Scalability and uncertainty of Gaussian processes

Hadji, M.A.

CHAPTER 1
Introduction

The main theme of this thesis is the theoretical study of Gaussian processes as a tool in Bayesian nonparametric statistics. We are interested in the frequentist properties of Bayesian nonparametric techniques in an asymptotic regime.

This chapter will give a general introduction to Bayesian nonparametric statistics. After a brief preamble on statistical science, we will compare the two main statistical approaches: frequentism and Bayesianism. Our work will consist of a combination of both, since we will study the Bayesian techniques from a frequentist perspective, specifically consistency, convergence rates, uncertainty quantification and adaptation. These properties will be studied in the context of non-parametric problems, that is to say models with few modeling constraints. We will focus on the non-parametric regression model and its idealized version, the signal-in-white-noise model.

One of the main theme of the thesis is scalability of Bayesian techniques. Indeed, these computation-hungry techniques rapidly become intractable as the number of observations grows. This issue led to the introduction of distributed Bayesian methods in order to decrease the computational complexity of the techniques. The later chapters of this thesis will focus on such distributed methods and their properties. They will also be briefly introduced in this chapter.

§1.1 Statistical science

Statistics can be defined as the study of data, which covers its collection, organization, analysis, interpretation and presentation. The main focus of this thesis will be mathematical statistics, which follows the framework of probability theory. Moreover, the theoretical results will be illustrated with simulated data.

In order to understand both probability theory and statistics, it is useful to start with a definition of modeling. A *model* describes and explains a system with the help of mathematical concepts and language. Even though models are helpful in most disciplines, they are all considered as oversimplification of the described system. As George Box (Box, 1976) famously said "all models are wrong, but some are useful". While in probability theory, one studies the behavior of data following a specific and known model, in theoretical statistics, one follows the opposite paths: one tries to gather as much information from the data in order to make inference on the model.

More formally, one can think of a probability distribution P as a model for generating an observation X. The classical statistical approach is to assume that our model $\mathcal{P}$ is a collection of probability distributions, and the end goal is to infer properties

of an element of $\mathcal{P}$ which could best describe the observation.

§1.2 Frequentist and Bayesian inference

Inference in statistics consists of establishing and evaluating propositions about the process by which the data is generated. Different paradigms of statistical inference coexist. This section will focus on two important paradigms: *frequentist inference* and *Bayesian inference*.

§1.2.1 Frequentist inference

Frequentist inference is associated with the frequentist interpretation of probability; we consider that the probability of an event is the frequency of appearance of this event if it were possible to repeat the same experiment independently ad infinitum. One of the main characteristics of frequentism is that frequentist modeling views the data X as the realization of a random variable following some fixed probability distribution P_0 belonging to some model $\mathcal{P}$. Often, the statistical model $\mathcal{P}$ is defined as

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where the probability distributions are indexed by a parameter θ and Θ is the set of all possible parameters, and we suppose the existence of a θ_0 such that $P_0 = P_{\theta_0}$. Basically, if θ_0 were known, the work of a frequentist statistician would be finished.

Among the most widely used tools of frequentist inference, one can think of the likelihood function. The *likelihood function* is a function of the parameter which measures how well a particular parameter of the model fits a given sample of data. Heuristically, it makes sense that parameters whose likelihoods are large are preferable, since they verify that the probability of the observed data occurring is high.

§1.2.2 Bayesian inference

The Bayesian approach is quite different from the frequentist one. It is based on the Bayesian interpretation of probability which stipulates that a probability is merely a quantification of a personal belief or a reasonable expectation. Bayesian statisticians generally believe that there is no fixed underlying parameter, but that the parameter itself is the realization of a random variable. More formally, the parameter θ has a probability distribution Π on the parameter space Θ called the prior distribution. This distribution represents the degree of belief the statistician attaches to possible parameters explaining the model. Then, the distribution P_θ describes the distribution of the data X conditionally on the value of the parameter θ. This belief is then updated once we have access to the data. If the model is dominated, which means all possible distributions P_θ are absolutely continuous with respect to a common measure μ and have corresponding densities p_θ, then this operation gives us the posterior distribution of the parameter

$$A \mapsto \Pi(\theta \in A|X) = \frac{\int_A p_\theta(X) d\Pi(\theta)}{\int_\Theta p_\theta(X) d\Pi(\theta)},$$

for a measurable set $A \subset \Theta$. The integrand in the numerator of the previous fraction represents the joint distribution of the parameter and the data, which can be obtained by multiplying the prior distribution of the parameter by the likelihood function, while the denominator represents the marginal distribution of the data.

Even though the Bayesian approach is quite different from the frequentist approach, it can be useful in certain situations. Indeed, Bayesian techniques benefit from conceptual simplicity in practice; after assigning a prior to the parameter, one needs only to compute the posterior to make relevant inference. Besides, previous expertise on a problem can be incorporated in the prior distribution of a parameter. Although this implied subjectivity can be seen as a drawback compared to the frequentist approach, subjectivity can never be completely absent from statistical modeling and Bayesian subjectivity might appear more natural. Moreover, once the statistician has access to the posterior distribution, all inference can be done exhaustively. Finally, from a decision theoretical point of view, there is a Bayesian procedure (or a limit of Bayesian procedures) which performs uniformly better than any other procedure according to the "complete class theorem", and only those procedures are admissible, see for instance (Balder et al., 1983) and references therein. For a more exhaustive introduction to Bayesian methods, see the books (Bernardo and Smith, 1994), (Robert, 2001), and (Ghosal and Van der Vaart, 2017).

