the distributed setting, as the data is distributed across machines.

This article will consider D&C methods. In particular, this report will focus on a Bayesian 'Signal-in-White-Noise' (SWN) model in a distributed setting and devise strategies and conditions to ensure attractive convergence properties of the posterior distribution, defined in section .1.2 of the appendix. The SWN model serves as a useful starting point for deriving theoretical results of the convergence of posteriors in a distributed setting; it is complicated enough to be interesting whilst also allowing for mathematical analysis.

Szabo and van Zanten (2017) shows that it is possible to achieve an optimal rate of convergence in the SWN model in a distributed case if an optimal hyperparameter α of the SWN model is known in advance, referred to as the 'oracle' case. Knapik et al. (2016) provides the theoretical backing of two 'adaptive' procedures which result in desired convergence rates of the SWN model, without knowing the hyperparameter, in the non-distributed case. In the distributed case, one of these adaptive procedures is shown by Szabo and van Zanten (2017) to be infeasible, and as far as we are aware, theoretical guarantees for the other method are yet to be given.

In this report we shall formally detail the SWN model in section 2; review existing approaches to the SWN model in sections 3 for the non-distributed case and in 4 for the distributed, oracle case. We shall then detail how the theorems from Knapik et al. (2016) can be translated into the distributed case in section 5 using computational approaches such as Markov Chain Monte Carlo and the surrogate likelihood method from Jordan et al. (2016). Finally, some extensions of the model using a uniform prior on the parameters and different methods of aggregation such as the Kullback-Leibler barycenter are considered and numerically tested in the Appendix.

2 The Signal-in-White-Noise Model

In a non-distributed setting, for unknown parameter $\theta = \{\theta_i\}_i$ and fixed σ^2, let observations $\{Y_i\}_i$ of the SWN model be distributed according to:

$$Y_i = \theta_i + \sqrt{\frac{\sigma^2}{n}}\tilde{Z}_i, \tag{1}$$

$$\tilde{Z}_i \sim \mathcal{N}(0,1), \tag{2}$$

In a distributed setting, let observations Y_i^j be grouped into m batches, where batch j consists of observations $Y^j = \{Y_1^j, Y_2^j, \ldots\}$. Each observation Y_i^j is distributed according to:

$$Y_i^j = \theta_i + \sqrt{\frac{\sigma^2 m}{n}}Z_i, \tag{3}$$

$$Z_i \sim \mathcal{N}(0,1).$$

The distributed observation model can be translated back to the non-distributed case via $Y_i = m^{-1}\sum_{j=1}^m Y_i^j$ and $\tilde{Z}_i = m^{-\frac{1}{2}}\sum_{j=1}^m Z_i^j$.

We assume independently distributed components of θ, with $\theta_i \sim \mathcal{N}(0, i^{-1-2\alpha})\ \forall i$, so for some hyper-parameter α, the prior on θ is:

$$\Pi(\cdot|\alpha) = \bigotimes_{i=1}^{\infty}\mathcal{N}(0, i^{-1-2\alpha}), \tag{4}$$

3 Non-Distributed Methods

Assume $Y_{1:n} = \{Y_i\}_{i=1}^n$ are generated from the SWM model for some fixed, 'true', θ_0. According to Szabo and van Zanten (2017), it is 'well known' that, if $\theta_0 \in \mathcal{H}_{\beta,M}$ for some 'regularity' parameters $\beta, M > 0$, see (5), then there exists an optimal rate of convergence (see appendix for definition) for estimators of θ_0 and this rate is proportional to $n^{-\beta/(1+2\beta)}$ with respect to the L_2-norm. Additionally, in the non-distributed case, there exists 'adaptive procedures' which achieve this optimal convergence, without knowledge of β or M .

$$\mathcal{H}_{\beta,M} = \{\theta \in \ell^2 : \sum_i i^{2\beta}\theta_i^2 \leq M^2\} \cup \{\theta \in \ell^2 : \sup_i(i^{1+2\beta}\theta_i^2) \leq M^2\} \tag{5}$$

In the Bayesian formulation with prior (4), Knapik et al. (2016) identified two approaches for achieving the optimal rate of convergence: $n^{-\beta/(1+2\beta)}$, for $\theta_0 \in \mathcal{H}_{\beta,M}$.
Firstly, one can set hyperparameter α using an empirical Bayes estimator, $\hat{\alpha}$ as below, to achieve the optimal convergence rate of the posterior for θ.

$$\hat{\alpha}_n = \underset{\alpha\in[0,\log n]}{\arg\max}\,\ell_n(\alpha), \tag{6}$$

$$\ell_n(\alpha) = \log\int p(Y_{1:n}|\theta)\Pi(d\theta|\alpha). \tag{7}$$

In a slight abuse of notation, let $p(\cdot|\theta)$ denote the density of sampling distribution $Y_{1:n}$ with respect to $\Pi(\cdot|\alpha)$. Szabo and van Zanten (2017) showed how one can approximate this empirical estimator in the distributed case, however also proved that this approximation fails to always achieve optimal convergence. In section 5, it is shown that the surrogate likelihood method of Jordan et al. (2016) can be used as an approximation, though further theoretical justification is required.

Secondly, Knapik et al. (2016) describes a hierarchical Bayesian procedure with prior λ on α. If λ satisfies Assumption 1 of Knapik et al. (2016), then the posterior for θ converges at optimal rate. This shall be extended for the distributed case in section 5.

4 Distributed Posterior Methods

In a distributed Bayesian setting each of m machines/nodes has access to a subset of the data, known as a 'shard' Scott et al. (2016). In a typical D&C approach, each machine computes a local posterior using its own subset. The results are then aggregated in a certain way on a central node to produce a global posterior.

