Oracle Posterior Rates in the White Noise Model

Oracle posterior rates in the White noise model
Aleksandra Babenko
Ph.D. Thesis Utrecht University, December 2010
ISBN 978-90-393-5437-7
Printed by Ridderprint BV, www.ridderprint.nl

Oracle Posterior Rates in the White Noise Model

Orakel-a-posteriorisnelheden
in het witteruismodel

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht op gezag van de rector magnificus,
prof. dr. J. C. Stoof, ingevolge het besluit van het college
voor promoties in het openbaar te verdedigen op maandag
6 december 2010 des ochtends te 10.30 uur

door

Aleksandra Aleksandrovna Babenko

geboren op 14 november 1983 te Rudny, Kazakhstan

Promotor:	Prof. dr. R. Fernandez
Co-promotor:	Dr. E. N. Belitser

This research was financially supported by the Netherlands Organisation for Scientific Research (NWO), project "Bayesian Adaptation in Nonparametric Models".

Contents

Chapter 1

Introduction

1.1 Nonparametric Bayes

The general statistical problem is, broadly speaking, to make some inference about the underlying probability measure, using the observed data coming from that probability measure. We observe a random element $X^{(\epsilon)}$ coming from some unknown probability distribution $P^{(\epsilon)}$, where $P^{(\epsilon)} \in \mathcal{P}^{(\epsilon)}$ and the class $\mathcal{P}^{(\epsilon)}$ of possible distributions is known:

$$X^{(\epsilon)} \sim P^{(\epsilon)} \in \mathcal{P}^{(\epsilon)}.$$

The parameter $\epsilon > 0$ is assumed to be known and reflects in some sense the "amount of the information" in the model $P_\theta^{(\epsilon)}$. It can be, for instance, the variance of an additive noise, or $\epsilon = n^{-1/2}$ with n as the number of observations in the sample. In an asymptotic setup one usually considers $\epsilon \to 0$ and tries to qualify the behavior of the statistical procedure in terms of parameter ϵ. We often drop the dependence on ϵ, for instance $X \sim P \in \mathcal{P}$, unless we want to emphasize this dependence for some reason.

The assumption that the class $\mathcal{P}$ is known results in the statistical model. If the class of distributions can be parameterized as

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

where Θ is a subset of a finite dimensional linear space, then the inference problem and the model are said to be *parametric*. The main drawback of a parametric modeling is the rigid restriction on the shape of the probability distribution, while the data might be of such a type that there is no suitable parametric family that gives a good fit.

In this thesis, we are concerned with *nonparametric statistical models*. Loosely speaking, nonparametric models are those that are characterized only by a qualitative description of the class of possible underlying probability measures. A way to describe a nonparametric model is to assume that the class of possible underlying probability measures $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, where Θ is a subset of an

infinite dimensional space. The statistical problem boils thus down to making an inference on the parameter θ. In the last few decades, the nonparametric models became the object of intensive investigation.

There are two main paradigms in statistical inference: frequentist and Bayes. In this thesis, we are not going to discuss their merits and pitfalls, but rather focus on the interplay between the both approaches. Namely, the basic idea which runs through this thesis is to apply Bayesian procedures to nonparametric models (in this thesis, the Gaussian white noise model) and study the performance of these procedures from the frequentist point of view, reconciling the both paradigms in some sense. First, we pretend that the unknown infinite dimensional parameter θ is a random element, $\theta \sim \pi$, with a prior distribution π, and the model P_θ is regarded as conditional distribution of the date X given θ. This leads to the posterior distribution $P(\theta|X)$, which depend on the data:

$$\begin{array}{c} X|\theta \sim P_\theta \\ \theta \sim \pi \end{array} \quad \Longrightarrow \quad P(\theta|X).$$

Next, we look at the posterior quantities (e.g. posterior mean, variance, median; posterior distribution itself) from the point of view of the measure P_{θ_0} for some "true" parameter values $\theta_0 \in \Theta$. This means that we, in fact, have taken a frequentist approach. The study of the asymptotic behavior of Bayesian procedures from the Bayes perspective (i.e. from the point of view of marginal probability measure of data X) is of interest on its own, especially in an adaptive context; see Belitser and Levit (2003).

An incomplete list of relevant references on nonparametric Bayesian inference includes Bayes (1763), Doob (1948), Schwartz (1965), Berger (1985), Diaconis and Freedman (1986), Cox (1993), Freedman (1999), Barron, Shervish and Wasserman (1999), Ghosal, Ghosh and van der Vaart (2000), Shen and Wasserman (2001), Walker and Hjort (2001), Ghosh and Ramamoorthi (2003), Hjort (2003), Walker (2004), Ghosal and van der Vaart (2010)

There is a problem here, namely, the choice of priors in nonparametric models is highly nontrivial; see Chapter 2 in Ghosal and van der Vaart (2010). In principle, we are not restricted in choice, we can take any prior we want. But, as it is well illustrated in a series of papers by Diaconis and Freedman, the Bayes approach may easily fail in infinite dimensional problems because of a wrong prior. We expect that the Bayesian strategy with the "right" prior should lead to a good asymptotic (and nonasymptotic, but this is even a harder problem) properties of resulting adaptation procedure. Generally, in order to be "right", the prior π should reflect adequately the requirement $\theta \in \Theta$. There are no general methods of choosing right priors. Belitser and Enikeeva (2008) proposed some heuristic guiding idea how to check whether a certain prior is not unreasonable, if only to make sure that this prior is potentially right. The underlying inference problem on θ should also be taken into account. Usually the underlying inference problems are associated with the appropriate performance criteria, e.g. the rate of convergence in the estimation problem. A particular prior leads to the corresponding Bayes procedure. We can look at its performance, according to the given criteria, from the two different perspectives: frequentist ($X \sim P_\theta$)

and Bayesian ($X \sim P_X$, marginal of X). So, a prior is not unreasonable (and potentially right) if it provides the same high performance, with respect to the given criteria, of the resulting Bayes procedure simultaneously under the both Bayesian and frequentist formulations. In other words, the Bayesian criterion of the Bayes solution to the specific statistical problem should match the corresponding frequentist criterion of the same Bayes solution to the same statistical problem. For instance, in the case of an estimation problem with convergence rate of the risk as performance criterion, if the prior is to be potentially right, both the (frequentist) risk of the corresponding resulting Bayes estimator and its Bayesian risk should have the same optimal convergence rate. This idea is proved to be working in Belitser and Ghosal (2003), Belitser and Enikeeva (2008), Babenko and Belitser (2009). The above approach is not to be taken as a precise prescription, but rather as a starting point in search of right priors in infinite dimensional statistical problems. After all, one will have to investigate the performance of the resulting Bayesian procedure in each particular statistical problem in order to claim that a certain prior is right for that problem. The choice of the prior should in general depend on the underlying inference problem on θ. For example, the right priors for the problems of signal estimation and estimation of a functional of the signal are different.

The following example demonstrates the above discussion.

Example 1. Consider a *Gaussian sequence model*, sometimes called a *white noise model* in the sequence form; more discussion about this model is below.

$$X_i = \theta_i + \epsilon\xi_i, \quad i = 1, 2, \ldots,$$

where $\theta = (\theta_i)_{i\in\mathbb{N}} \in \Theta_\beta \subset \ell_2$ is an unknown parameter of interest,

$$\Theta_\beta = \Theta_\beta(Q) = \big\{\theta \in \ell_2 : \sum_{i=1}^{\infty} i^{2\beta}\theta_i^2 \le Q\big\}$$

is the Sobolev ellipsoid of smoothness $\beta > 0$, the noise variables ξ_i are independent, identically distributed $\mathcal{N}(0,1)$ random variables, the parameter ϵ is a known noise level.

Consider the problem of estimating θ using the observations $X^{(\epsilon)} = X = (X_1, X_2, \ldots)$. The quality of an estimator $\widehat{\theta} = \widehat{\theta}(X)$ is measured by the ℓ_2-norm risk

$$R_\epsilon(\widehat{\theta}, \theta) = E_\theta\|\widehat{\theta} - \theta\|^2.$$

Suppose we apply a Bayesian approach to the problem of estimating θ. We take a prior π_β described as follows:

$$\theta_i \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2(\beta)), \quad \tau_i^2(\beta) = i^{-(2\beta+1)}, \quad i = 1, 2, \ldots.$$

We now illustrate that this prior a right one in a certain sense for the above model and inference problem. It is not difficult to derive the Bayes estimator

$$\widehat{\theta} = \widehat{\theta}(\beta) = \{\widehat{\theta}_i(\beta),\, i = 1, 2, \ldots\}, \quad \widehat{\theta}_i(\beta) = \frac{i^{-(2\beta+1)}X_i}{\epsilon^2 + i^{-(2\beta+1)}}.$$

The performance of this Bayesian procedure from the frequentist point of view is the maximal risk

$$\sup_{\theta\in\Theta_\beta} R_\epsilon(\widehat{\theta},\theta) = O\big(\epsilon^{4\beta/(2\beta+1)}\big), \quad \text{as} \quad \epsilon\to 0,$$

as is shown by Belitser and Enikeeva (2008) (with $\epsilon = n^{-1/2}$). It is also well known (Pinsker (1980), see also Belitser and Levit (1995)), that the minimax risk (we discuss the notion of minimaxity in the next section) for this estimation problem is also of order $\epsilon^{4\beta/(2\beta+1)}$. The performance from the Bayesian point of view is the Bayes risk

$$R_{\pi_\beta}(\widehat{\theta},\theta) = E_{\pi_\beta} R_\epsilon(\widehat{\theta},\theta) = O\big(\epsilon^{4\beta/(2\beta+1)}\big), \quad \text{as} \quad \epsilon\to 0.$$

We see that the both quantities are of the same order, which confirms the heuristic idea that our prior π_β correctly models the condition $\theta\in\Theta_\beta$ for this estimation problem.

Suppose we want to estimate a linear functional $\Phi = \Phi(\theta) = \sum_{i=1}^\infty b_i\theta_i$ for a $\theta\in\Theta_\beta\subset\ell_2$ with $\beta > 1/2$. For an estimator $\widehat{\Phi} = \widehat{\Phi}(X)$, the estimation quality is measured by the risk function

$$R_\epsilon(\widehat{\Phi},\Phi(\theta)) = E_\theta\big(\widehat{\Phi}-\Phi(\theta)\big)^2.$$

In this case, a correct prior λ_β turns out to be the same as π_β with the difference that now $\tau_i^2(\beta) = i^{-2\beta}$, $i = 1,2,\ldots$. Indeed, as is shown by Babenko and Belitser (2009),

$$\sup_{\theta\in\Theta_\beta} R_\epsilon(\widehat{\Phi},\Phi(\theta)) = O\big(\epsilon^{(2\beta-1)/\beta}\big),$$
$$R_\lambda(\widehat{\Phi},\Phi(\theta)) = E_{\lambda_\beta} R_\epsilon(\widehat{\Phi},\Phi(\theta)) = O\big(\epsilon^{(2\beta-1)/\beta}\big),$$

as $\epsilon\to 0$. It is also known that, under some assumptions on sequence b_i, the minimax risk in this problem is also of order $\epsilon^{(2\beta-1)/\beta}$ as $\epsilon\to 0$.

1.2 Optimality and adaptivity: minimax and oracle approach

Our prior knowledge about the model is formalized as follows: the unknown parameter (sometimes we call it signal or curve) $\theta\in\Theta_\beta$, $\beta\in\mathcal{B}$, where Θ_β is a subset of some linear space $\mathcal{T}$ equipped with a (semi-) metric $d(\cdot,\cdot)$. Parameter $\beta\in\mathcal{B}$ typically has a meaning of signal smoothness. If it is known that $\theta\in\Theta_\beta$ for a certain $\beta\in\mathcal{B}$ (single model situation), then apart from the model itself, the functional dependence P_θ on θ, which we assume to be fixed, only the complexity of the set Θ_β affects the quality of statistical inference on θ. At the moment we do not specify the statistical problem – it may be the problem of estimation of the signal or its functional, the problem of the construction of a confidence set or testing hypothesis problem.

