Journal of Machine Learning Research 20 (2019) 1-30 Submitted 11/17; Revised 10/18; Published 6/19

An asymptotic analysis of distributed nonparametric methods

Botond Szabó B.T.SZABO@MATH.LEIDENUNIV.NL
Mathematical Institute
Leiden University
2333 CA Leiden
The Netherlands

Harry van Zanten HVZANTEN@UVA.NL
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 105-107
1098 XG Amsterdam
The Netherlands

Editor: Manfred Opper

Abstract

We investigate and compare the fundamental performance of several distributed learning methods that have been proposed recently. We do this in the context of a distributed version of the classical signal-in-Gaussian-white-noise model, which serves as a benchmark model for studying performance in this setting. The results show how the design and tuning of a distributed method can have great impact on convergence rates and validity of uncertainty quantification. Moreover, we highlight the difficulty of designing nonparametric distributed procedures that automatically adapt to smoothness.

Keywords: distributed learning, nonparametric models, high-dimensional models, Gaussian processes, convergence rates

1. Introduction

Both in statistics and machine learning there has been substantial interest in the design and study of distributed statistical or learning methods in recent years. One driving reason is the fact that in certain applications datasets have become so large that it is often unfeasible, or computationally undesirable, to carry out the analysis on a single machine. In a distributed method the data are divided over a cluster consisting of several machines and/or cores. The machines in the cluster then process their data locally, after which the local results are somehow aggregated on a central machine to finally produce the overall outcome of the statistical analysis. Distributed methods are not only used for computational reasons, but are for instance also of interest in situations where privacy is important and it is undesirable that all data are handled at a single location. Moreover, there are applications in which data are by construction gathered at multiple locations and first processed locally, before being combined at a central location.

Over the last years a variety of distributed methods have been proposed. Recent examples include Consensus Monte Carlo (Scott et al. (2016)), WASP (Srivastava et al. (2015)), distributed GP's (Deisenroth and Ng (2015)), and methods proposed in Shang and Cheng (2015), Jordan et al. (2016), Lee et al. (2015), Volgushev et al. (2017), to mention but a few. Most papers on distributed methods do extensive experiments on simulated, benchmark and real data to numerically assess and compare the performance of the various methods. Some papers also derive a number of theoretical properties. Theoretical results on the performance of distributed methods are not yet widely available however and there is certainly no common theoretical framework in place that allows a clear theoretical comparison of methods and the development of an understanding of fundamental performance guarantees and limitations. Clearly we can not consider the complete list of all existing methods in this paper. We limit ourselves to a number of representative methods that are Bayesian in nature, allowing for a meaningful comparison.

Since a better theoretical understanding of distributed methods can help to pinpoint fundamental difficulties and opportunities, we develop a framework in this paper which allows us to study and compare the performance of various methods. We are in particular interested in high-dimensional, or nonparametric problems. It is by now well known that the performance of learning or statistical methods in such settings depends crucially on wether or not a method succeeds in realizing the correct bias-variance trade-off, or, in different terminology, succeeds in balancing under- and overfitting. For classical, non-distributed settings we have a rather well-developed understanding of how methods should be tuned to achieve a proper bias-variance trade-off. For distributed methods however, such theory is currently not yet available.

To be able to develop relevant theory we study an idealized model, which is a distributed version of the canonical "signal-in-white-noise" model that serves as an important benchmark model in mathematical statistics (see for instance Tsybakov (2009); Johnstone (2017); Giné and Nickl (2016)). The model is on the one hand rich enough to be interesting, in the sense that it is really distributed in nature and the unknown object that needs to be learned is truly infinite-dimensional. On the other hand it is tractable enough to allow detailed mathematical analysis. In the non-distributed case the signal-in-white-noise model is well known to be very closely related to other nonparametric models, such as nonparametric regression and density estimation. (This can be made very precise in the context of Le Cam's theory of limits of experiments (e.g. Le Cam (2012), Brown and Low (1996), Nussbaum (1996)).) Similarly, the distributed signal-in-white-noise model that we consider in this paper provides a unified framework to compare methods that were originally introduced in different settings. Therefore, although our theoretical results are all obtained in the context of this relatively simple "benchmark model" and we consider only a number of representative distributed methods, we believe that the conclusions that we draw are relevant more generally. We introduce the model in Section 2.

It is not difficult to see that if the number of machines m is relatively large with respect to the total sample size, or signal-to-noise-ratio n, then doing things completely naively in the distributed case leads to a sub-optimal bias-variance trade-off (see also the simulation example in Section 2). In particular, just computing the "usual" estimators on every local machine and then averaging them on the central machine typically leads to a global estimator with a bias that is too large. To achieve good performance, the trade-off has to be adjusted somehow. This can in principle be done in various ways. For instance by locally choosing the “wrong” settings for tuning parameters on purpose, or, in a Bayesian setting, by adjusting the likelihood (e.g. raising it to some power) or by adjusting the prior. In Section 3 we study to what degree various methods that have been proposed in the literature succeed in ultimately achieving the right trade-off. We will see that some are more successful than others in this respect.

An important observation that we make is that the methods that are shown to work well in Section 3 all use information on aspects of the true signal that are in principle unknown, such as its degree of regularity. A key question is whether in distributed settings it is fundamentally possible to set tuning parameters correctly in a purely data-driven way, without using such information. In the non-distributed setting it is well known that such adaptive methods indeed exist (e.g. Tsybakov (2009) or Giné and Nickl (2016)). In the distributed case that we study here however, this is much less clear. In Section 4 we show that using a distributed version of a standard adaptation method that is known to work in the non-distributed case, such as maximum marginal likelihood empirical Bayes, can lead to sub-optimal results in the distributed setting. We will argue that this seems to be a fundamental issue and that we expect that correct automatic setting of tuning parameters in distributed methods is fundamentally more challenging than in the classical, non-distributed case. We believe this is an important issue and want to highlight it as an important and interesting topic for future research.

To further study the fundamental potential and limitations of distributed methods, one should also take into account that there are typically computational and/or communication cost restrictions in such settings. In fact, without such restrictions the matter of obtaining good strategies is essentially non-existent, since we could simply communicate all data to a central machine and then apply an existing optimal non-distributed method. The main goal of this paper, however, is to study and compare the theoretical performance of a number of *existing* distributed methods. The methods we consider are all divide-and-conquer algorithms in which computational or communication limitations are not explicitly imposed, however such restrictions do motivate the methods and all methods implicitly or explicitly consider situations in which simply aggregating and handling all data in one machine is not an option. Formally taking into account communication restrictions in the theoretical analysis is an important next step, but turns out to be rather delicate. We recently obtained some first results in the follow-up paper Szabo and van Zanten (2019), see also Zhu and Lafferty (2018) for related work.

The remainder of the paper is organized as follows. In the next section we introduce the distributed version of the signal-in-white-noise model and provide a simple simulation example to show that in a distributed setting, naively combining inferences from local machines into a global estimator may produce misleading results. In Section 3 we study the performance of a number of Bayesian procedures for signal reconstruction in the distributed signal-in-white-noise model introduced in Section 2. We include a number of methods that have recently been proposed in the literature. We show that some succeed in obtaining the appropriate bias-variance trade-off, but others do not. Moreover, the ones that do produce good results are all non-adaptive, in the sense that they use knowledge of the smoothness of the unkown signal to set their tuning parameters. In the final Section 4 we consider the more realistic setting in which this smoothness is unknown. We study a distributed method that has been proposed for data-driven tuning of the hyperparameters and show that there exist "difficult signals", which this method can not recover in the distributed model at an optimal rate. We argue that this appears to be a fundamental issue, and that designing procedures that automatically adapt to smoothness is fundamentally more challenging in the distributed framework. Mathematical proofs are collected in appendix Sections A and B.

2. Distributed signal-in-white-noise model

Consider the problem of estimating a signal in Gaussian white noise. This is the continuous regression-type problem in which we observe a signal $X = (X_t : t \in [0,1])$ satisfying a stochastic differential equation

$$X_t = \int_0^t f(s)\,ds + \frac{\sigma}{\sqrt{n}} W_t,$$

where W is a Brownian motion, modelling the "white noise", and f is an unknown signal, modelled by a square integrable function, that needs to be recovered from the data. The natural number n is the signal-to-noise ratio. Its size affects the difficulty of the problem and n can be seen as playing the role of sample size in this problem.

By expanding the data in a fixed orthonormal basis of $L^2[0,1]$ (for instance the classical Fourier basis), it is seen that the statistical problem of recovering the signal f from the data X is equivalent to the problem of recovering the sequence of (Fourier) coefficients $\theta \in \ell^2$ from noisy observations $Y_1, Y_2, \ldots$, satisfying

$$Y_i = \theta_i + \sqrt{\frac{\sigma^2}{n}} Z_i, \qquad i = 1, 2, \ldots, \tag{2.1}$$

where the Z_i are independent standard normal variables. This is the usual setting in which there is a single observer that observes every coefficient θ_i with additive Gaussian noise with variance σ^2/n. See for instance Tsybakov (2009); Johnstone (2017); Giné and Nickl (2016) for more details on this classical model.

In the distributed version of the model we divide the "precision budget" n over m different observers, so that each one observes the signal in Gaussian noise with variance $\sigma^2 m/n$, independent of the others. In other words, observer j has data $Y_1^j, Y_2^j, \ldots$ satisfying

$$Y_i^j = \theta_i + \sqrt{\frac{\sigma^2 m}{n}} Z_i^j, \qquad i = 1, 2, \ldots, \tag{2.2}$$

where the Z_i^j are independent, standard Gaussian random variables. We call the m independent sub-problems in which the signal-to-noise ratio is $\sigma^2 m/n$ the "local" problems.

Note that the classical, non-distributed signal-in-white-noise model is obtained again from the distributed model by aggregating all the local data. Indeed, if for $j = 1, 2, \ldots$ we define $Y_i = m^{-1} \sum_{j=1}^m Y_i^j$, then (2.1) holds, with $Z_i = m^{-1/2} \sum_{j=1}^m Z_i^j$ independent standard normal variables. This model has been studied extensively in the literature, serving as a canonical model for understanding the performance of high-dimensional or nonparametric

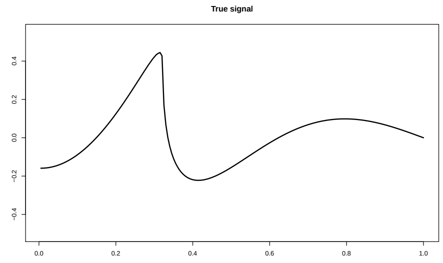

Figure 1: True signal.

statistical procedures. It is well known for instance that if the true signal θ belongs to an ellipsoid or a hyper rectangle of the form

$$\{\theta \in \ell^2 : \sum i^{2\beta}\theta_i^2 \leq M^2\} \text{ or } \{\theta \in \ell^2 : \sup_i (i^{1+2\beta}\theta_i^2) \leq M^2\}$$

for some $\beta, M > 0$, then the optimal rate of convergence of estimators (relative to the ℓ^2-norm) is of the order $n^{-\beta/(1+2\beta)}$. Moreover, there exist so-called adaptive estimators, which achieve this rate without using knowledge about the parameters β or M that describe the complexity, or regularity of the true signal. See, for instance, Tsybakov (2009) or Giné and Nickl (2016). Our central question is whether or not the same results can be obtained in the distributed setting in which each of the m different observers first separately make inference about the signal, and then the local estimates are aggregated into one joint estimator. For technical convenience we will quantify regularity using hyper rectangles in the remainder of the paper, but the analoguous results can be obtained for ellipsoids.