For the SWN model, Szabo and van Zanten (2017) showed that one can achieve the same optimal convergence to the non-distributed case under certain priors and aggregations of batches by assuming the regularity parameter β is known, the so-called oracle case. The local posteriors are Gaussian, hence the aggregation step can use a weighted average as in Scott et al. (2016) or a barycenter approach detailed further below. Further tools or approximations may be required for the general case.

4.1 Successful methods for creating a global posterior in the oracle case

We consider three methods for forming a global posterior: the naive way, the adjusted prior approach and Wasserstein Barycenter aggregation. Only the latter two methods have favourable properties in the limit.

The naive procedure is to average the local posteriors by taking the convolution of the rescaled local posteriors. The global posterior produced in such a way neither converges at the optimal rate, nor has coverage probabilities bounded away from zero.

Some adjustments are therefore needed to achieve performance comparable to that in the non-distributed case. Two methods examined by Szabo and van Zanten (2017) successfully achieve both optimal convergence rate and coverage. The first method is to use an adjusted prior:

$$\theta_i \sim \mathcal{N}(0, \tau i^{-1-2\alpha}) \quad ; \quad \tau = mn^{\frac{2(\alpha-\beta)}{1+2\beta}}. \tag{8}$$

In the oracle case, $\alpha = \beta$ so $\tau = m$, therefore the adjustment simply has the effect of raising the prior density to the power $1/m$. The local posteriors are combined into a global posterior through averaging, resulting in a procedure similar to that of 'Consensus Monte Carlo' proposed by Scott et al. (2016).

The second approach is the "WASP" method (Srivastava et al., 2015). The local posteriors are adjusted by raising each local likelihood to the power m, which has the effect of each local dataset having a larger influence on its local posterior. The local posteriors are then aggregated using their Wasserstein Barycenter. For m posterior probability measures $\mu_1, \ldots, \mu_m$, the 2-Wasserstein Barycenter is defined as the measure:

$$\bar{\mu}_W = \underset{\mu \in l^2}{\operatorname{argmin}} \frac{1}{m} \sum_{j=1}^{m} W_2^2(\mu, \mu_j)$$

where

$$W_2^2(\mu, \nu) = \inf_{\gamma} \int \int \|x - y\|_2^2 \gamma(dx, dy)$$

with the infimum over all couplings of μ and ν.

4.2 Failure of above methods in absence of the oracle

Now assume, as is more realistic, that the true regularity is unknown. Recall that in the non-distributed case, equation (6) provides a means of estimating the regularity, leading to an adaptive method with a posterior that contracts at the optimal rate. In the distributed case, Szabo and

van Zanten (2017) point out that the equivalent estimator proposed by is :

$$\hat{\alpha} = \arg\max_{\alpha} \log \int \prod_{j=1}^{m} \Big(p(Y^j|\theta) \Big) \Pi(d\theta|\alpha) \tag{9}$$

However, one may not calculate $\hat{\alpha}$ in (9) in the distributed case without all data on one compute node, which is inherently non-distributed. An approximation by Deisenroth and Ng (2015) is given by:

$$\tilde{\alpha} = \arg\max_{\alpha \in [0, \log n]} \sum_{j=1}^{m} \log \left(\int p(Y^j|\theta) \Pi(\theta|\alpha) d\theta \right), \tag{10}$$

which empirically performs well however there is a pathological case described by Szabo and van Zanten (2017) that shows it is not able to recover the true regularity β of a signal $\theta \in \mathcal{H}_{\beta,M}$. Therefore the methods considered above in the distributed, oracle case above no longer achieve either optimal rate convergence nor coverage with the approximate adaptive estimator for α. In particular, Szabo and van Zanten (2017) construct a series of elements $\{\theta_{0,n}\}_n$ of $\mathcal{H}_{\beta,M}$ such that

$$\mathbb{P}_{\theta_{0,n}}(\tilde{\alpha} \geq \beta + 1/2) \xrightarrow{n} 1$$

even as $n/m \to \infty$ and $m \to \infty$. The basic reason behind this is that each $\theta_{0,n}$ has a large number of 0 coefficients to begin with and hence local machines that only observe a fraction of the data will tend to overestimate the smoothness of the signal. The consequence of this is that the overestimation of β leads the Gaussian prior set on θ to be mis-specified, and this results in the Wasserstein method to lose the optimal rate convergence and coverage that it enjoyed in the oracle case.

5 Proposed Methods

5.1 Stronger assumptions on θ_0

As we saw in the previous section, the adaptive method proposed by Deisenroth and Ng (2015) does not entail the frequentist properties that hold in the oracle case. The counterexample shown by Szabo and van Zanten (2017) was designed to exploit weakness in the distributed case, and it has been suggested that stronger assumptions on θ_0, such as polished tail and self-similarity conditions, can ensure frequentist properties in the adaptive case. Here, we show that a stronger assumption and some simple arguments will enable us to estimate β and perform adaptive distributed inference successfully.