To compare different statistical procedures, several approaches have been developed: *minimax* (Pinsker (1980), Ibragimov and Hasminskii (1981), Stone (1980), Gill and Levit (1995), Ingster and Suslina (2003)), *oracle inequalities* (Donoho and Johnstone (1994), Laurent and Massart (2000), Birge and Massart (2001), Cavalier and Tsybakov (2001), Cavalier, Golubev, Picard and Tsybakov (2002)), *maxisets* (Kerkyacharian and Picard (2002)). Combining these possibilities with different observation models (white noise model, density, regression), risk types (pointwise, L_2, L_p, sup-norm, 0-1 loss function etc.), functional classes (Sobolev, Hölder, Besov) has led to a variety of results.

From now on we consider the following inference problem: we want to estimate a parameter θ in a minimax setup, measuring the quality of an estimator $\widehat{\theta} = \widehat{\theta}(X)$ by a risk function

$$R(\widehat{\theta}, \theta) = R_\epsilon(\widehat{\theta}, \theta) = E_\theta d(\widehat{\theta}, \theta).$$

Then a benchmark in this statistical problem is the *minimax risk*

$$r(\Theta_\beta) = r_\epsilon(\Theta_\beta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_\beta} R_\epsilon(\widehat{\theta}, \theta),$$

where the infimum is taken over all possible estimators $\widehat{\theta} = \widehat{\theta}(X)$, measurable functions of the data X. Recall that the set Θ_β is also known. The minimax risk expresses the least possible mean loss when the worst case happens and, in a way, reflects the complexity of the estimation problem over the class Θ_β. The goal is typically to find a so called minimax estimator $\widehat{\theta}$, i.e. the one attaining the minimax risk:

$$\sup_{\theta \in \Theta_\beta} R_\epsilon(\widehat{\theta}, \theta) \le C_\epsilon r_\epsilon(\Theta_\beta),$$

for some bounded $C_\epsilon \ge 1$ and the above inequality becomes stronger as C_ϵ gets closer to 1 (asymptotically as $\epsilon \to 0$ or uniformly in $\epsilon > 0$). For example, in case of the white noise model, a Sobolev ellipsoid Θ_β and the ℓ_2-norm risk, this problem was solved by Pinsker (1980), with $C_\epsilon \to 1$ as $\epsilon \to 0$ and the exact description of the asymptotic behavior of the minimax risk.

Suppose the smoothness parameter β is not known, that is we are given a family of models $\{\Theta_\beta, \beta \in \mathcal{B}\}$ (we call such a family of nonparametric classes by functional *scale*) and we only know that $\theta \in \Theta_\beta$ for some $\beta \in \mathcal{B}$. In fact, we assume $\theta \in \cup_{\beta \in \mathcal{B}} \Theta_\beta$, then the problem becomes in general more difficult and now the complexities of all Θ_β, $\beta \in \mathcal{B}$, and the complexity (e.g. cardinality) of set $\mathcal{B}$ matter. In this case one speaks of *adaptation* problem. Depending on the statistical problem at hand (for instance, signal estimation, functional estimation, testing hypothesis, confidence set), the general problem of adaptation is, loosely formulated, to design a method for solving this statistical problem which performs in a multiple model situation as good as in a single model, or, if this is not possible, with the smallest loss of quality.

For example, in the problem of *adaptive estimation*, it is assumed that the unknown parameter θ belongs to a *smoothness class* Θ_β (Sobolev, Hölder, Besov) where β is unknown and has the meaning of the smoothness of signal

θ. The goal is to find an estimator $\hat{\theta}$, independent of β, attaining the optimal behavior over the whole family $\{\Theta_\beta,\, \beta \in \mathcal{B}\}$.

Thus the problem of adaptation is to construct an adaptive minimax estimator $\widehat{\theta} = \widehat{\theta}(X)$ (i.e. without the knowledge of smoothness parameter β), if it exists:

$$\sup_{\theta \in \Theta_\beta} R_\epsilon(\widehat{\theta}, \theta) \le C_\epsilon r_\epsilon(\Theta_\beta), \tag{1.1}$$

for all $\beta \in \mathcal{B}$ and for some bounded $C_\epsilon \ge 1$. Again, the above inequality becomes stronger as C_ϵ gets closer to 1. The adaptation problem was first studied in Efromovich and Pinsker (1984) for the white noise model and by many authors afterwards.

In the last two decades, several adaptive estimation methods (optimal in one or other sense) have been developed: blockwise method (see e.g. Efromovich and Pinsker (1984), Cavalier and Tsybakov (2001), Cavalier, Golubev, Picard and Tsybakov (2002)), Lepski's method Lepski (1990), Lepski (1991), Lepski (1992), Lepski and Spokoiny (1997)), wavelet thresholding (Donoho and Johnstone (1994), later developed in many other papers), penalized estimators (Barron, Birge and Massart (1999) and further references therein, Laurent and Massart (2000), Birge and Massart (2001)). Some methods are designed for rather specific settings: e.g. blockwise method for the white noise sequence model with the mean squared risk. Some are more general, e.g. Lepski's method, which could be extended to different settings (various risk functions, multidimensional case) and even to different statistical problems: estimation of a functional of the signal, adaptive testing hypothesis problem (see Ingster and Suslina (2003) and further references therein).

There is a way to look at the adaptation problem from another perspective. Namely, a framework of the so called oracle inequalities has recently been developed. Suppose we are given a family of estimators $\widehat{\Theta} = \widehat{\Theta}(\mathcal{A}) = \{\widehat{\theta}(\alpha), \alpha \in \mathcal{A}\}$. Then, for an unknown true value θ, the best performance we can achieve is the *oracle risk*

$$r_\epsilon(\theta) = r_\epsilon(\widehat{\Theta}, \theta) = \inf_{\widehat{\theta} \in \widehat{\Theta}} R_\epsilon(\widehat{\theta}, \theta) = \inf_{\alpha \in \mathcal{A}} R_\epsilon(\widehat{\theta}(\alpha), \theta) = R_\epsilon(\widehat{\theta}(\alpha_o), \theta), \tag{1.2}$$

with possibly a sequence $\alpha_o^{(k)}$ (instead of α_o) for which the above infimum is attained. The optimal index α_o (or a sequence $\alpha_o^{(k)}$) is called *oracle* and the corresponding estimator $\widehat{\theta}(\alpha_o)$ is an *oracle estimator*, these are unobservable quantities.

Remark 1.1. Here and later we slightly abuse the notations: we use the same letter r to denote the minimax risk and the oracle risk. Later we will also use it for the posterior rate $r_\epsilon(\pi, \theta)$ with respect to a prior π and the posterior oracle rate $r_\epsilon(\Pi, \theta)$ corresponding to a family of priors Π. However it is always clear which of these is meant by looking at the arguments of the function r, for example, the minimax risk $r_\epsilon(\Theta_\beta)$ has always a nonparametric set as argument and the oracle risk $r_\epsilon(\widehat{\Theta}, \theta)$ is always a function of θ and a family of estimators $\widehat{\Theta}$.

The oracle risk is our new benchmark in this approach and the goal is to construct an estimator $\widehat{\theta} = \widehat{\theta}(\widehat{\alpha})$ for some $\widehat{\alpha} = \widehat{\alpha}(X)$ such that for some positive C_ϵ, the following *oracle inequality* is satisfied:

$$R_\epsilon(\widehat{\theta}(\widehat{\alpha}), \theta) \le C_\epsilon r_\epsilon(\widehat{\Theta}, \theta) \tag{1.3}$$

for every $\theta \in \Theta_0$, where the oracle risk $r_\epsilon(\widehat{\Theta}, \theta)$ is defined by (1.2). Certainly, $C_\epsilon \ge 1$ and the above oracle inequality becomes stronger as C_ϵ gets closer to 1 and the set Θ_0 gets bigger. Notice that $\widehat{\alpha}$ can be regarded as some kind of inference on the oracle α_o, this issue is discussed below. Actually, if one allows arbitrary estimators $\widehat{\theta}$, then C_ϵ may be smaller than 1. In this case, however, there is no direct inference on the oracle α_o. When properly motivated, one can study other more complicated forms of oracle inequalities, for example, for every $\theta \in \Theta_0$,

$$R_\epsilon(\widehat{\theta}(\widehat{\alpha}), \theta) \le C_\epsilon \inf_{\alpha \in \mathcal{A}} \Big\{ R_\epsilon(\widehat{\theta}(\alpha), \theta) + P_\epsilon(\alpha) \Big\} + \bar{P}_\epsilon$$

with some positive penalty terms $P_\epsilon(\alpha)$ and $\bar{P}_\epsilon$; see for example Cavalier and Tsybakov (2001), Cavalier, Golubev, Picard, and Tsybakov (2002), Golubev and Levit (2004).

In this thesis, we however stick to the basic oracle inequality of form (1.3) with $C_\epsilon \le C$, some absolute constant. Sometimes it makes sense to relax the inequality (1.3) a little by adding a penalty term $\bar{P}_\epsilon$ in the right hand side of the inequality, so that this resulting relaxed oracle inequality holds for some estimator $\widehat{\theta}(\widehat{\alpha})$ and for all $\theta \in \Theta_0$ with a much bigger set Θ_0. This is especially advantageous if the set Θ_0 becomes the whole underlying space, whereas the additional penalty term $\bar{P}_\epsilon$ is of a smaller (or the same) order, as $\epsilon \to 0$, than the oracle risk term $C_\epsilon r_\epsilon(\widehat{\Theta}, \theta)$. Commonly (and this is the case in this thesis), the additional penalty term $\bar{P}_\epsilon$ is of the parametric rate order, while the risk rate in typical nonparametric models is never better than the parametric rate.

Now the question arises how these two adaptation optimality frameworks, minimax over the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$ and oracle over the estimators family $\widehat{\Theta} = \widehat{\Theta}(\mathcal{A}) = \{\widehat{\theta}(\alpha), \alpha \in \mathcal{A}\}$, are related to each other. An important particular case is $\mathcal{A} = \mathcal{B}$, but in general the index sets $\mathcal{A}$ and $\mathcal{B}$ can be completely different. Suppose the family $\widehat{\Theta}(\mathcal{A})$ is sufficiently rich to contain the minimax estimators over all Θ_β, $\beta \in \mathcal{B}$, i.e. for any $\beta \in \mathcal{B}$, there exists an $\alpha = \alpha(\beta) \in \mathcal{A}$ such that $\widehat{\theta}(\alpha(\beta))$ is minimax over the class Θ_β:

$$\sup_{\theta \in \Theta_\beta} R_\epsilon(\widehat{\theta}(\alpha(\beta)), \theta) \le C'_\epsilon r_\epsilon(\Theta_\beta),$$

ideally with $C'_\epsilon = 1 + o(1)$ as $\epsilon \to 0$, otherwise with some uniform constant $1 \le C'_n = C' < \infty$. In this situation, we say that the family of estimators $\widehat{\Theta}(\mathcal{A})$ *covers* the functional scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$.