The specific examples of distributed procedures that we consider in this paper are about distributed Bayesian methods. These methods have in common that each local observer first chooses a prior distribution and computes the corresponding local posterior distribution using the local data (or an appropriate modification). In the next step the m local posteriors are somehow aggregated into a global posterior-type distribution, which is then used to produce an estimate of the signal and/or a quantification of the associated uncertainty. In general there is no guarantee that this "aggregated posterior" resembles the posterior distribution that would be obtained in the non-distributed setting, using all the data at once. In particular, it is not clear beforehand how a distributed Bayes method should be constructed in order to have good theoretical properties, like optimal convergence rates, reliable uncertainty quantification or adaptation properties. In this paper we investigate various distributed methods that have been proposed from this point of view.

To see that interesting things can happen it is exemplifying to compare the results of a distributed and a non-distributed (Bayesian) analysis of simulated data. Concretely, we consider a true signal θ consisting of the Fourier coefficients of the function shown in Figure 1. For this signal we simulate data according to (2.2), with $\sigma = 1$, $m = 40$ and $n = 120 \times 40 = 4800$. Then for every local observer a Bayesian procedure is carried out with a Gaussian prior on θ, postulating that the coordinates θ_i are independent and $N(0, i^{-1-2\alpha})$-distributed. The hyperparameter α, which describes the regularity of the prior,

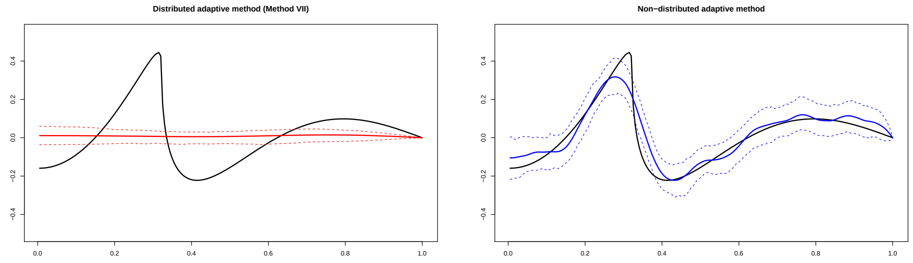

Figure 2: Signal reconstruction using the distributed method (left) and the non-distributed method (right).

is determined using a distributed version of maximum marginal likelihood, as described in Section 4. This analysis leads to $m = 40$ local posterior distributions. These are then combined to produce an overall posterior distribution for the signal. The precise procedure is described in Section 4. The resulting estimator for the signal, together with pointwise 95% credible intervals, is shown in the left plot in Figure 2. The corresponding non-distributed result is obtained by first aggregating all local data as in (2.1) and then carrying out the same Bayesian procedure on these complete data. The resulting non-distributed reconstruction of the signal is shown on the right in Figure 2.

The non-distributed version of this method was studied theoretically for instance in Knapik et al. (2016) and Szabó et al. (2015), where it was shown that the method is adaptive and rate-optimal. The simulation suggests however that an apparently reasonable distributed analogue of the method does not necessarily inherit these favourable properties. The procedure seems to be underfitting and the credible intervals appear to be too narrow. We will argue that this is in some sense a fundamental issue and in the next sections we will study various proposed distributed methods to investigate to what degree they succeed in avoiding or solving these problems.

3. Results for non-adaptive procedures

In this section we study the performance of a number of Bayesian procedures for signal reconstruction in the distributed signal-in-white-noise model introduced in Section 2. All methods involve putting a prior distribution on the unknown signal $\theta \in \ell^2$ in each local problem and then combining the resulting local posteriors into one global posterior-type distribution. To be able to compare the various methods we consider the same Gaussian process (GP) prior in every case, namely the prior

$$\Pi(\cdot|\alpha) = \bigotimes_{i=1}^{\infty} N(0, i^{-1-2\alpha}), \tag{3.1}$$

which postulates that the coefficients θ_i of the signal θ are independent and $N(0, i^{-1-2\alpha})$-distributed. The hyper parameter $\alpha > 0$ essentially controls the regularity of the prior, since it basically controls how fast the Fourier coefficients decrease. (Some of the methods we consider use exactly this prior, others modify it in a certain way with the aim of achieving better performance.) The global posterior-type distribution depends on all the data $\mathbf{Y} = (Y_i^j : j = 1, \dots, m; i = 1, 2, \dots)$ and is denoted by $\Pi(\cdot \,|\, \mathbf{Y})$. It is generally some type of average of the local posteriors, but its precise construction differs between proposed methods. We will see that this can have a significant effect on the performance.

We take an asymptotic perspective and investigate in every case the rate at which the global posterior contracts around the true signal as $n \to \infty$ relative to the ℓ^2-norm, which is as usual defined by $\|\theta\|_2^2 = \sum \theta_i^2$. For a sequence of positive numbers $\varepsilon_n \to 0$ we say that the global posterior contracts at the rate ε_n around the true signal θ_0 if for all sequences $M_n \to \infty$,

$$\mathrm{E}_{\theta_0} \Pi(\theta \in \ell^2 : \|\theta - \theta_0\|_2 > M_n \varepsilon_n \,|\, \mathbf{Y}) \to 0$$

as $n \to \infty$. This means that asymptotically, all posterior mass is concentrated in balls around the true signal θ_0 with ℓ^2-radius of the order ε_n.

Additionally, we study how well the posterior quantifies the remaining uncertainty. Specifically, we consider the coverage probabilities of credible balls around the global posterior mean. These credible sets are constructed by first computing the mean $\hat{\theta}$ of the global "posterior" $\Pi(\cdot \,|\, \mathbf{Y})$. Then for a level $\gamma \in (0, 1)$, the posterior is used to determine the radius r_γ such that the ball around $\hat{\theta}$ with radius r_γ receives $1 - \gamma$ posterior mass, i.e.

$$\Pi(\theta : \|\theta - \hat{\theta}\|_2 \le r_\gamma | \mathbf{Y}) = 1 - \gamma.$$

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

$$\hat{C}(L) = \{\theta : \|\theta - \hat{\theta}_n\|_2 \le L r_\gamma\}. \tag{3.2}$$

(The extra constant L gives some added flexibility, for $L = 1$ we obtain an exact $1 - \gamma$ credible set.) We are interested in the coverage probabilities $\mathrm{P}_{\theta_0}(\theta_0 \in \hat{C}(L))$. If this tends to 0 as $n \to \infty$, the credible sets are asymptotically not frequentist confidence sets, hence give a misleading quantification of the uncertainty. Ideally, the coverage probabilities stay bounded away from 0 as $n \to \infty$.

In the non-distributed case $m = 1$ it is well known that both the rate at which the posterior contracts around the truth and the behaviour of the coverage probabilities of credible sets depend crucially on how the hyper parameter α is tuned. The correct bias-variance trade-off is achieved if α is in accordance with the regularity of the unknown signal. To make this precise, we will consider signals belonging to hyper rectangles of the form

$$H^\beta(M) = \Big\{\theta \in \ell^2 : \sup_i (i^{1+2\beta} \theta_i^2) \le M^2\Big\} \tag{3.3}$$

for some $\beta, M > 0$. It is shown for instance in Knapik et al. (2011) for the non-distributed case that if $\theta_0 \in H^\beta(M)$ and we set $\alpha = \beta$, then the posterior contracts around θ_0 at the optimal rate $n^{-\beta/(1+2\beta)}$. Moreover, for L large enough it then holds that $\mathrm{P}_{\theta_0}(\theta_0 \notin \hat{C}(L)) \le \gamma$. Hence, in the non-distributed case it is optimal to choose the hyper parameter α in such a way that the regularity α of the prior matches the regularity β of the true signal.

In the remainder of this section we investigate distributed methods from this point of view. We will see that the different proposed methods lead to different behaviours in terms of contraction rates and coverage. We stress that the results in this section are non-adaptive, in the sense that we allow the tuning parameter α and other aspects of the constructions to use knowledge of the regularity β of the true signal. This is of course not realistic. It is however important to first understand for every method whether ideally, if the value of β is given to us by an oracle, it is possible to tune the method optimally. Whether this is also possible adaptively, without knowing β, is then the next natural question, which we address in Section 4.

3.1. Naive averaging of local posterior draws

Recall that we have m local observers that each have a dataset $\mathbf{Y}^j = (Y_1^j, Y_2^j, \ldots)$ of noisy coefficients satisfying (2.2). The aim is to recover the true sequence of coefficients θ.

As a starting point, and to have a baseline case to compare the other methods to, we analyse the naive distributed approach in which in every local problem we simply use the prior $\Pi(\cdot \,|\, \alpha)$ defined by (3.1), with $\alpha = \beta$ equal to the regularity of the true sequence θ, in the sense of (3.3). Every local observer then computes its corresponding local posterior, $\Pi^j(\cdot \,|\, \mathbf{Y}^j)$. By Bayes' formula this is given by

$$d\Pi^j(\theta \,|\, \mathbf{Y}^j) \propto p(\mathbf{Y}^j \,|\, \theta)\, d\Pi(\theta \,|\, \beta),$$

where the likelihood for the jth local problem is given by

$$p(\mathbf{Y}^j \,|\, \theta) \propto \prod e^{-\frac{1}{2}\frac{n(Y_i^j - \theta_i)^2}{\sigma^2 m}}. \tag{3.4}$$

Finally these local posteriors are combined into a global, "average posterior" $\mathbf{\Pi}_I(\cdot \,|\, \mathbf{Y})$ by postulating that a draw from this global posterior is generated by first drawing once from each local posterior and then averaging these m independent draws. (Formally, this means that the global "posterior" $\mathbf{\Pi}_I(\cdot \,|\, \mathbf{Y})$ is the convolution of the rescaled local posteriors $\Pi^1(m \times \cdot \,|\, \mathbf{Y}^1)$, …, $\Pi^m(m \times \cdot \,|\, \mathbf{Y}^m)$).