If we assume that $\theta_{0,i} \asymp i^{1+2\beta}$ i.e. $|\theta_{0,i}|/i^{1+2\beta}$ is bounded away from 0 and ∞ in the limit, and moreover that $n/m \to \infty$ then consider,

$$\tilde{\alpha}_1 = \arg\max_{\alpha \in [0, \log n/m]} \log \left(\int p(Y^1|\theta) \Pi(\theta|\alpha) d\theta \right)$$

in other words, $\tilde{\alpha}_1$ is the α value that is obtained using data purely from machine 1. Moreover, it is shown in Theorem 1 in Knapik et al. (2016) that with probability tending to 1, we have $\tilde{\alpha}_1 \in [\underline{\alpha}, \bar{\alpha}]$, where $[\underline{\alpha}, \bar{\alpha}]$ is an interval that contains β and has length $\mathcal{O}\Big(\frac{\log\log \frac{n}{m}}{\log \frac{n}{m}}\Big)$, under the assumption $\theta_{0,i} \asymp i^{1+2\beta}$. This means that $\alpha_1 \to \beta$ in probability.

It is clear from the proofs in Szabo and van Zanten (2017), that an underestimation $\hat{\alpha}$ of β is not as disastrous as overestimation, because $\mathcal{H}_{\beta,M} \subset \mathcal{H}_{\hat{\alpha},M}$: the only consequence is that we have rates pertaining to $\hat{\alpha}$ instead of β. So if we instead propose $\hat{\alpha}_n = \tilde{\alpha}_1 - \epsilon_n$ where $\epsilon_n \to 0$ and

$$\frac{\log\log \frac{n}{m}}{\log \frac{n}{m} \epsilon_n} \to 0$$

then we deduce that in the limit $n \to \infty$, $\hat{\alpha}_n$ will be an arbitrarily close underestimate of β with high probability, and hence we recover the desired frequentist properties if we use $\hat{\alpha}$ in the adaptive setting.

Of course we can replace $\tilde{\alpha}_1$ by $\frac{1}{k}\sum_{i=1}^{k}\tilde{\alpha}_i$ for a fixed k in order to obtain a lower variance estimator, and indeed it would be wasteful to only use the data from one machine. However, one must be careful in taking k to ∞ with n because more information must be known about the rate at which $\mathbb{P}\Big(\tilde{\alpha}_1 \in [\underline{\alpha}, \bar{\alpha}]\Big)$ converges to 1 .

5.2 Distributed adaptive method using the surrogate likelihood method

5.2.1 Surrogate loss function for parametric inference

Jordan et al. (2016) propose a framework for efficient distributed inference in the parametric case called *Communication-efficient Surrogate Likelihood*. This method can be used to find the Maximum Likelihood Estimator in regular parametric and penalized high-dimensional models, as well as to compute a global quasi-posterior distribution in Bayesian inference. If the parameter is $\theta \in \Theta \subset \mathbf{R}^d$ and the data is processed by m machines, the method only requires to communicate $O(md)$ bits. Moreover, for the MLE, the provided estimator converges at the same rate as the global likelihood estimator.

Let $\mathcal{L} : \Theta \times \mathcal{Z} \to \mathbf{R}$ be a loss function and the global and local versions:

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n/m}\mathcal{L}(\theta, z_i^j) \tag{11}$$

$$\mathcal{L}_j(\theta) = \frac{m}{n}\sum_{i=1}^{n/m}\mathcal{L}_j(\theta, z_i^j) \quad \text{for } 1 \leq j \leq m \tag{12}$$

The approximation of the surrogate function comes from a Taylor expansion of the global function $\mathcal{L}_n(\theta)$ at an initial estimator $\bar{\theta}$ in which the high-order derivatives are replaced by their equivalent for the local function $\mathcal{L}_1(\theta)$ at the central machine. Then, a second Taylor expansion of the latter at the same point $\bar{\theta}$ leads to the *surrogate loss function*:

$$\tilde{\mathcal{L}}(\theta) := \mathcal{L}_1(\theta) - < \theta, \nabla\mathcal{L}_1(\bar{\theta}) - \nabla\mathcal{L}_n(\bar{\theta}) > \tag{13}$$

The initial estimator can be chosen as the minimizer of one of the local functions: $\bar{\theta} = \arg\min_\theta \mathcal{L}_1(\theta)$, and the evaluation of the first derivative $\mathcal{L}_n(\bar{\theta})$ requires one communication round of $O(d)$ bits per machine. Under certain assumptions, with the resulting estimator $\tilde{\theta} = \arg\min_\theta \tilde{\mathcal{L}}_1(\theta)$, the authors provide some bounds on the estimation error $\|\tilde{\theta} - \theta_0\|$, or on the approximation error $\|\tilde{\pi}_n - \pi_n\|_1$ of the approximate posterior.

5.2.2 Adaptive estimation of the hyperparameter α

As evaluating $\hat{\alpha}$ in a distributed way is now a low-dimensional parametric problem, we can then use Jordan et al.'s trick to approximate the global marginal log likelihood with a surrogate version:

$$\tilde{\ell}(\alpha) := \ell_1(\alpha|Y^1) - \alpha \times (\ell_1'(\bar{\alpha}|Y^1) - \ell'(\bar{\alpha}|Y_{1:n})) \tag{14}$$

where $\bar{\alpha}$ is an initial estimator, for example $\bar{\alpha} = \arg\max_\alpha \ell_1(\alpha|Y^1)$