Suppose that the family $\widehat{\Theta}(\mathcal{A})$ covers the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$ and an estimator $\widehat{\theta}(\widehat{\alpha})$ satisfies the oracle inequality (1.3). Then, uniformly in $\theta \in \Theta_0 \cap \Theta_\beta$,

$$R_\epsilon(\widehat{\theta}(\widehat{\alpha}), \theta) \le C_\epsilon \inf_{\alpha \in \mathcal{A}} R_\epsilon(\widehat{\theta}(\alpha), \theta) \le C_\epsilon R_\epsilon(\widehat{\theta}(\alpha(\beta)), \theta) \le C_\epsilon C'_\epsilon r_\epsilon(\Theta_\beta).$$

This means that if the true $\theta \in \Theta_0 \cap \big(\cup_{\beta\in\mathcal{B}} \Theta_\beta\big)$, the oracle approach for an appropriately chosen family of estimators $\widehat{\Theta}(\mathcal{A})$ (that covers the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$) is stronger than the minimax in the sense that the oracle risk at each point $\theta \in \Theta_0 \cap \big(\cup_{\beta\in\mathcal{B}} \Theta_\beta\big)$ can only be smaller than a multiple of the minimax risk $r_\epsilon(\Theta_\beta)$. If the oracle inequality (1.3) holds for a set Θ_0 which contains the whole scale $\cup_{\beta\in\mathcal{B}}\Theta_\beta$, then the estimator $\widehat{\theta}(\widehat{\alpha})$ is adaptive minimax with respect to the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$. Indeed, for all $\beta \in \mathcal{B}$,

$$\begin{aligned}\sup_{\theta\in\Theta_\beta} R_\epsilon(\widehat{\theta}(\widehat{\alpha}),\theta) &\le C_\epsilon \sup_{\theta\in\Theta_\beta} \inf_{\alpha\in\mathcal{A}} R_\epsilon(\widehat{\theta}(\alpha),\theta)\\ &\le C_\epsilon \sup_{\theta\in\Theta_\beta} R_\epsilon(\widehat{\theta}(\alpha(\beta)),\theta) \le C_\epsilon C'_\epsilon r_\epsilon(\Theta_\beta).\end{aligned}$$

Clearly, to make the notion of oracle sensible, the family $\widehat{\Theta}(\mathcal{A})$ should neither be too poor, nor too rich. This trade-off is also discussed in Golubev (2004). Indeed, on the one hand, we want this family to contain some good estimators about which we know that they perform well over nonparametric classes Θ_β, $\beta \in \mathcal{B}$; for example, minimax estimators over Θ_β, $\beta \in \mathcal{B}$. On the other hand, if the family $\widehat{\Theta}(\mathcal{A})$ is too rich, then it may not be possible to find an estimator $\widehat{\theta}$ satisfying the relation (1.3) for a reasonable Θ_0. Instead, the relation (1.3) may hold only for a "thin" set Θ_0, while we would certainly like Θ_0 to be as big as possible, ideally containing all Θ_β, $\beta \in \mathcal{B}$.

To summarize, an oracle result of type (1.3) implies adaptive minimaxity results of type (1.1) over all functional scales which are contained in the set Θ_0 and covered by the family of estimators $\widehat{\Theta}(\mathcal{A})$. Thus the oracle approach is stronger than the adaptive minimax if the family of estimators $\widehat{\Theta}(\mathcal{A})$ is chosen appropriately. An oracle approach to optimality of estimators was probably first studied by Kneip (1994), although without referring to it by the term *oracle* at the time, within the estimators class of ordered linear smoothers and then developed in the series of works by Donoho and Johnstone. An oracle approach to the conditional density estimation is recently proposed by Efromovich (2010).

To illustrate the introduced notions, let us give an example.

Example 2. Consider a Gaussian sequence model as in Example 1: $X = X^{(\epsilon)} \sim P_\theta^{(\epsilon)}$, again with θ from the Sobolev ellipsoid $\Theta_\beta = \Theta_\beta(Q) \subseteq \ell_2$. We want to estimate θ using the observations X. The quality of an estimator $\widehat{\theta} = \widehat{\theta}(X)$ is measured by the ℓ_2-norm risk $R_\epsilon(\widehat{\theta},\theta) = E_\theta\|\widehat{\theta} - \theta\|^2$. The exact asymptotic behavior of the minimax risk is derived by Pinsker (1980) (see also Belitser and Levit (1995) for the second order asymptotic behavior):

$$r_\epsilon(\Theta_\beta) = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta_\beta} R_\epsilon(\widehat{\theta},\theta) = \epsilon^{4\beta/(2\beta+1)} P(\beta,Q)(1+o(1)), \quad \text{as} \quad \epsilon \to 0,$$

where the so called Pinsker constant

$$P(\beta,Q) = \big(Q(2\beta+1)\big)^{1/(2\beta+1)} \big(\beta/(\beta+1)\big)^{2\beta/(2\beta+1)}.$$

Consider a family of estimators $\widehat{\Theta}_1(\mathcal{N}) = \{\widehat{\theta}(N),\, N \in \mathcal{N}\}$: $\mathcal{N} = \{1, 2, \dots, \}$,

$$\widehat{\theta}(N) = (\widehat{\theta}_1(N), \widehat{\theta}_2(N), \dots), \quad \widehat{\theta}_i(N) = X_i I\{i \le N\}.$$

Note that the above family consists of the projection estimators parameterized by the cut-off parameter N. One more family of estimators $\widetilde{\Theta}_2(\mathcal{B}) = \{\widetilde{\theta}(\beta),\, \beta \in \mathcal{B}\}$: $\mathcal{B} = (0, +\infty)$,

$$\widetilde{\theta}(\beta) = (\widetilde{\theta}_1(\beta), \widetilde{\theta}_2(\beta), \dots), \quad \widetilde{\theta}_i(\beta) = \frac{i^{-(2\beta+1)} X_i}{\epsilon^2 + i^{-(2\beta+1)}}.$$

We note here that this family consists of the Bayes estimators corresponding to the family of priors $\{\pi_\beta,\, \beta \in \mathcal{B}\}$, with π_β described as follows: $\theta_i \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2(\beta))$, $i = 1, 2, \dots$, with $\tau_i^2(\beta) = i^{-(2\beta+1)}$.

The families of estimators $\widehat{\Theta}$ and $\widetilde{\Theta}$ both cover the scale $\{\Theta_\beta,\, \beta \in \mathcal{B}\}$. Indeed, for any $\beta \in \mathcal{B}$ take $N_\beta = \lfloor c\epsilon^{-2/(2\beta+1)} \rfloor \in \mathcal{N}$, then

$$\begin{aligned}
\sup_{\theta \in \Theta_\beta} R_\epsilon(\widehat{\theta}(N_\beta), \theta) &= \sup_{\theta \in \Theta_\beta} \Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^{\infty} \theta_i^2 \Big) \\
&\le \epsilon^2 N_\beta + \sup_{\theta \in \Theta_\beta} \Big(\sum_{i=N_\beta+1}^{\infty} \frac{\theta_i^2 i^{2\beta}}{N_\beta^{2\beta}} \Big) \\
&\le \epsilon^2 N_\beta + \frac{Q}{N^{2\beta}} = C\epsilon^{4\beta/(2\beta+1)}
\end{aligned}$$

and, as is shown in Belitser and Enikeeva (2008) with $\epsilon = n^{-1/2}$,

$$\sup_{\theta \in \Theta_\beta} R_n(\widetilde{\theta}(\beta), \theta) \le C\epsilon^{4\beta/(2\beta+1)}.$$

This implies that the oracle risks over the both families is of estimators can only be better than the minimax risk at each point $\theta \in \Theta_\beta$:

$$\begin{aligned}
r_\epsilon(\widehat{\Theta}, \theta) &= \inf_{N \in \mathcal{N}} R_\epsilon(\widehat{\theta}(N), \theta) \le \inf_{N \in \mathcal{N}} \sup_{\theta \in \Theta_\beta} R_\epsilon(\widehat{\theta}(N), \theta) \le C r_\epsilon(\Theta_\beta), \\
r_\epsilon(\widetilde{\Theta}, \theta) &= \inf_{\beta \in \mathcal{B}} R_\epsilon(\widetilde{\theta}(\beta), \theta) \le \inf_{\beta \in \mathcal{B}} \sup_{\theta \in \Theta_\beta} R_\epsilon(\widetilde{\theta}(\beta), \theta) \le C r_\epsilon(\Theta_\beta).
\end{aligned}$$

Thus, an oracle result of type (1.3) for an estimator $\widehat{\theta}$ and the family $\widehat{\Theta}$ with a set Θ_0 would imply that this estimator can only have a better (oracle risk) convergence rate of than the minimax risk rate for all $\theta \in \Theta_0 \cap \Theta_\beta$. The same applies to the family $\widetilde{\Theta}$.

1.3 Bayesian adaptation and model selection

Now consider a Bayesian approach to the adaptation problem. Suppose the smoothness parameter β is not known, that is we are given a family of models

$\{\Theta_\beta, \beta \in \mathcal{B}\}$ and we only know that $\theta \in \Theta_\beta$ for some $\beta \in \mathcal{B}$. One can apply two Bayesian adaptation approaches to nonparametric models: *pure Bayes* and *empirical Bayes* approaches. In both cases we start by putting a prior π_β on $\theta \in \Theta_\beta$, $\beta \in \mathcal{B}$, where index β labels the model $\{P_\theta, \theta \in \Theta_\beta\}$.

In the classical empirical Bayes approach (due to Robbins (1955)), the unknown β in prior π_β is not random. Suppose that in single model situation, i.e. $\theta \in \Theta_{\beta_0}$ for some known $\beta_0 \in \mathcal{B}$, we would know how to resolve our original statistical problem. For instance, some standard (optimal in some sense) Bayesian method is available in this case. Then the idea is make a data-dependent choice of the model $\hat{\beta} \in \mathcal{B}$ and once this choice is made, we can proceed in making the statistical inference using the selected model and our standard method as in the single model case. A common procedure is to estimate β by $\hat{\beta} = \hat{\beta}(X)$ by using the marginal likelihood P_β of X, for example, as a value which (nearly) maximizes the marginal likelihood of X. Interestingly, in the estimation problem there is a connection of the empirical Bayes approach with penalized estimators: the marginal likelihood can be related to the penalized criterion.

According to the (pure, sometimes called full) Bayesian approach, we regard π_β as the conditional distribution of θ given β, and β itself is distributed on $\mathcal{B}$ according to the prior λ. Thus, we design a two-level hierarchical prior π on the pair (θ, β): $\theta|\beta \sim \pi_\beta$, $\beta \sim \lambda$. Notice that the resulting hierarchical prior on θ is a mixture of the appropriate priors for different models indexed by the structural parameter $\beta \in \mathcal{B}$. This construction leads to the posterior distributions $P(\theta|X)$ and $P(\beta|X)$ and all the inference on θ and β is based on these two posterior distributions. Notice that the prior π can of course be designed in a different way without relating to any stochastic element β, but the above construction of hierarchical prior comes naturally from the Bayesian way of thinking and it has another appealing feature. Namely, we can make some statistical inference on β by looking at the resulting posterior $P(\beta|X)$, which can be regarded as Bayesian model selection. Of course, one has to be specific about how to measure the quality of such inference. In case of Sobolev classes in the white noise model, this problem has been studied in Belitser and Enikeeva (2008) in the context of testing hypothesis about the smoothness parameter, where also the uniformity issue has been further elaborated.

As we already mentioned above when describing two Bayesian approaches to adaptation, besides the original statistical inference problem, we can consider an attendant problem of a data-based choice $\hat{\beta} = \hat{\beta}(X)$ for structural parameter $\beta \in \mathcal{B}$ which marks the model. One can thus regard this attendant problem as the model selection problem. Notice that the both pure Bayes and empirical Bayes described above can be used to provide model selectors $\hat{\beta}$. A basic idea to design an adaptive procedure is then as follows. A good selection model method combined with corresponding good nonadaptive methods (which are proved to be good for single model situation) should lead in principle to good adaptation methods, simply by plugging the selected model in those nonadaptive methods. Clearly to be able to compare model selection methods and to warrant their quality, one has to relate the attendant model selection problem to the original statistical problem.