This distributed method is conceptually very simple, but it turns out that neither from the point of view of contraction rates, nor from the point of view of uncertainty quantification it performs very well. The reason is basically that although the choice $\alpha = \beta$ of the tuning parameter of the prior correctly matches squared bias, variance and posterior spread in the local problems, the averaging procedure results in a global "posterior" for which the spread and the variance of the mean are too small relative to the squared bias. The following theorem asserts that for every smoothness level $\beta > 0$ there exist β-regular truths for which the contraction rate of the posterior deteriorates substantially and for which the uncertainty quantification by the credible sets (3.2) constructed from the global posterior $\mathbf{\Pi}_I(\cdot \,|\, \mathbf{Y})$ is useless, no matter how far they are blown up by a constant $L > 0$.

Theorem 1 (naive averaging) *For every $\beta, M > 0$ there exists a $\theta_0 \in H^\beta(M)$ such that for small enough $c > 0$,*

$$E_{\theta_0}\mathbf{\Pi}_I(\theta : \|\theta - \theta_0\|_2 \leq c m^{\frac{\beta}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}} | \beta, \mathbf{Y}) \to 0$$

as $m \to \infty$ and $n/m \to \infty$. Furthermore, for all $L > 0$ it holds that

$$P_{\theta_0}\big(\theta_0 \in \hat{C}(L)\big) \to 0.$$

Proof The proof of the theorem is given in Section A.1. ■

In the literature several less naive distributed strategies have been proposed. These methods either change the local likelihoods in a certain way, and/or the priors that are locally used, and/or the way that the local posteriors are aggregated. In the next few sections we investigate whether such strategies can improve the bad asymptotic performance of the naive averaging method.

3.2. Adjusted local likelihoods and averaging

One perspective on the bad performance of the naive method is to say that since the "sample size" n/m in the local problems is too small, the influence of the data on the local posterior is too small, resulting in a variance (and spread) that is too small relative to the squared bias in the global posterior. A possible way to remedy this that has been proposed in several papers is to raise the local likelihoods to the power m, in order to mimic the situation that we have sample size n in the local problems. This generalized Bayesian approach for the local problems has for instance been considered in the distributed context by Srivastava et al. (2015). They combine it with a different aggregation method however, which we consider in Section 3.4. In this section we still consider the simple averaging scheme, in order to isolate the effect of adjusting the local likelihoods.

So in method II all local observers use the prior $\Pi(\cdot\,|\,\alpha)$ again, with $\alpha = \beta$ equal to the regularity of the truth. They now each compute a generalized local posterior $\tilde{\Pi}^j(\cdot\,|\,\mathbf{Y}^j)$, defined by

$$d\tilde{\Pi}^j(\theta\,|\,\mathbf{Y}^j) \propto \Big(p(\mathbf{Y}^j\,|\,\theta)\Big)^m\, d\Pi(\theta\,|\,\beta).$$

As before the global "posterior" $\mathbf{\Pi}_{II}(\cdot\,|\,\mathbf{Y})$ is defined by postulating that a draw from this global posterior is generated by first drawing once from each local generalized posterior and then averaging these m independent draws.

The following theorem states that this method indeed improves the naive approach of Section 3.1. The global posterior now contracts at the optimal rate for every β-regular truth. Unfortunately, the bad behaviour of the credible sets has not been remedied. For this approach the uncertainty quantification is in fact misleading for all β-regular truths.

Theorem 2 (adjusted likelihoods + averaging) *For all $\beta, M > 0$ and all sequences $M_n \to \infty$,*

$$\sup_{\theta_0\in H^\beta(M)} \mathrm{E}_{\theta_0}\mathbf{\Pi}_{II}(\theta:\, \|\theta-\theta_0\|_2 \geq M_n n^{-\frac{\beta}{1+2\beta}}|\mathbf{Y}) \to 0$$

as $n, m \to \infty$. However, for all $\theta_0 \in H^\beta(M)$ and all $L > 0$ it holds that

$$\mathrm{P}_{\theta_0}\big(\theta_0 \in \hat{C}(L)\big) \to 0.$$

Proof The proof is given in Section A.2. ■

3.3. Adjusted priors and averaging

Adjusting the likelihood as in the preceding section resulted in a correct trade-off between the bias and the variance of the global posterior mean, yielding an optimal posterior contraction rate. The spread of the posterior remained too small in comparison however, resulting in credible sets with zero asymptotic coverage. Instead of raising the local posteriors to the power m, as considered in the preceding section, we could alternatively raise the prior density to the power $1/m$. This has for instance been proposed in the context of the "Consensus Monte Carlo" approach by Scott et al. (2016), in combination with simple averaging of the local posteriors. In this section we investigate the performance of this method in terms of posterior contraction and uncertainty quantification in our distributed signal-in-white-noise model.

The prior $\Pi(\cdot\,|\,\alpha)$ that we use in the local problems is again a product of centered Gaussians with variance $i^{-1-2\alpha}$. Raising the corresponding densities to the power $1/m$ has the effect of multiplying the ith prior variance by m. Hence, in our case raising the prior density to the power $1/m$ is the same as multiplicative rescaling, postulating that θ is a-priori distributed according to $\Pi(\cdot\,|\,\alpha, m)$, where

$$\Pi(\cdot|\alpha,\tau) = \bigotimes_{i=1}^{\infty} N(0, \tau i^{-1-2\alpha}) \tag{3.5}$$

for $\alpha, \tau > 0$. Rescaled GPs have also been considered by Shang and Cheng (2015), who have used them in the distributed setting to construct global credible sets from local ones.

Using rescaling we can actually obtain good results if the prior regularity α is not exactly equal to the true regularity β. By using a scaling different from $\tau = m$ we can somehow compensate for the mismatch between α and β, at least in the range $\beta \leq 1 + 2\alpha$. In the non-distributed setting this is a well-known phenomenon, see for instance van der Vaart and van Zanten (2007); Knapik et al. (2011); Szabó et al. (2013).

The distributed procedure that we consider in this section then takes the following form. Every local observer uses the rescaled prior $\Pi(\cdot|\alpha,\tau)$ defined by (3.5), with $\alpha > 0$ and

$$\tau = mn^{\frac{2(\alpha-\beta)}{1+2\beta}},$$

where β is the regularity of the truth. Next the (normal, unadjusted) corresponding posteriors are computed and they are averaged into a global "posterior" $\mathbf{\Pi}_{III}(\cdot\,|\,\mathbf{Y})$ as in the preceding sections. (Note that if in the local problems the prior regularity $\alpha = \beta$ is used, then $\tau = m$, so the method corresponds to raising the prior density to the power $1/m$.)

The following theorem gives the posterior contraction and coverage results for this method.

Theorem 3 (adjusted priors + averaging) *Suppose $\beta, M > 0$ and $\beta \leq 1 + 2\alpha$. Then for all sequences $M_n \to \infty$,*

$$\sup_{\theta_0 \in H^\beta(M)} \mathrm{E}_{\theta_0} \mathbf{\Pi}_{III}(\theta : \|\theta - \theta_0\|_2 > M_n n^{-\frac{\beta}{1+2\beta}} | \mathbf{Y}) \to 0$$

as $n \to \infty$. Moreover, for all $\gamma \in (0,1)$ it holds that

$$\sup_{\theta_0 \in H^\beta(M)} \mathrm{P}_{\theta_0}\Big(\theta_0 \notin \hat{C}(L)\Big) \leq \gamma$$

for large enough $L > 0$.

Proof See Section A.3. ■

So adjusting the prior in this way actually works better than adjusting the likelihood. Not only do we get optimal contraction rates, but the credible sets that this method produces have asymptotic frequentist coverage too. The proof shows that the credible sets have optimal radius of the order $n^{-\beta/(1+2\beta)}$ as well.

3.4. Adjusted local likelihoods and Wasserstein barycenters

In Section 3.2 we saw that raising the local likelihoods to the power m and then averaging the corresponding generalized posteriors yields optimal contraction rates, but can produce badly performing credible sets. In this section we study the approach considered by Minsker et al. (2014); Srivastava et al. (2015) in the context of their "WASP" method, which consists in aggregating the local posteriors not by simple averaging, but by computing their Wasserstein barycenter.

The generalized local posteriors $\tilde{\Pi}^j(\cdot \,|\, \mathbf{Y}^j)$, as defined in Section 3.2, are (Gaussian) measures on ℓ^2. The 2-Wasserstein distance $W_2(\mu,\nu)$ between two probability measures μ and ν on ℓ^2 is defined by

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

where the infimum is over all measures γ on $\ell^2 \times \ell^2$ with marginals μ and ν. The corresponding 2-Wasserstein barycenter of m probability measures $\mu_1,\ldots,\mu_m$ on ℓ^2 is then defined by

$$\bar{\mu} = \operatorname*{argmin}_\mu \frac{1}{m}\sum_{j=1}^m W_2^2(\mu,\mu_j),$$

where the minimum is over all probability measures on ℓ^2 with finite second moments. There exist effective algorithms to compute Wasserstein barycenters in many cases, see for instance Cuturi and Doucet (2014) and the references therein.

Having this notion at our disposal the distributed method we consider in this section proceeds as follows. In every local problem the prior $\Pi(\cdot \,|\, \alpha)$ is used, with $\alpha = \beta$ equal to the regularity of the truth. Next, the corresponding generalized posteriors $\tilde{\Pi}^j(\cdot \,|\, \mathbf{Y}^j)$ are computed locally, which involves raising the likelihood to the power m as described in Section 3.2. Finally, the global "posterior" $\mathbf{\Pi}_{IV}(\cdot \,|\, \mathbf{Y})$ is constructed as the 2-Wasserstein barycenter of the local measures $\tilde{\Pi}^1(\cdot \,|\, \mathbf{Y}^1),\ldots,\tilde{\Pi}^m(\cdot \,|\, \mathbf{Y}^m)$.

The following theorem asserts that this method results in optimal posterior contraction rates and correct quantification of uncertainty.

Theorem 4 (adjusted likelihoods + barycenters) *For all $\beta, M > 0$ and all sequences $M_n \to \infty$,*

$$\sup_{\theta_0 \in H^\beta(M)} \mathrm{E}_{\theta_0} \mathbf{\Pi}_{IV}(\theta: \|\theta - \theta_0\|_2 > M_n n^{-\frac{\beta}{1+2\beta}} | \mathbf{Y}) \to 0$$

as $n \to \infty$. Moreover, for all $\gamma \in (0,1)$ it holds that

$$\sup_{\theta_0 \in H^\beta(M)} \mathrm{P}_{\theta_0}\Big(\theta_0 \notin \hat{C}(L)\Big) \leq \gamma$$

for large enough $L > 0$.