In our setting with a Gaussian prior $\theta_i|\alpha \sim \mathcal{N}(0, \tau i^{-1-2\alpha})$ and a Gaussian local likelihood $Y_i^j|\theta_i \sim \mathcal{N}(\theta_i, \sigma^2 m/n)$, the local log marginal likelihood for each machine j has a closed form:

$$\ell_j(\alpha|Y^j) = \log\int_\theta \prod_{i=1}^{\infty} p(Y_i^j|\theta_i)\Pi(\theta_i|\alpha)d\theta_i$$

$$\ell_j(\alpha|Y^j) = \sum_{i=1}^{\infty} \log \int_{-\infty}^{+\infty} p(Y_i^j|\theta_i)\Pi(\theta_i|\alpha)d\theta_i$$

$$\ell_j(\alpha|Y^j) = \sum_{i=1}^{\infty} \log \ell_j^i(\alpha|Y_i^j)$$

and

$$\ell_j^i(\alpha|Y_i^j) = -\frac{1}{2}\frac{{Y_i^j}^2}{\frac{\sigma^2 m}{n} + \frac{\tau}{i^{1+2\alpha}}} - \frac{1}{2}\log(2\pi(\frac{\sigma^2 m}{n} + \frac{\tau}{i^{1+2\alpha}}))$$

The derivative of the local surrogate marginals also have a closed form and we can approximate the derivative of the global marginal likelihood at the initial point using the technique of Deisenroth and Ng (2015):

$$\ell'(\bar{\alpha}|Y_{1:n}) \approx \frac{1}{m}\sum_{j=1}^{m} \ell_j'(\bar{\alpha}|Y^j)$$

with

$$\ell_j'(\alpha|Y^j) = \sum_{i=1}^{\infty} \frac{\log i}{1 + \frac{\sigma^2 m i^{1+2\alpha}}{\tau n}}(1 - \frac{{Y_i^j}^2}{\frac{\sigma^2 m}{n} + \frac{\tau}{i^{1+2\alpha}}})$$

The hyperparameter of the prior is then set to be the maximizer of the surrogate likelihood function $\tilde{\alpha} = \arg\max_\alpha \tilde{\ell}(\alpha)$. However, as our version of the surrogate likelihood adds one more level of approximation compared to Jordan et al's framework, theoretical guarantees on its convergence properties are not established yet.

5.2.3 Simulation

We test our adaptive method with the following signal (Figure 1) with regularity $\beta = 1$:

$$f(x) = \sum_{i=1}^{\infty} \frac{\sin i}{i^{3/2}} \times \cos(\pi(i - \frac{1}{2})x) \quad \text{and} \quad \theta_i = \frac{\sin i}{i^{3/2}}$$

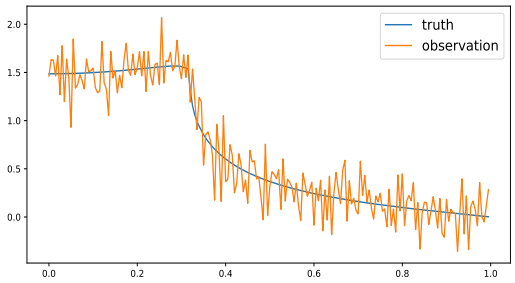

Figure 1: Signal and one noisy observation

For computational purposes, we truncate the infinite sum to $N = 300$ coefficients and divide the $n = 4000$ observations into $m = 40$ machines. The local, non-distributed (global) and surrogate marginal likelihood functions are represented on Figure 2.
We compare our estimator to the existing methods: the MLE of the sum of the local marginal likelihoods (Szabo and van Zanten (2017)), the MLE of the global likelihood and the MLE of the local likelihood on the central machine. Our results are summarized in Figure 3. The surrogate likelihood seems to have very similar performances to the (local) central one from which it is derived. In particular, it also has a much larger variance than the sum of local marginals method - the latter slightly underestimating the global function's one. Besides, the sum of local marginal functions method has a much higher communication complexity during optimization, as it requires repeated evaluations of the function and the gradient on the local nodes. In comparison, our method only requires one round of communication - for the derivative of the local marginal likelihood at the initial estimator.

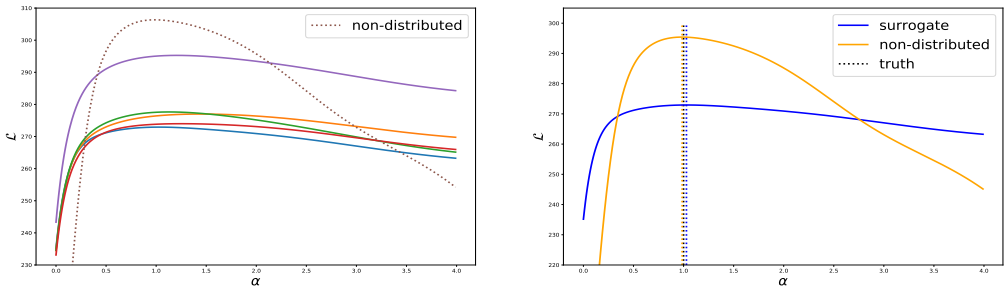

Figure 2: First 5 local marginal likelihoods (left panel) and surrogate marginal with its estimator (right panel) compared to the global marginal likelihood

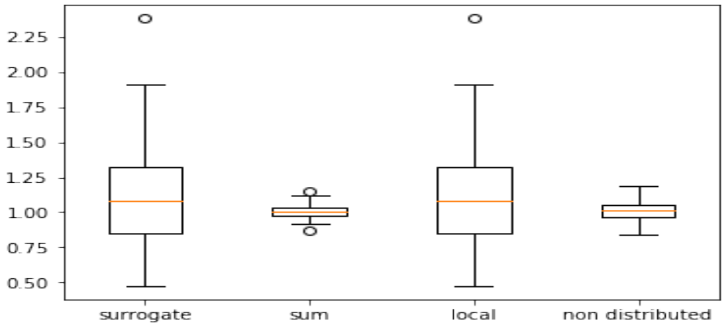

Figure 3: Comparison of the surrogate estimator with the sum of locals, the local and the global (non-distributed) estimators