Another related aspect is the assessment of the performance of an adaptive procedure. The adaptive procedures may perform in the best way in some sense, but what is their actual performance? For example, in the problem of signal estimation, the optimal convergence rate is usually expressed in terms of unknown β. So, we know that our estimator attains the optimal rate but we do not know what it is. Adaptive estimators do not say anything about the actual accuracy of the estimation. The problem essentially remains even if we knew that $\theta \in \Theta_{\beta_0}$ for some known $\beta_0 \in \mathcal{B}$. Indeed, in this case we would be able to specify the optimal (global, typically minimax) rate in terms of the known parameter β_0. However, since the optimal rate is of a global nature, there is always a room for an improvement for the individual θ's. Namely, for certain $\theta \in \Theta_{\beta_0}$ an adaptive procedure can pick the model with smoothness β_1, leading to a better convergence rate. It is not clear how to characterize the convergence rate pertinent to the particular signal θ in this case.

The above discussion is closely related to the problem of construction of adaptive confidence sets. We need to infer on the smoothness of the signal, i.e. select the best model, to be able to construct a confidence region with smallest possible size. It is an important problem to elaborate on the optimality issue for the confidence sets. In fact, it is the combination of two factors, the asymptotic level and the asymptotic size of the corresponding confidence region, that is interesting to investigate. References on this topic are few in number: Low (1997), Picard and Tribouley (2000), Hoffmann and Lepski (2002), Robins and van der Vaart (2006); there is still a great deal to learn in this area.

When constructing the adaptive procedure for θ, some adaptation methods (Bayes methods, Lepski's method, penalized estimators) do provide some kind of estimator $\hat{\beta}$ for the smoothness β in passing as a result of selecting the model. The question is whether one can use these estimators $\hat{\beta}$, possibly modified in an appropriate way, in constructing adaptive confidence regions. One methodological advantage of the Bayes approach in this respect is that, no matter how complex the setting of the problem is, there is always a reasonable candidate for adaptive confidence set, namely, the credible set constructed by using the corresponding posterior distribution of θ. Another possibility is to use a (empirical) Bayes estimator for θ and a (empirical) Bayes estimator for model index β. The idea is to establish the tightness of the risk function normalized by random convergence rate (cf. random normalizing factors in Hoffmann and Lepski (2002)) where the estimator $\hat{\beta}$ is used instead of β.

1.4 Oracle posterior convergence rate

In this section we introduce the notions of posterior and posterior oracle rates. According to the Bayesian approach, the statistical model $X \sim P_\theta = P_\theta^{(\epsilon)}$ and a prior $\theta \sim \pi$ lead to the posterior distribution $P(\theta|X)$, the main quantity of interest for a Bayesian analysis. A Bayesian procedure is considered to have good asymptotic frequentist properties if the corresponding posterior distribution $P(\theta|X)$ concentrates around θ_0 as $\epsilon \to 0$ under the measure P_{θ_0}. To

characterize the quality of Bayesian procedures, we look at the rate at which the neighborhood of θ_0 may decrease, while still capturing the most of the posterior mass, from the point of view of the measure P_{θ_0}. To be more precise, a positive sequence r_ϵ is called the *posterior rate* if for any $M_\epsilon \to \infty$

$$P\{r_\epsilon^{-1} d(\theta, \theta_0) \geq M_\epsilon | X\} \to 0 \quad \text{as } \epsilon \to 0 \tag{1.4}$$

in P_{θ_0}-probability. We will also use the terms *posterior convergence rate* or *posterior concentration rate*. Note that the posterior convergence rate r_ϵ may be a global or local quantity, the latter in the sense that it may depend on θ_0.

Clearly, it is desirable that the relation (1.4) holds for a sequence r_ϵ converging to zero as fast as possible, for all $\theta_0 \in \Theta_0$ with set Θ_0 as big as possible and preferably uniformly over $\theta_0 \in \Theta_0$. If these requests can not be satisfied simultaneously, one should try to find a good balance among them. A non-asymptotic (and more informative) version of (1.4) is as follows

$$E_{\theta_0} P\{r_\epsilon^{-1} d(\theta, \theta_0) \geq M | X\} \leq \phi(M), \tag{1.5}$$

for some monotonically decreasing to zero function $\phi(M)$ as $M \to \infty$ and all $\theta_0 \in \Theta_0$.

For example, if we know that $\theta_0 \in \Theta_\beta$ for nonparametric class Θ_β with smoothness parameter $\beta \in \mathcal{B}$, then the typical benchmark for the posterior rate r_ϵ in (1.4) is the minimax risk $r_\epsilon(\Theta_\beta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_\beta} R(\widehat{\theta}, \theta)$. For a prior $\theta \sim \pi_\beta$, we want the resulting posterior to satisfy the relation (1.4) with the posterior rate equal to the minimax rate $r_\epsilon(\Theta_\beta)$, preferably uniformly over $\theta_0 \in \Theta_0 \supseteq \Theta_\beta$.

Recall the adaptive Bayesian approach which we discussed in the previous section: we are given a family of models $\{\Theta_\beta, \beta \in \mathcal{B}\}$ and we only know that $\theta \in \Theta_\beta$ for some $\beta \in \mathcal{B}$. We model this uncertainty by putting a prior on β. Namely, given a family of priors $\{\pi_\beta,\ \beta \in \mathcal{B}\}$ and a prior λ on $\beta \in \mathcal{B}$, we can design a two-level hierarchical prior π on the pair (θ, β): $\theta|\beta \sim \pi_\beta$, $\beta \sim \lambda$. Prior π leads thus to the posterior $P(\theta|X)$ and if the true $\theta_0 \in \Theta_\beta$, then we want the relation (1.4) to hold for the posterior rate equal to the minimax rate $r_\epsilon(\Theta_\beta)$, preferably uniformly over $\theta_0 \in \Theta_\beta$. In case of Sobolev classes Θ_β, $\beta \in \mathcal{B}$, this problem has been studied by Belitser and Ghosal (2003), where also the issue of uniformity was discussed.

The first fundamental result on posterior convergence rates is due to Ghosal, Ghosh and van der Vaart (2000), since then this has become a rapidly developing area in contemporary mathematical statistics. Van der Vaart and van Zanten (2008) studied the posterior convergence rate for density estimation using Gaussian process prior and they derived, with appropriately chosen Gaussian prior (for example, an appropriately fractionally integrated Brownian motion), the minimax rate $n^{-\alpha/(2\alpha+1)}$ (up to a log factor) for an α-smooth true density.

Now we are ready to introduce an oracle optimality framework for the Bayes approach. Suppose we are given the model $X \sim P_\theta$ and a family of priors $\Pi = \Pi(\Gamma) = \{\pi_\gamma,\ \gamma \in \Gamma\}$. This determines the corresponding family of the posterior concentration rates $\{r_\epsilon(\pi_\gamma, \theta),\ \gamma \in \Gamma\}$. For a fixed unknown "true" θ there is an *oracle prior* π_{γ_o}, pertinent to an *oracle* $\gamma_o = \gamma_o(\theta)$ (or a corresponding

oracle sequence $\gamma_o^{(k)}$), corresponding to the *oracle posterior rate*

$$r_\epsilon(\Pi, \theta) = \inf_{\gamma \in \Gamma} r_\epsilon(\pi_\gamma, \theta) = r_\epsilon(\pi_{\gamma_o}, \theta).$$

Again, we abuse slightly the notations keeping the same letter r for the posterior concentration rate and the oracle posterior rate. The goal is now to design an adaptive prior π (without knowledge of the oracle $\gamma_o(\theta)$) to mimic the performance of the oracle prior, i.e. the concentration rate of the resulting adaptive posterior should be within a constant factor of the oracle posterior rate. This means that the relation (1.4) holds with $r_\epsilon(\pi, \theta)$ instead of r_ϵ and

$$r_\epsilon(\pi, \theta) \leq C_\epsilon r_\epsilon(\Pi, \theta) = C_\epsilon \inf_{\gamma \in \Gamma} r_\epsilon(\pi_\gamma, \theta)$$

for all $\theta \in \Theta_0$, ideally with a set Θ_0 as big as possible and with some positive bounded $C_\epsilon \leq C < \infty$. In this case, we say that the adaptive prior π mimics the oracle π_{γ_o} over the family of priors $\Pi = \Pi(\Gamma) = \{\pi_\gamma,\, \gamma \in \Gamma\}$, for all $\theta \in \Theta_0$. The bigger the set Θ_0, the stronger this property is. A stronger property would also be to have a sufficiently large constant $M > 0$ in relation (1.4) instead of sequence $M_\epsilon \to \infty$ as $\epsilon \to 0$. An even more stronger property would be a non-asymptotic version of the form (1.5) with $r_\epsilon(\pi, \theta)$ instead of r_ϵ. Notice that oracle approach is intrinsically adaptive and local (the oracle rate depends on the true parameter value θ). One can use a fully Bayesian approach by putting a prior on γ to obtain the resulting hierarchical mixture prior, or an empirical Bayes approach, or a hybrid version of those; see the next chapter of this thesis and Babenko and Belitser (2010).

However, in general there is an issue of assessment of the quality of the oracle in this setting. The point is that the family of priors Π under consideration can be uninteresting in the sense of a poor frequentist behavior of the corresponding posteriors. Even if we establish that our adaptive posterior attains the oracle over a given family of priors, i.e. it has the best (oracle) rate $r_\epsilon(\Pi, \theta)$ among the family of posterior rates $\{r_\epsilon(\pi_\gamma, \theta),\, \gamma \in \Gamma\}$, we would also like to know how "good" this whole family is. This can be done by relating the resulting family of posterior rates to the risks of some families of estimators, which cover some interesting scales of nonparametric classes. In doing so, one solves the adaptation problem as well.

A way to relate a family of priors to a family of estimators is by comparing of the posterior rates $r_\epsilon(\pi_\gamma, \theta)$ with the rates $R_\epsilon(\widehat{\theta}(\alpha), \theta)$ of the estimators from $\widehat{\Theta}(\mathcal{A})$. We say that the family of the posterior concentration rates $\{r_\epsilon(\pi_\gamma, \theta),\, \gamma \in \Gamma\}$ (corresponding to the family of priors $\Pi(\Gamma) = \{\pi_\gamma,\, \gamma \in \Gamma\}$) covers the family of estimators $\widehat{\Theta}(\mathcal{A}) = \{\widehat{\theta}(\alpha),\, \alpha \in \mathcal{A}\}$ on a set Θ_0 if for any $\alpha \in \mathcal{A}$, there exists an $\gamma = \gamma(\alpha) \in \Gamma$ such that

$$r_\epsilon(\pi_{\gamma(\alpha)}, \theta) \leq C_\epsilon R_\epsilon(\widehat{\theta}(\alpha), \theta)$$

for all $\theta \in \Theta_0$, ideally with a set Θ_0 as big as possible and with some bounded $C_\epsilon \leq C < \infty$. Thus, in this situation an oracle posterior rate result over the family of the posterior concentration rates $\{r_\epsilon(\pi_\gamma, \theta),\, \gamma \in \Gamma\}$ can only be better

than an oracle posterior rate result over any family of estimators $\widehat{\Theta}(\mathcal{A})$ covered by the family of the posterior concentration rates $\{r_\epsilon(\pi_\gamma, \theta),\, \gamma \in \Gamma\}$.

Remark 1.2. In general, one can start with an arbitrary family of rates $\mathcal{R}(\Gamma) = \{r_\epsilon(\gamma, \theta), \gamma \in \Gamma\}$ and try to design an adaptive prior π to mimic the oracle $r_\epsilon(\Gamma, \theta) = \inf_{\gamma\in\Gamma} r_\epsilon(\gamma, \theta)$ over the family $\mathcal{R}(\Gamma)$ in the sense that

$$r_\epsilon(\pi, \theta) \le C_\epsilon r_\epsilon(\Gamma, \theta) = C_\epsilon \inf_{\gamma\in\Gamma} r_\epsilon(\gamma, \theta)$$

for all $\theta \in \Theta_0$, ideally with a set Θ_0 as big as possible and with some positive bounded $C_\epsilon \le C < \infty$. Of course, the issue of assessing the quality of the obtained oracle remains. In a similar manner as above, one can introduce the notion that a family $\mathcal{R}_1(\Gamma_1)$ covers a family $\mathcal{R}_2(\Gamma_2)$.

