Asymptotic results in nonparametric Bayesian function estimation

Kirichenko, A.

Chapter 1

Bayesian nonparametric statistics

In this chapter we give an introduction to nonparametric statistical analysis with the focus on the theoretical assessment of Bayesian methods.

1.1 Statistical inference

Consider an observation X which might be any mathematical object, such as a collection of numbers, vectors, matrices, functions, etc. Let X belong to some space $\mathscr{X}$ that consist of other potential observations that we think we could have seen instead of X, meaning that we assume that X has been a result of a random choice from the elements of $\mathscr{X}$. Denote by P the distribution on $\mathscr{X}$ according to which this random choice has been made. This distribution is not known to us and usually is the object of interest. We restrict our possibilities by letting P belong to a statistical model $\mathscr{P}$. The statistical model $\mathscr{P}$ is a collection of probability distributions that is usually parameterised using some parameter space Θ

$$\mathscr{P} = \{P_\theta, \theta \in \Theta\}.$$

The assumptions about the data are incorporated in the model through the parameter θ. The goal of a statistical inference is to learn different aspects of the probability distribution P_θ from the model that fits the data X the best.

The parameter set Θ might be equal to a subset of Euclidean space, or it might be some infinite-dimensional space, for example, a space of functions. In the latter case we call the problem nonparametric.

1.2 Nonparametric models

In this thesis we focus on studying nonparametric models. Although, finite–dimensional models are simpler and attractive for computational convenience, they are often too restrictive. As more is assumed, when the assumptions are not correct, parametric models fail to deliver accurate estimates. Nonparametric models aim to make an estimation using fewer assumptions. These models are known to be more flexible and robust, which makes them attractive for practical applications. A broader introduction to nonparametric statistics can be found for instance in Wasserman [2004].

In function estimation problems it is common to consider nonparametric models with some assumption on the smoothness or regularity of the functional parameter $\theta \in \Theta$, which can be defined in different fashions. In this thesis we use two kinds of smoothness spaces: the Sobolev and the Hölder spaces. Here we present the definitions of these spaces for functions defined on the interval $[0, 1]$. However, the notions can be generalised to functions on other domains as well.

Sobolev spaces are defined as follows. Let ψ_j denote an orthonormal basis of $L^2[0, 1]$ and consider functions in $L^2[0, 1]$ of the form

$$\theta = \sum_{j=1}^{\infty} \theta_j \psi_j,$$

where the θ_j are the coefficients of θ with respect to the basis ψ_j. The Sobolev ball $H^\beta(Q)$ of radius $Q > 0$ and regularity parameter $\beta > 0$ is given by

$$H^\beta(Q) = \{\theta \in L^2[0, 1] : \sum_{j=1}^{\infty} j^{2\beta}\theta_j^2 \leq Q^2\}.$$

When β is an integer and ψ_j is the classical Fourier basis, the Sobolev class is equal to the set of functions $\theta : [0, 1] \to \mathbb{R}$ with $\beta - 1$ absolute continuous derivatives and with the βth derivative $\theta^{(\beta)}$ satisfying

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

The Hölder space of order β on $[0, 1]$ consists of functions in $C[0, 1]$ such that they have continuous derivatives up to order $k = \lfloor\beta\rfloor$ and their kth derivative satisfies the Hölder condition with the exponent $\alpha = \beta - k$. More precisely, the Hölder ball $C^\beta(Q)$ of radius $Q > 0$ and the smoothness parameter β is defined as follows

$$C^\beta(Q) = \{\theta : \sup_{x,y\in[0,1]} |\theta^{(k)}(x) - \theta^{(k)}(y)| \leq Q|x - y|^\alpha, \max_{j\leq k} \|\theta^{(j)}\|_\infty \leq Q\}.$$

where $\theta \in C[0, 1]$.

In order to provide a better insight into typical nonparametric problems, we provide examples of the most basic nonparametric models: the Gaussian white noise model, nonparametric regression, and density function estimation. For a more detailed description and the discussion of the models see, for instance, Tsybakov [2009].