One of the advantages of the Bayesian framework is the simplicity with which we can quantify uncertainty. As the parameter is a random variable whose posterior distribution depends on the data, we can construct sets containing the parameter with high posterior probability. It is interesting to note that this is a probability statement about the unknown parameter; thus, those sets are quite different from frequentist confidence sets which provide a probability statement about the sets themselves.

After the introduction of both paradigms, the following section will be focused on a hybrid approach where Bayesian methods are used, but given a frequentist interpretation.

§1.2.3 Frequentist Bayes

Since Bayesian inference always starts with the choice of a prior which represents subjective belief of an expert, if two experts do not share the exact same subjective belief, the outcomes of their analyses can vary even if they shared the methodology. This situation seldom occurs within the frequentist community, because of the shared assumption that there exists a single true distribution explaining how the data has been generated. Although the Bayesian paradigm differs significantly from the frequentist one, it can be interesting to see Bayesian inference as a mere different frequentist method. Indeed, one can still assume that the data X follows a fixed probability distribution P_0 belonging to some model $\mathcal{P}$, and then study the posterior distribution $\Pi(\cdot|X)$ of the parameter θ as a random measure which depends on the prior and the "true" distribution P_0. On one hand, this hybrid view allows statisticians to enjoy the relative simplicity of implementation of Bayesian methods, for instance the built-in uncertainty quantification; on the other hand, it gives frequentist guarantees for those same methods.

In the parametric case, where the parameter θ describing the model has a finite dimensional structure, the well-known Bernstein-von Mises theorem establishes that

under mild regularity conditions Bayesian inference is asymptotically equivalent to frequentist inference. More precisely, the theorem states that the posterior becomes asymptotically normal centered around the maximum likelihood or any efficient estimator with variance equal to the inverse of the Fisher information. Despite the strength of this result, its limitation to the parametric and semi-parametric case requires statisticians to find other ways to study non-parametric Bayesian methods.

§1.3 Non-parametric asymptotic statistics

As established, non-parametric statistical techniques are based on models with as few assumptions as possible. Technically, a *non-parametric model* does rely on a parameter θ, but this parameter is infinite dimensional; it can for instance be an infinite sequence or a function. Non-parametric models are attractive thanks to their flexibility and robustness. The books (Wasserman, 2006), (Tsybakov, 2009) and (Giné and Nickl, 2016) provide a more detailed introduction to non-parametric statistics.

In non-parametric statistics, it is common to consider some assumption on the smoothness or regularity of the parameter $\theta \in \Theta$. In this thesis, we consider frequently used smoothness spaces: the Sobolev and the Hölder spaces. These spaces will be defined for functions defined on $[0,1]$, but the idea can be expanded to functions on other domains as well.

Let $(\phi_i)_{i\geq 1}$ denote an orthonormal basis of $L^2([0,1])$ (i.e. measurable functions that are square-integrable in $[0,1]$) and consider functions on $L^2([0,1])$ in the form

$$\theta = \sum_{i=1}^{+\infty} \theta_i \phi_i,$$

where θ_i is the ith coefficient in the series expansion. The Sobolev ball $S^\beta(M)$ with smoothness parameter $\beta > 0$ and radius $M > 0$ is defined as

$$S^\beta(M) = \left\{ \theta \in L^2([0,1]) : \sum_i \alpha_i \theta_i^2 \leq M^2 \right\},$$

where $\alpha_i \asymp i^{2\beta}$. In this thesis, we will also consider a related regularity space of function: hyper-rectangles

$$\Theta^\beta(M) = \{\theta \in L^2([0,1]) : \sup_i i^{2\beta}\theta_i^2 \leq M\}.$$

For integer regularity parameters β and the classical Fourier basis, it is also possible to represent the Sobolev class as the set of $\beta - 1$ times differentiable functions $[0,1]$ with absolute continuous derivatives and with a βth derivative satisfying

$$\int_0^1 (\theta^\beta(t))^2 dt \leq M\pi^{2\beta}.$$

The Hölder space of smoothness $\beta > 0$ on $[0,1]$ denoted by $C^\beta([0,1])$, on the other hand, contains all functions θ with $\lfloor\beta\rfloor$ (the integer part of β) derivatives and their last continuous derivatives verifies the Hölder condition

$$\sup_{x,y\in[0,1]} |\theta^{\lfloor\beta\rfloor}(x) - \theta^{\lfloor\beta\rfloor}(y)| \leq M|x-y|^\alpha$$

with $\alpha = \beta - \lfloor\beta\rfloor$ and $M > 0$. The corresponding Hölder ball with radius $M > 0$ is defined as

$$H^\beta(M) = \left\{ \begin{array}{ll} \theta \in C^\beta([0,1]) : & \sup_{x,y\in[0,1]} |\theta^{\lfloor\beta\rfloor}(x) - \theta^{\lfloor\beta\rfloor}(y)| \leq M|x-y|^{\beta-\lfloor\beta\rfloor}, \\ & \sup_{k\leq\beta} \|\theta^{(k)}\|_\infty \leq M \end{array} \right\}.$$

The asymptotic approach to assess the performance of a statistical technique can be described intuitively. Indeed, the outcomes of any statistical procedure can be arranged in a sequence indexed by the size of the sample used during this procedure. The theoretical study of this sequence allows to understand the behavior of the procedure when the size of the data grows, and ideally the larger our data sample is, the better the procedure performs. Mathematically, if we observe the data $X^{(n)}$, generated from the distribution P_{θ_0}, and our outcome is a random variable called point estimator $T(X^{(n)})$, given by a measurable function T, then we are interested in the study of the sequence $(T(X^{(n)}))_{n\in\mathbb{N}}$. Since this thesis focuses on the frequentist Bayes approach, we will introduce some frequentist properties of Bayesian asymptotics: consistency, contraction rates and coverage of credible sets.

