Adaptation Using Spatially Distributed Gaussian Processes

Szabo, Botond; Hadji, Amine; van der Vaart, Aad

DOI
10.1080/01621459.2025.2501717

Publication date
2025

Document Version
Final published version

Published in
Journal of the American Statistical Association

Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Adaptation Using Spatially Distributed Gaussian Processes

Botond Szabo, Amine Hadji & Aad van der Vaart

To cite this article: Botond Szabo, Amine Hadji & Aad van der Vaart (10 Jul 2025): Adaptation Using Spatially Distributed Gaussian Processes, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2501717

To link to this article: https://doi.org/10.1080/01621459.2025.2501717

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2025.2501717

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Adaptation Using Spatially Distributed Gaussian Processes

Botond Szabo[a], Amine Hadji[b], and Aad van der Vaart[c]

[a]Department of Decision Sciences and BIDSA, Bocconi University, Milano, Italy; [b]Mathematical Institute, Leiden University, Leiden, The Netherlands; [c]Delft Institute of Applied Mathematics, DIAM, Delft University of Technology, Delft, The Netherlands

ABSTRACT

We consider the accuracy of an approximate posterior distribution in nonparametric regression problems by combining posterior distributions computed on subsets of the data defined by the locations of the independent variables. We show that this approximate posterior retains the rate of recovery of the full data posterior distribution, where the rate of recovery adapts to the smoothness of the true regression function. As particular examples we consider Gaussian process priors based on integrated Brownian motion and the Matérn kernel augmented with a prior on the length scale. Besides theoretical guarantees we present a numerical study of the methods both on synthetic and real world data. We also propose a new aggregation technique, which numerically outperforms previous approaches. Finally, we demonstrate empirically that spatially distributed methods can adapt to local regularities, potentially outperforming the original Gaussian process. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received March 2024
Accepted April 2025

KEYWORDS
Adaptation; Bayesian asymptotics; Distributed computation; Gaussian processes; Nonparametric Bayes

1. Introduction

Gaussian processes (GPs) are standard tools in statistical and machine learning. They provide a particularly effective prior distribution over the space of functions and are routinely used in regression and classification tasks, amongst others. The monograph Rasmussen and Williams (2006) gives an in-depth overview of the foundations and practical applications of this approach. However, GPs scale poorly with the sample size n. For instance, in regression the computational complexity and memory requirements are of the orders $O(n^3)$ and $O(n^2)$, respectively. This limitation has triggered the development of various approximation methods, including sparse approximations of the empirical covariance matrices Gibbs, Poole Jr, and Stockmeyer (1976), Saad (1990), and Quiñonero-Candela and Rasmussen (2005), variational Bayes approximations Titsias (2009), David, Rasmussen, and van der Wilk (2019), Nieman, Szabo, and Van Zanten (2022) or distributed methods. We focus on the latter method in this article.

In distributed (or divide-and-conquer) methods, the computational burden is reduced by splitting the data over "local" machines (or servers, experts, or cores). Next the computations are carried out locally, in parallel to each other, before transmitting the outcomes to a "central" server or core, where the partial, local results are combined, forming the final outcome of the procedure. This distributed architecture occurs naturally when data is collected and processed locally and only a summary statistic is transmitted to a central server. Besides speeding up the computations and reducing the memory requirements, distributed methods can also help in protecting privacy, as the data do not have to be stored in a central data base, but are processed locally.

In the literature various distributed methods were proposed to speed up Bayesian computation, in particular in the context of Gaussian Processes. One can distinguish two main strategies depending on the data-splitting technique. The first approach is to partition the data randomly over the servers, computing a posterior distribution on each server and finally aggregating these local distributions by some type of averaging. Examples include Consensus Monte Carlo Scott et al. (2016), WASP Srivastava et al. (2015), Generalized Product of Experts Cao and Fleet (2014), Bayesian Committee Machine Tresp (2000), Deisenroth and Ng (2015), Distributed Bayesian Varying Coefficients Guhaniyogi et al. (2022) and Distributed Kriging Guhaniyogi (2017). The second approach takes advantage of the spatial structure of the data and splits the observations based on a partition of the design space. Each machine is assigned a specific region of the space, a local posterior distribution is computed using the data in this region, and these are glued together to form the final answer. This approach is referred to as the Naive-Local-Experts model, see Kim, Mallick, and Holmes (2005) and Vasudevan et al. (2009). We discuss various methods of combining the local outputs in Section 4.2, including a new proposal, which outperforms its competitors in the case that the length scale of the priors is determined from the data.

A number of papers in the literature studied the randomly split data approach, deriving theoretical guarantees, but also limitations, for a range of methods and models. Under the assumption that the regularity of the underlying functional parameter is

CONTACT Botond Szabo botond.szabo@unibocconi.it Department of Decision Sciences and BIDSA, Bocconi University, Via Guglielmo Roentgen 1, Milano, Italy.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

known, minimax rate-optimal contraction rates and frequentist coverage guarantees for Bayesian credible sets were derived in the context of the Gaussian white noise Szabó and van Zanten (2019) and nonparametric regression models Guhaniyogi (2017), Shang and Cheng (2015), and Hadji, Hesselink, and Szabó (2022). However, in practice the latter regularity is typically not known, but inferred from the data in some way. In the mentioned references it is shown that with randomly split data, standard adaptation techniques necessarily lead to highly sub-optimal inference, see Szabó and van Zanten (2019). In the numerical analysis in the present article we observe this on both synthetic and real world datasets.

In contrast, spatially partitioned distributed approaches have received little theoretical attention, despite their popularity in applications. In this article we aim to fill this gap in the literature. We derive general contraction rate theorems under mild assumptions and apply them in the context of the nonparametric regression and classification models. We also consider two specific GP priors: the rescaled integrated Brownian motion and the Matérn process and show that both priors (augmented with an additional layer of prior on the scale parameter) lead to rate-adaptive posterior contraction rates. This is in sharp contrast to the randomly split data framework, which necessarily results in sub-optimal estimation. Thus, we provide the first adaptive distributed Bayesian method with theoretical guarantees. We also demonstrate the superior performance of spatially distributed methods on synthetic and real world datasets. Furthermore, we propose a novel aggregation technique, which numerically outperforms its close competitors, especially in the realistic situation that the length scales of the local posteriors are adapted to the data. Finally, we also demonstrate numerically that the spatially distributed GP methods can adapt to different local regularities in contrast to the original GP. Therefore, spatially distributing the data can not only speed up the computations, but can potentially improve the accuracy of the GP method.

The article is organised as follows. In Section 2 we introduce the spatially distributed general framework with GP priors and recall the regression and classification models, considered as examples in our article. Then in Sections 2.1 and 2.2 we derive general contraction rate results under mild conditions in the non-adaptive and adaptive frameworks, respectively, using the hierarchical Bayesian method in the latter one. As specific examples we consider the rescaled integrated Brownian motion and the Matérn process in Sections 3.1 and 3.2, respectively. For both priors we derive rate-adaptive contraction rates in the regression and classification models using the fully Bayesian approach. The theoretical guarantees are complemented with a numerical analysis. In Section 4.1 we discuss various aggregation techniques. We investigate their numerical properties compared to benchmark distributed and non-distributed methods on synthetic and real world datasets in Sections 4.2 and 4.3, respectively. We discuss our results and future directions in Section 5. The proofs for the general theorems are given in Section B and for the specific examples in Section C. A collection of auxiliary lemmas is presented in Section D. Finally, additional numerical analysis on real and simulated datasets are provided in Sections E and F, respectively. We highlight that local adaptation of the process is investigated in Section F.3 of the supplement.

We write $C^\beta([a,b])$ for the Hölder space of order $\beta > 0$: the space of functions $f : [a,b] \to \mathbb{R}$ that are b times differentiable, for b the largest integer strictly smaller than β, with highest order derivative $f^{(b)}$ satisfying $|f^{(b)}(x)-f^{(b)}(y)| \lesssim |x-y|^{\beta-b}$, for every $x, y \in [a,b]$. We also write $H^\beta([a,b])$ for the Sobolev space of order β.

2. Spatially Distributed Bayesian Inference with GP Priors

We consider general nonparametric regression models. The observations are independent pairs $(x_1, Y_1), \dots, (x_n, Y_n)$, where the covariates x_i are considered to be fixed and the corresponding dependent variables Y_i random. We state our abstract theorems in this general setting, but next focus on two commonly used models: nonparametric regression with Gaussian errors and logistic regression.

In the standard nonparametric regression model the observed data $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n) \in \mathbb{R}^n$ satisfy the relation

$$Y_i = f_0(x_i) + Z_i, \qquad Z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \qquad i = 1, \dots, n. \quad (1)$$

The goal is to estimate the unknown regression function f_0, which is assumed to be smooth, but not to take a known parametric form. In the logistic regression model the observed data $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n) \in \{0,1\}^n$ are binary with likelihood function

$$\Pr(Y_i = 1|x_i) = \psi\big(f_0(x_i)\big), \qquad i = 1, \dots, n, \quad (2)$$

where $\psi(x) = 1/(1+e^{-x})$ denotes the logistic function. Additional examples include Poisson regression, binomial regression, etc.