In the Gaussian white noise model the goal is to recover the function $\theta_0 \in L^2[0,1]$ from an observed sample path $X^{(n)} = \{X(t), t \in [0,1]\}$ satisfying the following stochastic differential equation

$$dX(t) = \theta_0(t)dt + \frac{1}{\sqrt{n}}dW(t), \quad t \in [0,1], \tag{1.1}$$

where W is the standard Brownian motion.

The regression model can be defined in various ways. In univariate nonparametric regression problem we observe n independent pairs of random variables (X_i, Y_i) satisfying

$$Y_i = \theta_0(X_i) + \xi_i,$$

where $X_i \in [0,1]$, ξ_i given X_i, say, are independent identically distributed random variables with mean zero, and θ_0 is the function of interest. A particular case of this problem is regression with fixed design, where the X_i are deterministic points in $[0,1]$, for instance, the regular grid points i/n. In the case of random X_i the problem is called regression with random design. It is commonly assumed for the true function to be smooth, for example, for it to belong to a Sobolev space.

In the density function estimation model we observe n identically distributed real-valued random variables with a common distribution that is absolutely continuous, for example, with respect to the Lebesgue measure on $[0,1]$. Let $\rho_0 : \mathbb{R} \to [0,\infty)$ be the density function of this distribution. If we know a-priori that ρ_0 belongs to a parametrisable family with finite-dimensional parameter the problem becomes parametric. However, in many application cases there is no prior information on this function. Hence, it is common to have relatively weak restrictions, such as assuming that the density belongs some Hölder class, making the model nonparametric.

In the next section we introduce the asymptotic approach to comparing estimation procedures by looking at the maximum risks of the estimators. We also present minimax rates for the three models discussed above.

1.3 Frequentist asymptotics

One of the ways to assess the performances of the estimation methods is to take an asymptotic approach. Any statistical procedure can be indexed by the size of the sample used to calculate it, therefore producing a sequence of estimators. Properties of this sequence describe the behaviour of the estimation procedure when the sample size increases. An intuitively reasonable requirement for an estimation procedure is (asymptotic) consistency, which asks the outcome of the procedure with unlimited data to identify the underlying truth.

In order to define it mathematically we consider an observation $X = X^{(n)}$ indexed by n, where n is for instance the number of observations, or, as in the Gaussian white noise model, the signal to noise ratio. Assume that there exists a true distribution P_{θ_0} according to which the data is generated. Let d be a semi–distance on the parameter space Θ. A sequence of estimators $\hat{\theta}_n$ is called (asymptotically) consistent if

$$d(\hat{\theta}_n, \theta_0) \xrightarrow{P_{\theta_0}} 0, \text{ as } n \to \infty ,$$

where the convergence is in probability with respect to P_{θ_0}. We can measure the performance of an estimator $\hat{\theta}_n$ of θ using the maximum risk of this estimator on the set Θ, which is defined as follows

$$r(\hat{\theta}_n) = \sup_{\theta \in \Theta} \mathbb{E}_\theta l(\hat{\theta}_n, \theta),$$

where $\mathbb{E}_\theta$ is the expectation with respect to the probability measure P_θ and $l : \Theta \times \Theta \to \mathbb{R}$ is a loss function. This function quantifies the amount by which the estimator deviates from the value of the true parameter. The choice of a meaningful loss function depends on the studied model and is far from obvious. We consider the commonly used loss function $l = d^2$.

The minimax risk associated with the statistical model $\{P_\theta, \theta \in \Theta\}$ is given by

$$R_n = \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta l(\hat{\theta}_n, \theta),$$

where the infimum is taken over all possible estimators. For a wide range of problems it is possible to establish an asymptotic lower and an upper bound on the minimax risk. An estimator attaining the lower bound up to a constant is called a rate optimal estimator. In problems where a smooth function is being estimated the minimax rate is influenced by the regularity of the true function. Pinsker's theorem (see for instance Theorem 3.1 in Tsybakov [2009]) asserts that the minimax rate for the Gaussian white noise model is equal to $n^{-2\beta/(2\beta+1)}$ with respect to the squared L^2 loss function, where β is the Sobolev smoothness of the target function.