§1.3.1 Consistency

The first frequentist property one could check when applying Bayesian techniques is the consistency of the posterior. *Posterior consistency* simply means that the posterior distribution $\Pi(.|X^{(n)})$ puts most of its mass around a smaller and smaller neighborhood of the true parameter θ_0. Formally, it means that for all $\varepsilon > 0$,

$$\Pi(\theta : d(\theta, \theta_0) < \varepsilon | X^{(n)}) \overset{P_{\theta_0}}{\to} 1,$$

where d is a given metric and the convergence is in probability under the true parameter. Posterior consistency is a necessary frequentist property to assess the performance of a Bayesian method, but it is relatively weak and non-informative.

In non-parametric models, posterior consistency is not always satisfied even if the prior distribution covers the true parameter; it is possible that some posterior distribution does not concentrate around the true parameter even when any neighborhood of this parameter has positive mass, see for instance (Freedman, 1963), (Diaconis and Freedman, 1986) and (Kim and Lee, 2001). Fortunately, Doob's theorem and Schwartz's theorem provide a robust result to assess the consistency of non-parametric Bayesian models. Doob's consistency theorem (Doob, 1949) states that if the true parameter θ_0 is not in some null-set of the prior, then the posterior distribution is consistent. While topologically, prior null-sets can be large, Schwartz's theorem (Schwartz, 1965) gives a more extensive result. It assesses that posterior consistency over the whole parameter space is verified under two conditions: the prior mass condition which requires that the prior assigns sufficiently large probabilities to neighborhoods of the true parameter, and the test condition which necessitates the existence of a sequence of tests which separates the true parameter from the complements of neighborhoods around the parameter.

Although consistency is important in evaluating non-parametric Bayesian techniques, since it is not considered strong enough, we will focus on a more descriptive property: rate of contraction.

§1.3.2 Contraction rates

When estimating a parameter θ, the frequentist asymptotic way to evaluate the estimation $\hat{\theta}_n$, which is a measurable function of the data, is not only to verify that the sequence $(\hat{\theta}_n)_{n\in\mathbb{N}}$ converges (in probability or almost surely) to the true value θ_0 uniformly over the possible "truths", but also to evaluate the risk of this estimator. Mathematically, we first choose a loss function $L : \Theta \times \Theta \to (0, \infty)$ and then define the maximum risk of our estimator as

$$r(\hat{\theta}_n) = \sup_{\theta\in\Theta} E_\theta L(\hat{\theta}_n, \theta),$$

where the expectation E_θ is taken with respect to the probability measure P_θ. This function represents the amount an estimator deviates from the true parameter in the worst case scenario. The choice of a loss function impacts directly the choice of the best estimator. Once the maximum risk over the class Θ has been defined, the goal is to find an estimator which minimizes this risk. Such an estimator is called the minimax estimator, and the risk associated to it is the minimax risk

$$R_n = \inf_{\hat{\theta}_n} r(\hat{\theta}_n),$$

where the infimum is taken over all measurable functions of the data. Thanks to Markov's and Chebyshev's inequality, it can be shown that the minimax estimator $\hat{\theta}^*_n$ has the best convergence rate $\varepsilon_n^2 := R_n$, which means that

$$\sup_{\theta\in\Theta} P_\theta \left(d(\hat{\theta}^*_n, \theta) > M_n \varepsilon_n \right) \to 0,$$

for any $M_n \to +\infty$, and that for all δ_n such that $\delta_n = o(\varepsilon_n)$ there exists at least one $\theta^* \in \Theta$ such that

$$P_{\theta^*} \left(d(\hat{\theta}^*_n, \theta^*) > L\delta_n \right) \nrightarrow 0$$

for all $L > 0$. The minimax risk and estimators have been studied for a large variety of problems. We refer to (Lehmann and Casella, 1998) for more details about the topic.

The analogous property to convergence rate for Bayesian techniques is the rate of contraction. The *contraction rate* ε_n of a Bayesian procedure is a sequence converging to 0 in n such that

$$\Pi_n \left(\theta : d(\theta, \theta_0) \leq \varepsilon_n | X^{(n)} \right) \overset{P_{\theta_0}}{\to} 1.$$

This rate quantifies how quickly the posterior mass concentrates around the true parameter θ_0. In regular parametric models, the posterior contraction rate can always achieve the same order of the optimal frequentist convergence rate. This result follows from the Bernstein-von Mises theorem. However, it is not yet clear if such a result could also exist in the non-parametric case.

(Ghosal et al., 2000), (Ghosal and van der Vaart, 2007), (van der Vaart and van Zanten, 2008) and (Ghosal and Van der Vaart, 2017) address this issue. Let us consider the iid case where $X_1, ..., X_n$ are n random variable distributed according to P_{θ_0}. We consider the Hellinger metric d on the parameter space given by

$$d^2(\theta, \theta') = \int (\sqrt{p_\theta} - \sqrt{p_{\theta'}})^2 d\mu,$$

where p_θ and $p_{\theta'}$ are the densities of P_θ and $P_{\theta'}$ with respect to the measure μ. We also define the entropy number (denoted by $N(\varepsilon, B, d)$), which represents the number of ε-radius balls required to cover the set B with respect to a given metric d.