Proof See Section A.4. ■

3.5. Product of Gaussian process experts

The proofs of the theorems presented so far show that since in our context the global "posterior" is always a Gaussian measure, the behaviour of the procedure can be understood by analyzing three central quantities: the bias of the posterior mean, the variance of the posterior mean, and the spread of the posterior. Depending on how these quantities are related we have found different behaviours: sub-optimal posterior contraction and bad coverage of credible sets (Section 3.1), optimal posterior contraction but bad coverage of credible sets (Section 3.2), and optimal posterior contraction and also good coverage of credible sets (Sections 3.3 and 3.4).

In principle it is now straightforward to analyze different methods as well, provided the three central quantities can be controlled. As an illustration we consider in this section the single-layer version of the product-of-Gaussian-process-expert (PoE) model, introduced in Ng and Deisenroth (2014) and a generalization proposed in Cao and Fleet (2014). An interesting fact is that we will encounter a combination of behaviours that we have not seen yet: sub-optimal contraction rates, but good coverage of credible sets. These methods were introduced to deal with the distributed non-parametric regression model, but for the sake of comparison we analyze them in the context of our distributed signal-in-white-noise model, which can be thought of as an idealized version of the regression model.

The idea of the basic version of the Gaussian PoE model is to employ a Gaussian prior in every local machine, compute the corresponding posterior densities and approximate the global posterior density by multiplying and normalizing these. In our infinite-dimensional setting this does not make sense strictly speaking, since we can not express priors and posteriors on ℓ^2 in terms of densities with respect to some generic dominating measure. We could remedy this by considering a truncated version of our distributed model, where we assume we only observe the first n noisy coefficients Y_i^j in every machine, say, and focus on making inference about the first n true coefficients θ_i. This would make the setting finite-dimensional, allowing us to write prior and posterior densities with respect to the Lebesgue measure. Alternatively, we can stay in the infinite-dimensional setting of the paper and just reason formally and still arrive at a well-defined global PoE "posterior". This is the approach we follow here.

Indeed, say that as before we use the prior $\Pi(\cdot\,|\,\alpha)$ given by (3.1) in every local machine, with $\alpha = \beta$ equal to the regularity of the true signal. This prior has formal "density" proportional to

$$\theta \mapsto \prod e^{-\frac{1}{2}\frac{\theta_i^2}{i^{-1-2\beta}}}.$$

By completing the square we see that the product of this expression with the local likelihood given by (3.4) is, still formally, proportional to

$$\theta \mapsto \prod e^{-\frac{1}{2}\frac{\theta_i^2}{i^{-1-2\beta}}} e^{-\frac{1}{2}\frac{n(Y_i^j-\theta_i)^2}{\sigma^2 m}} \propto \prod e^{-\frac{1}{2}\theta_i^2(i^{1+2\beta}+\frac{n}{m\sigma^2})+\theta_i\frac{nY_i^j}{m\sigma^2}}.$$

Taking the product over j we then obtain the formal density of the PoE posterior, which is proportional to

$$\theta \mapsto \prod e^{-\frac{1}{2}\theta_i^2(mi^{1+2\beta}+\frac{n}{\sigma^2})+\theta_i\frac{n\sum_{j=1}^m Y_i^j}{m\sigma^2}}.$$

Now this last expression is, up to a constant, the density of a product of Gaussians with means $\hat{\theta}_i$ and variances t_i^2 given by

$$\hat{\theta}_i = \frac{nm^{-1}\sum Y_i^j}{n+\sigma^2 mi^{1+2\beta}}, \qquad t_i^2 = \frac{\sigma^2}{n+\sigma^2 mi^{1+2\beta}}.$$

The latter is in fact a well-defined Gaussian measure on ℓ^2, so we can now simply define the global PoE "posterior" $\mathbf{\Pi}_V(\cdot\,|\,\mathbf{Y})$ as the latter measure.

We see that the expressions for the global mean and spread are in fact the same as what we found in Section A.1 for the naive averaging method. As a consequence, the negative result of Theorem 1 holds for the basic version of the Gaussian PoE model as well.

Theorem 5 (product of Gaussian experts) *For every $\beta, M > 0$ there exists a $\theta_0 \in H^\beta(M)$ such that for small enough $c > 0$,*

$$\mathrm{E}_{\theta_0}\mathbf{\Pi}_V(\theta:\, \|\theta-\theta_0\|_2 \le cm^{\frac{\beta}{1+2\beta}}n^{-\frac{\beta}{1+2\beta}}|\mathbf{Y}) \to 0$$

as $m \to \infty$ and $n/m \to \infty$. Furthermore, for all $L > 0$ it holds that

$$\mathrm{P}_{\theta_0}\big(\theta_0 \in \hat{C}(L)\big) \to 0.$$

One can generalize the PoE model by raising the local posterior densities to some power before multiplying and normalizing them, as proposed in Cao and Fleet (2014). In the subsequent analysis we consider the choice $1/m$ for the power, as suggested in Deisenroth and Ng (2015). Adapting the preceding analysis for the ordinary PoE model we see that for this generalized PoE model the global "posterior" $\mathbf{\Pi}_{VI}(\cdot\,|\,\mathbf{Y})$ is in our setting again a product of Gaussians, but now with means and variances given by

$$\hat{\theta}_i = \frac{nm^{-1}\sum Y_i^j}{n+\sigma^2 mi^{1+2\beta}}, \qquad t_i^2 = \frac{\sigma^2 m}{n+\sigma^2 mi^{1+2\beta}}.$$

So the global posterior mean is unaltered compared to the basic PoE model, but the global posterior spread has been blown up by a factor m. As a result, there still exists the same

Method	Description	Optimal rate	Coverage
I	naive averaging	no	no
II	adjusted likelihoods, averaging	yes	no
III	adjusted priors, averaging	yes	yes
IV	adjusted likelihoods, barycenter	yes	yes
V	product of experts	no	no
VI	generalized product of experts	no	yes

Table 1: Performance of the various non-adaptive methods.

class of truths as in Section A.1 for which the squared bias and the variance of the posterior mean will be incorrectly balanced, resulting in a sub-optimal rate of posterior contraction. However, the larger posterior spread ensures that we do have asymptotic coverage of credible sets. It should be noted however that these sets have a diameter that is sub-optimal, i.e. they are too conservative.

Theorem 6 (generalized product of Gaussian experts) *For every $\beta, M > 0$ there exists a $\theta_0 \in H^\beta(M)$ such that for small enough $c > 0$,*

$$\mathrm{E}_{\theta_0}\mathbf{\Pi}_{VI}(\theta : \|\theta-\theta_0\|_2 \leq c m^{\frac{\beta}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}} |\mathbf{Y}) \to 0$$

as $m \to \infty$ and $n/m \to \infty$. However, for all $\gamma \in (0,1)$ it holds that

$$\sup_{\theta_0 \in H^\beta(M)} \mathrm{P}_{\theta_0}\Big(\theta_0 \notin \hat{C}(L)\Big) \leq \gamma$$

for large enough $L > 0$.

Proof The proof of the theorem can be found in Section A.5. ■

3.6. Summary of results for non-adaptive methods

We have seen that the various methods for aggregation of the local posteriors can give quite different results. The methods we considered produce different global “posterior” measures. Depending on the relation between the bias and variance of the global posterior mean and the spread of this global posterior, the posterior contraction rate and coverage probabilities of credible sets can have different behaviours. We summarize our findings in Table 1. This is certainly not meant to be an exhaustive list of methods, but rather an illustration of how the design of distributed procedures can affect their fundamental performance.

Simulations further illustrate the theoretical results. We have considered a true signal θ consisting of the Fourier coefficients of the function shown in the left panel of Figure 3. This is a signal which has regularity $\beta = 1$ in the sense of (3.3). For this signal we simulated data according to (2.2), with $\sigma = 1$, $n = 4800$ and $m = 40$, i.e. we considered a distributed setting with $m = 40$ machines. For the sake of comparison, the right panel of Figure 3 shows the

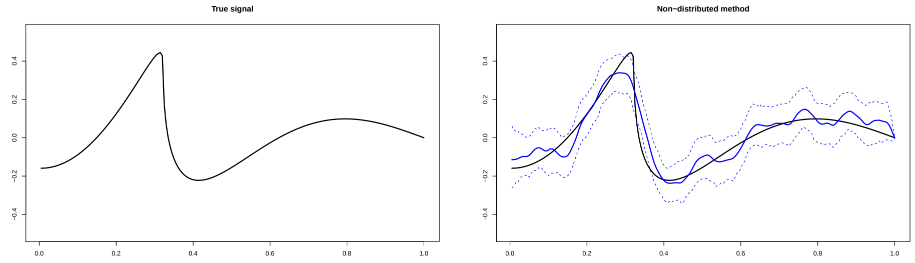

Figure 3: Left: true signal. Right: posterior mean (blue solid curve) and 95% pointwise credible bands (dashed blue curves) for the non-distributed method.

signal reconstruction and uncertainty quantification for the non-distributed method which first aggregates all data in a single machine and then computes the posterior corresponding to the prior $\Pi(\cdot\,|\,\alpha)$ defined by (3.1), with $\alpha = \beta$. This is a method which is known to have an optimal convergence rate and correct quantification of uncertainty. This classical, non-distributed result should be compared to Figure 4, which visualizes the "posteriors" generated by each of the distributed methods I–VI.

In accordance with our theoretical results, we see that the results of methods III and IV are comparable with the non-distributed method. Methods I, V and VI have worse signal reconstruction. The posterior mean of Method II is comparable to that of the optimal methods, but the uncertainty is underestimated.

An important observation to make is that the methods that achieve the same optimal performance as non-distributed methods, all use information about the regularity β of the unkown signal, mostly through the setting of tuning parameters in the priors. In that sense, they are non-adaptive. They serve as useful results that indicate what is possible in principle if we have certain oracle knowledge about the truth we are trying to learn. To understand what realistic procedures can achieve this has to be combined with insight into what can be learned about this oracle knowledge from the data. In the next section we address this issue in the context of our distributed signal-in-white-noise model.

4. Results for adaptive procedures

In the non-distributed case it is well known that there exist adaptive methods that achieve the same optimal performance as non-adaptive procedures, without using knowledge of the regularity β of the unkown signal. These methods somehow succeed in correctly trading off bias, variance (and spread in Bayesian methods) in a purely data-driven manner. For several such result in the context of the signal-in-white-noise model, see, for instance, Giné and Nickl (2016) and the references therein. For distributed methods the issue of adaptation appears to be a lot more subtle. In this paper we only have a first, negative result on adaptive properties of distributed methods.