Theorem 1.3.1 (Pinsker's theorem)**.** *Let $\beta, Q > 0$. Then for the Gaussian white noise model the minimax risk satisfies*

$$\lim_{n \to \infty} \inf_{\hat{\theta}_n} \sup_{\theta \in H^\beta(Q)} \mathbb{E}_\theta n^{2\beta/(2\beta+1)} \|\hat{\theta}_n - \theta\|_2^2 = C^\star,$$

where the infimum is taken over all estimators and

$$C^\star = Q^{2/(2\beta+1)} (2\beta+1)^{1/(2\beta+1)} \left(\frac{\beta}{\pi(\beta+1)} \right)^{2\beta/(2\beta+1)} .$$

Observe that in this case it is possible to determine the exact minimax rate including the constant in front of the exponent. Also, note that the theorem provides a lower and an upper bound on the risk. The latter is done by devising a projection estimator (called Pinsker's estimator) that attains the minimax rate. To construct

the estimator we transfer the signal into a sequence of Fourier coefficients by taking the inner product of the sample path X with the elements ψ_j of the orthonormal basis of L^2. Given the model (1.1) the following infinite sequence of Gaussian observations is available to the statistician

$$X_j = \int_0^1 \psi_j(t) dX(t) = \theta_j + \frac{1}{\sqrt{n}} \xi_j,$$

where ξ_j are independent standard Gaussian and the θ_j are the Fourier coefficients of θ_0 with respect to the basis ψ_j. Then we can estimate the Fourier coefficients of the true parameter with the first $n^{1/(2\beta+1)}$ coefficients of the observed signal.

From Brown and Low [1996] and Nussbaum [1996] we know that both the nonparametric regression model and the density function estimation model are asymptotically equivalent to the Gaussian white noise model. In fact, for the regression or density function θ_0 in the Sobolev class with a smoothness parameter $\beta > 0$ the minimax rate in those models is equal to $n^{-\beta/(2\beta+1)}$ up to a constant with respect to the L^2–norm (for more details see, for instance, Tsybakov [2009] and the references therein).

1.4 Bayesian approach

Contrary to the frequentist perspective, the Bayesian approach to statistics is based on the belief that there is no true fixed underlying parameter, but the parameter θ is random itself. It uses probability distribution on the parameter set to represent the belief of the statistician about the structure of the data. Consider a prior distribution Π to be a probability measure on the parameter space Θ. Then in the Bayesian setting the measure P_θ describes the conditional distribution of X given the parameter value θ. The prior can be interpreted as the degree of belief attached to subsets of the model before any observations has been made. Central in the Bayesian framework is the conditional distribution of θ given X. It is called the posterior distribution and can be viewed as a data–amended version of the prior after the observations are incorporated in the estimation procedure. If every distribution P_θ in the model $\mathscr{P}$ admits a density p_θ with respect to some dominating measure, the posterior distribution can be described using Bayes formula

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

for any measurable set $A \subset \Theta$. However, in many applications the straightforward computation of the posterior is not possible. A major reason for the popularity of Bayesian methods is the availability of sampling algorithms such as Markov chain Monte Carlo (MCMC) methods. These methods are designed for sampling from a probability distribution P without having an explicit formula for it. They are based on the construction of Markov chains with the equilibrium distribution P.

They help to draw from a posterior in the settings when it cannot be computed explicitly. More details on MCMC methods can be found, for instance, in the books Gilks et al. [1995] and Brooks et al. [2011].

Apart from philosophical reasons, Bayesian techniques are commonly used in practical applications due to their conceptual simplicity. Additionally, an attractive feature of such techniques for applications is the ability to incorporate knowledge about the parameter into a prior distribution. Moreover, Bayesian procedures can be considered as a form of regularisation which is performed to avoid overfitting. Specifically, a Bayesian method can give higher probabilities to the functions that are considered to be more likely, for example because of their smoothness properties. Finally, Bayesian procedures can be appealing from point of view of decision theory because of the complete class theorems. These theorems state that under mild conditions for any procedure there is a better (or at least not worse) Bayes procedure, and only those procedures are admissible (for more details see Ferguson [1967]). For a broader introduction to Bayesian methods see e.g. the books of Berger [1993], Ghosh and Ramamoorthi [2003], or Ghosal and van der Vaart [2017].