We consider the distributed version of these models. We assume that the data is spatially distributed over m machines in the following way. The kth machine, for $k \in \{1, \dots, m\}$, receives the observations Y_i with design points x_i belonging to the kth subregion $\mathcal{D}^{(k)}$ of the design space $\mathcal{D}$, that is $x_i \in \mathcal{D}^{(k)}$. In the examples in Section 3 we take the domain of the regression function f_0 to be the unit interval $[0,1]$ and split it into equidistant sub-intervals $I^{(k)} = (\frac{k-1}{m}, \frac{k}{m}]$. We use the shorthand notations $\mathbf{x}^{(k)} = \{x_i : x_i \in \mathcal{D}^{(k)}\}$, and $\mathbf{Y}^{(k)} = \{Y_i : x_i \in \mathcal{D}^{(k)}\}$. However, our results hold more generally, also in the multivariate setting. For simplicity of notation, we assume that $|\mathbf{x}^{(k)}| = n/m$, but in general it is enough that the number of points in each subregion is proportional to n/m (with respect to a universal constant).

We endow the functional parameter f_0 in each machine with a Gaussian Process prior $(G_t^{(k)} : t \in \mathcal{D})$, with sample paths supported on the full covariate space, identical in distribution but independent across the machines. Gaussian prior processes often depend on regularity and/or scale hyperparameters, which can be adjusted to achieve bigger flexibility. Corresponding local posteriors are computed based on the data $\mathbf{Y}^{(k)} = \{Y_i : x_i \in \mathcal{D}^{(k)}\}$ available at the local machines, independently across the machines. These define stochastic processes (supported on the full covariate space), which we aggregate into a single one by restricting them to the corresponding subregions and pasting them together, that is a draw f from the "aggregated posterior" is defined as

$$f(x) = \sum_{k=1}^{m} 1_{\mathcal{D}^{(k)}}(x) f^{(k)}(x), \tag{3}$$

where $f^{(k)}$ is a draw from the kth local posterior. Formally, an "aggregated posterior measure" is defined as

$$\Pi_{n,m}(B|\mathbf{Y}) = \prod_{k=1}^{m} \Pi^{(k)}(B_k|\mathbf{Y}^{(k)}), \tag{4}$$

where B is a measurable set of functions, $\Pi^{(k)}(\cdot|\mathbf{Y}^{(k)})$ is the posterior distribution in the kth local problem corresponding to the prior $\Pi^{(k)}(\cdot)$, and B_k is the set of all functions whose restriction to $\mathcal{D}^{(k)}$ agrees with the corresponding restriction of some element of B, that is

$$B_k = \Big\{\vartheta : [0,1] \to \mathbb{R} \Big| \exists f \in B \text{ such that } \vartheta(x) = f(x),\ \forall x \in \mathcal{D}^{(k)}\Big\}.$$

2.1. Posterior Contraction for Distributed GP for Independent Observations

We investigate the contraction rate of the aggregate posterior $\Pi_{n,m}(\cdot|\mathbf{Y})$ given in (4). Our general result is stated in terms of a local version of the concentration function originally introduced in van der Vaart and van Zanten (2008) for the non-distributed model. This local concentration function is attached to the restriction of the Gaussian prior process to the subregion $\mathcal{D}^{(k)}$. For $k = 1, \dots, m$, let $\|f\|_{\infty,k} = \sup_{x\in\mathcal{D}^{(k)}} |f(x)|$ denote the L_∞-norm restricted to $\mathcal{D}^{(k)}$, and define

$$\phi_{f_0}^{(k)}(\varepsilon) = \inf_{h\in\mathbb{H}^{(k)}:\|f_0-h\|_{\infty,k}\le\varepsilon} \|h\|^2_{\mathbb{H}^{(k)}} - \log \Pi^{(k)}\left(f : \|f\|_{\infty,k} < \varepsilon\right), \tag{5}$$

where $\|\cdot\|_{\mathbb{H}^{(k)}}$ is the norm corresponding to the Reproducing Kernel Hilbert Space (RKHS) $\mathbb{H}^{(k)}$ of the Gaussian process $(G_t^{(k)} : t \in \mathcal{D}^{(k)})$ and $\Pi^{(k)}$ denotes the law of this process.

We consider contraction rates relative to the semimetric d_n, whose square is given by

$$d_n(f,g)^2 = \frac{1}{n}\sum_{i=1}^{n} h_i(f,g)^2, \tag{6}$$

with

$$h_i(f,g)^2 = \int \left(\sqrt{p_{f,i}} - \sqrt{p_{g,i}}\right)^2 d\mu_i,$$

where $p_{f,i}$ denotes the density of Y_i given x_i and f with respect to some dominating measure μ_i, for $i = 1, \dots, n$. The semimetrics d_n are convenient for general theory, but as we discuss below, our results can be extended to other semimetrics as well, for instance to the empirical L_2-distance $\|f-g\|_n$, for $\|f\|_n^2 = n^{-1}\sum_{i=1}^{n} f^2(x_i)$, in the case of the nonparametric regression model.

The following standard and mild assumption relate the supremum norm to the d_n semimetric and Kullback-Leibler divergence and variation.

Assumption 1. For all bounded functions f, g,

$$\max\left\{h_i(f,g)^2, K(p_{f,i},p_{g,i}), V(p_{f,i},p_{g,i})\right\} \lesssim \|f-g\|^2_{\infty,k},$$

where $K(p_{f,i},p_{g,i}) = \int \log(p_{f,i}/p_{g,i}) p_{f,i}\, d\mu_i$ and $V(p_{f,i},p_{g,i}) = \int \log(p_{f,i}/p_{g,i})^2 p_{f,i}\, d\mu_i$.

For instance, in the case of the nonparametric regression model, the maximum in the left hand side is bounded above by a multiple of $(f(x_i) - g(x_i))^2 \le \|f-g\|^2_{\infty,k}$, for any $x_i \in \mathcal{D}^{(k)}$; see for example p. 214 of Ghosal and van der Vaart (2007). The condition also holds for the logistic regression model; see for instance the proof of Lemma 2.8 of Ghosal and van der Vaart (2017).

The preceding assumption suffices to express the posterior contraction rate with the help of the local concentration functions. The proof of the following theorem can be found in Section B.1.

Theorem 2. Let f_0 be a bounded function and assume that there exists a sequence $\varepsilon_n \to 0$ with $(n/m^2)\varepsilon_n^2 \to \infty$ such that $\phi_{f_0}^{(k)}(\varepsilon_n) \le (n/m)\varepsilon_n^2$, for $k = 1, \dots, m$. Then under Assumption 1, the aggregated posterior given in (4) contracts around the truth with rate ε_n, that is

$$\mathbb{E}_0\, \Pi_{n,m}\big(f : d_n(f,f_0) \ge M_n\varepsilon_n|\mathbf{Y}\big) \to 0,$$

for arbitrary $M_n \to \infty$. In the distributed nonparametric regression model (1) or the classification model (2), the condition $(n/m^2)\varepsilon_n^2 \to \infty$ may be relaxed to $m = o(n\varepsilon_n^2/\log n)$.

Remark 3. We note that the frequentist contraction rate guarantees for the spatially distributed methods, given in Theorem 2, hold more generally, beyond Gaussian Process priors. The general results can be stated with the help of the appropriately adapted versions of the prior small ball, remaining mass and entropy conditions to the distributed setting. We provide such a general result in Theorem 4, for the adaptive, hierarchical choice of the prior.

2.2. Adaptation

It is common practice to tune the prior GP by changing its "length scale" and consider the process $t \mapsto G_t^\tau := G_{\tau t}$, for a given parameter τ instead of the original process. Even though the qualitative smoothness of the sample paths does not change, a dramatic impact on the posterior contraction rate can be observed when $\tau = \tau_n$ tends to infinity or zero with the sample size n. A length scale $\tau > 1$ entails shrinking a process on a bigger time set to the time set $[0,1]$, whereas $\tau < 1$ corresponds to stretching. Intuitively, shrinking makes the sample paths more variable, as the randomness on a bigger time set is packed inside a smaller one, whereas stretching creates a smoother process. We show in our examples that by optimally choosing the scale hyper-parameter, depending on the regularity of the true function f_0 and the GP, one can achieve rate-optimal contraction for the aggregated posterior. We also show that this same rate-optimal contraction (up to an arbitrary level set by the user) is achieved in a data-driven way by choosing the length scale from a prior, without knowledge of the regularity of the underlying function f_0. Thus, we augment the model with another layer of prior, making the scale parameter τ a random variable, in a fully Bayesian approach.

Each local problem, for $k = 1, \dots, m$, receives its own scale parameter, independently of the other problems, and a local posterior is formed using the local data of each problem independently across the local problems. For simplicity we use the same prior for τ in each local problem. If this is given by a Lebesgue density g and $\Pi^{\tau,(k)}$ is the prior on f with scale τ used in the kth local problem, then the hierarchical prior for f in the kth local problem takes the form

$$\Pi^{g,(k)}(\cdot) = \int \Pi^{\tau,(k)}(\cdot) g(\tau)\, d\tau. \tag{7}$$

After forming a local posterior using this prior and the corresponding local data in each local problem, an aggregated posterior is constructed as in the non-adaptive case, that is a draw f from the aggregated posterior is given in (3). The corresponding aggregated posterior measure takes the form

$$\Pi^g_{n,m}(B|\mathbf{Y}) = \prod_{k=1}^{m} \Pi^{g,(k)}(B_k|\mathbf{Y}^{(k)}), \tag{8}$$

for $\Pi^{g,(k)}(\cdot|\mathbf{Y}^{(k)})$ the kth posterior distribution corresponding to the prior (7).