So now we do not assume that we know the true regularity β of the unknown signal. As before we employ the prior $\Pi(\cdot\,|\,\alpha)$ in the local machines. To tune the regularity parameter α

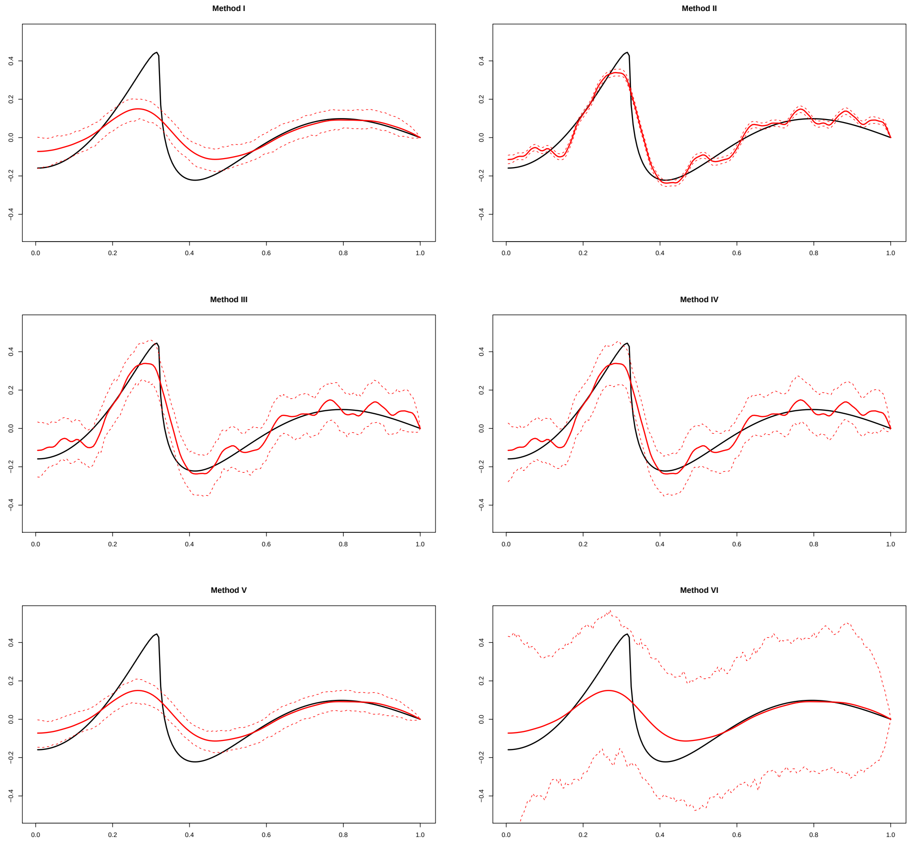

Figure 4: Global posterior mean (solid red curve) and 95% pointwise credible bands (dashed red curves) for each of the methods I–VI.

of the prior we consider a distributed version of maximum marginal likelihood estimator, as proposed by Deisenroth and Ng (2015). The usual, non-distributed version of that method would use the maximizer of the map

$$\alpha \mapsto \log \int \Big(\prod_{j=1}^{m} p(\mathbf{Y}^j \,|\, \theta) \Big)\, \Pi(d\theta \,|\, \alpha)$$

as tuning parameter. Maximizing this function however requires having all data available in a central machine. In the distributed setting, Deisenroth and Ng (2015) argue that this map is well approximated by the map

$$\alpha \mapsto \sum_{j=1}^{m} \log \Big(\int p(\mathbf{Y}^j \,|\, \theta)\, \Pi(d\theta \,|\, \alpha) \Big).$$

Now every term in the sum just depends on one of the local machines and this function can be maximized on the central machine by repeatedly asking the local machines for function evaluations and gradients of the local log-marginal likelihoods

$$\log \int p(\mathbf{Y}^j \,|\, \theta)\, \Pi(d\theta \,|\, \alpha).$$

The resulting estimator is denoted by $\hat{\alpha}$, i.e.

$$\hat{\alpha} = \underset{\alpha \in [0, \log n]}{\operatorname{argmax}} \sum_{j=1}^{m} \log \Big(\int p(\mathbf{Y}^j \,|\, \theta)\, \Pi(d\theta \,|\, \alpha) \Big).$$

(We maximize over a compact interval to ensure that the maximizer exists.)

It turns out that in the distributed setting, the local machines are in general not able to learn enough about the true signal regularity β. The following lemma asserts that there exist "difficult" signals for which the estimator $\hat{\alpha}$ overestimates the regularity.

Lemma 7 *For $\beta, M > 0$, consider a signal $\theta_0 \in \ell^2$ such that*

$$\theta_{0,i}^2 = \begin{cases} M^2 i^{-1-2\beta} & \text{if } i \geq (n/(\sigma^2 \sqrt{m}))^{1/(1+2\beta)}, \\ 0 & \text{else.} \end{cases} \tag{4.1}$$

Then $\theta_0 \in H^\beta(M)$ and if M is small enough, then

$$\mathrm{P}_{\theta_0}(\hat{\alpha} \geq \beta + 1/2) \to 1 \tag{4.2}$$

if $n/m \to \infty$ and $m \to \infty$.

Proof The proof is given in Section B.1. ■

In view of Lemma 7 it is perhaps not surprising that if the approximated maximum marginal likelihood estimator $\hat{\alpha}$ is used to tune the local prior that is used in every machine, sub-optimal performance is obtained for certain truths. Intuitively this is because due to the smaller signal-to-noise ratio, or "sample size" in the local machines, certain truths may appear more regular than they really are. It turns out that using the estimator $\hat{\alpha}$ in combination with any of the methods considered in the preceding section indeed leads to sub-optimal rates and bad coverage probabilities for certain truths. As an illustration we present a rigorous statement for the method of Section 3.4, but similar results can be derived for the others methods as well.

So suppose that in every local problem the prior $\Pi(\cdot \,|\, \alpha)$ is used, the corresponding generalized posterior $\tilde{\Pi}^j(\cdot \,|\, \mathbf{Y}^j)$ is computed locally (which involves raising the local likelihood to the power m), and then the tuning parameter α is substituted by the estimator $\hat{\alpha}$ defined above. In the central machine, the global "posterior" $\mathbf{\Pi}_{VII}(\cdot \,|\, \mathbf{Y})$ is constructed as the 2-Wasserstein barycenter of the local "posterior" measures $\tilde{\Pi}^1(\cdot \,|\, \hat{\alpha}, \mathbf{Y}^1), \ldots, \tilde{\Pi}^m(\cdot \,|\, \hat{\alpha}, \mathbf{Y}^m)$.

Theorem 8 *For $\beta, M > 0$ and θ_0 as in Lemma 7 we have, for some $c > 0$,*

$$\mathrm{E}_{\theta_0}\mathbf{\Pi}_{VII}(\theta :\, \|\theta - \theta_0\|_2 \leq c(n/\sqrt{m})^{-\frac{\beta}{1+2\beta}} |\mathbf{Y}) \to 0$$

as $m \to \infty$ and $n/m \to \infty$. Furthermore, for all $L > 0$ it holds that

$$\mathrm{P}_{\theta_0}\big(\theta_0 \in \hat{C}(L)\big) \to 0.$$

Proof See Section B.2. ■

A simulation illustrating the theoretical result of the theorem is given in Figure 2. The left panel visualizes the "posterior" generated by method VII, in the same distributed setting, and using the same simulated data as considered in Section 3.6.

So when combined with a data-driven tuning method like the distributed version of maximum marginal likelihood considered here, even the distributed methods that perform well in the non-adaptive setting loose their favourable properties. None of the methods yields a procedure that automatically adapts to regularity and achieves the optimal non-distributed rate. This does not imply of course that such an adaptive method does not exist. We expect however that the matter is delicate and that fundamental limitations exist.

The issue appears to be similar to that of the existence of adaptive confidence sets. To achieve adaptation in our distributed setting the local machines must be able to learn the "global" regularity of the signal from the limited local data that they have available. Analogous to the adaptive confidence problem we expect that this is in general only possible under additional assumptions on the true signal, like the self-similarity or polished tail conditions proposed for instance in Giné and Nickl (2010), Bull (2012), Szabó et al. (2015), Nickl and Szabó (2016), Belitser et al. (2017). Making these admittedly somewhat loose claims mathematically precise takes considerably more effort. Recent work shows that in a distributed setting with communication restrictions, some degree of adaptation to smoothness is in principle possible, but requires different kinds of algorithms (see Szabo and van Zanten (2019)). Many open questions remain at the moment however. It is for instance unclear how the possibility of adaptation, or purely data-driven tuning, is related the degree of communication or the amount of central computation allowed. It would in particular be interesting to better understand the theoretical performance of distributed methods which allow multiple rounds of communication (e.g. Shamir et al. (2014), Heinze et al. (2016), Wang et al. (2017a,b), Lu et al. (2016)).

Appendix A. Proofs for Section 3

A.1. Proof of Theorem 1

By completing the square we see that under the local posterior $\Pi^j(\cdot\,|\,\mathbf{Y}^j)$ the coefficients θ_i are independent and Gaussian, with mean $\hat\theta_i^j$ and variance s_i^2 given by

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

Hence the global "posterior" $\mathbf{\Pi}_I(\cdot\,|\,\mathbf{Y})$ is Gaussian as well, and under that measure the coefficients θ_i are independent and have mean $\hat\theta_i$ and variance t_i^2 given by

$$\hat\theta_i = \frac{1}{m}\sum_{j=1}^m \hat\theta_i^j, \qquad t_i^2 = \frac{s_i^2}{m}.$$

For the global posterior mean we have, for every $\theta_0 \in \ell^2$,

$$\mathrm{E}_{\theta_0}\hat\theta_i - \theta_{0,i} = \frac{-\sigma^2 m i^{1+2\beta}}{n+\sigma^2 m i^{1+2\beta}}\theta_{0,i}, \qquad \mathrm{Var}_{\theta_0}\hat\theta_i = \frac{\sigma^2 n}{(n+\sigma^2 m i^{1+2\beta})^2},$$

and hence,

$$\mathrm{E}_{\theta_0}\|\hat\theta-\theta_0\|_2^2 = \sum \frac{\sigma^4 m^2 i^{2+4\beta}}{(n+\sigma^2 m i^{1+2\beta})^2}\theta_{0,i}^2 + \sum \frac{\sigma^2 n}{(n+\sigma^2 m i^{1+2\beta})^2}.$$