Theorem 1.3.1 (Theorem 2.1 of (Ghosal et al., 2000))**.** *Suppose that for a sequence ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$, a constant $C > 0$ and subsets $\Theta_n \subset \Theta$ of the parameter set, we have*

$$\begin{aligned}
&\log N(\varepsilon_n, \Theta_n, d) \leq n\varepsilon_n^2, \\
&\Pi(\theta \notin \Theta_n) \leq \exp(-(C+4)n\varepsilon_n^2), \\
&\Pi\Big(\theta : P_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}} \leq \varepsilon_n^2, P_{\theta_0} \log^2 \frac{p_\theta}{p_{\theta_0}} \leq \varepsilon_n^2\Big) \geq \exp(-Cn\varepsilon_n^2).
\end{aligned}$$

Then for sufficiently large M, we have that $\Pi(\theta : d(\theta, \theta_0) \geq M\varepsilon_n | X^{(n)}) \to 0$ in P_{θ_0}-probability.

The conditions can be understood as follows: the first condition, the entropy condition, implies that the sets Θ_n are not too large; the second condition, the remaining mass condition, implies that prior distribution puts enough mass on these sets; the last condition, the prior mass condition, requires the prior to put enough mass in a small neighborhood of the true distribution.

We consider that the contraction rate is optimal if it is equal to the minimax convergence rate. It is relatively easy to construct an estimator $\hat{\theta}$ from the posterior which attains the same convergence rate as the posterior construction. Indeed, we can for instance take the center of the smallest ball with posterior mass greater than $1/2$, see (Ghosal et al., 2000) and (Ghosal and Van der Vaart, 2017). Therefore, finding a posterior achieving optimal contraction rate leads directly to a Bayesian estimator with minimax convergence rate.

§1.3.3 Bayesian uncertainty quantification

Once the posterior distribution is computed, one can derive Bayesian point estimates. Those are simply functions of the data which minimize an integrated loss function with respect to the posterior measure. Despite their usefulness, point estimates do not provide information about the uncertainty of our inference on θ_0. For that reason, it is also interesting to generate a set of possible values for our parameter.

In the frequentist case, these sets are known as confidence sets. *Confidence sets* are set-valued estimators $\hat{C}_n(X^{(n)})$ to which the true parameter is most likely to belong, according to the data. More formally, we define a confidence set $\hat{C}_n$ of level $1-\alpha$ as

$$\inf_{\theta \in \Theta} P_\theta(\theta \in \hat{C}_n) \geq 1 - \alpha.$$

Asymptotically, confidence sets should also verify

$$\liminf_{n \to \infty} \inf_{\theta \in \Theta} P_\theta(\theta \in \hat{C}_n) \geq 1 - \alpha.$$

These sets have asymptotically uniform confidence level $1-\alpha$. In other words, they have the same confidence level asymptotically even in the worst case scenario for the true parameter.

Nonetheless, the construction of confidence sets is not easy when the model is computationally complex. In the Bayesian approach, we talk instead of *credible sets*, which accumulate most of the posterior probability mass. Since it is possible to sample from the posterior, the construction of approximate credible sets thanks to Markov chain Monte-Carlo (MCMC) methods is relatively simpler than the construction of confidence sets. Formally, credible sets $\tilde{C}_n$ of level $1-\alpha$ are sets such that

$$\Pi_n(\theta \in \tilde{C}_n | X^{(n)}) \geq 1 - \alpha.$$

Contrary to confidence sets, credible sets are subjective in nature because of the influence of the prior distribution in the posterior. Although confidence sets and credible sets are asymptotically equivalent in the parametric case under some regularity assumptions thanks to the Bernstein-von Mises theorem, the question of their equivalence in the non-parametric case has not been entirely answered.

Actually, it is known through several negative examples that Bayesian credible sets do not converge automatically to frequentist confidence sets when the model is infinite-dimensional, see for instance (Cox, 1993). This phenomenon encourages a further analysis of the behavior of credible sets from a frequentist point of view. It appears in (Knapik et al., 2011), (Castillo and Nickl, 2013), (Szabo et al., 2015), (Sniekers and van der Vaart, 2015b), (Belitser, 2017) and (Szabo et al., 2017) that in some situations, Bayesian credible sets ensure good coverage of the true parameter, that is to say that the credible sets are asymptotically likely to contain the true parameter according to the true distribution. Moreover, (Rousseau and Szabo, 2020) studied the asymptotic coverage of Bayesian credible sets in a general non-parametric framework. The general intuition is that Bayesian credible sets have good coverage when the prior generates parameters that are slightly less smooth than the true parameter.

§1.4 Gaussian processes

The examination of Bayesian techniques from a frequentist perspective requires a good understanding of the prior distribution because it influences heavily the outcome of a Bayesian statistical procedure. In this work, we mainly treat Gaussian process priors. *Gaussian processes* are stochastic processes, i.e. collections of random variables indexed by time or space which can be viewed as random elements in a function space verifying the following condition: every finite collection of the evaluation of the process at different times has a multivariate normal distribution. We refer to (Rasmussen and Williams, 2006) for a more comprehensive introduction to Gaussian processes and their applications.