Theorem 4. Let f_0 be a bounded function and assume that there exist measurable sets of functions $B^{(k)}_{n,m}$ such that for all local hierarchical priors $\Pi^{g,(k)}$ given in (7) and $\varepsilon_n \to 0$ such that $(n/m^2)\varepsilon_n^2 \to \infty$, it holds that, for some $c, C > 0$,

$$\Pi^{g,(k)}(f : f \notin B^{(k)}_{n,m}) \le e^{-4(n/m)\varepsilon_n^2}, \tag{9}$$

$$\Pi^{g,(k)}(f : \|f - f_0\|_{\infty,k} \le \varepsilon_n) \ge e^{-(n/m)\varepsilon_n^2}, \tag{10}$$

$$\log N(c\varepsilon_n, B^{(k)}_{n,m}, \|\cdot\|_{\infty,k}) \le C(n/m)\varepsilon_n^2. \tag{11}$$

Then under Assumption 1, the aggregated hierarchical posterior given in (8), contracts around the truth with rate ε_n, that is

$$\mathbb{E}_0\, \Pi^g_{n,m}\big(d_n(f, f_0) \ge M_n\varepsilon_n|\mathbf{Y}\big) \to 0, \tag{12}$$

for arbitrary $M_n \to \infty$. In the distributed nonparametric regression model (1) or the classification model (2), the condition $(n/m^2)\varepsilon_n^2 \to \infty$ may be relaxed to $m = o(n\varepsilon_n^2/\log n)$.

The proof of the theorem is deferred to Section B.2.

Remark 5. One can consider adaptation to other type of hyperparameters as well, for instance choosing the regularity or truncation parameters in a data driven way. Our results can be extend to these cases as well in a straightforward way. However, such approaches are, typically, computationally substantially more expensive and hence less popular in practice than rescaling the process. Therefore, we have refrained from including such cases in our analysis.

3. Examples

In this section we apply the general results of the preceding section to obtain (adaptive) minimax contraction rates for regression and classification, with priors built on integrated Brownian motion and the Matérn process.

3.1. Rescaled Integrated Brownian Motion

The "released" ℓ-fold integrated Brownian motion is defined as

$$G_t := B \sum_{j=0}^{\ell} \frac{Z_j t^j}{j!} + (I^\ell W)_t, \quad t \in [0,1], \tag{13}$$

with $B > 0$, and iid standard normal random variables $(Z_j)_{j=0}^{\ell}$ independent from a Brownian motion W. The functional operator I^ℓ denotes taking repeated indefinite integrals and has the purpose of smoothing out the Brownian motion sample paths. Formally we define $(If)_t = \int_0^t f(s)\, ds$ and next $I^1 = I$ and $I^\ell = I^{\ell-1}I$ for $\ell \ge 2$. Because the sample paths of Brownian motion are Hölder continuous of order almost $1/2$ (almost surely), the process $t \mapsto (I^\ell W)_t$ and hence the process $t \mapsto G_t$ has sample paths that are ℓ times differentiable with ℓth derivative Hölder of order almost $1/2$. The polynomial term in $t \mapsto G_t$ allows this process to have nonzero derivatives at zero, where the scaling by B of this fixed-dimensional part of the prior is relatively inessential. The prior process $t \mapsto G_t$ in (13) is an appropriate model for a function that is regular of order $\ell + 1/2$: it is known from van der Vaart and van Zanten (2008) that the resulting posterior contraction rate is equal to the minimax rate for a β-Hölder function f_0 if and only if $\beta = \ell + 1/2$. For $\beta \ne \ell + 1/2$, the posterior still contracts, but at a suboptimal rate. To remedy this, we introduce additional flexibility by rescaling the prior.

Because the integrated Brownian motion is self-similar, a time rescaling is equivalent to a space rescaling with another coefficient. We consider a time rescaling and introduce, for a fixed $\tau > 0$,

$$G_t^{\tau,(k)} := B_n \sum_{j=0}^{\ell} \frac{Z_j^{(k)}(\tau t)^j}{j!} + (I^\ell W^{(k)})_{\tau t}, \quad t \in [0,1]. \tag{14}$$

The $(Z_j^{(k)})$ and $W^{(k)}$ are standard normal variables and a Brownian motion, as in (13), but independently across the local problems. This process has been studied in van der Vaart and van Zanten (2007) (or see sec. 11.5 of Ghosal and van der Vaart 2017) in the non-distributed nonparametric regression setting. The authors demonstrated that for a given $\beta \le \ell + 1$, the scale parameter $\tau := \tau_n = n^{(\ell+1/2-\beta)/((\ell+1/2)(2\beta+1))}$ leads to the optimal contraction rate in the minimax sense at a β-regular function f_0, that is

$$\Pi^{\tau_n}\Big(f : \|f - f_0\|_n \ge M_n n^{-\beta/(2\beta+1)}|\mathbf{Y}\Big) \to 0,$$

for arbitrary M_n tending to infinity. Our first result shows that this same choice of length scale in the local prior distributions leads to the same contraction rate for the distributed, aggregated posterior distribution.

Corollary 6. Consider the distributed nonparametric regression model (1) or the classification model (2) with a function $f_0 \in C^\beta([0,1])$, for $\beta > 1/2$. In each local problem endow f with the rescaled integrated Brownian motion prior (14) with $\ell + 1/2 \ge \beta$ with $\tau = \tau_n \asymp n^{(\ell+1/2-\beta)/((\ell+1/2)(2\beta+1))}$ and $\exp\{n^{1/(1+2\beta)}/m\} \ge B_n^2 \ge n^{\frac{-1+2(\ell-\beta)\vee 0}{1+2\beta}} m$. Then for $m =$

$o(n^{1/(2\beta+1)}/\log n)$, the aggregated posterior (4) achieves the minimax contraction rate, that is

$$\mathbb{E}_0\, \Pi_{n,m}\Big(f : d_n(f,f_0) \geq M_n n^{-\beta/(2\beta+1)}|\mathbf{Y}\Big) \to 0,$$

for arbitrary $M_n \to \infty$. In case of the regression model (1), the pseudo-metric d_n can be replaced by the empirical L_2-metric $\|\cdot\|_n$.

Thus, the aggregated posterior contracts at the optimal rate, provided that the number of machines does not increase more than a certain polynomial in the number of data points.

Unfortunately, the corollary employs a scaling rate τ_n that depends on the smoothness β of the true function, which is typically unknown in practice. To remedy this, we consider a data-driven procedure for selecting τ. In each local problem we choose a random scale factor τ, independently from the variables $(Z_j^{(k)})$ and $W^{(k)}$ and independently across the problems, from a hyper-prior distribution with Lebesgue density $g_{\ell,n,m}$ satisfying, for every $\tau > 0$,

$$\begin{aligned} &C_1 \exp\{-D_1 n^{\frac{1}{2(\ell+1)}}\tau^{\frac{\ell+1/2}{\ell+1}}/m\} \\ &\quad\leq g_{\ell,n,m}(\tau) \leq C_2 \exp\{-D_2 n^{\frac{1}{2\ell+2}}\tau^{\frac{\ell+1/2}{\ell+1}}/m\}, \end{aligned} \tag{15}$$

where C_1, D_1, C_2, D_2 are positive constants. The following corollary shows that this procedure results in rate-optimal recovery of the underlying truth.

Corollary 7. Consider the distributed nonparametric regression model (1) or the classification model (2) with a function $f_0 \in C^\beta([0,1])$, for $\beta > 1/2$. In each local problem endow f with the hierarchical prior (7) built on the randomly rescaled integrated Brownian motion prior given in (14) with $\ell + 1/2 \geq \beta$ and $\exp\{n^{1/(2+2\ell)}/m\} \geq B_n^2 \geq n^{(\ell-1)\vee 0}m$ and hyper-prior density $g_{\ell,n,m}$ satisfying (15). Then, for $m = o(n^{1/(2\ell+2)}/\log n)$, the aggregated posterior (8) adapts to the optimal minimax contraction rate, that is

$$\mathbb{E}_0\, \Pi^g_{n,m}\Big(f : d_n(f,f_0) \geq M_n n^{-\beta/(2\beta+1)}|\mathbf{Y}\Big) \to 0,$$

for arbitrary $M_n \to \infty$. In case of the regression model (1), the pseudo-metric d_n can be replaced by the empirical L_2-metric $\|\cdot\|_n$.

The corollary shows that the aggregated posterior with randomly rescaled local priors contracts at the optimal rate for a true function of given Hölder smoothness level, as long as the hyper-prior and the number of experts are chosen appropriately.

Proofs for the results in this section are given in Section C.2.