By Lemma 9 the second, variance term is of the order

$$m^{-1/(1+2\beta))} n^{-2\beta/(1+2\beta)},$$

as $n/m \to \infty$. For $\theta_{0,i}^2 = M i^{-1-2\beta}$, by the same lemma, the first, squared bias term is proportional to $(n/m)^{-2\beta/(1+2\beta)}$. For the global spread, again in view of Lemma 9, we have

$$\sum t_i^2 = \sum \frac{\sigma^2}{n+\sigma^2 m i^{1+2\beta}} \asymp m^{-1/(1+2\beta))} n^{-2\beta/(1+2\beta)}. \tag{A.1}$$

By the triangle inequality we have

$$\begin{aligned}&\mathbf{\Pi}_I(\|\theta-\theta_0\|_2 \le c m^{\frac{\beta}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}}|\mathbf{Y})\\ &\le \mathbf{\Pi}_I(\theta:\ \|\mathrm{E}_{\theta_0}\hat\theta-\theta_0\|_2 - c m^{\frac{\beta}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}} - \|\hat\theta-\mathrm{E}_{\theta_0}\hat\theta\|_2 \le \|\theta-\hat\theta\|_2|\mathbf{Y}).\end{aligned}$$

It follows from the bounds on the variance and squared bias of the posterior mean that for θ_0 as chosen above, the quantity

$$\|\mathrm{E}_{\theta_0}\hat\theta-\theta_0\|_2 - c m^{\frac{\beta}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}} - \|\hat\theta-\mathrm{E}_{\theta_0}\hat\theta\|_2$$

appearing in the posterior probability is with P_{θ_0}-probability tending to one bounded from below by $c m^{\frac{\beta}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}}$ for $c > 0$ small enough. Then by the upper bound for the posterior spread and Chebyshev's inequality we obtain the first statement of the theorem.

For the coverage we note that the radius r_γ of the credible set is a multiple of $m^{-1/(2+4\beta)}n^{-\beta/(1+2\beta)}$, which follows from the Gaussianity of the posterior and (A.1). Then by similar computations as above we get that for the same truth θ_0,

$$\begin{aligned}
\mathrm{P}_{\theta_0}(\theta_0 \in \hat{C}(L)) &= \mathrm{P}_{\theta_0}(\|\hat\theta - \theta_0\|_2 \leq L r_\gamma) \\
&\leq \mathrm{P}_{\theta_0}\big(\|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2 \geq \|\mathrm{E}_{\theta_0}\hat\theta - \theta_0\|_2 - L r_\gamma\big) \\
&\leq \mathrm{P}_{\theta_0}\big(\|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2 \geq c m^{\frac{\beta}{1+2\beta}} n^{-\frac{\beta}{1+2\beta}}\big) \\
&\lesssim m^{\frac{-2\beta}{1+2\beta}} n^{\frac{2\beta}{1+2\beta}} \mathrm{E}_{\theta_0}\|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2^2 \lesssim m^{-1} \to 0.
\end{aligned}$$

This completes the proof of the theorem.

A.2. Proof of Theorem 2

Raising the local likelihood (3.4) to the power m makes it proportional to

$$\prod e^{-\frac{1}{2}\frac{n(Y_i^j - \theta_i)^2}{\sigma^2}},$$

which is the likelihood for the case $m = 1$. It follows that under the generalized local posterior $\tilde\Pi^j(\cdot \,|\, \mathbf{Y}^j)$ the coefficients θ_i are independent and Gaussian, with mean $\hat\theta_i^j$ and variance s_i^2 given by

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

Hence the global "posterior" $\mathbf{\Pi}_{II}(\cdot \,|\, \mathbf{Y})$ is again Gaussian, and under this global measure the coefficients θ_i are independent and have mean $\hat\theta_i$ and variance t_i^2 given by

$$\hat\theta_i = \frac{1}{m}\sum_{j=1}^m \hat\theta_i^j, \qquad t_i^2 = \frac{s_i^2}{m}.$$

For the global posterior mean we have in this case, for every $\theta_0 \in \ell^2$,

$$\mathrm{E}_{\theta_0}\,\hat\theta_i - \theta_{0,i} = \frac{-\sigma^2 i^{1+2\beta}}{n + \sigma^2 i^{1+2\beta}}\theta_{0,i}, \qquad \mathrm{Var}_{\theta_0}\,\hat\theta_i = \frac{\sigma^2 n}{(n + \sigma^2 i^{1+2\beta})^2},$$

and hence,

$$\mathrm{E}_{\theta_0}\|\hat\theta - \theta_0\|_2^2 = \sum \frac{\sigma^4 i^{2+4\beta}}{(n + \sigma^2 i^{1+2\beta})^2}\theta_{0,i}^2 + \sum \frac{\sigma^2 n}{(n + \sigma^2 i^{1+2\beta})^2}.$$

For all $\theta_0 \in H^\beta(M)$, in view of Lemma 9, the squared bias term is bounded by

$$M^2 \sum \frac{\sigma^4 i^{1+2\beta}}{(n + \sigma^2 i^{1+2\beta})^2} \lesssim M^2 n^{-2\beta/(1+2\beta)}$$

for large n, and the variance term behaves like a constant times $n^{-2\beta/(1+2\beta)}$ as well. The global spread $\sum t_i^2$ is of the order $m^{-1}n^{-2\beta/(1+2\beta)}$ for large n, again following from Lemma 9.

For $M_n \to \infty$ and $\theta_0 \in H^\beta(M)$ we now have, by the triangle inequality,

$$\begin{aligned}&\mathbf{\Pi}_{II}(\theta :\, \|\theta - \theta_0\|_2 \geq M_n n^{-\beta/(1+2\beta)} \,|\, \mathbf{Y})\\ &\leq \mathbf{\Pi}_{II}(\theta :\, \|\theta - \hat\theta\|_2 \geq M_n n^{-\beta/(1+2\beta)} - \|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2 - \|\theta_0 - \mathrm{E}_{\theta_0}\hat\theta\|_2 \,|\, \mathbf{Y}).\end{aligned}$$

By the bounds on the bias and the variance of the posterior mean derived above the quantity on the right of the inequality in the last posterior probability is bounded from below by $(M_n/2)n^{-\beta/(1+2\beta)}$ with P_{θ_0}-probability tending to one as $n, m \to \infty$, uniformly in $\theta_0 \in H^\beta(M)$. By Chebychev's inequality, and the bound on the posterior spread, we conclude that the first statement of the theorem holds.

For the second statement we first note that by Chebychev's inequality and by the upper bound on the posterior spread the radius r_γ of the credible set is for large n bounded by $Cm^{-1/2}n^{-\beta/(1+2\beta)}$ for some $C > 0$. Hence, since the posterior mean is Gaussian and the Gaussian measure of a ball of a fixed size is maximal if the ball is centered at the mean (a consequence of Anderson's inequality, e.g. Lifshits (1995), Section 11), we have

$$\begin{aligned}\mathrm{P}_{\theta_0}\big(\theta_0 \in \hat C(L)\big) &\leq \mathrm{P}_{\theta_0}\big(\|\hat\theta - \theta_0\|_2 \leq CLm^{-1/2}n^{-\beta/(1+2\beta)}\big)\\ &\leq \mathrm{P}_{\theta_0}\big(\|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2 \leq CLm^{-1/2}n^{-\beta/(1+2\beta)}\big).\end{aligned}$$

By Chebychev's inequality,

$$\mathrm{P}_{\theta_0}\Big(\|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2^2 \leq \sum \sigma_i^2 - a\sqrt{2\sum \sigma_i^4}\Big) \leq \frac{1}{a^2}$$

for all $a > 0$, where $\sigma_i^2 = \mathrm{Var}_{\theta_0}\hat\theta_i$. Above we saw that $\sum \sigma_i^2 \asymp n^{-2\beta/(1+2\beta)}$. Similarly, it is easily seen that $\sum \sigma_i^4 \asymp n^{(-1-4\beta)/(1+2\beta)}$. Hence by taking $a = n^{(1/4)/(1+2\beta)}$, for instance, we see that for $c > 0$ small enough,

$$\mathrm{P}_{\theta_0}\big(\|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2 \leq cn^{-\beta/(1+2\beta)}\big) \to 0$$

as $n \to \infty$. But then also

$$\mathrm{P}_{\theta_0}\big(\|\hat\theta - \mathrm{E}_{\theta_0}\hat\theta\|_2 \leq CLm^{-1/2}n^{-\beta/(1+2\beta)}\big) \to 0$$

as $m, n \to \infty$.

A.3. Proof of Theorem 3

In this case the jth local posterior is a product of Gaussians with means and variances given by

$$\hat\theta_i^j = \frac{n}{n + \sigma^2 m\tau^{-1} i^{1+2\alpha}} Y_i^j, \qquad s_i^2 = \frac{\sigma^2 m}{n + \sigma^2 m\tau^{-1} i^{1+2\alpha}}.$$

As before the global "posterior" is Gaussian as well, and under that measure the coefficients θ_i are independent and have mean $\hat\theta_i$ and variance t_i^2 given by

$$\hat\theta_i = \frac{1}{m}\sum_{j=1}^m \hat\theta_i^j, \qquad t_i^2 = \frac{s_i^2}{m}.$$

For the global posterior mean we have, for every $\theta_0 \in \ell^2$,

$$\mathrm{E}_{\theta_0}\hat{\theta}_i - \theta_{0,i} = \frac{-\sigma^2 m\tau^{-1} i^{1+2\alpha}}{n+\sigma^2 m\tau^{-1} i^{1+2\alpha}}\theta_{0,i}, \qquad \mathrm{Var}_{\theta_0}\hat{\theta}_i = \frac{\sigma^2 n}{(n+\sigma^2 m\tau^{-1} i^{1+2\alpha})^2},$$

and hence

$$\mathrm{E}_{\theta_0}\|\hat{\theta}-\theta_0\|_2^2 = \sum \frac{\sigma^4 m^2\tau^{-2} i^{2+4\alpha}}{(n+\sigma^2 m\tau^{-1} i^{1+2\alpha})^2}\theta_{0,i}^2 + \sum \frac{\sigma^2 n}{(n+\sigma^2 m\tau^{-1} i^{1+2\alpha})^2}.$$

Then in view of Lemma 9, we see that for $\beta < 1+2\alpha$ and uniformly for $\theta_0 \in H^\beta(M)$, the squared bias term is bounded by a constant times

$$M^2(\tau n/m)^{-2\beta/(1+2\alpha)}.$$

Similarly, the variance term and the posterior spread $\sum t_i^2$ both behave like a constant times

$$(\tau/m)^{1/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)}$$

as $n \to \infty$. The choice $\tau = mn^{2(\alpha-\beta)/(1+2\beta)}$ balances these quantities, so that all three are of the order $n^{-2\beta/(1+2\beta)}$.