A Gaussian process $W = \{W_t : t \in T\}$ indexed by a set T is characterized by its mean $\mu : T \to \mathbb{R}$ and its covariance function $K : T \times T \to \mathbb{R}$ given by $\mu(t) = EW_t$ and $K(s,t) = \mathrm{Cov}(W_t, W_s)$, respectively. Generally, the set T is a subset of $\mathbb{R}^d$ so that the function $t \mapsto W_t$ belongs to the space of real-valued functions on T. As priors, we typically take centered Gaussian processes, i.e. their mean $\mu(t)$ is set to be zero; thus, their behaviors can be understood entirely through their covariance functions. Moreover, thanks to Mercer's theorem (Mercer, 1909), we can see that the covariance

function of a Gaussian process can be represented as follows

$$K(s,t) = \sum_{j=1}^{+\infty} \lambda_j \phi_j(s)\phi_j(t),$$

because the covariance is a symmetric non-negative definite kernel. Here, the ϕ_j's are eigenfunctions representing an orthonormal basis of $L^2(T)$, the λ_j's are the corresponding non-negative eigenvalues and the convergence of the series is absolute and uniform when the kernel is continuous and T is compact. In addition to Mercer's theorem, the Karhunen-Loève theorem, (Karhunen, 1947) and (Loève, 1978), provides a nice representation of the random variable W_t for any $t \in T$:

$$W_t = \sum_{j=1}^{+\infty} \sqrt{\lambda_j} Z_j \phi_j(t),$$

where the Z_j's are independent standard normal random variables and the convergence is almost sure.

In this thesis, we will focus on the following specific Gaussian processes (GP):

- the GP with a **Matérn covariance kernel**

  $$K(s,t) = \frac{2^{1-\alpha}}{\Gamma(\alpha)} \Big(\frac{|s-t|\sqrt{2\alpha}}{a} \Big)^{\alpha} K_\alpha \Big(\frac{|s-t|\sqrt{2\alpha}}{a} \Big),$$

  where Γ is the gamma function, K_α is the modified Bessel function, a is a rescaling parameter and α is the smoothness parameter. The sample paths $t \mapsto W_t$ of a GP with Matérn covariance function are $\lfloor \alpha \rfloor$ times differentiable.

- the GP with a **squared-exponential covariance kernel**

  $$K(s,t) = \exp\Big(\frac{(s-t)^2}{2a^2} \Big),$$

  where a is a scaling parameter. The squared-exponential covariance kernel can be seen as the limit of Matérn covariance kernels with α going to infinity. A GP with squared-exponential covariance function has infinitely differentiable sample paths.

- the **Brownian motion and its primitives**, the Brownian motion has covariance kernel

  $$K(s,t) = \min(s,t).$$

  Since the sample paths of a Brownian motion are Lipschitz continuous of any order $\alpha < 1/2$ but not differentiable anywhere, it is common to integrate this process k times in order to have a GP with a smoothness order $k + 1/2$. Moreover, the k-fold integrated Brownian motion has vanishing derivatives at zero; hence, it is common to "release" this process by adding a polynomial process $t \mapsto \sum_{j=0}^{k} Z_j t^j / j!$ where the Z_j's are independent standard normal random variables. Thus, the k-fold integrated Brownian motion can be written as

  $$G_t = \sum_{j=0}^{k} \frac{Z_j t^j}{j!} + I^k W_t,$$

where $I^k = I^{k-1}I$ and $If(t) = \int_0^t f(s)ds$.

Both the GP with Matérn and with squared-exponential covariance kernel are stationary; $\mathrm{Cov}(W_t, W_s)$ only depends on the distance between the two points t and s, and not on their position. On the other hand the Brownian motion, while not stationary, has independent and stationary increments, which means the consecutive increments of this process over the same distance are iid random variables.

Gaussian processes appear to be good priors of choice in the statistical models studied in this thesis. For more detailed investigations on these processes, we give a non-exhaustive list of work and references therein: (Tokdar and Ghosal, 2005), (Choudhuri et al., 2007), (van der Vaart and van Zanten, 2007), (van der Vaart and van Zanten, 2008) and (van der Vaart and van Zanten, 2009b).

§1.5 Models

The main non-parametric models investigated in this thesis are the signal in Gaussian white noise model and the non-parametric regression model.

§1.5.1 Signal-in-white-noise model

Let the random function $Y^{(n)}(t)$ be defined as follows

$$Y^{(n)}(t) := \int_0^t \theta_0(s)ds + \frac{1}{\sqrt{n}}W_t, \qquad t \in [0,1],$$

where $\theta_0 \in L^2[0,1]$ is the parameter of interest and W_t is the Brownian motion. This problem can also be interpreted as observing n independent realizations of a random variable defined as

$$Y_j(t) := \int_0^t \theta_0(s)ds + W_{j,t}, \qquad j \in \{1, ..., n\}, \qquad t \in [0,1],$$

where the $W_{j,t}$ are n independent Brownian motions.

It is usual for practicality to convert the problem into the spectral domain using an orthonormal basis $(\phi_i)_{i\geq 1}$ of $L^2[0,1]$. The random function $Y^{(n)}(t)$ becomes a sequence $(Y_i)_{i\geq 1}$ defined as

$$Y_i = \theta_{0,i} + \frac{1}{\sqrt{n}}Z_i, \qquad i \in \mathbb{N},$$

where $(\theta_{0,i})_{i\geq 1} \in \ell^2$ is an infinite sequence and $(Z_i)_{i\geq 1}$ are independent standard normal random variables. The relative simplicity of the model and its relation to non-parametric regression (Brown and Low, 1996) and (Nussbaum, 1996) makes it an important benchmark model in the literature; see for instance (Donoho, 1994), (Tsybakov, 2009) and (Giné and Nickl, 2016).