3.2. Matérn Process

The Matérn process is a popular prior, particularly in spatial statistics (see e.g., Rasmussen and Williams 2006, p. 84). It is a stationary mean zero Gaussian process with spectral density

$$\rho_{\alpha,\tau}(\lambda) = C_{\alpha,d}\tau^{2\alpha}(c_{\alpha,d}\tau^2 + \|\lambda\|^2)^{-\alpha-d/2}, \tag{16}$$

where $\alpha, \tau > 0$ are parameters, d is the dimension (we shall restrict to $d = 1$) and $c_{\alpha,d}, C_{\alpha,d} > 0$ are constants. The sample paths of the Matérn process are Sobolev smooth of order α, and τ is a scale parameter: if $t \mapsto G_t$ is Matérn with parameter $\tau = 1$, then $t \mapsto G_{\tau t}$ is Matérn with parameter τ. (For consistency of notation we took $\tau = 1/\ell$ in Rasmussen and Williams (2006, p. 84)) The present time-rescaled Matérn process is different from the space-rescaled version $t \mapsto \tau^\alpha G_t$ (for any α) and has been less studied. In Section C.5 we derive bounds on its small ball probability and the entropy of the unit ball of its reproducing kernel Hilbert space. These quantities, in their dependence on τ, are important drivers of posterior contraction rates, and of independent interest. For computation the Matérn process can be spatially represented with the help of Bessel functions.

First we consider the non-adaptive setting where the regularity parameter $\beta > 0$ of the unknown function of interest f_0 is assumed to be known. We choose each local prior equal to a Matérn process with regularity parameter α satisfying $\beta \leq \alpha$, scaled by $\tau_n = n^{\frac{\alpha-\beta}{\alpha(1+2\beta)}}$ to compensate for the possible mismatch between α and β. It is known that the Matérn prior gives minimax optimal contraction rates if used on the full data van der Vaart and van Zanten (2011). The following corollary asserts that, in the distributed setting, the aggregated posterior corresponding to this choice of prior also achieves the minimax contraction rate.

Corollary 8. Consider the distributed nonparametric regression model (1) or the classification model (2) with a function $f_0 \in C^\beta([0,1])$, for $\beta > 1/2$. In each local problem endow f with the rescaled Matérn process prior with regularity parameter α satisfying $\alpha \geq \beta$ and $\alpha + 1/2 \in \mathbb{N}$ and scale parameter $\tau_n = n^{(\alpha-\beta)/(\alpha(1+2\beta))}$. Then for $m = o(n^{1/(1+2\beta)}/\log n)$, the corresponding aggregated posterior (4) achieves the minimax contraction rate $n^{-\beta/(1+2\beta)}$, that is

$$\mathbb{E}_0\, \Pi_{n,m}\big(f : d_n(f,f_0) \geq M_n n^{-\beta/(2\beta+1)}|\mathbf{Y}\big) \to 0,$$

for arbitrary $M_n \to \infty$. In case of the regression model (1), the pseudo-metric d_n can be replaced by the empirical L_2-metric $\|\cdot\|_n$.

Next we consider the local hierarchical priors (7) with hyper-prior density satisfying, for every $\tau > 0$,

$$\begin{aligned} &c_1 \exp\{-d_1 n^{\frac{1}{2\alpha+1}}\tau^{\frac{\alpha}{\alpha+1/2}}/m\} \\ &\quad\leq g_{\alpha,n,m}(\tau) \leq c_2 \exp\{-d_2 n^{\frac{1}{2\alpha+1}}\tau^{\frac{\alpha}{\alpha+1/2}}/m\}, \end{aligned} \tag{17}$$

where c_1, d_1, c_2, d_2 are positive constants. The priors in the local problems are then chosen to be Matérn with random scales drawn from $g_{\alpha,n,m}$, and the aggregated distributed posterior follows our general construction in (8). The following corollary shows that using Matérn processes yields rate-optimal contraction over a range of regularity classes, similarly to the integrated Brownian motion prior case.

Corollary 9. Consider the distributed nonparametric regression model (1) or the classification model (2) with a function $f_0 \in C^\beta([0,1])$, for $\beta > 1/2$. In each local problem endow f with the hierarchical prior built on the randomly rescaled Matérn Process with regularity parameter α satisfying $\alpha \geq \beta$ and $\alpha + 1/2 \in \mathbb{N}$ and scale drawn from a density satisfying (17). Then for $m =$

$o(n^{1/(1+2\alpha)}/\log n)$ the aggregated posterior (8) adapts to the optimal minimax contraction rate, that is

$$\mathbb{E}_0\, \Pi^{g}_{n,m}\left(f:\, d_n(f,f_0) \geq M_n n^{-\beta/(2\beta+1)}|\mathbf{Y}\right) \to 0,$$

for arbitrary $M_n \to \infty$. In case of the regression model (1), the pseudo-metric d_n can be replaced by the empirical L_2-metric $\|\cdot\|_n$.

Proofs for the results in this section are given in Section C.4.

4. Numerical Analysis

In this section we investigate the distributed methods numerically by simulation and illustrate it on a real data problem. We start by a discussion of more refined aggregation techniques than (3). Our numerical analysis was carried out using the MatLab package `gpml`.

4.1. Aggregation Techniques

In spatially distributed GP regression a draw from the aggregated posterior takes the form (3), where the $\mathcal{D}^{(k)}$ are the subregions into which the design points are partitioned and $f^{(k)} \sim \Pi^{(k)}(\cdot|\mathbf{Y}^{(k)})$. The output can be considered a weighted average of the local posteriors, with the indicator functions $1_{\mathcal{D}^{(k)}}(x)$ as weights. Although the procedure provides optimal recovery of the underlying truth, as shown in the preceding sections, the sample functions (3) are discontinuous at the boundaries of the regions $\mathcal{D}(k)$. The optimality implies that the discontinuities are small, but they are visually unappealing.

Various approaches in the literature palliate this problem. In the Patched GP method neighboring local GPs are constrained to share nearly identical predictions on the boundary, see Park and Huang (2016) and Park and Apley (2018). In Tresp (2001), Rasmussen and Ghahramani (2002), and Meeds and Osindero (2006) a two-step mixture procedure was proposed, following the mixture of experts architecture of Jacobs et al. (1991). A prediction at a given point is drawn from an expert (local posterior) that is selected from a pool of experts by a latent variable, which is endowed with a prior to provide a dynamical, Bayesian procedure.

Another method, more closely related to (3), is to consider continuous weights instead of the discontinuous indicators $1_{\mathcal{D}^{(k)}}(x)$. Since the pointwise variances of a local posterior is smaller in the region where the local data lies than outside of this region, inverse pointwise variances are natural as weights. Following this idea, Ng and Deisenroth (2014) introduced as aggregation technique

$$f(x) = \sum_{k=1}^{m} \frac{f^{(k)}(x)}{\sigma_k^2(x)} \Big/ \sum_{k=1}^{m} \frac{1}{\sigma_k^2(x)}, \qquad (18)$$

where $\sigma_k^2(x)$ is the variance of $f^{(k)}(x)$ if $f^{(k)} \sim \Pi^{(k)}(\cdot|\mathbf{Y}^{(k)})$. This approach provides data-driven and continuous weights. However, as shown in our numerical analysis, this leads to suboptimal behavior in the adaptive setting, where the scale parameter is tuned to the data. Perceived local smoothness in the data in region $\mathcal{D}^{(k)}$ will induce a small variance in the induced local posterior distribution, due to the adaptive bandwidth choice. This posterior variance will then also be relatively small outside the local region, where the local posterior is not informed by data, no matter the nature of the data in this region. The inverse variance weights then lead to overly large weights even outside of the experts' domain. That is to say that an expert will be overly confident about their knowledge of the true function in the whole space when this function is particularly smooth in this expert's own domain.

In view of these observations we propose a new approach, which introduces more severe shrinkage outside of the local domain. As samples from the aggregated posterior, consider the weighted average

$$f(x) = \sum_{k=1}^{m} w_k(x) f^{(k)}(x) \Big/ \sum_{k=1}^{m} w_k(x), \qquad (19)$$

with weights, for c_k being the geometric center of $\mathcal{D}^{(k)}$,

$$w_k(x) = \frac{e^{-\rho m^2 (x-c_k)^2}}{\sigma_k^2(x)},$$

for some $\rho > 0$. These weights are also continuous and data-driven and impose an exponential shrinkage, which depends on the distance from the subregion. Furthermore, we note, that by choosing $\rho = C\log n$ and considering the one dimensional, unit interval case $\mathcal{D}^{(k)} = I^{(k)}$, the proportion of the exponential weights for points in the kth region and away from it can be bounded by $e^{-\rho m^2 (y-c_k)^2}/e^{-\rho m^2 (x-c_k)^2} \leq n^{-C}$, for $x \in I^{(k)}$ and $y \in I^{(j)}$, with $|j-k| > 2$. Hence, the contribution of the local posteriors built on datasets not in the neighbourhood of the kth interval is negligible for $x \in I^{(k)}$. This approach provides therefore a continuous aggregated posterior which at the same time better resembles the localization properties of the standard, glue together approach than the one proposed in Ng and Deisenroth (2014). We show below numerically, both on synthetic and real world datasets, that this new aggregation technique substantially improves the performance of the distributed GP procedure, especially when the scale hyper-parameter is selected in a data-driven way.