By exactly the same reasoning as in Section A.2, the fact that the squared bias bound and the variance and spread are of the same order implies the first statement of the theorem. For the coverage statement we first note that the squared credible set radius r_γ^2 is the $1-\gamma$ quantile of the distribution of $\sum t_i^2 Z_i^2$, with t_i^2 as above and Z_i independent standard normals. This distribution has mean $\sum t_i^2 \asymp n^{-2\beta/(1+2\beta)}$ and variance

$$2\sum t_i^4 \asymp \frac{1}{n} n^{-2\beta/(1+2\beta)},$$

following from Lemma 9. As the standard deviation is of smaller order than the mean, it follows from Chebychev's inequality that $r_\gamma \geq cn^{-\beta/(1+2\beta)}$ for some $c > 0$. For the coverage probability we then have

$$\mathrm{P}_{\theta_0}\Big(\theta_0 \notin \hat{C}(L)\Big) \leq \mathrm{P}_{\theta_0}\Big(\|\hat{\theta}-\theta_0\|_2 \geq cLn^{-\beta/(1+2\beta)}\Big) \leq \frac{n^{2\beta/(1+2\beta)}}{c^2L^2}\mathrm{E}_{\theta_0}\|\hat{\theta}-\theta_0\|_2^2.$$

By the bounds on the bias and variance of the posterior mean the right-hand side is smaller than γ for L large enough, uniformly for $\theta_0 \in H^\beta(L)$.

A.4. Proof of Theorem 4

As we saw in Section A.2, the jth local generalized posterior is a product of Gaussians with means $\hat{\theta}_i^j$ and variances s_i^2 given by

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

In other words, the jth local measure is a Gaussian measure on ℓ^2 with mean $\hat{\theta}^j = (\hat{\theta}_i^j)_i$ and (diagonal) covariance operator $R: \ell^2 \to \ell^2$ given by $(Rx)_i = s_i^2 x_i$, which is the same for every local machine. The Wasserstein barycenter of a finite collection of Gaussian measures is a Gaussian measure again (e.g. Agueh and Carlier (2011)). By Theorem 3.5 of Gelbrich (1990) the squared 2-Wasserstein distance between the jth local measure and a Gaussian measure on ℓ^2 with mean μ and covariance operator K is given by

$$\|\hat{\theta}^j - \mu\|_2^2 + \mathrm{tr}(R) + \mathrm{tr}(K) - 2\mathrm{tr}\sqrt{R^{1/2}KR^{1/2}}.$$

It follows that the barycenter $\mathbf{\Pi}_{IV}(\cdot \,|\, \mathbf{Y})$ of the local generalized posteriors is the Gaussian measure on ℓ^2 with mean $\hat{\theta}$ equal to the average of the local means $\hat{\theta}^j$ and covariance operator equal to R. In other words, the global "posterior" is a product of Gaussians with means and variances given by

$$\hat{\theta}_i = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_i^j, \qquad t_i^2 = s_i^2.$$

So the global posterior mean is the same as in Section A.2 and the posterior spread $\sum t_i^2$ is a factor m larger. It then follows from the considerations in Section A.2 that the squared bias of the global posterior mean is bounded by a constant times $M^2 n^{-2\beta/(1+2\beta)}$, uniformly for $\theta_0 \in H^\beta(M)$. Moreover, the variance term $\sum s_i^2$ and the posterior spread $\sum t_i^2$ behave like a multiple of $n^{-2\beta/(1+2\beta)}$ as well. As was explained in Section A.3, this leads to the statement of the theorem.

A.5. Proof of Theorem 6

The proof of the first statement is the same as in Section A.1, since the mean of the global "posterior" is the same as for the naive averaging method.

For the second statement, we observe that for $\theta_0 \in H^\beta(M)$, the squared bias term for the posterior mean satisfies

$$\sum \frac{\sigma^4 m^2 i^{2+4\beta}}{(n+\sigma^2 m i^{1+2\beta})^2}\theta_{0,i}^2 \le M^2 \sum \frac{\sigma^4 m^2 i^{1+2\beta}}{(n+\sigma^2 m i^{1+2\beta})^2} \lesssim M^2 (n/m)^{-2\beta/(1+2\beta)}.$$

for $n/m \to \infty$. As was shown in Section A.1 the variance of the posterior mean behaves as $m^{-1/(1+2\beta)} n^{-2\beta/(1+2\beta)}$. Since the spread $\sum t_i^2$ of the posterior is a factor m larger than in Section A.1, it is of the same order $(n/m)^{-2\beta/(1+2\beta)}$ as the squared bias term. Since squared bias and spread are of the same order, the variance is of smaller order, and

$$\sqrt{\sum t_i^4} \asymp \Big(\frac{n}{m}\Big)^{\frac{-1/2}{1+2\beta}} \sum t_i^2$$

is of lower order than $\sum t_i^2$, the coverage statement can be proved as in Section A.3.

A.6. Technical Lemma

Lemma 9 *For $s, t > 0$ with $st > 1$ and $r < st - 1$ consider the function $f(x) = x^r(x^s+1)^{-t}$ and set $f_\nu = \sum_{k=1}^{\infty} \nu^{-1} f(k/\nu)$. Then as $\nu \to \infty$,*

(i) If $r > -1$, then $f_\nu \asymp \int_0^\infty f(x)dx$.

(ii) If $r=-1$, then $f_\nu \asymp \log\nu$.

(iii) If $r<-1$, then $f_\nu \asymp \nu^{r-1}\sum_{k=1}^\infty k^{-r}$.

Proof Assertions (ii) and (iii), along with (i) for $-1<r\leq 0$ are proved in Lemma A.1 of Szabó et al. (2013), hence it remains to verify assertion (i) for $0<r<st-1$. Note that

$$\sum_{k=1}^{N\nu}\nu^{-1}f(k/\nu) < f_\nu < \sum_{k=1}^{N\nu}\nu^{-1}f(k/\nu)+\sum_{k=N\nu+1}^{\infty}\nu^{-1}f(k/\nu). \tag{A.2}$$

Since the function $f(x)$ is continuous on $[0,N]$ it is Riemann integrable (see for instance Theorem 6.8 of Rudin (1976)), hence the right Riemann sum converges to the integral, i.e. $\sum_{k=1}^{N\nu}\nu^{-1}f(k/\nu)\to\int_0^N f(x)dx$ as $\nu\to\infty$ (for simplicity assume that $\nu\in\mathbb{N}$). Furthermore, for every $\nu>0$,

$$\sum_{k=N\nu+1}^{\infty}\nu^{-1}f(k/\nu)\leq \nu^{st-r-1}\sum_{k=N\nu+1}^{\infty}k^{r-st}\leq \nu^{st-r-1}\int_{N\nu}^\infty x^{r-st}dx\leq\frac{1}{st-r-1}N^{r+1-st}$$

and $\int_{N+1}^\infty f(x)dx\leq\int_{N+1}^\infty x^{r-st}dx\leq N^{-st+r+1}/(st-r-1)$. Therefore by choosing N sufficiently large, both sides of (A.2) gets arbitrarily close to $\int_0^\infty f(x)dx$ as $\nu\to\infty$, concluding the proof of the statement. ■

Appendix B. Proofs for Section 4

B.1. Proof of Lemma 7

The estimator $\hat\alpha$ is the maximizer of the random map $\alpha\mapsto\sum_j \ell_j(\alpha)$, where

$$\ell_j(\alpha)=\log\int p(\mathbf{Y}^j\,|\,\theta)\,\Pi(d\theta\,|\,\alpha).$$

The asymptotic behaviour of the local log-marginal likelihood ℓ_j has been studied in Knapik et al. (2016). Denote the derivative of ℓ_j with respect to α by $\dot\ell_j$ and let $k=n/(\sigma^2 m)$ be the local "sample size". Moreover, for $l>0$, define

$$\underline{\alpha}=\inf\{\alpha>0: h_k(\alpha)>l\}\wedge\sqrt{\log k},$$

where

$$h_k(\alpha)=\frac{1+2\alpha}{k^{1/(1+2\alpha)}\log k}\sum_i\frac{k^2 i^{1+2\alpha}\theta_{0,i}^2\log i}{(k+i^{1+2\alpha})^2}.$$

Note that the expectation $\mathrm{E}_{\theta_0}\dot\ell_j(\alpha)$ does not depend on j. It is proved in Section 5.3 of Knapik et al. (2016) that if l is smaller than some universal threshold, then for every j

$$\liminf_{k\to\infty}\inf_{\alpha\leq\underline{\alpha}}\frac{1+2\alpha}{k^{1/(1+2\alpha)}\log k}\mathrm{E}_{\theta_0}\dot\ell_j(\alpha)=\delta>0,$$

$$\mathrm{E}_{\theta_0} \sup_{\alpha \leq \underline{\alpha}} \frac{1+2\alpha}{k^{1/(1+2\alpha)}\log k} |\dot{\ell}_j(\alpha) - \mathrm{E}_{\theta_0}\dot{\ell}_j(\alpha)| \lesssim e^{-C\sqrt{\log k}}$$

for constants $\delta, C > 0$. But then we also have

$$\liminf_{k\to\infty} \inf_{\alpha\leq\underline{\alpha}} \frac{1+2\alpha}{k^{1/(1+2\alpha)}\log k} \mathrm{E}_{\theta_0} \sum_j \dot{\ell}_j(\alpha) > m\delta$$

and

$$\mathrm{E}_{\theta_0} \sup_{\alpha\leq\underline{\alpha}} \frac{1+2\alpha}{k^{1/(1+2\alpha)}\log k} \Big| \sum_j \dot{\ell}_j(\alpha) - \mathrm{E}_{\theta_0} \sum_j \dot{\ell}_j(\alpha) \Big| \lesssim m e^{-C\sqrt{\log k}}.$$

By Markov's inequality, it follows that with probability at least $1 - C_1 \exp(-C_2\sqrt{\log k})$ the map $\alpha \mapsto \sum_j \ell_j(\alpha)$ is strictly increasing on the interval $[0, \underline{\alpha}]$. Hence, on that event we have $\hat{\alpha} \geq \underline{\alpha}$.