In this thesis (see also Babenko and Belitser (2010)) we deal with the situation when the family of the risks of estimators from $\widehat{\Theta}$ coincides (up to a constant) with the family of the concentrations rates of the posteriors corresponding to the family of priors $\{\pi_\gamma,\, \gamma \in \Gamma\}$. Then the both approaches are in essence the same and it is just a matter of terminology in defining the oracle posterior rate, depending on the starting point, either with the family of estimators or the family of priors. Recall that $r_\epsilon(\theta) = r_\epsilon(\widehat{\Theta}, \theta) = \inf_{\widehat{\theta}\in\widehat{\Theta}} R_\epsilon(\widehat{\theta}, \theta)$ is the oracle risk at θ over a family of estimators $\widehat{\Theta}$ as defined by (1.2). Then $r_\epsilon(\theta_0)$, $\theta_0 \in \Theta_0$, is said to be the *posterior oracle rate* with respect to the estimators class $\widehat{\Theta}$ if for any $M_\epsilon \to \infty$

$$P\big\{r_\epsilon^{-1}(\theta_0) d(\theta, \theta_0) \ge M_\epsilon | X\big\} \to 0 \quad \text{as} \quad \epsilon \to 0 \tag{1.6}$$

in P_{θ_0}-probability, preferably (if possible) uniformly over $\theta_0 \in \Theta_0$. Again, the bigger the set Θ_0, the stronger this property is. A non-asymptotic version of the form (1.5) with $r_\epsilon(\theta_0)$ instead of r_ϵ is a stronger claim too. Thus it makes sense to take a good class of estimators (the one that covers some interesting scales) as a starting point and then try to design such a prior that the resulting posterior concentration rate is bounded by a multiple of the oracle risk over that class of estimators. Of course, if one looks at the problem from the purely Bayesian point of view (the true θ is intrinsically random), then we deal only with Bayes risks and the above arguments are not relevant.

Similar to the estimation problem discussed in Section 1.2, it is important to understand how the posterior oracle rate over a family of posterior rates $\{r_\epsilon(\pi_\gamma, \theta),\, \gamma \in \Gamma\}$ (generated by a family of priors $\Pi = \{\pi_\gamma,\, \gamma \in \Gamma\}$) is related to the posterior minimax rate over a functional scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$. In fact, the same ideas apply in this situation as for the estimation problem. In principle, it all depends on the following two ingredients: whether the first family covers the second and how massive, with respect to the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$, the set Θ_0 is on which an oracle inequality over the family $r_\epsilon(\pi_\gamma, \theta)$ is established.

However, relating functional scales to families of estimators (instead of families of posterior rates) is more natural. Therefore we look at posterior oracle rates with respect to families of estimators rather than to families of posterior rates. Suppose that the family $\widehat{\Theta}(\mathcal{B})$ covers the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$. Then

clearly, in the same manner as for the estimation problem, the result on posterior oracle convergence rate over the family $\widehat{\Theta}(\mathcal{B})$ is stronger than the result on the adaptive posterior convergence with the minimax rate, at least for all $\theta_0 \in \Theta_0 \cap \big(\cup_{\beta\in\mathcal{B}} \Theta_\beta \big)$. It would be desirable to obtain an oracle posterior rate result for a set Θ_0 which contains the whole scale $\cup_{\beta\in\mathcal{B}}\Theta_\beta$. In this thesis, our Θ_0 will be the whole space ℓ_2 so that our results are always stronger than any results on adaptive posterior convergence rate with the minimax rate over any scale $\{\Theta_\beta, \beta \in \mathcal{B}\} \subseteq \ell_2$ covered by the estimators family $\widehat{\Theta}(\mathcal{B})$. Thus, another appealing feature of the proposed oracle approach to the posterior convergence quality is that the family $\widehat{\Theta}(\mathcal{B})$ can cover many scales at once. Therefore, all the results on adaptive posterior convergence rate with the minimax rate over all these scales follow immediately from the results on oracle posterior rate corresponding to the estimators family $\widehat{\Theta}(\mathcal{B})$. Moreover, all these results will also be uniform over all scales contained in Θ_0, provided the oracle result (1.6) is uniform over Θ_0.

A result of the type (1.4) (or a non-asymptotic version (1.5)) claims in essence that the posterior concentration rate is at most r_ϵ, this does not forbid it to be smaller. A result of the type saying that the posterior concentration rate is at least r_ϵ is called a lower bound result for the posterior rate. Together with an upper bound result (1.5), such a lower bound result specifies the exact posterior concentration rate. To be precise, besides (1.4) (or (1.5)), we want to derive a result of the type

$$E_{\theta_0} P\big\{ r_\epsilon^{-1} d(\theta, \theta_0) \le \delta | X \big\} \le \psi(\delta), \tag{1.7}$$

for some function ψ monotonically decreasing to zero as $\delta \to 0$ and all $\theta_0 \in \Theta_0$. This means that the oracle rate r_ϵ is sharp for the posterior concentration and can not be improved, at least for the given prior. This also implies that the posterior mass is mainly concentrated in the annulus $\{\theta : \delta r_\epsilon \le d(\theta, \theta_0) \le M r_\epsilon\}$ for a sufficiently small $\delta > 0$ and a sufficiently large $M > 0$. The most interesting and the finest results of this type concern the local rates $r_\epsilon = r_\epsilon(\theta_0)$, i.e. depending on the true θ_0, especially when $r_\epsilon(\theta_0)$ is of an oracle kind, an oracle rate with respect to a family of priors, or estimators, or an arbitrary interesting family of posterior rates.

The first results on lower bounds for the posterior rate are obtained in Castillo (2008) in case of a Gaussian prior in an asymptotic setting. The method is essentially based on the so called concentration function of the Gaussian process and the small ball probabilities for a Gaussian process; see Lifshits (1995), van der Vaart and van Zanten (2008) and further references therein.

Let us describe an approach to lower bounds for an oracle posterior rate based on a family of priors. Suppose we start with a family of priors $\{\pi_\gamma, \gamma \in \Gamma\}$, then we design an adaptive prior and establish an upper bound result (1.4) or (1.5) for it, i.e. the corresponding adaptive posterior mimics the oracle over the family of the posteriors coming from the family of priors we started with. A good candidate for an adaptive prior is a mixture over the priors from the family $\{\pi_\gamma, \gamma \in \Gamma\}$. If the adaptive prior is a mixture, then the resulting adaptive posterior is a (random) mixture of posteriors. If we want to establish the lower

bound (1.7) with the oracle rate for this adaptive posterior, we can proceed as follows. Clearly, the smallest lower bound among the members of the mixture will certainly be a lower bound for the resulting adaptive posterior. Therefore for each prior π_γ from the family of priors we derive the corresponding family of posterior rates $\{r_\epsilon(\pi_\gamma, \theta),\ \gamma \in \Gamma\}$ and then show that all these posterior rates are sharp in the sense of lower bound (1.7), including the oracle rate $r_\epsilon(\pi_{\gamma_o}, \theta)$. Since the oracle rate is the smallest among them, the resulting adaptive posterior must satisfy the lower bound with the oracle rate. In fact, this is exactly how the proofs of the results on the lower bounds in this thesis go. It is a challenging problem to derive this kind of results for a wide class of priors instead of specific ones.

1.5 Gaussian sequence model

Consider a *Gaussian sequence model*, as is introduce earlier in Example 1,

$$X_k = \theta_k + \epsilon\xi_k, \quad k \in \mathcal{N}, \tag{1.8}$$

where $\theta \in \ell_2$ is an unknown parameter of interest, $\mathcal{N} = \{1, 2, \ldots\} = \mathbb{N}$, the random noise variables ξ_k's are independent standard Gaussian, the parameter ϵ is a known noise level. Recall that this models is sometimes called *white noise model* in the sequence form; more discussion about the continuous version of the white noise model, which is equivalent to the Gaussian sequence model, is below. If the set $\mathcal{N}$ is finite or of a growing cardinality $|\mathcal{N}_\epsilon| = M_\epsilon \to \infty$ as $\epsilon \to 0$, this models is also called *many normal means model*.

Although the model (1.8) seems to be a very specific, unpractical and mathematically idealized object, it turned out to play a significant role in nonparametric statistics as much (if not more) as finite dimensional normal models in parametric statistics. Looking innocently simple (probably one of the simplest nonparametric models one can think of), this model "captures many of the conceptual issues associated with non-parametric estimation, with a minimum of technical complication" (Johnstone (2003)). Moreover, this model, besides being of interest and widely used on its own right, for example, in communication theory and signal transfer (see Kotelnikov (1959), serves as a purified approximation (*principle of equivalence*, see Efromovich (1999)) to some other statistical models, such as probability density estimation, regression model, spectral function estimation. We will not go into more detail on the equivalence principle, the problem of establishing the equivalence in a precise sense is a delicate and in general difficult task. Some authors undertook a serious effort to make the notion of equivalence mathematically precise: see Brown and Low (1996) for regression model, Nussbaum (1996) and Claeskens and Hjort (2004) for the density estimation.

To summarize, the Gaussian sequence model captures the statistical essence of the asymptotically equivalent nonparametric models and preserves its traits in a pure form. The statistical inference results for the "generic" model (1.8) can therefore be conveyed, under some appropriate assumptions, to other models

according to the above mentioned equivalence principle. This is why this model gave rise to many important, beautiful and surprising results whose influence extends well beyond the formal mathematical constructions into the practical world of data analysis. The model (1.8) was studied in different settings by many authors, for example, Ibragimov and Khasminskii (1981), Belitser and Levit (1995), Efromovich (1999), Johnstone (2003), and further references therein.

In the rest of this section, we illustrate the equivalence of the Gaussian sequence model to the white noise model and an approximate equivalence to the nonparametric regression model.

1.5.1 Gaussian white noise model

Suppose we observe a stochastic process $X^{(\epsilon)}(t) = X(t)$, satisfying the stochastic differential equation, the "signal in Gaussian white noise" model:

$$dX(t) = f(t)dt + \epsilon dW(t), \quad t \in [0,1], \tag{1.9}$$

where $W(t)$ is a standard Wiener process and $f(t) \in L_2[0,1]$ is an unknown function of interest called *signal* from the space of the square integrable functions on the unit interval, $\epsilon > 0$ is a known noise level parameter. The model is called the Gaussian white noise model, probably first introduced by Kotelnikov (1959), see also Ibragimov and Khasminskii (1981).

Let $\{\varphi_k,\, k \in \mathbb{N}\}$ be an orthonormal basis for $L_2[0,1]$ and expand signal f as $f(t) = \sum_{k=1}^{\infty} \theta_k \varphi_k(t)$ in the L_2-sense, with an isomorphic correspondence between signal f and its Fourier coefficients $\theta = (\theta_k,\, k \in \mathbb{N})$. Then the equivalence of models (1.9) and (1.8) is seen by the expansion of (1.9) in the basis $\{\varphi_k,\, k \in \mathbb{N}\}$:

$$X_k = \int_0^1 \varphi_k(t) dX(t), \quad \theta_k = \int_0^1 f(t)\varphi_k(t)dt, \quad \xi_k = \int_0^1 \varphi_k(t) dW(t),$$

so that (1.8) follows

$$X_k = \theta_k + \epsilon \xi_k, \quad k \in \mathbb{N}.$$

The last integral is a Wiener-Ito stochastic integral. By the elementary properties of stochastic integrals, ξ_k's are independent standard Gaussian random variables.