In the case $\theta_0 \in S^\beta$ with regularity parameter $\beta > 0$, the minimax convergence rate to estimate the parameter is proportional to $n^{-\beta/(1+2\beta)}$ with respect to the L_2-loss; see (Tsybakov, 2009) and reference therein. One way to achieve this rate is to

simply estimate the Fourier coefficients of θ_0 with the first $n^{1/(1+2\beta)}$ coefficients of $(Y_i)_{i\geq 1}$. Furthermore, it has also been shown in (Knapik et al., 2011) and (Ghosal and Van der Vaart, 2017) that using an appropriately smooth GP prior (for instance a squared-exponential GP with rescaling $a_n := n^{1/(1+2\beta)}/\log^{2/(1+2\beta)} n$) leads to optimal posterior contraction rates (up to a multiplicative logarithmic factor).

§1.5.2 Non-parametric regression

Let $(X_i, Y_i)_{i=1}^n$ be n iid pairs of random variables such that

$$Y_i = \theta_0(X_i) + Z_i, \quad i \in \{1, ..., n\}$$

where the X_i's are in a set $\mathcal{X}$ (e.g. $[0,1]$), the Z_i's are iid centered random variables (generally taken as standard normal) and $\theta_0 \in L^2$ is the functional parameter of interest. When X_i is deterministic, we talk about regression with fixed design, for instance the regular grid $X_i = i/n$ on $\mathcal{X} = [0,1]$, while in the case of random X_i, we talk about regression with random design. Generally, θ_0 is assumed to belong to a certain class of smoothness with regularity $\beta > 0$, for instance a Sobolev or Hölder class. The corresponding minimax rate for both classes is, up to a constant multiplier, $n^{-\beta/(1+2\beta)}$ with respect to the empirical L_2-norm in the fixed design case and the usual L_2-norm in the uniformly random design case.

Among the estimators achieving the minimax convergence rate, we can present the Nadaraya-Watson estimator introduced in (Nadaraya, 1964) and (Watson, 1964)

$$\hat{\theta}_n(x) = \frac{\sum_{i=1}^n K(\frac{x-X_i}{h})Y_i}{\sum_{i=1}^n K(\frac{x-X_i}{h})},$$

which uses a kernel K, a non-negative integrable function, as a weight function and $h > 0$ is a smoothing parameter called the bandwidth. In (Bickel and Doksum, 1977), we can see that if the bandwidth is well chosen, namely $h \asymp n^{-1/(2\beta+1)}$, then both the point-wise and integrated mean-square errors are proportional to the optimal convergence rate $n^{-\beta/(1+2\beta)}$. In this model as well, (van der Vaart and van Zanten, 2008), (Knapik et al., 2011), (Ghosal and Van der Vaart, 2017) and (Bhattacharya et al., 2017) have shown that using a GP prior can result in a posterior density which contracts around the true parameter at an optimal rate in L_2-norm up to a possible logarithmic factor.

§1.6 Adaptation

In the previous section, we have seen that some estimators are minimax for the discussed models. However, these estimators require the knowledge of the exact regularity of the true parameter. In practice, this might not be realistic since the true regularity is seldom known in advance; thus, it is desirable to have procedures which do not rely on the true smoothness hyper-parameters, but adapt to them. In mathematical words, we can consider that the parameter of interest θ_0 belongs to a parameter space Θ seen as a collection of subspaces Θ^β indexed by a hyper-parameter

β. We would like to find an estimator $\hat{\theta}_n \in \Theta$ which attains the minimax rate $R_{n,\beta}$ corresponding to Θ^β whatever the true smoothness β is

$$\sup_{\theta \in \Theta^\beta} E_\theta L(\hat{\theta}_n, \theta) \leq C_\beta R_{n,\beta},$$

where $C_\beta > 0$ is a positive constant depending only on β and $R_{n,\beta}$ is defined as

$$R_{n,\beta} = \inf_{\hat{\theta}_n \in \Theta^\beta} \sup_{\theta \in \Theta^\beta} E_\theta L(\hat{\theta}_n, \theta).$$

Although the theory of adaptation is fairly developed in frequentist statistics; see for instance (Bickel, 1982), (Lepski and Spokoiny, 1997), and (Goldenshluger and Nemirovski, 1997), its Bayesian counterpart is relatively recent, and has been getting more attention, see the monograph (Ghosal and Van der Vaart, 2017) and reference therein.

The idea is to construct a prior distribution reaching the optimal contraction rate for a set of possible regularities. In the non-parametric case, the oracle choice, the tuning parameter of the prior associated with the true smoothness, is impossible to select in practice. This encourages us in most cases to make a data-driven selection of this parameter. The two adaptive Bayesian techniques discussed in this thesis are the empirical Bayes and the hierarchical Bayes methods.