It remains to show that $\underline{\alpha} \geq \beta + 1/2$. To that end it suffices to prove that $h_k(\alpha) \leq l$ for all $\alpha \leq \beta + 1/2$. To see this, suppose first that $\alpha < \beta$. Define $N_\beta = (n/(\sigma^2\sqrt{m}))^{1/(1+2\beta)}$ and $M_\alpha = k^{1/(1+2\alpha)}$. By definition of θ_0 we then have

$$\begin{aligned}
h_k(\alpha) &= \frac{M^2}{M_\alpha \log M_\alpha} \sum_{i=N_\beta}^{\infty} \frac{k^2 i^{2\alpha-2\beta}\log i}{(i^{1+2\alpha}+k)^2} \\
&\leq \frac{M^2}{M_\alpha \log M_\alpha} \sum_{i=N_\beta}^{M_\alpha} i^{2\alpha-2\beta}\log i + \frac{M^2k^2}{M_\alpha \log M_\alpha} \sum_{i=M_\alpha}^{\infty} i^{-2-2\alpha-2\beta}\log i \\
&\leq M^2 \frac{M_\alpha N_\beta^{2\alpha-2\beta}\log M_\alpha}{M_\alpha \log M_\alpha} + M^2 \frac{k^2 M_\alpha^{-1-2\alpha-2\beta}\log M_\alpha}{M_\alpha \log M_\alpha} \\
&\lesssim M^2
\end{aligned}$$

for n, m large enough. Hence, if M is small enough, then $h_k \leq l$ for $\alpha < \beta$. For $\beta \leq \alpha \leq \beta + 1/2$ we have

$$\begin{aligned}
h_k(\alpha) &\leq \frac{M^2k^2}{M_\alpha \log M_\alpha} \sum_{i=N_\beta}^{\infty} i^{-2-2\alpha-2\beta}\log i \\
&\leq \frac{M^2k^2 N_\beta^{-1-2\alpha-2\beta}\log N_\beta}{M_\alpha \log M_\alpha} \\
&= M^2 (n/\sigma^2)^{\frac{2\alpha}{1+2\alpha}-\frac{2\alpha}{1+2\beta}} m^{-2+\frac{1}{1+2\alpha}+\frac{1+2\alpha+2\beta}{2(1+2\beta)}} \frac{\log N_\beta}{\log M_\alpha} \lesssim m^{-\frac{2\alpha}{1+2\alpha}} \log m
\end{aligned}$$

for n/m large enough. Together, this shows that if both n/m and m are large enough, then indeed $h_k(\alpha) \leq l$ for all $\alpha \leq \beta + 1/2$.

B.2. Proof of Theorem 8

In view of the proof of Theorem 4 the jth local generalized posterior is a product of Gaussians with means $\hat{\theta}_i^j$ and variances s_i^2 given by

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

Using again that the Wasserstein barycenter of a finite collection of Gaussian measures is a Gaussian measure in combination with the explicit expression for the 2-Wasserstein distance between Gaussians (see Section A.4) we see that the global "posterior" is a product of Gaussians with means $\hat{\theta}_i$ and variances t_i^2 given by

$$\hat{\theta}_i = \frac{1}{m}\sum_{j=1}^m \hat{\theta}_i^j, \qquad t_i^2 = s_i^2.$$

The posterior mean can be written as $\hat{\theta} = \hat{\theta}(\hat{\alpha})$, where $\hat{\theta}(\alpha)$ is the estimator with a fixed choice α for the hyperparameter, i.e.

$$\hat{\theta}_i(\alpha) = \frac{1}{m}\sum_{j=1}^m \frac{n}{n+\sigma^2 i^{1+2\alpha}} Y_i^j.$$

For fixed α we also define the corresponding expectation $E(\alpha) = \mathrm{E}_{\theta_0}\hat{\theta}(\alpha)$. Then by the triangle inequality,

$$\|\hat{\theta} - \theta_0\|_2 \geq \|E(\hat{\alpha}) - \theta_0\|_2 - \|E(\hat{\alpha}) - \hat{\theta}(\hat{\alpha})\|_2.$$

We have the explicit expressions

$$\|E(\alpha) - \theta_0\|_2^2 = \sum_i \frac{\sigma^4 i^{2+4\alpha}\theta_{0,i}^2}{(n+\sigma^2 i^{1+2\alpha})^2}$$

and

$$\|E(\alpha) - \hat{\theta}(\alpha)\|_2^2 = \sum_i \frac{\sigma^2 n}{(n+\sigma^2 i^{1+2\alpha})^2}\Big(\frac{1}{\sqrt{m}}\sum_{j=1}^m Z_i^j\Big)^2.$$

Since the first expression is increasing in α and the second one is decreasing, we see that on the event $A = \{\hat{\alpha} \geq \beta + 1/2\}$ it holds that

$$\|\hat{\theta} - \theta_0\|_2 \geq \sqrt{\sum_i \frac{\sigma^4 \theta_{0,i}^2 i^{4+4\beta}}{(n+\sigma^2 i^{2+2\beta})^2}} - \sqrt{\sum_i \frac{\sigma^2 n}{(n+\sigma^2 i^{2+2\beta})^2}\Big(\frac{1}{\sqrt{m}}\sum_{j=1}^m Z_i^j\Big)^2}.$$

By definition of θ_0, the square of the first term on the right is bounded from below by

$$M^2 \sum_{i \geq \big((n/(\sigma^2\sqrt{m}))\big)^{1/(1+2\beta)}} \frac{\sigma^4 i^{3+2\beta}}{(n+\sigma^2 i^{2+2\beta})^2}.$$

In view of Lemma 9 we see that it is of the order $M^2(n/\sqrt{m})^{-2\beta/(1+2\beta)}$. The square of the second term can be written as

$$\sum_i \frac{\sigma^2 n}{(n+\sigma^2 i^{2+2\beta})^2} U_i^2,$$

with the U_i independent and standard normal under P_{θ_0}. Again in view of Lemma 9 it is easily seen that the mean and variance of this sum behave as $n^{-(1+2\beta)/(2+2\beta)}$ and $n^{-(3+4\beta)/(2+2\beta)}$, respectively. Hence the standard deviation is of smaller order than the mean for large n, so that by Chebychev's inequality the square of the second term is of stochastic order $n^{-(1+2\beta)/(2+2\beta)}$. Since this is of smaller order than $(n/\sqrt{m})^{-2\beta/(1+2\beta)}$, we conclude that for the global "posterior" mean we have, for some constant $c > 0$,

$$\mathrm{P}_{\theta_0}(\|\hat{\theta} - \theta_0\|_2 \geq c(n/\sqrt{m})^{-\beta/(1+2\beta)}) \to 1$$

as $n/m \to \infty$ and $m \to \infty$. The spread $\sum t_i^2$ of the global posterior is on the event A bounded by

$$\sum_i \frac{1}{n + i^{2+2\beta}},$$

which is of the order $n^{-(1+2\beta)/(2+2\beta)} \ll (n/\sqrt{m})^{-2\beta/(1+2\beta)}$ as well, see Lemma 9. The conclusions of the theorem now follow.

Acknowledgments

The research leading to these results has received funding from the Netherlands Science foundation NWO and from the European Research Council under ERC Grant Agreement 320637.

References

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Eduard Belitser et al. On coverage and local radial rates of credible sets. *The Annals of Statistics*, 45(3):1124–1151, 2017.

Lawrence D. Brown and Mark G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.

Adam D. Bull. Honest adaptive confidence bands and self-similar functions. *Electron. J. Statist.*, 6:1490–1516, 2012.

Y. Cao and D. J. Fleet. Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions. *ArXiv e-prints*, October 2014.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

Marc Deisenroth and Jun Wei Ng. Distributed Gaussian processes. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1481–1490, Lille, France, 07–09 Jul 2015. PMLR.

Matthias Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Evarist Giné and Richard Nickl. Confidence bands in density estimation. *Ann. Statist.*, 38 (2):1122–1170, 04 2010. doi: 10.1214/09-AOS738.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press. 2016.

Christina Heinze, Brian McWilliams, and Nicolai Meinshausen. Dual-loco: Distributing statistical estimation using random projections. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 875–883, Cadiz, Spain, 09–11 May 2016. PMLR.

I. M. Johnstone. *Gaussian estimation: Sequence and wavelet models*. Book draft, 2017.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-Efficient Distributed Statistical Inference. *ArXiv e-prints*, May 2016.

B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657, 10 2011.

B. T. Knapik, B. T. Szabó, A. W. Vaart, and J. H. Zanten. Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164(3):771–813, 2016.

Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer, 2012.

J. D. Lee, Y. Sun, Q. Liu, and J. E. Taylor. Communication-efficient sparse regression: a one-shot approach. *ArXiv e-prints*, March 2015.

Mikhail Anatolevich Lifshits. *Gaussian random functions*. Springer, 1995.

J. Lu, G. Cheng, and H. Liu. Nonparametric Heterogeneity Testing For Massive Data. *ArXiv e-prints*, January 2016.

S. Minsker, S. Srivastava, L. Lin, and D. B. Dunson. Robust and scalable Bayes via a median of subset posterior measures. *ArXiv e-prints*, March 2014.

J. W. Ng and M. P. Deisenroth. Hierarchical Mixture-of-Experts Model for Large-Scale Gaussian Process Regression. *ArXiv e-prints*, December 2014.

Richard Nickl and Botond Szabó. A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and their Applications*, 126(12):3913–3934, 2016.

Michael Nussbaum. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6):2399–2430, 1996.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976.

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*, 11:78–88, 2016.

Ohad Shamir, Nathan Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1000–II–1008. JMLR.org, 2014.

Z. Shang and G. Cheng. A Bayesian splitotic theory for nonparametric models. *ArXiv e-prints*, August 2015.

Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 912–920, San Diego, California, USA, 09–12 May 2015. PMLR.

B. Szabo and H. van Zanten. Adaptive distributed methods under communication constraints. *ArXiv e-prints*, 2019.

B. T. Szabó, A. W. van der Vaart, and J. H. van Zanten. Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Statist.*, 7:991–1018, 2013.

Botond Szabó, A. W. van der Vaart, and J. H. van Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.*, 43(4):1391–1428, 08 2015.

Alexandre B. Tsybakov. *Introduction to nonparametric estimation.* Springer, New York, 2009.

Aad van der Vaart and J. H. van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electron. J. Statist.*, 1:433–448, 2007.

S. Volgushev, S.-K. Chao, and G. Cheng. Distributed inference for quantile regression processes. *ArXiv e-prints*, January 2017.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3636–3645, International Convention Centre, Sydney, Australia, 06–11 Aug 2017a. PMLR.

Jialei Wang, Jason Lee, Mehrdad Mahdavi, Mladen Kolar, and Nati Srebro. Sketching Meets Random Projection in the Dual: A Provable Recovery Algorithm for Big and High-dimensional Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1150–1158, Fort Lauderdale, FL, USA, 20–22 Apr 2017b. PMLR.

Y. Zhu and J. Lafferty. Distributed Nonparametric Regression under Communication Constraints. *ArXiv e-prints*, March 2018.