4.2. Synthetic Datasets

In this section we investigate the performance of various distributed Gaussian process regression methods on synthetic datasets, and compare them to the benchmark: the non-distributed approach that computes the posterior distribution on all data.

We consider recovery and confidence statements for the functional parameter f_0 based on n independent data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the model

$$Y_i = f_0(X_i) + Z_i, \qquad Z_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2), \quad X_i \overset{\text{iid}}{\sim} U(0,1).$$

We simulated the data with noise standard deviation $\sigma = 1$ and the function f_0 defined by coefficients $f_{0,j}$ relative to the cosine basis $\psi_j(x) = \sqrt{2}\cos(\pi(j-1/2)x), j = 1, 2, \ldots$.

Next to the true posterior distribution, based on all data, we computed distributed posterior distributions, using four methods. Method 1 (M1) is the consensus Monte Carlo method proposed by Scott et al. (2016) and applied to Gaussian Processes in Szabó and van Zanten (2019) and Hadji, Hesselink, and Szabó (2022). This method splits the data randomly between the machines (i.e., the kth machine receives a random subset of size $n_k = n/m$ from the observations) and compensates for working with only partial datasets by scaling the priors in the local machines by a factor $1/m$. A draw from the aggregated posterior is constructed as the average $f(x) = m^{-1}\sum_{k=1}^{m} f^{(k)}(x)$ of independent draws $f^{(k)}$ from each (modified) local posterior. Methods 2–4 all split the data spatially (i.e., the kth machine receives the pairs (X_i, Y_i) for which $X_i \in I^{(k)} = \left(\frac{k-1}{m}, \frac{k}{m}\right]$), and differ only in their aggregation technique. Method 2 (M2) uses the standard "glue together" approach displayed in (3), Method 3 (M3) uses the inverse variance weighted average (18), and Method 4 (M4) uses the exponential weights (19).

All distributed methods were carried out on a single core, drawing sequentially from the local posteriors. Parallelizing them over multiple cores or machines would have shortened the reported run times substantially, approximately by a factor m.

First we considered the Matérn covariance kernel. We studied both a version with sample paths rescaled deterministically by the optimal length scale τ_n for the given true function f_0 and versions with data-based rescaling (via both empirical and hierarhical Bayes approaches) that do not use any information about f_0. While for the oracle choice (depending on the typically unknown regularity β of the underlying function f_0) of the scaling parameter all methods performed similarly well, for the data driven choices of the hyper-parameter there were substantial differences between the distributed Methods 1–4. Spatially distributing the data (Methods 2 and 4) clearly outperformed random distribution (Method 1). This is in agreement with the theory, and can be explained by the inability to determine suitable scale parameters from the data in the randomly distributed case. However, it was also observed that the benefits of spatial distribution can be destroyed when smoothing out the inherent spatially discontinuities using aggregation weights that depend on the local length scales in the wrong way (Method 3).

Then we investigated whether the methods can adapt to different local regularities. We considered a true function f_0, which is rough in the first half of the co-domain and smooth in the second half. It is well known that stationary Gaussian processes are not appropriate for picking up different local behavior as they localize the signal at the spectral not at the spatial domain. This can be also observed in our numerical analyzis for the non-distributed and the randomly distributed methods (BM and M1). However, by spatially dividing the data over the local machines one can pick up different local behaviors and can achieve substantially better estimations and uncertainty quantification for the smoother part of the signal than using the standard, non-distributed approach. Hence, in addition to significantly speeding up the computations, spatially distributed methods have the additional advantage of better learning the local properties of the signal by adapting to the local regularity. We deferred the corresponding numerical analysis to the supplement.

Finally, we have also investigated the popular squared exponential covariance kernel with data driven rescaling hyper-parameter, using both the empirical and hierarchical Bayes methods. Although this prior is not explicitly covered in our examples, we observe similar phenomenas as for the Matérn covariance kernel. The corresponding simulation studies are deferred to the supplement.

As mentioned earlier, for adaptation we have used both the hierarchical and empirical Bayes procedures. In the empirical Bayes method we took the maximum marginal likelihood estimator (MMLE) of the scale parameter τ, while in hierarchical Bayes we have endowed it with another layer of prior distribution. The (MMLE) empirical Bayes method has been shown to behave similarly to the hierarchical Bayes method, considered in our theoretical study (see for instance Szabó, van der Vaart, and van Zanten 2013; Sniekers and van der Vaart 2015; Rousseau and Szabo 2017; Sniekers and van der Vaart 2020). In both approaches we computed first the marginal log-likelihood function on a (fine enough) grid using the `gpml` Matlab package. Then in the empirical Bayes method we selected the maximizer of this likelihood. In the hierarchical Bayes approach we used an exponential hyper-prior distribution on τ (which was approximated by a truncated geometric distribution on the chosen grid) and derived the corresponding marginal posterior of the hyper-parameter. Alternatively, one could also use the `minimize` function built in the `gpml` package for estimating the hyper-parameters of the GP prior. However, this Matlab function approximates simultaneously various additional hyper-parameters as well. Since in our theoretical framework we have tuned only the length scale parameter τ, for better connection to the preceding sections, we have refrained from using this built in optimizer in the synthetic dataset.simulation study.

To assess the quality of the recovery we report the L_2 error of estimating f_0 with the posterior mean. As a measure of the size of L_2-credible balls we report twice the root average posterior variance

$$r = 2\sqrt{\int_0^1 \sigma^2(x|\mathbf{X},\mathbf{Y})\,dx}.$$

We consider the true function to be in the credible ball if its L_2-distance to the posterior mean is smaller than r. Furthermore, we also investigate the point-wise behavior of the posterior. We report both the length of the 95% confidence interval $4\sigma(x)$ for some selected points x and the corresponding local coverage probabilities. In case of the hierarchical Bayes approach we report the average credible bands with respect to the hyper-posterior.

4.2.1. Matérn Kernel

In our study with the Matérn prior, we used this kernel with regularity hyper-parameter $\alpha = 3$ (see (16)), and generated the data from the true parameter f_0 with coefficients $f_{0,j} = 1.5\sin(j)j^{-1/2-\beta}$, with $\beta = 1$, for $j \geq 4$ and $f_{0,j} = 0$ for $j \leq 3$, with respect to the cosine basis. This function f_0 is essentially β smooth: f_0 belongs to the Sobolev space $H^{\gamma}([0,1])$ for all $\gamma < \beta$. The optimal length scale parameter of the prior is then $\tau_n = n^{(\alpha-\beta)/(1+2\beta)/\alpha}$, as seen in Section 3.2.

We considered pairs (n, m) of sample sizes and numbers of machines equal to $(2000, 10)$, $(5000, 20)$ and $(10{,}000, 50)$. In all test cases we repeated the experiment 100 times, except in the adaptive settings with $n \geq 5000$, where we considered

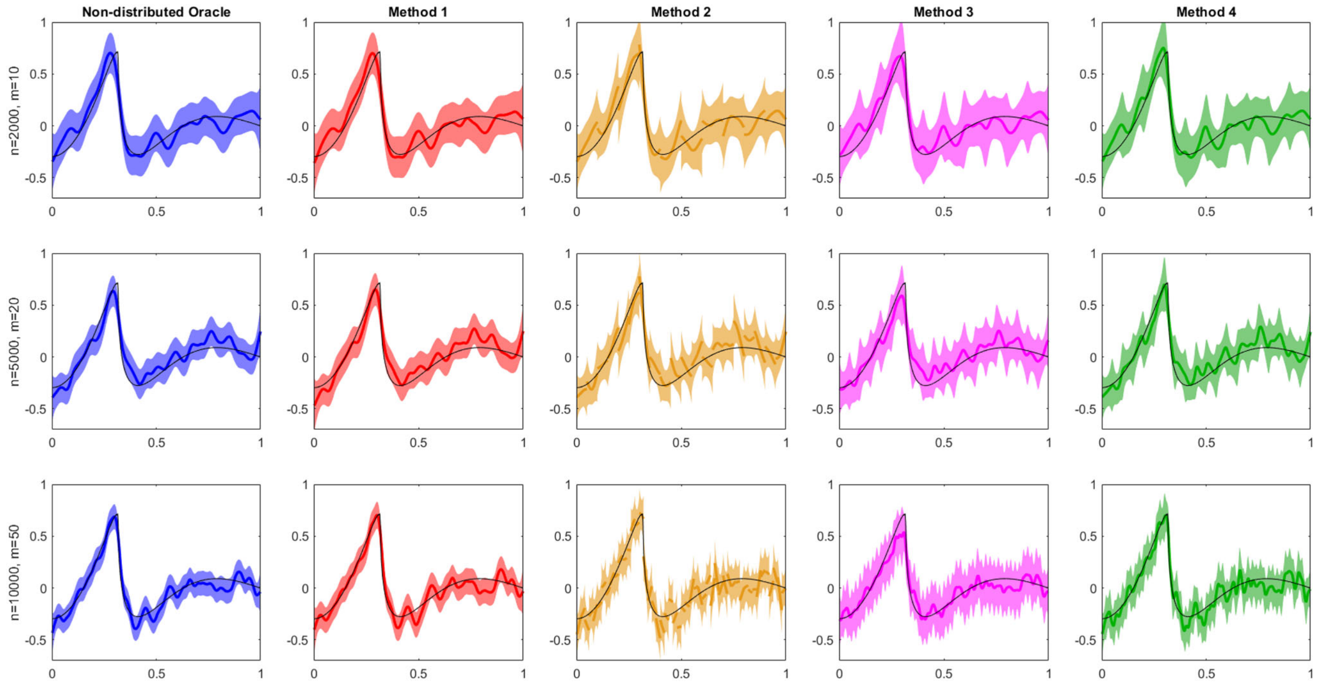

Figure 1. Deterministic (oracle) rescaling of the Matérn process prior ($\alpha = 3$). Benchmark and distributed GP posteriors. True function $f_0(x) = \sum_{j=4}^{\infty} 1.5j^{-3/2} \sin(j)\psi_j(x)$ drawn in black. Posterior means drawn by solid lines, surrounded by 95% point-wise credible sets shaded between two dotted lines. The five columns correspond (left to right) to the non-distributed method, the distributed method with random partitioning, and the distributed methods with spatial partitioning without smoothing, with inverse variance weights and with exponential weights. From top to bottom the sample sizes are $n = 2000, 5000, 10{,}000$ and the number of experts $m = 10, 20, 50$.