5.3 Hierarchical Prior

Consider an additional prior λ on hyperparameter α detailed in equation 8. This generates the following hierarchical Bayesian model illustrated in the plate diagram in figure 4 and fully described below $\forall i \in \mathbb{N}, \forall j \in \{1, \ldots, m\}$:

$$\begin{aligned} \alpha &\sim \lambda(\cdot|a), \\ \theta &\sim \bigotimes_{i=1}^{\infty} \mathcal{N}(0, \tau i^{-1-2\alpha}) \\ y_i^j &\sim \mathcal{N}\left(\theta, \sigma^2 \frac{n}{m}\right). \end{aligned}$$

This essentially places a hierarchical prior on θ or a mixture prior given by $\int \Pi(\cdot|\alpha)\lambda(d\alpha|a)$ and substitutes the overhead in optimising hyperparameter α as in equation equation 6, which is difficult to compute in the distributed setting as it requires access to the whole subset of data, with the added computation required to handle the randomness of α or the unwieldiness of the mixture prior.

Theorem 3 of Knapik et al. (2016) shows that one may use a prior satisfying Assumption 1 of Knapik et al. (2016) and also achieve the same optimal convergence properties as the optimised adaptive hyperparamter given by equation 6. Priors such as the exponential, gamma and inverse gamma all satisfy Assumption 1.

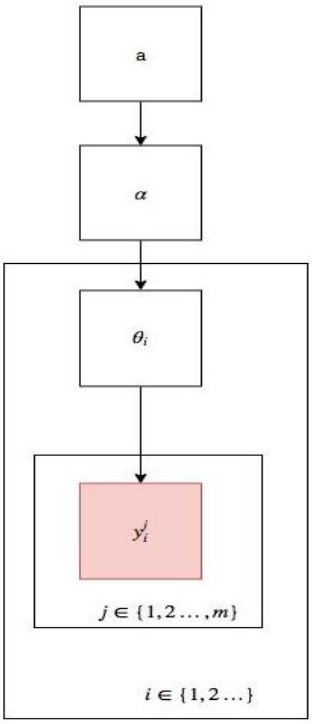

Figure 4: Plate diagram: hierarchical Bayesian model with prior on α with hyperparameter a

Because α is conditionally independent on the data, $\{Y_j\}_j$, given θ, Theorem 3 will also hold in the distributed case and any posterior will have the desired asymptotic properties.

5.4 Implementation Considerations

There exists a Markov Chain Monte Carlo (MCMC) procedure using Gibbs sampling, see Roberts et al. (2004), that can asymptotically sample from the desired distribution in a distributed fashion. In an idealised Gibbs sampler, at each iteration one may record the samples:

Step 1: Draw $\alpha|\theta$

Step 2: Draw $\theta|\alpha, \{Y^j\}_{j=1}^m$

Step 2 may be performed via a divide and conquer strategy as described above and in Scott et al. (2016). The sub-posteriors have the same variance for each i, hence global posterior can be computed using equation (15) with $\tau = m$.

$$\theta_i^j|\alpha, y_i^j \sim \mathcal{N}(\mu, s^2) \qquad \theta_i = m^{-1}\sum_{j=1}^m \theta_i^j \tag{15}$$

$$\mu = \frac{n y_i^j}{n + \sigma^2 m \tau^{-1} i^{1+2\alpha}} \qquad s^2 = \frac{\sigma^2 m}{n + \sigma^2 m \tau^{-1} i^{1+2\alpha}} \tag{16}$$

Step 1 is difficult if at all defined due to θ being an infinite dimensional vector and due to no obvious conjugate prior that abides by the condition of Assumption 1 in Knapik et al. (2016). In practical terms observed Y^j will however be finite dimensional, say dimension N. This means only a finite dimensional subset of θ will be 'active'. Let $\lambda(\cdot|a)$ be the exponential prior with rate a for notational simplicity. Given that for any finite α and N, the summation in the exponent in equation 18 is finite then the posterior for α can be sampled by, for example, rejection sampling with the Gumbel distribution as a proposal.

$$\Pi(\theta_{i=1}^N, \alpha|\{y_j\}_{j=1}^m) = \int_{i>N} \Pi(\theta_{i=1}^\infty, \alpha|\{y_j\}_{j=1}^m) d\theta_{i>N} \propto \prod_i^N \left[\prod_{j=1}^m \left[p(y_i^j|\theta_i)\right]\Pi(\theta_i|\alpha)\right]\lambda(\alpha|a) \tag{17}$$

$$\Pi(\alpha|\theta_{i=1}^N) \propto \prod_i^N \left[\Pi(\theta_i|\alpha)\right]\lambda(\alpha|a) \propto \exp\left\{\sum_{i=1}^N \left[\frac{-i^{1+2\alpha}\theta_i^2}{2\tau}\right] - a\alpha\right\} \tag{18}$$

The cost of this procedure is the per iteration communication of α to the m worker machines and then the communication of the N-dimensional θ from each of the workers to the master. In addition, the speed of the MCMC convergence is unknown.

6 Discussion

This report reviews existing literature and proposes two methods for performing inference on the distributed Signal in White Noise model. Further work is however required to theoretically justify the surrogate likelihood approximation and its comparison to the existing approximate adaptive estimate for the hyperparameter.

At a more abstract level, we have considered a particular case of a Bayesian non-parametric model with distributed observations but shared, possibly infinite dimensional, random parameters. We have used aggregation methods that assume Gaussian batch posteriors in all of the described divide and conquer strategies. This Gaussianity assumptions limits generalisation, however, recent work by Dai et al. (2019) on 'Monte Carlo Fusion' provides some promising machinery for approaching this aggregation step without the Gaussianity assumption. Similarly, other interesting extensions are to investigate asymptotic properties using non-Gaussian priors and other aggregation methods, as briefly explored in the appendix with the uniform prior and other barycenter approaches.