1.5 Bayesian asymptotics

In view of the ongoing debate between frequentist and Bayesian statisticians, it is of great interest to study the theoretical performance of Bayesian methods from the frequentist perspective. To assess the performance of Bayesian procedure we again take an asymptotic approach. A posterior distribution is called asymptotically consistent if most of its mass is concentrated in an arbitrary small neighbourhood U of the truth for large n, i.e.

$$\Pi_n(U \mid X^{(n)}) \xrightarrow{P_{\theta_0}} 1,$$

where the convergence is in probability.

A classical result about posterior consistency is the Doob's theorem (see Doob [1949]) stating that in the i.i.d. framework a Bayesian procedure with a prior Π is Π–almost surely consistent. However, in practical applications it is not known whether the specific true parameter θ_0 belongs to the prior–null set for a certain prior. A discussion and examples of inconsistency of posteriors can be found, for example, in Freedman [1963, 1965] and Diaconis and Freedman [1986a,b]. Among other things those results show that the null-set of inconsistency can be quite big.

Therefore, it is important to get sufficient conditions for consistency for a given parameter θ. Such a result is presented in Schwartz [1965]. The paper studies the case of statistically separable models for which there exists an exponentially powerful test for $\theta = \theta_0$ against the hypothesis $\theta \notin U$ for every neighbourhood U of the true parameter θ_0. The result shows that if in that case the prior gives positive mass to any small Kullback–Leibler neighbourhood of the true parameter, then the posterior distribution is consistent for this true parameter.

A more descriptive property of the posterior distribution is the rate of contraction

around the true parameter. The Bayesian procedure Π_n has a contraction rate ε_n if for $M > 0$ large enough

$$\Pi_n(\theta : d(\theta, \theta_0) \leq M\varepsilon_n \mid X^{(n)}) \xrightarrow{P_{\theta_0}} 1.$$

The contraction rate quantifies how quickly the mass of the posterior distribution concentrates around the true parameter θ_0. Results on the rate of convergence for posterior mean in parametric models can be found for example in Le Cam [1973] and Ibragimov and Has′minskiĭ [1981]. They show that under some regularity conditions, Bayesian procedures achieve the optimal rate of convergence $1/\sqrt{n}$ in that case.

However, matters are more complicated for nonparametric models. Recent results on the convergence rates of Bayesian procedures include Ghosal et al. [2000], Ghosal and van der Vaart [2007], and Shen and Wasserman [2001]. We discuss two theorems from Ghosal et al. [2000] and Ghosal and van der Vaart [2007] that allow to determine the rates of convergence based on information about the concentration of the prior and the complexity of the parameter space.

First, consider the case of independent and identically distributed observations. Let $X_1, \ldots, X_n$ be distributed according to some distribution P_{θ_0} with density p_{θ_0} with respect to some measure μ on the space $\mathscr{X}$. Let Π_n be a sequence of prior probability measures on the set Θ. Consider the Hellinger metric d on Θ given by

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

For two densities $f, g : \mathscr{X} \to [0, \infty)$ on a measurable space $(\mathscr{X}, \mu)$ define the Kullback–Leibler divergence to be

$$K(f, g) = \int f \log(f/g) d\mu.$$

Additionally, let

$$V_k(f, g) = \int f |\log(f/g)|^k d\mu, \qquad V_{k,0}(f, g) = \int f |\log(f/g) - K(f, g)|^k d\mu.$$

Consider the following neighbourhoods of the true parameter

$$B(\theta_0, \varepsilon) = \left\{\theta \in \Theta : K(p_{\theta_0}, p_\theta) \leq \varepsilon_n^2, V_2(p_{\theta_0}, p_\theta) \leq \varepsilon_n^2\right\}.$$

For a set $\Theta_0 \subset \Theta$ define $N(\varepsilon, \Theta_0, d)$ to be the covering number, which is the minimal number of balls of radius ε required to cover the set Θ_0 with respect to the metric d. By Ghosal et al. [2000] the following theorem holds.