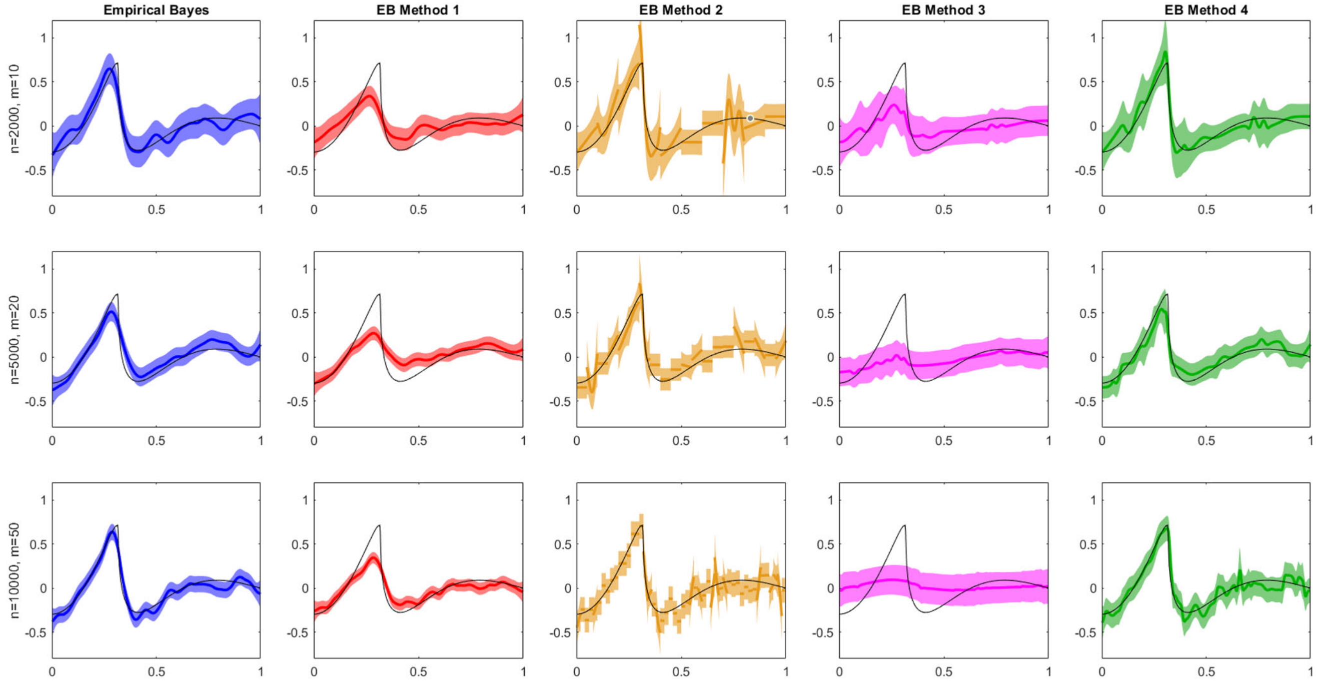

Figure 2. Empirical Bayes (MMLE) approach for the rescaled Matérn process prior ($\alpha = 3$). Benchmark and distributed GP posteriors. True function $f_0(x) = \sum_{j=4}^{\infty} 1.5j^{-3/2} \sin(j)\psi_j(x)$ drawn in black. Posterior means drawn by solid lines, surrounded by 95% point-wise credible sets, shaded between two dotted lines. The five columns correspond (left to right) to the non-distributed method, the distributed method with random partitioning, and the distributed methods with spatial partitioning without smoothing, with inverse variance weights and with exponential weights. From top to bottom the sample sizes are $n = 2000, 5000, 10{,}000$ and the number of experts is $m = 10, 20, 50$.

only 20 repetitions, due to the overly slow non-distributed approach. Posterior means and 95% point-wise credible bands for a single experiment are visualized in Figures 1–3, for the oracle (deterministic, optimal rescaling), empirical Bayes and hierarchical Bayes scaling, respectively. The average L_2-errors, the sizes and frequentist coverages of the L_2 credible sets and the run times are reported in Tables 1–2 and Table F.1 in the supplement for the deterministic scaling, in Tables 3–5 for the empirical Bayes method, and in Tables 6–8 for the hierarchical Bayes approach. Due to space restriction

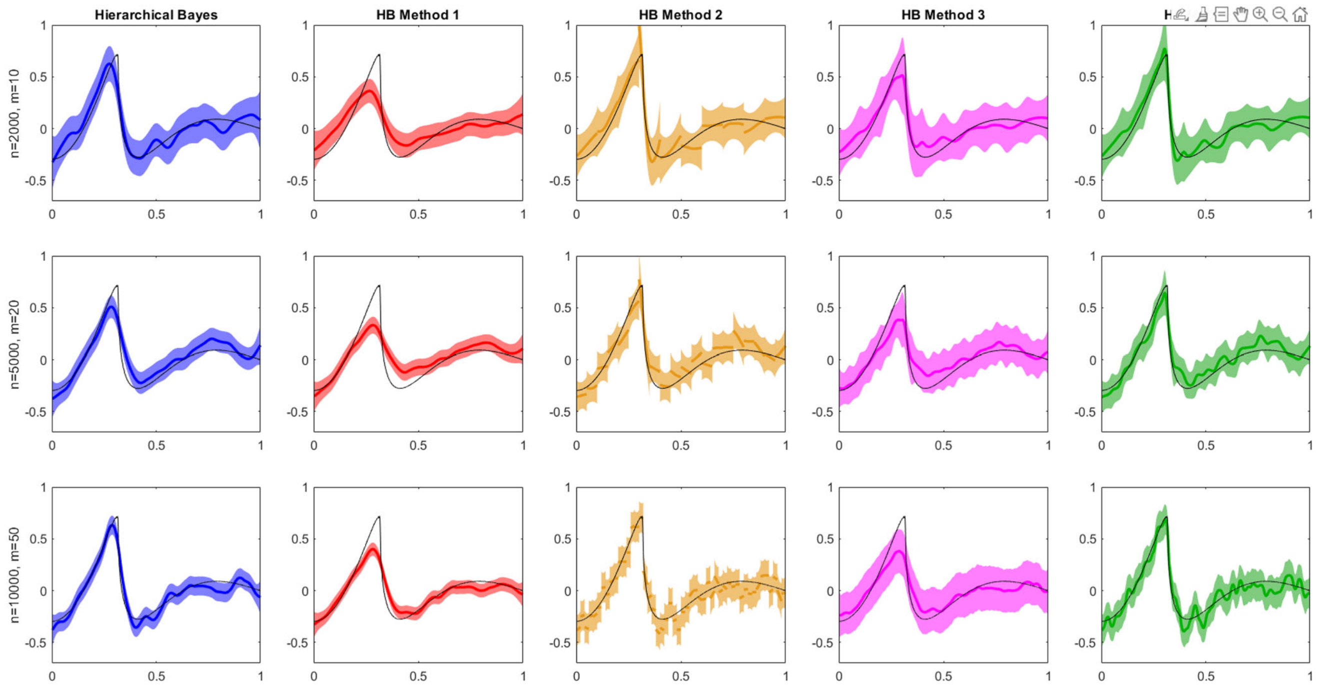

Figure 3. Hierarchical Bayes methods for the rescales Matérn process prior ($\alpha = 3$). Benchmark and distributed GP posteriors. True function $f_0(x) = \sum_{j=4}^{\infty} 1.5j^{-3/2} \sin(j)\psi_j(x)$ drawn in black. Posterior means drawn by solid lines, surrounded by 95% point-wise credible sets, shaded between two dotted lines. The five columns correspond (left to right) to the non-distributed method, the distributed method with random partitioning, and the distributed methods with spatial partitioning without smoothing, with inverse variance weights and with exponential weights. From top to bottom the sample sizes are $n = 2000, 5000, 10{,}000$ and the number of experts is $m = 10, 20, 50$.

Table 1. Average L_2-distance between f_0 and posterior mean for deterministic (oracle) rescaling of the Matérn process prior (with $\alpha = 3$).