References

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On markov chain monte carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. pages 685–693, 2014.

Hongsheng Dai, Murray Pollock, and Gareth Roberts. Monte carlo fusion. *arXiv preprint arXiv:1901.00139*, 2019.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

Marc Peter Deisenroth and Jun Wei Ng. Distributed gaussian processes. *arXiv preprint arXiv:1502.02843*, 2015.

Michael Jordan, Jason Lee, and Yun Yang. Communication-efficient distributed statistical inference. *arXiv preprint arXiv:1711.03149*, 2016.

BT Knapik, BT Szabó, AW van der Vaart, and JH van Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164(3-4):771–813, 2016.

Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

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. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920, 2015.

Botond Szabo and Harry van Zanten. An asymptotic analysis of distributed nonparametric methods. *arXiv preprint arXiv:1711.03149*, 2017.

7 Appendix

.1 Explanatory Material

.1.1 Notation and Terminology

In a distributed Bayesian context, the terms 'sub-posterior', 'batch' and 'local' posterior refer to the posteriors using subsets of the data and 'global' posterior refers to the idealised posterior whereby all the data is used. The objective of the 'Divide and Conquer' approach is to combine local posteriors to obtain the correct global posterior.

.1.2 Criteria for assessing the asymptotic behaviour of the posterior

The focus of this report is a Bayesian model of a data generating process, however this report shall examine asymptotic frequentist properties of the posterior distribution, as the number of observations, n, tends to infinity. Two such properties used to examine the behaviour of a posterior distribution are: the rate at which it converges around the true signal and how successfully it quantifies uncertainty.

Definition 1. ***Convergence Rate***
Under some data generating process with 'true' parameter θ_0 for observations $Y_{1:n} = \{Y_i\}_{i=1}^n$, estimator $\hat{\theta}_n$ for θ_0 with distribution $\Pi_n(\cdot)$ has a convergence rate of at least $A_n \downarrow 0$ with respect to the L2−norm, $\left\|(\cdot)\right\|_2$, if $\forall B_n$ such that $B_n < A_n$:

$$\Pi_n(\{\hat{\theta}_n : \left\|\hat{\theta}_n - \theta_0\right\|_2 \geq B_n\}) \underset{n\to\infty}{\longrightarrow} 0$$

In a Bayesian setting, let unknown parameter θ correspond to θ_0 in the data generating process, with prior $\Pi(\cdot)$. Given observations $Y_{1:n} = \{Y_i\}_{i=1}^n$, let the posterior distribution for θ be denoted $\Pi(\cdot|Y_{1:n})$. The convergence rate for the posterior of θ is at least $A_n \downarrow 0$ with respect to the L^2-norm, $\left\|(\cdot)\right\|_2$, if $\forall B_n$ such that $B_n < A_n$:

$$\mathbb{E}_{\theta_0}\left[\Pi(\{\theta : \|\theta - \theta_0\|_2 \geq B_n\}|Y_{1:n})\right] \underset{n\to\infty}{\longrightarrow} 0$$

Definition 2. ***Credible set***
For a level $\gamma \in (0,1)$, let r_γ be the radius such that the ball around $\hat{\theta}$ with radius r_γ receives $1-\gamma$ posterior mass:

$$\Pi_n(\{\theta : \left\|\hat{\theta} - \theta\right\|_2 \leq r_\gamma\}) = 1 - \gamma$$

For $L > 0$, the credible set $\hat{C}(L)$ is defined as:

$$\hat{C}(L) = \{\theta : \left\|\theta - \hat{\theta}_n\right\|_2 \leq L r_\gamma\}$$

Definition 3. ***Coverage***
The coverage probability of $\hat{C}(L)$ is defined as $\mathbb{P}_{\theta_0}(\theta_0 \in \hat{C}(L))$.

The second criterion is for the coverage probabilities to remain bounded away from 0 as $n \to \infty$. This ensures the credible sets are asymptotically frequentist confidence sets. In this paper, for the sake of ease, these criteria will be referred to as 'optimal rate' and 'coverage' respectively.

.2 Model and Divide & Conquer Extensions

.2.1 Uniform prior

As the choice of prior may have a strong influence on the posterior in the distributed case, one can choose a uniform prior instead of the Gaussian one. As previously, the θ_i's are supposed to be independent and:

$$\theta_i \sim \mathcal{U}[-\frac{\tau}{i^{\alpha+1/2}}, \frac{\tau}{i^{\alpha+1/2}}] \tag{19}$$

where, similarly, $tau > 0$ is the scale parameter and $\alpha > 0$. With this prior, the (local) posterior distribution of each θ_i at node j is a truncated normal:

$$\Pi^j(\theta_i|Y_i) \sim \mathcal{TN}(Y_i^j, \frac{\sigma^2 m}{n}, -\frac{\tau}{i^{\alpha+1/2}}, \frac{\tau}{i^{\alpha+1/2}}) \tag{20}$$

The same methods for aggregating the local posteriors into a global one can then be used:

- simple averaging of samples of the local posteriors
- adjusted likelihoods, raised at the power m
- adjusted priors, raised at the power $1/m$

As the simple averaging of the local samples corresponds to a global posterior which is the convolution of the rescaled local posteriors, it does not have a closed form with the uniform prior. We thus only propose here a numerical simulation to have an insight of the global posterior's behaviour.