1.5.2 Regression model into Gaussian sequence model

Suppose that we observe $Y_1, \ldots, Y_n$ in the nonparametric *regression model*

$$Y_i = f(i/n) + Z_i, \quad i = 1, 2, \ldots n, \tag{1.10}$$

where the Z_i's are independent $\mathcal{N}(0,1)$ random variables, f is an unknown function (sometimes called *curve*) from $L_2[0,1]$. Let $\{\varphi_k,\, k \in \mathbb{N}\}$ be an orthonormal basis in $L_2[0,1]$. We can expand the regression function f in a Fourier series $f(x) = \sum_{k=1}^{\infty} \theta_k \varphi_k(x)$ in the L_2-sense, where the Fourier coefficients $\theta_k = \int_0^1 f(x)\varphi_k(x)dx$, $k \in \mathbb{N}$. Due to Parseval's identity, the problem of

estimating the regression f with the L_2-risk can be converted into a problem of estimating the parameter $\theta = \{\theta_k,\, k \in \mathbb{N}\}$ with the ℓ_2-risk.

To illustrate the equivalence of the models, for $k \in \mathbb{N}$, we define $\epsilon = n^{-1/2}$,

$$\begin{aligned} X_k &= \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_k(i/n), \\ \widetilde{\theta}_k &= \frac{1}{n}\sum_{i=1}^{n} f(i/n)\varphi_k(i/n), \\ \xi_k &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i \varphi_k(i/n). \end{aligned}$$

Model (1.10) together with the above relations yields

$$X_k = \widetilde{\theta}_k + \epsilon\xi_k, \quad k \in \mathbb{N}. \tag{1.11}$$

Next, notice that the ξ_k's are normal with zero means,

$$\widetilde{\theta}_k = \frac{1}{n}\sum_{i=1}^{n} f(i/n)\varphi_k(i/n) \approx \int_0^1 f(x)\varphi_k(x)dx = \theta_k, \quad k \in \mathbb{N},$$

$$\mathrm{Cov}(\xi_k, \xi_m) = \frac{1}{n}\sum_{i=1}^{n} \varphi_k(i/n)\varphi_m(i/n) \approx \int_0^1 \varphi_k(x)\varphi_m(x)dx = \delta_{km}, \quad k, m \in \mathbb{N},$$

where $\delta_{km} = 1$ if $k = m$ and zero otherwise. Thus, the ξ_k's are approximately, as n is sufficiently large, independent standard Gaussian random variables. We note here that if the basis $\{\varphi_k,\, k \in \mathbb{N}\}$ is standard trigonometrical, then $\mathrm{Cov}(\xi_k, \xi_m) = \delta_{km}$ $k, m = 1, \ldots, n$; see Belitser (2000), Tsybakov (2008). Hence, for sufficiently large n, the model (1.11) approximates the Gaussian sequence model (1.8). Note the correct calibration of the information parameter $\epsilon = n^{-1/2}$ in the model (1.8) with respect to the sample size n from the model (1.10). Of course, here we used only heuristic arguments, but with a fair piece of effort they can be made rigorous.

1.5.3 White noise model into regression model

Consider the process $X(t) = X^{(\epsilon)}(t)$, $t \in [0, 1]$, in the Gaussian white noise model (1.9). Notice, that the model can be rewritten in the equivalent form:

$$X(t) = X(0) + \int_0^t f(s)ds + \epsilon W(t), \quad t \in [0, 1],$$

where $X(0)$ is the initial value of the stochastic differential equation (1.9), assumed to be a random variable, independent of $W(t)$, $t \in [0, 1]$. The increments of the process $X(t)$ are

$$X(t_i) - X(t_{i-1}) = \int_{t_{i-1}}^{t_i} f(s)ds + \epsilon\big(W(t_i) - W(t_{i-1})\big), \quad t_i = \frac{i}{n}, \quad i = 1, \ldots, n.$$

The Wiener process has independent increments with distribution $W(t_i) - W(t_{i-1}) \sim \mathcal{N}(0, n^{-1})$.

Again, take the calibration $\epsilon = n^{-1/2}$. For large n and a continuous f, the integral $\int_{t_{i-1}}^{t_i} f(s)ds$ is approximately $n^{-1}f(t_i)$. Multiplying both sides of the last equation by n and defining the random variables $X_i = n\big(X(t_i) - X(t_{i-1})\big)$ and $Z_i = \sqrt{n}\big(W(t_i) - W(t_{i-1})\big)$ for $i = 1, \ldots, n$, we obtain

$$X_i \approx f(i/n) + Z_i, \quad i = 1, \ldots, n.$$

Thus, according to our heuristic arguments, the Gaussian white noise model (1.9) is asymptotically, for sufficiently large n, equivalent to the regression model (1.10).

1.6 Scope of the thesis

All the results about posterior rates obtained until now are related to the optimal (minimax) rates for the estimation problem over the corresponding nonparametric smoothness classes, i.e. of a global nature; cf. Ghosal, Ghosh and van der Vaart (2000), Ghosal and van der Vaart (2007), Ghosal, Lember and van der Vaart (2008), van der Vaart and van Zanten (2007), van der Vaart and van Zanten (2008). In the meantime, a new local approach to optimality has been developed within the estimation framework, namely, the oracle approach.

The main goal of this thesis is to develop and study the oracle approach to the posterior concentration performance in the Gaussian sequence model, which includes the equivalent models as well (see the previous section), although some more technical work is needed to make the exact transfer of the results to the asymptotically equivalent models. Instead of the minimax risk, a new benchmark for the posterior concentration rate is proposed, the so called posterior oracle rate, which is the smallest possible rate over a family of posterior rates corresponding to an appropriately chosen family of priors.

In this thesis we start with a family of priors modeling the estimation oracle in the sense that the family of resulting posterior rates essentially coincides with the family of risks of the estimators from some sufficiently "rich" family of estimators, containing some good estimators, e.g. minimax over some interesting nonparametric classes. The quality of the oracle posterior rate is then characterized by the best (oracle) risk over that family of estimators. The oracle posterior rate is an intrinsically local quantity since it depends on the true parameter value and it is supposed to be good for "good" parameter values. Besides, the results on the oracle (with respect to a certain estimators family) posterior rate will always be stronger than any result about posterior convergence with the minimax rate over all scales which are covered by that estimators family.

At this moment we are not aware of any result on oracle posterior rate other than Babenko and Belitser (2010) which is essentially contained in Chapter 2 of the present thesis. In Chapter 2, we apply the Bayes approach to the problem of projection estimation of an unknown parameter observed in the Gaussian

sequence model and we study the rate at which the posterior distribution concentrates about the true parameter from the space ℓ_2, as the information in observations tends to infinity. The projection estimators projects the data on the infinite family of embedded finite dimensional subsets of the space ℓ_2. This corresponds to the so called ordered variable selection problem as described in Birge and Massart (2001). Under an appropriate hierarchical prior, we study the performance of the resulting (appropriately adjusted by the empirical Bayes approach) posterior distribution and establish that the posterior concentrates about the true parameter with the oracle projection convergence rate. We also construct a Bayes estimator based on the posterior and show that it satisfies an oracle inequality. The results are nonasymptotic and uniform over ℓ_2.

Another important feature of our approach is that our results on the oracle projection posterior rate are always stronger than any result about posterior convergence with the minimax rate over any nonparametric class if our family of projection estimators contains a minimax estimator over that class. If this holds for a functional scale, then our result implies the adaptive minimaxity over that scale. For example, the results of Chapter 2 (or Babenko and Belitser (2010)) are stronger than those in Belitser and Ghosal (2003) where the Sobolev scale was considered since the Sobolev scale is covered by the family of the projection estimators we consider. Actually, the Sobolev scale is just one particular functional scale for which the results on the posterior convergence with the minimax rate follow. Other scales are exponential ellipsoids, hyperrectangles and tail classes. Besides, the uniformity of the result on adaptive minimax convergence rate of the posterior over all these scales also follows immediately from the uniformity of the results in Babenko and Belitser (2010) over the whole space ℓ_2. We also study implications for the model selection problem, namely we propose a Bayes model selector and assess its quality in terms of the so called false selection probability.

In Chapter 3, we complement the upper bound results on the posterior concentration rate obtained in Chapter 2 by a lower bound result of type (1.7) for the oracle rate. It turns out that the rates of the upper and lower bounds coincide with the oracle projection convergence rate. This implies that the oracle posterior rate is sharp and all of the posterior mass concentrates in some annulus around the true parameter value. In our case the prior is a mixture (as it handles the adaptation problem) of normals, thus the resulting posterior is also a mixture of normals, which enables us in essence, as in Castillo (2008), to use the small ball probabilities technique for Gaussian distributions. The lower bound results are non-asymptotic and uniform over ℓ_2.

In Chapter 4, we study the posterior distributions in the many normal means model using numerical simulations and compare the results with those previously obtained theoretically in the Gaussian sequence model. To illustrate our findings, we construct credible bands for the signal function in the equivalent white noise model formulation (or approximate discrete regression model) on the basis of posterior distribution of the signal given data. Empirical performance of the methods is illustrated on test examples.

In Chapter 5, we first consider the problem of Bayes estimation of a linear

functional of the signal in the continuous Gaussian white noise model, under the assumption that the unknown signal is from a Sobolev smoothness class. We propose a simple family of conjugate (in fact Gaussian) priors for the signal, which makes the explicit calculations possible. Next we claim that the proposed prior models adequately the smoothness requirement on signal in the underlying estimation problem in that the resulting Bayes estimator of the linear functional of the signal is proved to be rate minimax over the Sobolev smoothness class from the both frequentist and Bayesian perspectives. Using this, we establish the result on the convergence rate of the posterior distribution of the linear functional of the signal, which turns out to be optimal in the minimax sense over the Sobolev class.

In the second part of Chapter 5, we consider an adaptive filtering problem. We assume that the signal is distributed according a prior parameterized by an unknown "true smoothness" parameter. Then the Bayes estimator of the functional cannot be used, because this parameter is unknown. This Bayes estimator of the functional is regarded as a Bayesian oracle and its risk becomes our benchmark, which we call the oracle Bayes risk. By applying the empirical Bayes approach, which is due to Robbins (1956), we construct a marginal likelihood (an empirical Bayes) estimator for the smoothness parameter and show that the resulting plug-in adaptive procedure for the estimator of the functional mimics the Bayesian oracle. Actually we show that our adaptive estimator of the functional is asymptotically sharp, i.e. its risk coincides asymptotically with the oracle Bayes risk. The accompanying problem of estimating the smoothness parameter by the empirical Bayes procedure can be seen as a Bayesian counterpart of the inference problem on the smoothness parameter. In the minimax frequentist setting, it is impossible to estimate the smoothness in any meaningful sense, while it is a well defined problem from the Bayesian point of view, merely as the problem of estimating a parameter of the prior distribution.

Chapter 2

Oracle convergence rate of posterior under projection prior and Bayesian model selection

We apply the Bayes approach to the problem of projection estimation of an unknown parameter observed in the Gaussian white noise model and we study the rate at which the posterior distribution concentrates about the true parameter.

Suppose we observe $X = (X_1, X_2, \ldots)$, where

$$X_i = \theta_i + \frac{\xi_i}{\sqrt{n}}, \quad i = 1, 2, \ldots, \tag{2.1}$$

$\theta = (\theta_i)_{i\in\mathbb{N}} = (\theta_1, \theta_2, \ldots) \in \ell_2$ is an unknown parameter of interest, the noise variables ξ_i are independent, identically distributed $\mathcal{N}(0, 1)$ random variables, the parameter n is the noise intensity and reflects the increase of information in the data X as $n \to \infty$. The goal is to make an inference on θ on the basis of the observed data $X = X^{(n)}$. Many quantities actually depend on parameter n, but for the sake of notational simplicity we will often skip this dependence.

Model (2.1) is known to be a Gaussian white noise model or many Normal means model, and it arises in various statistical settings, for example, in nonparametric regression model or in the continuous white noise model, which is widely used in communication theory and signal transmission. The equivalence of the models was already discussed in the general introduction (Chapter 1).