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
BM	0.091 (0.014)	0.068 (0.008)	0.054 (0.007)
M1	0.093 (0.014)	0.070 (0.008)	0.057 (0.007)
M2	0.105 (0.016)	0.086 (0.009)	0.080 (0.007)
M3	0.090 (0.015)	0.070 (0.008)	0.069 (0.007)
M4	0.094 (0.015)	0.075 (0.008)	0.065 (0.007)

NOTE: BM: Benchmark, Non-distributed method. M1: Random partitioning, M2: Spatial partitioning, M3: Spatial partitioning with inverse variance weights, M4: Spatial partitioning with exponential weights. Average values over 100 replications of the experiment with standard error in brackets.

Table 2. Deterministic (oracle) rescaling of the Matérn process prior ($\alpha = 3$).

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
Benchmark	0.897s (0.406s)	9.379s (3.986s)	53.86s (15.49s)
Random	0.121s (0.081s)	0.260s (0.120s)	0.46s (0.43s)
Spatial	0.114s (0.084s)	0.235s (0.098s)	0.44s (0.45s)

NOTE: Average run time for computing the posterior. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning.

Table 3. Empirical (MMLE) Bayes rescaling of the Matérn Gaussian process prior.

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
BM	0.092 (0.013)	0.067 (0.008)	0.053 (0.006)
M1	0.136 (0.027)	0.109 (0.026)	0.118 (0.015)
M2	0.102 (0.018)	0.083 (0.009)	0.084 (0.009)
M3	0.184 (0.019)	0.197 (0.011)	0.205 (0.003)
M4	0.091 (0.017)	0.069 (0.009)	0.057 (0.006)

Average L_2-distance between f_0 and posterior mean.

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
BM	0.160 (0.010)	0.118 (0.007)	0.093 (0.005)
M1	0.125 (0.020)	0.090 (0.010)	0.067 (0.006)
M2	0.182 (0.011)	0.151 (0.006)	0.155 (0.003)
M3	0.187 (0.008)	0.162 (0.002)	0.176 (0.001))
M4	0.172 (0.010)	0.143 (0.006)	0.149 (0.002)

Average radius of the L_2-credible ball.

NOTE: BM: Benchmark, Non-distributed method. M1: Random partitioning, M2: Spatial partitioning, M3: Spatial partitioning with inverse variance weights, M4: Spatial partitioning with exponential weights.

Table 4. Empirical (MMLE) Bayes rescaling of the Matérn Gaussian process prior.

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
BM	1.00	1.00	1.00
M1	0.49	0.25	0.00
M2	0.98	1.00	1.00
M3	0.45	0.00	0.00
M4	0.96	1.00	1.00

NOTE: Proportion of experiments when the true function f_0 was inside in the L_2-credible ball.

we report the point-wise analysis of these approaches in the supplement.

In the non-adaptive setting, where the GP was optimally rescaled, all methods performed similarly well. They all resulted in good estimators and reliable uncertainty statements. The run time of the distributed algorithms were similar and substantially shorter than for the non-distributed counterpart (on average below 1s in all cases).

In the adaptive setting we considered both the empirical and hierarchical Bayes approaches. In the first method we estimate the scaling hyper-parameter with the MMLE, while in the second one we endow it with another layer of prior, resulting in a fully Bayesian, hierarchical procedure. In the latter case, as hyper-prior, we chose the exponential distribution with parameter $\lambda = 1/5$ and approximated the hyper-posterior on a fine enough grid. One can observe that both data driven Bayesian

Table 5. Empirical (MMLE) Bayes rescaling of the Matérn Gaussian process prior.

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
Benchmark	90.32s (17.51s)	895.1s (86.5s)	4767s (611s)
Random	6.31s (1.95s)	14.4s (3.3s)	22.5s (4.0s)
Spatial	6.36s (2.49s)	13.8s (3.0s)	21.2s (4.9s)

NOTE: Average run time for computing the posterior. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning.

Table 6. Hierarchical Bayes rescaling of the Matérn Gaussian process prior.

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
BM	0.093 (0.016)	0.067 (0.008)	0.056 (0.006)
M1	0.189 (0.019)	0.095 (0.016)	0.096 (0.007)
M2	0.120 (0.020)	0.075 (0.009)	0.074 (0.007)
M3	0.225 (0.001)	0.087 (0.011)	0.099 (0.005)
M4	0.100 (0.021)	0.060 (0.007)	0.055 (0.007)
	Average L_2-distance between f_0 and posterior mean.		
(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
BM	0.183 (0.025)	0.117 (0.005)	0.091 (0.004)
M1	0.122 (0.069)	0.091 (0.006)	0.066 (0.002)
M2	0.194 (0.031)	0.147 (0.011)	0.148 (0.001)
M3	0.100 (0.003)	0.184 (0.001)	0.211 (0.001)
M4	0.157 (0.019)	0.151 (0.001)	0.152 (0.001)
	Average radius of the L_2-credible ball		

NOTE: BM: Benchmark, Non-distributed method. M1: Random partitioning, M2: Spatial partitioning, M3: Spatial partitioning with inverse variance weights, M4: Spatial partitioning with exponential weights.

Table 7. Hierarchical Bayes rescaling of the Matérn Gaussian process prior.

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
BM	1.00	1.00	1.00
M1	0.18	0.45	0.00
M2	0.98	1.00	1.00
M3	0.00	1.00	1.00
M4	0.96	1.00	1.00

Table 8. Hierarchical Bayes rescaling of the Matérn Gaussian process prior.

(n,m)	(2000, 10)	(5000, 20)	(10,000, 50)
Benchmark	35.85s (6.48s)	1292.2s (84.9s)	10,000s (1362s)
Random	4.97s (1.32s)	26.7s (0.8s)	61.4s (12.8s)
Spatial	5.25s (2.49s)	25.8s (1.7s)	58.6s (12.1s)

NOTE: Average run time for computing the posterior. Benchmark: Non-distributed method. Method 1: Random partitioning, Method 2: Spatial partitioning.

methods performed similarly. In case of randomly distributing the data to local machines (M1) the aggregated posterior is oversmoothed and provides too narrow, overconfident uncertainty quantification. The standard spatially distributed approach (M2) performed well, but produced visible discontinuities. The aggregation approach (M3) provided poor and overconfident estimator using empirical Bayes method as is very evident in Figure 2. Using hierarchical Bayes, Method 3 performed better, but the estimation accuracy and the size of the credible sets were still sub-optimally large, see Figure 3 and the corresponding tables. Our approach (M4) combined the best of both worlds: it provided continuous sample paths and maintained (and even improved) the performance of the standard glue-together spatial approach (M2), while substantially reducing the computational burden compared to the non-distributed approach.

Here again we note, that by parallelized implementation of the algorithms the run time could be further reduced by a factor of m. For instance, in the last scenario of the hierarchical Bayes approach with $(n, m) = (10{,}000, 50)$ this would reduce the computation time of 10,000 sec needed for the non-distributed method to around 1 sec.

Table 9. RMSE and runtime of non-distributed (BM) and distributed (M1–M4) GP regression with squared exponential covariance kernel for the Superconductivity dataset Hamidieh (2018b).

Methods	BM	M1	M2	M3	M4
RMSE	12.63 (0.21)	15.58 (0.19)	13.14 (0.35)	16.63 (1.40)	12.69 (0.32)
Runtime	18,740s (1720s)	248s (82s)	220s (23s)	220s (23s)	220s (23s)

4.3. *Real World Dataset: Superconductivity*

Superconducting materials lose their resistance when they are cooled down below a certain temperature, called critical temperature, and as a consequence can conduct current with zero resistance. Materials with this property are used for instance in magnetic resonance imaging (MRI) and nuclear magnetic resonance (NMR) applications. Therefore, predicting the critical temperature is an important problem of wide interest.

In our analysis we consider the Superconductivity dataset Hamidieh (2018a, 2018b). It contains 81 covariates describing the superconductor's elemental properties and the goal is to predict the critical temperature based on them. In total 21,263 measurement points were collected. In our analysis we have divided the data randomly into a training and testing dataset consisting of 15,000 and 6263 measurements, respectively. We have repeated the experiment 10 times to measure the variability of the result. We compared the different distributed GP approaches (M1–M4) to the benchmark non-distributed approach (BM). In the spatially distributed methods we have split the data amongst the machines with respect to the wtd_mean_atomic_mass variable. We have considered the squared exponential covariance kernel and selected the hyper-parameters with the minimize subroutine built in the gpml MATLAB package. In Method 4 we set the weight parameter $\rho = 4$. The results are reported in Table 9.

One can conclude that the naive (M2) and the exponentially re-weighted (M4) spatially distributed methods performed the best, similarly well to the benchmark non-distributed approach. At the same time, the product of experts method (M1) and the spatially distributed method with aggregation weights proportional to the inverse of the posterior variance (M3) performed sub-optimally, providing around 30% bigger error. At the same time the distributed methods were around two magnitudes faster than the non-distributed counterpart, and their speed could further increase by parallelizing the computations instead of sequentially executing them, as in our analysis. Finally, we have also considered splitting the data in the spatially distributed methods with respect to other covariates as well. In view of Table 10, methods M2 and M4 are robust with respect to the splitting approach and provide similarly accurate predictions. The only requirement is that the feature used for splitting does not contain too many repetitions, which would result in imbalanced group sizes.