For our illustration, we use the same signal as in Section 5.2.3 represented on Figure 1 with the corresponding coefficients (and regularity $\beta = 1$):

$$f(x) = \sum_{i=1}^{\infty} \frac{\sin i}{i^{3/2}} \times \cos(\pi(i - \frac{1}{2})x) \quad \text{and} \quad \theta_i = \frac{\sin i}{i^{3/2}}$$

For computational purposes, we truncate the infinite sum to $N = 300$ coefficients and divide the $n = 4000$ observations into $m = 40$ machines. The hyperparameters of the prior are set to $\alpha = \beta = 1$ and $\tau = 1$. We compare the distributed version with the non-distributed one, as well as the two counterparts using Gaussian priors (Figure 7).

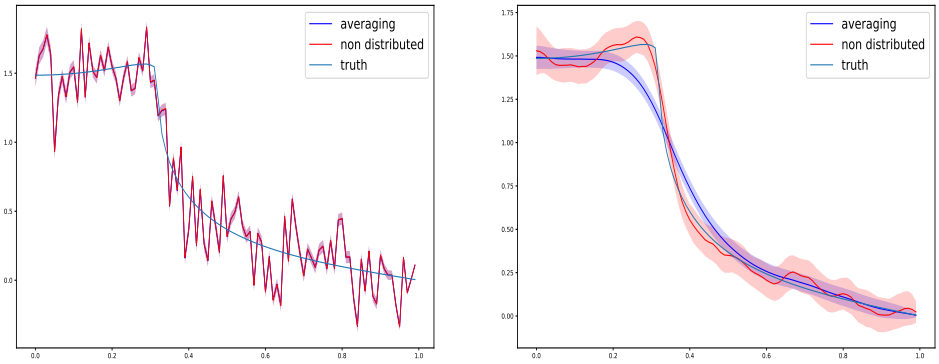

Figure 5: Estimated signal and pointwise 95% credible intervals using a simple averaging aggregation or a non-distributed method, with a uniform prior (left panel) and a Gaussian prior (right panel)

We note that the uniform prior gives a much more erratic estimation of the signal compared to the Gaussian prior, and both methods give too narrow credible intervals. Moreover, the simple averaging method gives roughly similar results than the non-distributed case for the uniform prior. With the adjusted likelihood, the latter seems to have more relevant credible intervals than for the Gaussian prior. Finally, both priors give very similar results using the adjusted prior technique (Figure 6).

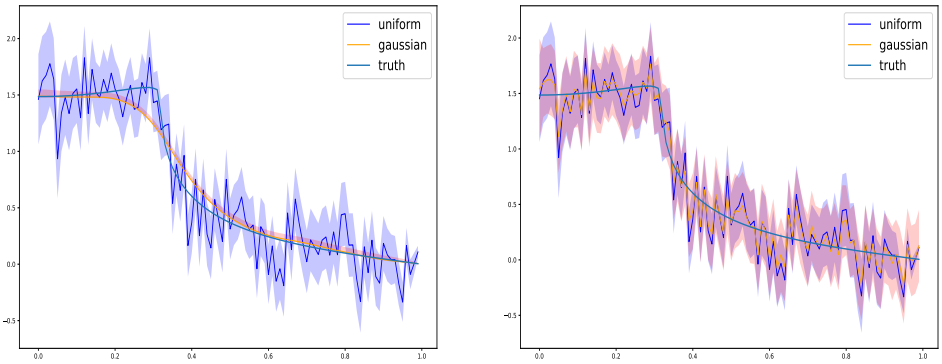

Figure 6: Estimated signal and pointwise 95% credible intervals with a uniform prior and a Gaussian prior using adjusted likelihoods (left panel) and adjusted priors (right panel)

Moreover, we noted that the choice of the scale parameter of the prior τ has a big influence on the posterior distribution, as illustrated on Figure 7. The bigger τ is, the closer to a gaussian distribution (the likelihood function) the posterior is.

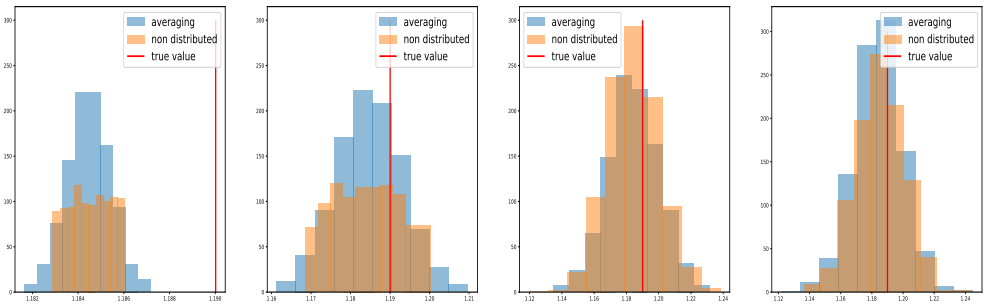

Figure 7: Posterior distribution of the first parameter θ_1 with the simple averaging method and the non-distributed setting for $\tau = 0.1, 1, 10, 100$

.2.2 Kullback-Leibler Barycenter

In the Gaussian case, W_2^2 has an explicit form, and this motivated us to consider the similarly defined Kullback-Leibler Barycenter:

$$\bar{\mu}_{KL} = \underset{\mu}{\operatorname{argmin}} \frac{1}{m} \sum_{j=1}^{m} KL(\mu_j, \mu)$$

where KL is the Kullback-Leibler divergence and the infimum is taken over all univariate Gaussian distributions. For two Gaussian distributions, $\mu_1 \sim \mathcal{N}(\nu_1, \sigma_1^2)$ and $\mu_2 \sim \mathcal{N}(\nu_2, \sigma_2^2)$, it is well known that:

$$KL(\mu_1, \mu_2) = \frac{1}{2}\Big(\frac{\sigma_1^2 + (\nu_1 - \nu_2)^2}{\sigma_2^2} - 1 + \log\Big(\frac{\sigma_2^2}{\sigma_1^2}\Big)\Big)$$