In this chapter, we introduce the notion of oracle projection convergence rate, propose an appropriate hierarchical prior, construct several different estimators using the posterior, study their oracle properties, establish the oracle projection convergence rate for the resulting (appropriately adjusted by the empirical Bayes approach) posterior distribution and finally address the problem

of Bayesian model selection. The class of projection estimators $\widehat{\Theta}(\mathcal{N})$ is parameterized by the so called cut-off parameter $N \in \mathbb{N}$, which can be thought of as model selector. According to Bayesian paradigm, we designed a two-level hierarchical prior on (θ, N): conditionally $\theta|N \sim \pi_N$ and $N \sim \lambda$. Conditional priors π_N's are all normal, which makes it possible to compute many quantities. Discrete prior λ on the cut-off parameter N satisfies certain conditions, which we will indicate later. Despite of its simplicity, this projection estimators family turns out to be a very good choice. Indeed, we established the oracle relations of types (1.3) and (1.6) uniformly over $\Theta_0 = \ell_2$. This implies that our results are always stronger than any results on adaptive posterior convergence rate with the minimax rate over any scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$ covered by the estimators family $\widehat{\Theta}(\mathcal{N})$. On the other hand, this estimators family is rich enough to cover some important scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ such as Sobolev ellipsoids, exponential ellipsoids, hyperrectangles and tail classes; see Section 2.3 below. This is another appealing feature of this approach: the family $\widehat{\Theta}(\mathcal{N})$ can cover many scales at once. Besides, all the results are nonasymptotic and uniform over ℓ_2.

Recall that our hierarchical prior includes a prior on parameter N, which enables us to perform the Bayesian model selection to mimic the oracle N_o (in our case N_o is the dimension of the oracle model) by using the posterior $P(N|X)$; for example, by taking the maximum a posteriori probability selector $\widehat{N}_{\text{MAP}}$ or by simply generating from $P(N|X)$ the posterior randomized selector. To assess the quality of model selectors, we introduce the notion of *false selection probability* $\text{FSP}(\widehat{N}, \tau, \theta_0) = P_{\theta_0}\{\widehat{N} \notin \mathcal{N}(\tau, \theta_0)\}$, where $\mathcal{N}(\tau, \theta_0) \subseteq \mathbb{N}$ is the index set of the so called admissible models, τ is a predetermined *tolerance* parameter that describes how lenient the definition of admissible model is. We investigate the performance of the proposed Bayes model selectors with respect to this criterion and establish that the false selection probability at point θ_0 is small if, roughly speaking, the oracle risk at point θ_0 is not "too parametric", in the sense that $r_n(\theta_0) \geq Cn^{-1}$ for some big constant C.

The proposed methodology can, in principle, be extended to other statistical models. In this chapter, we, however, consider the simplest Gaussian infinite-dimensional framework, which allows us to illustrate the main ideas of the proposed approach without too much technicalities and to derive closed expressions of many quantities and constants involved. As to the results on the oracle posterior rate, we are not aware of other studies of such kind. To the best of our knowledge, all the results about posterior rates obtained until now are actually global, typically such results are related to the minimax rates for the estimation problem over some nonparametric smoothness classes; cf. Ghosal, Ghosh and Van der Vaart (2000). The point is that the existing methods for the derivation of posterior concentration rates are often based on some global quantities, like entropy characteristics and existence of uniform tests, and lead therefore to global posterior rates. Oracle rates are on the contrary intrinsically local: the oracle rate depends on the true value θ_0 and it is typically small for "good" parameter values. The derivation of the results on oracle concentration rates requires therefore the development of new techniques and the chapter in this thesis is intended to make a first step in this direction.

2.1 Preliminaries

Denote the probability measure of X from the model (2.1) by $P_\theta = P_\theta^{(n)}$. In case θ is a stochastic element as in the Bayesian analysis, denote by P_θ the conditional probability measure of X given θ, by $P(\theta|X)$ the conditional distribution of θ given X and by P the joint probability measure of (X, θ). The same notations apply to the expectation operation. Denote by $I\{S\}$ the indicator function of the set S and by $|S|$ the cardinality of the set S, $\mathbb{N} = \{1, 2, \ldots\}$, i.e. zero is not included in $\mathbb{N}$. For a constant $c \in \mathbb{R}$, $c\theta = (c\theta_i)_{i\in\mathbb{N}}$.

Introduce the class of projection estimators $\{\widehat{\theta}(N),\, N \in \mathbb{N}\}$ which is parameterized by a so called cut-off parameter N. It is the dimension of the approximating linear subspace $S_N = \{(s_k)_{k\in\mathbb{N}} : s_k = 0,\, k > N\}$, on which we project the data X to obtain the corresponding projection estimator: for an $N \in \mathbb{N}$,

$$\widehat{\theta}(N) = \Pi_{S_N} X = (\widehat{\theta}_i(N))_{i\in\mathbb{N}}, \quad \widehat{\theta}_i(N) = X_i I\{i \le N\}, \quad i \in \mathbb{N}. \tag{2.2}$$

The estimation quality of the estimator $\widehat{\theta}(N)$ at point θ is measured by the risk function

$$\mathcal{R}_n(N) = \mathcal{R}_n(N, \theta) = R(\widehat{\theta}(N), \theta) = E_\theta \|\widehat{\theta}(N) - \theta\|^2.$$

It can easily be checked that

$$\begin{aligned} \mathcal{R}_n(N, \theta) &= E_\theta \sum_{i=1}^{\infty} (\widehat{\theta}_i(N) - \theta_i)^2 = E_\theta \Big[\sum_{i=1}^{N} \frac{\xi_i^2}{n} + \sum_{i=N+1}^{\infty} \theta_i^2 \Big] \\ &= \frac{N}{n} + \sum_{i=N+1}^{\infty} \theta_i^2. \end{aligned} \tag{2.3}$$

For a fixed $\theta \in \ell_2$, we define the oracle $N_o = N_o(\theta) \in \mathbb{N}$ and the *oracle projection risk* (or just the *oracle risk*) $r_n = r_n(\theta) = \mathcal{R}_n(N_o, \theta)$ by the following relation:

$$r_n = r_n(\theta) = \mathcal{R}_n(N_o, \theta) = \frac{N_o}{n} + \sum_{i=N_o+1}^{\infty} \theta_i^2 = \min_{N\in\mathbb{N}} \mathcal{R}_n(N, \theta). \tag{2.4}$$

As is easy to see, the oracle $N_o(\theta)$ is well defined for any $\theta \in \ell_2$ and it is not an estimator since it depends on the unknown θ. Recall that zero is not included in $\mathbb{N}$, which ensures that oracle risk is always positive, in fact $\mathcal{R}_n(N_o, \theta) \ge n^{-1}$ and $N_o \ge 1$. It is not restrictive since n^{-1} is the best (parametric) rate in case at least one coordinate $\theta_i \ne 0$. We want our problem to be at least as difficult as the parametric one, so we will have to avoid the trivial case $\theta = (0, 0, \ldots)$ anyway. For example, if we allow $N = 0$, then our results will hold only for $\theta \in \Theta_0 = \{\theta \in \ell_2 : \mathcal{R}_n(N_o, \theta) \ge n^{-1}\}$. Alternatively, we can add a penalty term $1/n$ to the oracle risk and take the resulting sum as our new oracle benchmark.

Now our goal is to propose a two-level hierarchical prior on (θ, N) in the manner we described in the introduction, so that we can study the resulting

posterior $P(\theta|X)$ and the properties of an estimator $\widehat{\theta}$ constructed by using the posterior distribution $P(\theta|X)$ and $P(N|X)$, under the probability measure P_{θ_0} for some fixed ("true") $\theta_0 \in \ell_2$.

First, consider the case of a fixed cut-off parameter N. Let $\{\pi_N(\theta),\, N \in \mathbb{N}\}$, be a family of priors on θ defined as follows:

$$\theta_i|N \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2(N)), \quad \tau_i^2(N) = n^{-1} I\{i \le N\}, \quad i \in \mathbb{N}, \tag{2.5}$$

with the convention from now on that if $Z \sim \mathcal{N}(c, 0)$ then $Z = c$ with probability 1. $\pi_N(\theta)$ is the product measure of normal distributions $\mathcal{N}(0, \tau_i^2(N))$, $i \in \mathbb{N}$. Recall that $X_i|\theta \overset{ind}{\sim} \mathcal{N}(\theta_i, n^{-1})$, so that the posterior distribution $P_N(\theta|X)$ corresponding to the prior $\pi_N(\theta)$ defined by (2.5) is readily obtained:

$$\theta_i|(X, N) \overset{ind}{\sim} \mathcal{N}\Big(\frac{X_i}{2} I\{i \le N\}, \frac{n^{-1}}{2} I\{i \le N\}\Big), \quad i \in \mathbb{N}, \tag{2.6}$$

Here we used the elementary fact that if $Z|Y \sim \mathcal{N}(Y, \sigma^2)$ and $Y \sim \mathcal{N}(\mu, \tau^2)$, then

$$Y|Z \sim \mathcal{N}\Big(\frac{Z\tau^2 + \mu\sigma^2}{\tau^2 + \sigma^2}, \frac{\tau^2\sigma^2}{\tau^2 + \sigma^2}\Big).$$

The Bayes estimator with respect to the prior $\pi_N(\theta)$ is $\tilde{\theta} = E_N(\theta|X) = (\tilde{\theta}_i)_{i\in\mathbb{N}}$. Notice that $\tilde{\theta} = \frac{\widehat{\theta}(N)}{2}$, that is $\tilde{\theta}_i = \frac{\widehat{\theta}_i(N)}{2}$, $i \in \mathbb{N}$, where $\widehat{\theta}(N)$ is the projection estimator (2.2) with cut-off N. It follows therefore that the Bayes estimator $\tilde{\theta}$ converges to $\theta_0/2$ with the rate $\mathcal{R}_n(N, \theta_0)$ (up to a constant). Besides, it is not so difficult to see (for example, by applying the conditional Chebyshev inequality) that the posterior $P_N(\theta|X)$ concentrates about $\theta_0/2$ as $n \to \infty$ in P_{θ_0}-probability with the rate $\mathcal{R}_n(N, \theta_0)$. There is nothing special about factor $1/2$, in fact we can get any other factor in $(0, 1)$ by taking the variances in the product prior π_N as $\tau_i^2(N) = Cn^{-1} I\{i \le N\}$, $i \in \mathbb{N}$, for some constant $C > 0$. The reason for this is that our prior π_N is normal with zero mean so that the corresponding Bayes estimator is always a shrinkage estimator and thus it is always going to be a fraction of the projection estimator $\widehat{\theta}(N)$. To fix this problem, instead of (2.5), take the prior $\pi_{N,\mu}$ defined as

$$\theta_i|N \overset{ind}{\sim} \mathcal{N}(\mu_i(N), \tau_i^2(N)), \quad \mu_i(N) = \mu_i I\{i \le N\},\ \tau_i^2(N) = n^{-1} I\{i \le N\},$$

with the idea that $\mu(N) = (\mu_i(N))_{i\in\mathbb{N}}$ would model a possible shift. The corresponding posterior is easy to derive in the same way as for (2.6):

$$\theta_i|(X, N) \overset{ind}{\sim} \mathcal{N}\Big(\frac{X_i + \mu_i}{2} I\{i \le N\}, \frac{n^{-1}}{2} I\{i \le N\}\Big), \quad i \in \mathbb{N}. \tag{2.7}$$

For now μ_i's are some parameters. To estimate these, apply the empirical Bayes approach, i.e. we use the marginal (of X) maximum likelihood estimates for μ_i's. Marginal distribution of X: $X_i|\mu(N) \overset{ind}{\sim} \mathcal{N}\big(\mu_i(N), n^{-1} + \tau_i^2(N)\big)$, $i \in \mathbb{N}$, so that trivially $\widehat{\mu}_i = X_i$, $i \in \mathbb{N}$. Thus, as posterior we take the conditional distribution $P'_N(\theta|X)$ defined as

$$\theta_i|(X, N) \overset{ind}{\sim} \mathcal{N}\Big(X_i I\{i \le N\}, \frac{n^{-1}}{2} I\{i \le N\}\Big), \quad i \in \mathbb{N}. \tag{2.8}$$

Now we can make all our inference on the basis of this empirical Bayes posterior $P'_N(\theta|X)$. Clearly, the corresponding (empirical) Bayes estimator coincides with the projection estimator $\widehat{\theta}(N)$ and this empirical Bayes posterior distribution concentrates about θ_0 with projection rate $\mathcal{R}_n(N,\theta_0)$ in P_{θ_0}-probability.