§1.6.1 Empirical Bayes

One possible way to proceed is to estimate the tuning parameter from the data. Although not fully Bayesian, since the parameter of interest is estimated using frequentist techniques, this method can be computationally convenient in some statistical settings. This approach is known as the empirical Bayes method in the literature. Basically, if the prior distribution is tuned by a parameter α such that we can denote it by Π_α, we first need to estimate this parameter by maximizing the marginal likelihood $\int_\Theta P_\theta(X) d\Pi_\alpha(\theta)$ seen as a function of α to obtain an estimator $\hat{\alpha}$ with X being our observation, then plug $\hat{\alpha}$ into the posterior used in our inference. In other words, the maximum marginal likelihood estimator (MMLE) is

$$\hat{\alpha} = \arg\max_\alpha \int_\Theta P_\theta(X) d\Pi_\alpha(\theta),$$

and the corresponding posterior would be $\Pi_{\hat{\alpha}}(\cdot|X) = \Pi_\alpha(\cdot|X)\Big|_{\alpha=\hat{\alpha}}$.

We consider the following toy-example to demonstrate the idea. Let $X^{(n)}$ be a sample of n iid observation from a Bernoulli distribution with unknown mean θ. We endow θ with a Beta prior with parameters $a, b > 0$, which are also unknown. We will first estimate the hyper-parameter a and b by maximizing the marginal likelihood

$$(\hat{a}_n, \hat{b}_n) = \arg\max_{a,b>0} \frac{1}{B(a,b)} \theta^{a+\sum_i X_i - 1} (1-\theta)^{b+n-\sum_i X_i - 1},$$

where $B(a,b)$ is the normalizing Beta function. Then, we plug $\hat{a}_n$ and $\hat{b}_n$ in the posterior distribution of the parameter θ

$$\Pi_{\hat{a}_n, \hat{b}_n}(\cdot|X^{(n)}) = \Pi_{a,b}(\cdot|X^{(n)})\Big|_{a=\hat{a}_n, b=\hat{b}_n} \sim \text{Beta}\big(\hat{a}_n + \sum_i X_i, \hat{b}_n + n - \sum_i X_i\big).$$

The approach benefits from the fact that once the estimators of the hyper-parameters are computed, the posterior distribution generally becomes simple to compute. Furthermore, the choice of the hyper-parameters makes sense intuitively for the frequentist community, since it removes part of the subjectivity tied to the prior. For a more comprehensive overview of the use of Empirical Bayes in practice and theoretical properties thereof, see (Johnstone and Silverman, 2005), (Belitser and Enikeeva, 2008), (Jiang and Zhang, 2009) (Szabo et al., 2013) and (Rousseau and Szabo, 2017).

§1.6.2 Hierarchical Bayes

On the other hand, the hierarchical Bayes method is more appealing to Bayesian statisticians. In this approach we treat the hyper-parameter α as a random variable, similarly to the parameter θ, and we endow it with a hyper-prior distribution. Formally,

$$X|\theta \sim P_\theta, \quad \theta|\alpha \sim \Pi_\alpha, \quad \alpha \sim \Lambda,$$

where Λ is a hyper-prior distribution, Π_α is the prior distribution of the parameter of interest θ conditionally on α and P_θ is the distribution of our data conditionally on the parameter θ. This creates a multilevel, hierarchical Bayesian procedure. We are then interested in the marginal posterior distribution $\Pi(\cdot|X) = \int \Pi_\alpha(\cdot|X)d\Lambda(\alpha)$. This fully Bayesian method has been studied in different models, and it has been shown in, among other papers, (Huang, 2004), (Lember and van der Vaart, 2007), (de Jonge and van Zanten, 2009), (van der Vaart and van Zanten, 2009a) and (Arbel et al., 2013) that if the hyper-prior is chosen appropriately, the posterior distribution contracts at an optimal rate adaptively.

Considering our previous toy-example, in a hierarchical Bayes procedure, we would put a hyper-prior on $a, b > 0$, for instance two independent Gamma distributions with parameters $k, \gamma > 0$. Formally, we would have

$$X^{(n)}|\theta \sim \mathrm{Ber}(\theta), \quad \theta|a,b \sim B(a,b), \quad a, b \overset{\text{ind}}{\sim} \Gamma(k,\gamma).$$

In some cases, it is possible to compute the marginal posterior $\theta|X^{(n)}$ straightforwardly; however, in most cases, only approximation techniques (like MCMC methods) can be applied.

One of the reasons hierarchical Bayes procedures are popular is that they ensue directly from the Bayesian philosophy. Indeed, once a suitable sampling method is found for the parameters of interest, most questions about the data-generating process can be answered from a Bayesian point of view.

§1.7 Distributed computation

An asymptotic assessment of a statistical procedure requires the size n of the data to be large; paradoxically, the optimal procedures suffer from having larger and larger data set, consequently increasing computation time. On one hand, a larger sample size is appreciated because it leads to more precise statistical statements, and on the other hand, increasing the number of observations results in more computational burden. Among the statistical procedures showcasing this phenomenon, we can give

the GP non-parametric regression as an example. In this statistical model, we endow the functional parameter θ with a GP prior $\theta \sim \mathrm{GP}(0, K)$ where K is a covariance kernel. When the noise is also Gaussian with $(Z_i)_{i=1}^n \overset{\text{ind}}{\sim} N(0, \sigma^2)$, the corresponding posterior becomes easy to compute and we have $\theta|(X_i, Y_i)_{i=1}^n \sim \mathrm{GP}(\hat{\theta}, \hat{K})$, where the parameters $\hat{\theta}$ and $\hat{K}$ have closed forms:

$$\hat{\theta}(x) = K(x, \mathbf{X})^T (\mathbf{K} + \sigma^2 I_n)^{-1} \mathbf{Y}, \tag{1.7.1}$$

$$\hat{K}(x, x') = K(x, x') - K(x, \mathbf{X})^T (\mathbf{K} + \sigma^2 I_n)^{-1} K(x', \mathbf{X}), \tag{1.7.2}$$

where $\mathbf{Y} = (Y_i)_{i=1}^n$, $K(x, \mathbf{X}) = (K(x, X_i))_{i=1}^n$ and $\mathbf{K} = (K(X_i, X_j))_{i,j=1}^n$ for all $x, x' \in \mathcal{X}$ where $\mathcal{X}$ is a compact space (generally $[0,1]^d$). Even though it is possible to compute the parameters of the posterior directly, they require the inversion of a $n \times n$ matrix, which scales the computational complexity to $O(n^3)$ and requires a memory of order $O(n^2)$.