From Szabo and van Zanten (2017) we know that the j^{th} local generalised posterior is a product of Gaussians with means $\hat{\theta}_i^j$ and variances s_i^2 given by:

$$\hat{\theta}_i^j = \frac{n}{n + \sigma^2 i^{1+2\beta}} Y_i^j, \qquad s_i^2 = \frac{\sigma^2}{n + \sigma^2 i^{1+2\beta}}$$

From here, it can be easily deduced that the KL-Barycenter for the i^{th} component of $\bar{\mu}_{i,KL} \sim \mathcal{N}(\hat{\theta}_i, t_i^2)$ where $\hat{\theta}_i = \frac{1}{m}\sum_j \hat{\theta}_i^j$ and $t_i^2 = s_i^2 + \frac{1}{m}\sum_{j=1}^m (\hat{\theta}_i^j - \hat{\theta}_i)^2$, and each component is independent as before.

For ease of notation, let us denote $\delta_i^2 = \frac{1}{m}\sum_{j=1}^m (\hat{\theta}_i^j - \hat{\theta})^2$. Then we have $\delta_i^2 \sim \mathrm{Var}_{\theta_0}(\hat{\theta}_i^1)\frac{\chi^2_{m-1}}{m} = \mathcal{O}_p(\mathrm{Var}_{\theta_0}(\hat{\theta}_i^1))$. Unfortunately, in this case $\mathrm{Var}_{\theta_0}(\hat{\theta}_i^1) = \frac{nm\sigma^2}{(n+\sigma^2 i^{1+2\beta})^2}$ and moreover we know from Theorem 3.2 of Szabo and van Zanten (2017) that $\sum_i \frac{n\sigma^2}{(n+\sigma^2 i^{1+2\beta})^2}$ behaves like a constant times $n^{-2\beta/(1+2\beta)}$ for large n, and thus the extra m factor will render the posterior spread to be of order larger than $n^{-2\beta/(1+2\beta)}$. Thus, we do not recover the optimal rate and convergence frequentist properties for the KL Barycenter, unlike in the Wasserstein case. The contrast between these two methods can be seen in the left hand plot of Figure 8, and we clearly see that the posterior spread in the KL Barycenter case is too conservative compared to the Wasserstein. One interesting thing to note is that reversing the measures so that we consider $KL(\mu, \mu_j)$ instead of $KL(\mu_j, \mu)$ recovers the same posterior distribution as in the Wasserstein case.

.2.3 Sinkhorn Barycenters

In practice, Wasserstein Barycenters are computationally expensive to evaluate. One method that has been introduced in Cuturi and Doucet (2014) to counter this is to use Sinkhorn distances, which add an entropic regularisation term to the Wasserstein distance that is determined by a

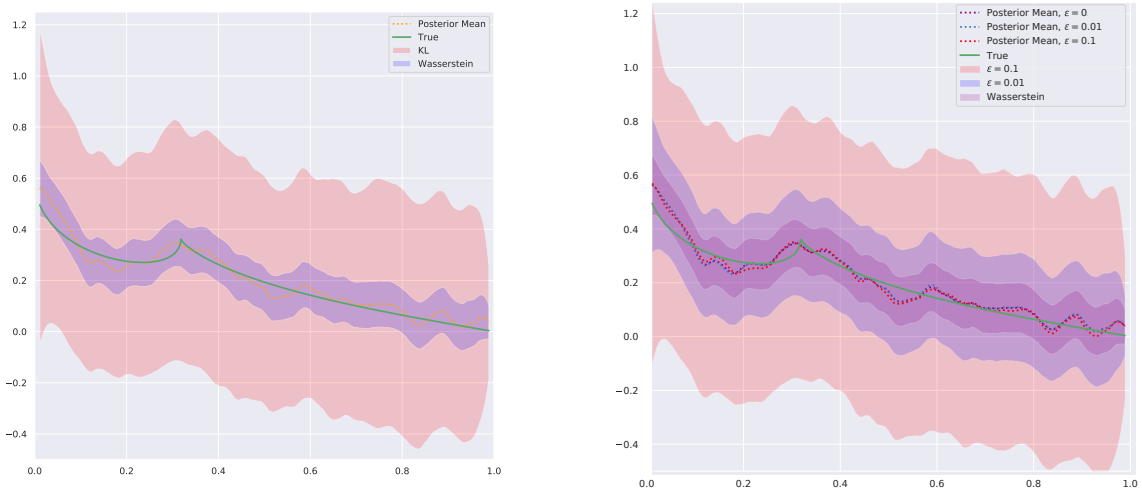

Figure 8: (Left) Comparison between the posterior pointwise 95% credible intervals for the Wasserstein and KL Barycenter methods. (Right) Posterior credible intervals as regularisation parameter increases in Sinkhorn Barycenters.

parameter ϵ, and have been shown to be computationally more efficient. Figure 8 shows us that increasing ϵ has little effect on the posterior mean, but rather increases the posterior spread, which is expected as the regularisation encourages the posterior to have larger variance. Thus, it is apparent that a trade-off must be struck between improved computational efficiency and increased posterior spread.