Theorem 1.5.1. *Suppose for a sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \to \infty$, a constant*

$C > 0$, and sets $\Theta_n \subset \Theta$ the following conditions hold

$$\log N(\varepsilon_n, \Theta_n, d) \leq n\varepsilon_n^2; \tag{1.2}$$

$$\Pi_n(\Theta \setminus \Theta_n) \leq e^{-n\varepsilon_n^2(C+4)}; \tag{1.3}$$

$$\Pi_n(B(\theta_0, \varepsilon_n)) \geq e^{-Cn\varepsilon_n^2}. \tag{1.4}$$

Then $\Pi_n(\theta : d(\theta, \theta_0) \geq M\varepsilon_n \,|\, X_1, \ldots, X_n) \xrightarrow{P_{\theta_0}} 0$ for some $M > 0$, as $n \to \infty$.

The remaining mass condition (1.3) can be understood as expressing that Θ_n is almost as big as the support of the prior. The first and the third conditions of the theorem are the essential ones. Condition (1.2) ensures that the submodel Θ_n is not too complex. The prior mass condition (1.4) requires the prior distribution to put a sufficient amount of mass in a small neighbourhood of the true parameter.

One of the generalisations of this theorem is studied in Ghosal and van der Vaart [2007], where the result is stated for the case of data being sampled independently, without the additional assumption of being identically distributed. Consider the observation vector $X = (X_1, \ldots, X_n)$ of independent observations X_i. In this case we take the measures P_θ to be equal to the product measures $\otimes_{i=1}^n P_{\theta,i}$ on a product space $\otimes_{i=1}^n (\mathscr{X}_i, \mathcal{A}_i)$. We assume that the distribution $P_{\theta,i}$ of the ith component X_i has a density $p_{\theta,i}$ relative to a σ-finite measure μ_i on $(\mathscr{X}_i, \mathcal{A}_i)$, $i = 1, \ldots, n$. Consider the average Hellinger distance d_n given by

$$d_n^2(\theta, \theta') = \frac{1}{n} \sum_{i=1}^n \int \left(\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}}\right)^2 d\mu_i.$$

Additionally, consider a neighbourhood of the true parameter θ_0 defined as follows

$$B_n(\theta_0, \varepsilon, k) = \left\{ \theta \in \Theta : \frac{1}{n} \sum_{i=1}^n K(p_{\theta_0,i}, p_{\theta,i}) \leq \varepsilon^2, \frac{1}{n} \sum_{i=1}^n V_{k,0}(p_{\theta_0,i}, p_{\theta,i}) \leq C_k \varepsilon^k \right\}.$$

The following theorem of Ghosal and van der Vaart [2007] is an extension of Theorem 1.5.1.

Theorem 1.5.2. *Let P_θ be the product measures and d_n be the semi-metric defined above. Suppose that for a sequence $\varepsilon_n \to 0$ such that $n\varepsilon_n^2 \to \infty$, some $k > 1$, all sufficiently large j and sets $\Theta_n \subset \Theta$, the following conditions hold*

$$\log N(\varepsilon_n/36, \Theta_n, d_n) \leq n\varepsilon_n^2, \tag{1.5}$$

$$\frac{\Pi(\Theta \setminus \Theta_n)}{\Pi(B_n(\theta_0, \varepsilon_n, k))} = o\left(e^{-2n\varepsilon_n^2}\right), \tag{1.6}$$

$$\frac{\Pi(\theta \in \Theta_n : j\varepsilon_n < d_n(\theta, \theta_0) \leq 2j\varepsilon_n)}{\Pi(B_n(\theta_0, \varepsilon_n, k))} \leq e^{n\varepsilon_n^2 j^2/4}. \tag{1.7}$$

Then $\Pi(\theta : d_n(\theta, \theta_0) \geq M\varepsilon_n \,|\, X^{(n)}) \xrightarrow{P_{\theta_0}} 0$ for some $M > 0$, as $n \to \infty$.