Table 10. RMSE for spatially distributed GP regression methods with squared exponential covariance kernel using different feature variables (in all cases we omitted the "_atomic_mass" from their names) for splitting the data in the Superconductivity dataset Hamidieh (2018b).

Methods	wtd_range	mean	wtd_mean	gmean	wtd_entropy
M2	13.72 (0.51)	13.46 (0.35)	13.14 (0.35)	13.69 (0.64)	13.49 (0.31)
M3	19.25 (1.28)	24.34 (0.88)	16.63 (1.40)	27.57 (1.50)	28.16 (1.14)
M4	13.27 (0.42)	13.14 (0.42)	12.69 (0.32)	13.10 (0.41s)	12.91 (0.23)

5. Discussion

The article provides the first theoretical guarantees for the method of spatial distribution applied to Gaussian processes. Our general results show that the resulting approximation to the posterior provides optimal recovery (both in the case of known and unknown regularity parameter) of the underlying functional parameter of interest in a range of models, including the nonparametric regression model with Gaussian errors and the logistic regression model. As specific examples of priors we considered the popular Matérn process and integrated Brownian motion, but in principle other GP priors could be covered as well. The theoretical findings are complemented with a numerical analysis both on synthetic and real world datasets, where we also proposed a novel aggregation technique for aggregating the local posteriors together, which empirically outperformed the close competitors.

The main advantage of spatial distribution of the data is the ability to adapt the length scale of the prior in a data-driven way, which was highlighted by both theory and numerical illustration. The latter showed that the combination technique of the local posteriors is highly important and can substantially influence the performance of the method. We also demonstrated that spatially distributed GPs can adapt to different local regularities of the true function, hence, can potentially outperform the original GP.

Our results, although formulated for Gaussian processes, in principle rely on general Bayesian nonparametric techniques, adapted to the spatially distributed architecture and hence can be potentially extended to other classes of priors. Also, in the theoretical results univariate functional parameters were considered for simplicity, but the results could be extended to higher dimensional covariates. Another interesting extension is to derive theoretical guarantees for the proposed aggregation techniques beyond the "glue together" approach covered by this article. These extensions, although of interest, are left for future work.

Supplementary Materials

The supplement contains appendices A–F. A: Comparison with other scalable approximations. B, C, D: Proofs of main results and examples and auxiliary results. E and F: Details on synthetic and real-world numerical experiments and additional experiments.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Co-funded by the European Union (ERC, BigBayesUQ, project number: 101041064). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This research was partially funded by a Spinoza grant of the Dutch Research Council (NWO).

References

Burt, D. R., Rasmussen, C. E., and van der Wilk, M. (2019), "Rates of Convergence for Sparse Variational Gaussian Process Regression," in *International Conference on Machine Learning*, pp. 862–871, PMLR. [1]

Cao, Y., and Fleet, D. J. (2014), "Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions," arXiv e-prints. [1]

Deisenroth, M., and Ng, J. W. (2015), "Distributed Gaussian Processes," *Proceedings of Machine Learning Research*, 37, 1481–1490. [1]

Ghosal, S., and van der Vaart, A. (2007), "Convergence Rates of Posterior Distributions for Noniid Observations," *The Annals of Statistics*, 35, 192–223. [3]

Ghosal, S., and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge: Cambridge University Press. [3,4]

Gibbs, N. E., Poole Jr, W. G., and Stockmeyer, P. K. (1976), "An Algorithm for Reducing the Bandwidth and Profile of a Sparse Matrix," *SIAM Journal on Numerical Analysis*, 13, 236–250. [1]

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2017), "A Divide-and-Conquer Bayesian Approach to Large-Scale Kriging," arXiv preprint arXiv:1712.09767. [1,2]

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2022), "Distributed Bayesian Varying Coefficient Modeling Using a Gaussian Process Prior," *The Journal of Machine Learning Research*, 23, 3642–3700. [1]

Hadji, A., Hesselink, T., and Szabó, B. (2022), "Optimal Recovery and Uncertainty Quantification for Distributed Gaussian Process Regression," arXiv preprint arXiv:2205.03150. [2,7]

Hamidieh, K. (2018a), "A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor," *Computational Materials Science*, 154, 346–354. [10]

——— (2018b), "Superconductivty Data," UCI Machine Learning Repository. DOI:10.24432/C53P47. [10,11]

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive Mixtures of Local Experts," *Neural Computation*, 3, 79–87. [6]

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005), "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes," *Journal of the American Statistical Association*, 100, 653–668. [1]

Meeds, E., and Osindero, S. (2006), "An Alternative Infinite Mixture of Gaussian Process Experts," in *Advances in Neural Information Processing Systems* (Vol. 18), eds. Y. Weiss, B. Schölkopf, and J. Platt, MIT Press. [6]

Ng, J. W., and Deisenroth, M. P. (2014), "Hierarchical Mixture-of-Experts Model for Large-Scale Gaussian Process Regression," arXiv e-prints. [6]

Nieman, D., Szabo, B., and Van Zanten, H. (2022), "Contraction Rates for Sparse Variational Approximations in Gaussian Process Regression," *The Journal of Machine Learning Research*, 23, 9289–9314. [1]

Park, C., and Apley, D. W. (2018), "Patchwork Kriging for Large-Scale Gaussian Process Regression," *Journal of Machine Learning Research*, 19, 7:1–7:43. [6]

Park, C., and Huang, J. Z. (2016), "Efficient Computation of Gaussian Process Regression for Large Spatial Data Sets by Patching Local Gaussian Processes," *Journal of Machine Learning Research*, 17, 1–29. [6]

Quiñonero-Candela, J., and Rasmussen, C. E. (2005), "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research*, 6, 1939–1959. [1]

Rasmussen, C., and Ghahramani, Z. (2002), "Infinite Mixtures of Gaussian Process Experts," in *Advances in Neural Information Processing Systems* (Vol. 14), eds. T. Dietterich, S. Becker, and Z. Ghahramani, MIT Press. [6]

Rasmussen, C. E., and Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, Boston: MIT Press. [1,5]

Rousseau, J., and Szabo, B. (2017), "Asymptotic Behaviour of the Empirical Bayes Posteriors Associated to Maximum Marginal Likelihood Estimator," *The Annals of Statistics*, 45, 833–865. [7]

Saad, Y. (1990), "Sparskit: A Basic Tool Kit for Sparse Matrix Computations." RIACS Technical Report 90.20, Institute for Advanced Computer Science, NASA Ames Research Center. [1]

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2016), "Bayes and Big Data: The Consensus Monte Carlo Algorithm," *International Journal of Management Science and Engineering Management*, 11, 78–88. [1,7]

Shang, Z., and Cheng, G. (2015), "A Bayesian Splitotic Theory for Nonparametric Models," ArXiv e-prints. [2]

Sniekers, S., and van der Vaart, A. (2015), "Adaptive Bayesian Credible Sets in Regression with a Gaussian Process Prior," *Electronic Journal of Statistics*, 9, 2475–2527. [7]

Sniekers, S., and van der Vaart, A. (2020), "Adaptive Bayesian Credible Bands in Regression with a Gaussian Process Prior," *Sankhya A*, 82, 386–425. [7]

Srivastava, S., Cevher, V., Dinh, Q., and Dunson, D. (2015), "WASP: Scalable Bayes via Barycenters of Subset Posteriors," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of Proceedings of Machine Learning Research, eds. G. Lebanon and S. V. N. Vishwanathan, pp. 912–920, San Diego, California, USA, 09–12 May 2015, PMLR. [1]

Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2013), "Empirical Bayes Scaling of Gaussian Priors in the White Noise Model," *Electronic Journal of Statistics*, 7, 991–1018. [7]

Szabó, B., and van Zanten, H. (2019), "An Asymptotic Analysis of Distributed Nonparametric Methods," *Journal of Machine Learning Research*, 20, 1–30. [2,7]

Titsias, M. (2009), "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *Artificial Intelligence and Statistics*, pp. 567–574. [1]

Tresp, V. (2000), "The Generalized Bayesian Committee Machine," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 130–139, New York, NY, USA. Association for Computing Machinery. [1]

Tresp, V. (2001), "Mixtures of Gaussian Processes," in *Advances in Neural Information Processing Systems* (Vol. 13), eds. T. Leen, T. Dietterich, and V. Tresp, Cambridge, MA: MIT Press. [6]

van der Vaart, A., and van Zanten, H. (2007), "Bayesian Inference with Rescaled Gaussian Process Priors," *Electronic Journal of Statistics*, 1, 433–448. [4]

——— (2011), "Information Rates of Nonparametric Gaussian Process Methods," *Journal of Machine Learning Research*, 12, 2095–2119. [5]

van der Vaart, A. W., and van Zanten, J. H. (2008), "Rates of Contraction of Posterior Distributions based on Gaussian Process Priors," *The Annals of Statistics*, 36, 1435–1463. [3,4]

Vasudevan, S., Ramos, F., Nettleton, E., and Durrant-Whyte, H. (2009), "Gaussian Process Modeling of Large-Scale Terrain," *Journal of Field Robotics*, 26, 812–840. [1]