Against the background of the high computational complexity, methods based on distributed computation have emerged. *Distributed methods* partition the data over several machines called experts, see for instance (Jordan and Jacobs, 1994), (Minsker et al., 2014), (Ng and Deisenroth, 2014), (Cao and Fleet, 2014), (Srivastava et al., 2015) and (Scott et al., 2016). The experts process their share of the data locally, solving smaller versions of the problem. Then, all the local results are aggregated on a central machine to produce a final outcome of the statistical analysis. To formalize these ideas, let us partition the data $(X_i, Y_i)_{i=1}^n$ into m batches $(X_i^j, Y_i^j)_{i=1}^{n_j}$ with $j \in \{1, ..., m\}$, such that $\sum_j n_j = n$. So as to keep the simplicity of the GP non-parametric regression problem, we assume that each machine endows the parameter with a GP prior. Once the local posterior distributions $\theta|(X_i^j, Y_i^j)_{i=1}^{n_j}$ are computed, we can sample from the global posterior by aggregating each sample from the local posterior distributions. There exist different types of distributed computation methods based on the manner the data is partitioned, the way the local posterior or its modification is computed and the aggregation technique used in the process. We provide here a short description of these methods, but more details will be given in Chapters 3 and 4.

§1.7.1 Uniformly random partitioning

The data can be partitioned uniformly randomly among the machines. For simplicity, we assume that $n \equiv 0 \bmod m$. Each machine will receive n/m data points chosen randomly among $(X_i, Y_i)_{i=1}^n$ such that no data point simultaneously belongs to two machines or more.

Although in the classical Bayesian approach, the posterior is merely proportional to the likelihood multiplied by the prior, it is sometimes beneficial in distributed computation to modify the local posteriors. For instance, in order to tone down the effect of the prior, it can be useful to raise the prior to a power decreasing in m, the number of machines, for instance $1/m$. Another technique would be to increase the effect of the likelihood by raising it to a higher power, for instance m.

Furthermore, the aggregation of the local posteriors also affects the quality of the global posterior. In the case the data has been partitioned uniformly randomly, one

can either simply average the local posterior distributions or compute their Wasserstein barycenter, which is based on the Wasserstein distance between probability measures.

§1.7.2 Spatial partitioning

It is also possible to partition the data spatially. In this case, we partition the design space $\mathcal{X}$ into m sub-regions called $\mathcal{D}_j$ with $j \in \{1, ..., m\}$, and the jth machine will deal with the data points $\{(X_i, Y_i)_{i=1}^n, X_i \in \mathcal{D}_j\}$.

In this scenario, we will see that no modification of the local posteriors is needed. Moreover, even though each machine only receives observations with X_i in $\mathcal{D}_j$, it can produce a posterior mean and covariance for all $x \in \mathcal{X}$. Then, it is possible to draw from the local posterior and compute a weighted average of these draw, subsequently constructing a global posterior. Mathematically, if θ_j is a sample from the local posterior Π_j, then

$$\theta = \sum_{j=1}^{m} \omega_j \theta_j,$$

is a draw from the global posterior Π^*. The functions ω_j are defined on $\mathcal{X}$ such that $\sum_{j=1}^m \omega_j(x) = 1$ for all $x \in \mathcal{X}$. The most naive approach would be to take $\omega_j = \mathbf{1}_{\mathcal{D}_j}$, "gluing" in a sense the local GP posteriors. However, this approach leads to discontinuous samples from the global posterior. In order to avoid discontinuities, continuous data-driven weight functions concentrating around $\mathcal{D}_j$ are favored.

§1.8 Overview

This thesis focuses on frequentist properties of Bayesian techniques. The first chapter examines how adaptation affects uncertainty quantification using exponentially decaying covariance kernel. The two following chapters (Chapter 3-4) focuses on distributed computation in the Bayesian non-parametric regression model. In Chapter 3, we derive contraction rates and coverage for uniformly randomly partitioned distributed methods when the smoothness of the true function is known. On the other hand, Chapter 4 deals with adaptive optimal recovery for spatially partitioned distributed methods. Finally, the last chapter is an extensive comparative simulation study of the aforementioned distributed methods.

§1.9 Notations

For two positive sequences a_n, b_n we use the notation $a_n \lesssim b_n$ if there exists an universal positive constant C such that $a_n \leq C b_n$. Along the lines $a_n \asymp b_n$ denotes that $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold, simultaneously. For $\theta \in L_2[0,1]$ we denote the standard L_2-norm as $\|\theta\|_2^2 = \int_0^1 \theta(x)^2 dx$ and let $\mathrm{diam}(S)$ denote the ℓ_2-diameter of the set $S \subset \ell_2$. Throughout the thesis, c and C denote global constants whose value may change one line to another. The dependence of the constants c, C on the model parameters we denote by sub-indexes, e.g. c_β, $C_{\beta,m,M}$.