It is known that on the basis of the posterior distribution one can construct a frequentist point estimator that has a convergence rate at least of the same order as the posterior contraction rate. For example, according to Belitser and Ghosal [2003] an estimator equal to the centre of the smallest ball with posterior mass at least 3/4 satisfies the requirement. Moreover, for a bounded metric d with a convex square the posterior mean typically attains the convergence rate as well (see Ghosal et al. [2000]). Hence, using Theorems 1.5.1, 1.5.2 one can establish whether the posterior distribution achieves the optimal contraction rate around the true parameter. Results on this topic can be found, for instance, in Ghosal and van der Vaart [2017].

1.6 Adaptation

In Section 1.3 we encountered a minimax estimator that is constructed based on the knowledge of the regularity level of the target function. Such estimators are not particularly useful for practical applications, since often the regularity of the true function is not known in advance. Hence, it is desirable to develop procedures that are more flexible and that can attain the minimax rates across a wide range of regularity of parameters.

As we have seen, nonparametric classes are commonly characterised by a few hyperparameters that quantify different properties of the underlying function, such as its level of smoothness. In Theorem 1.3.1, as in many other cases, the minimax estimator relies on the knowledge of the value of such hyperparameters. For practical applications one would like to have an adaptive estimator that can attain minimax rates for a broad collection of the values of the hyperparameters. Such procedures have been thoroughly studied in the frequentist setting, see for example Bickel [1982], Efromovich and Pinsker [1996], Lepski and Spokoiny [1997].

The development of Bayesian procedures that are adaptive has only started relatively recently. To illustrate one of the ideas behind the construction of adaptive Bayesian estimators we present a random scaled squared exponential process prior studied, for example, in van der Vaart and van Zanten [2009]. The procedure employs a scale of priors indexed by a bandwidth parameter and adapts by making a data-dependent choice of the bandwidth. The squared exponential process is the centred Gaussian process $W = \{W_t, t \in \mathbb{R}\}$ with covariance function

$$\mathbb{E}W_s W_t = e^{-|t-s|^2}.$$

This process is known to have a version with infinitely smooth sample paths. As a prior distribution for a function $\theta \in C[0,1]$ we consider the law of the process

$$\{W_{At}, t \in [0,1]\},$$

where A is an independent random variable with a Gamma distribution. The inverse $1/A$ of the variable A can be viewed as a bandwidth or length scale parameter.

The random variable A plays the role of a scaling parameter that makes the prior suitable for less regular functions, when A is large enough. The paper of van der Vaart and van Zanten [2009] shows that this Bayesian procedure attains optimal convergence rates (up to a log factor) in various settings for the parameter class $C^\beta(Q)$ with $Q, \beta > 0$. Observe that the prior itself does not depend on β which makes the procedure fully rate-adaptive.

In general, adaptive Bayesian methods automatically choose the setting of hyperparameters by using either hierarchical or empirical Bayesian techniques. In this thesis we focus on hierarchical Bayesian methods that endow the hyperparameters with hyperpriors, making the Bayesian procedure multilevel. Examples of such procedures can be found in Belitser and Ghosal [2003], where an adaptive method is developed in the context of Gaussian white noise model. de Jonge and van Zanten [2010, 2012] studies adaptive procedure for the nonparametric regression problem. Full Bayesian methods in the context of density estimation are discussed, for example, in Ghosal et al. [2008], Lember and van der Vaart [2007], Kruijer et al. [2010], Rousseau [2010], and van der Vaart and van Zanten [2008a].

In this thesis we devise adaptive Bayesian procedures for two different nonparametric settings: function estimation on graphs and Poisson intensity estimation. We present hierarchical Bayesian procedures and study their asymptotic performance, showing that they attain (almost) optimal rates of convergence.