The next theorem shows that the posteriors $P_N(\theta|X)$ and $P'_N(\theta|X)$, defined by (2.6) and (2.8), are concentrated about $\theta_0/2$ and θ_0 respectively with the rate $\mathcal{R}_n(N,\theta_0)$.

Theorem 2.1. *Let $\mathcal{R}_n(N,\theta_0)$ be defined by (2.3). Then there exist constants $c_1, c_2 > 0$ such that, for any $\theta_0 \in \ell_2$ and any $M > 0$,*

$$E_{\theta_0} P_N\Big\{ \big\|\theta - \theta_0/2\big\|^2 \ge M\mathcal{R}_n(N,\theta_0) \,\big|X\Big\} \le \frac{c_1}{M},$$
$$E_{\theta_0} P'_N\Big\{ \big\|\theta - \theta_0\big\|^2 \ge M\mathcal{R}_n(N,\theta_0) \,\big|X\Big\} \le \frac{c_2}{M}.$$

Thus, if $N = N_o$ then the posteriors $P_N(\theta|X)$ and $P'_N(\theta|X)$ are concentrated with the oracle rate $\mathcal{R}_n(N_o,\theta_0)$. Clearly, if we knew the oracle N_o defined by (2.4), we would of course use the prior π_{N_o}. Since the oracle is not known we step to the next level in our Bayesian analysis by putting a prior on N as well. Formally, we now regard the distribution $\pi_N(\theta)$ to be the conditional distribution on θ given N, and next put a prior $\lambda = \lambda_\alpha$ on $N \in \mathbb{N}$, $N \sim \lambda$:

$$P(N = k) = \lambda_k = c(\alpha)e^{-\alpha k}, \quad k \in \mathbb{N},$$

where $c(\alpha) = (1 - e^{-\alpha})e^{\alpha}$ is the normalizing constant and the constant $\alpha > 0$ is to be specified later. Note that the prior λ is a geometric distribution $\lambda_k = p(1-p)^{k-1}$, $k \in \mathbb{N}$, with parameter $p = 1 - e^{-\alpha}$. In doing so, we introduced a two-level hierarchical prior $\pi = \pi_\alpha$ on (θ, N):

$$(\theta, N) \sim \pi \quad \Longleftrightarrow \quad \theta|N \sim \pi_N, \quad N \sim \lambda_\alpha, \tag{2.9}$$

with λ_α given above and π_N defined by (2.5). This leads to the posterior distributions $P(\theta|X) = P_\alpha(\theta|X)$ and $P(N|X) = P_\alpha(N|X)$. Although many quantities (like π_α, λ_α and $P_\alpha(\theta|X)$) depend on the parameter $\alpha > 0$, we typically skip this subscript to avoid complicated notations, unless we occasionally want to emphasize this dependence. There is a small notational abuse: π_α is the mixture of π_N's over $N \sim \lambda_\alpha$.

The posterior distribution $P(\theta|X)$ can be expressed as follows (for a measurable set $S \subset \ell_2$):

$$P(\theta \in S|X) = \sum_{k \in \mathbb{N}} P(\theta \in S|X, N = k) P(N = k|X) \tag{2.10}$$

where the conditional distribution $P(\theta \in S|X, N = k)$ is defined by (2.6) with $N = k$.

Recall that the family of priors π_N, $N \in \mathbb{N}$, makes the resulting posterior concentrate about $\theta_0/2$ in P_{θ_0}-probability. The same phenomenon occurs for the posterior $P(\theta|X)$ corresponding to the prior π_α, which is not surprising since π_α

is merely a mixture of π_N's. We fix this problem by using the empirical Bayes approach as we did this above for the prior π_N. Namely, instead of posterior (2.10) we base the inference on the empirical Bayes posterior distribution

$$P'(\theta \in S|X) = \sum_{k\in\mathbb{N}} P'(\theta \in S|X, N = k)P(N = k|X), \tag{2.11}$$

where the conditional distribution $P'(\theta|X, N = k)$ is defined by (2.8) with $N = k$.

Using the above posterior distributions, we finally construct some estimators for the signal θ. From now on denote by $\widehat{\theta} = (\widehat{\theta}_i)_{i\in\mathbb{N}}$ an estimator of signal θ which is in general a measurable function of X, α and n: $\widehat{\theta} = \widehat{\theta}(X, \alpha, n)$. First define a version of empirical Bayes estimator:

$$\widehat{\theta} = E'(\theta|X), \tag{2.12}$$

with the conditional expectation E' taken with respect to the empirical Bayes posterior distribution (2.11). The next estimator is in fact a version of the projection estimator with a posterior-randomized cut-off parameter $\widehat{N}$:

$$\widehat{N} \sim P(N|X), \qquad \widehat{\theta} = \widehat{\theta}(\widehat{N}) = \widehat{\theta}(\widehat{N}, X, \alpha), \tag{2.13}$$

where the random variable $\widehat{N} = \widehat{N}(X, \alpha)$ is drawn from the posterior distribution $P(N|X) = P_\alpha(N|X)$, $\widehat{\theta}(N)$ is the projection estimator defined by (2.2) with cut-off N. Notice that $\widehat{\theta}$ defined by (2.13) can be expressed as $\widehat{\theta} = 2E(\theta|X, N = k)\big|_{k=\widehat{N}}$.

We conclude this section with a couple of remarks.

Remark 2.1. Notice that as $\theta \in \ell_2$, then $\widehat{\theta}(N)$ with $N > n$ are all inadmissible for sufficiently large n. Indeed, assume that $N > n$, then $\mathcal{R}_n(N, \theta) > 1$ while $\mathcal{R}_n\big(\lfloor\sqrt{n}\rfloor, \theta\big) \to 0$ as $n \to \infty$. Therefore, one could in principle restrict the set of possible values for the cut-off parameter N to the set $\mathcal{N}_n = \{1, \dots, n\}$ instead of $\mathbb{N}$. The set $\mathcal{N}_n$ of possible values for the cut-off parameter N would arise by itself in case we had a high-dimensional version the model (2.1) instead of infinite-dimensional, i.e. $i = 1 \dots, n$ in (2.1), the observation $X = (X_1, \dots, X_n)$ and the unknown parameter $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$. Typically, there is not much difference between the above described high-dimensional and our infinite-dimensional models and actually many papers deal with the high-dimensional case rather than with the infinite dimensional one. However, there is a significant difference between these two situations in case when the intensity of noise n^{-1} is unknown. The high-dimensional situation becomes problematic, while in the infinite dimensional model one can base an estimate of n^{-1} on the observations $X_{l+1}, X_{l+2}, \dots, X_{l+p}$ for sufficiently large integers l, p, which are approximately $\mathcal{N}(0, n^{-1})$-distributed for large l. In fact, one can achieve an arbitrary precision in estimating n^{-1}, uniformly over some Sobolev ellipsoid $\Theta_\beta(Q)$ for some fixed $\beta, Q > 0$. In this chapter we however assume parameter n to be known and thus consider the more general infinite dimensional case $N \in \mathbb{N}$.

Remark 2.2. The above framework can be generalized as follows. Instead of $\{\widehat{\theta}(N),\, N \in \mathbb{N}\}$, we can consider the following more general estimators family $\{\widehat{\theta}(I_k),\, k \in \mathbb{N}\}$:

$$\widehat{\theta}(I_k) = \Pi_{S_{I_k}} X = (\widehat{\theta}_i(I_k))_{i\in\mathbb{N}} \quad \text{with} \quad \widehat{\theta}_i(I_k) = X_i I\{i \in I_k\},\ i \in \mathbb{N}.$$

the family of the projectors of the data on the subspaces $S_{I_k} = \{(s_i)_{i\in\mathbb{N}} : s_i \in \mathbb{R},\, i \in I_k;\, s_i = 0,\, i \notin I_k\}$. Here $\{I_k\}_{k\in\mathbb{N}}$ is a countable family of embedded subsets of $\mathbb{N}$: $I_k \subset \mathbb{N}$, $I_k \subset I_{k+1}$, $|I_k| = M_k$. Finite families can be considered as well. So the corresponding nested family of finite dimensional linear subspaces $\{S_{I_k}\}_{k\in\mathbb{N}}$ (i.e. $S_{I_k} \subset S_{I_{k+1}}$, $\dim(S_{I_k}) = M_k$, $k \in \mathbb{N}$) can be interpreted as a collection of finite dimensional models, where model S_{I_k} corresponds to the choice $\{\theta_i,\, i \in I_k\}$ as "most significant" variables. We can assume without loss of generality that $I_k = \{1, \dots, M_k\}$. Indeed, we can always rearrange the coordinates of signal θ in such a way that this holds. Next, notice that the particular case $M_k = k$ is exactly the one we study in this chapter. It is related to the so called ordered variable selection problem, cf. Birge and Massart (2001). However this case is most important since it corresponds to the most detailed slicing of the space ℓ_2 by a family $\{S_{I_k}\}_{k\in\mathbb{N}}$ of embedded pieces $S_{I_k} \subset S_{I_{k+1}}$, $k \in \mathbb{N}$, and thus the most difficult one to handle. Results and their proofs for any less detailed slicing case ($M_{k+1} > M_k + 1$ for some $k \in \mathbb{N}$) can be obtained along similar lines.

Another generalization concerns the high-dimensional case $\theta \in \mathbb{R}^n$ described in the previous remark. One can apply our approach to the situation corresponding to the complete variable selection, namely the family of linear subspaces $\{S_I,\, I \subseteq \mathcal{N}_n\}$, $S_I = \{(s_i)_{i\in\mathcal{N}_n} \in \mathbb{R}^n :\, s_i = 0, i \notin I\}$, $\mathcal{N}_n = \{1, \dots, n\}$. The conjecture is that there will be a price for the complete variable selection in the posterior convergence rate: as compared to the oracle risk, it is expected to be slower by a $\log n$ factor. This problem will be considered elsewhere.

Remark 2.3. As to the estimating the parameter μ in (2.7), instead of the empirical Bayes approach one could also use pure Bayesian approach by putting some prior on μ. We believe this approach should work as well, but its theoretical treatment seems to be more difficult since it adds one more hierarchy level to the Bayesian analysis.

Remark 2.4. Apart from the empirical Bayes approach, one can propose two more ways to fix the posterior: shifting the posterior and rescaling the posterior by an appropriate factor. By the first approach, construct an estimator $\widehat{\theta}$ by using the posterior $P(N|X)$ which converges to θ_0 with oracle rate and then shift the posterior $P(\theta|X)$ by the factor $\widehat{\theta}/2$ to make the resulting posterior distribution concentrate about the true value θ_0 with the oracle rate in P_{θ_0}-probability.

The second approach is based on rescaling the posterior distributions $P(\theta|X, N) = P_N(\theta|X)$ in (2.10) by factor 2. First consider nonadaptive case when the parameter N is fixed. Introduce the conditional distribution $P''_N(\theta|X)$:

$$\theta_i|X \overset{ind}{\sim} \mathcal{N}\Big(X_i I\{i \le N\}, 2n^{-1} I\{i \le N\}\Big), \quad i \in \mathbb{N}. \tag{2.14}$$

Notice that the distribution $P''_N(\theta|X)$ is simply the distribution of $2\theta|X$ with $\theta|X \sim P_N(\theta|X)$, i.e. a rescaled version of $P_N(\theta|X)$ defined by (2.6). Then this newly defined posterior distribution will move towards θ_0 in the sense that its Bayes estimator will converge to θ_0 and this posterior itself will concentrate about θ_0 with the rate $\mathcal{R}_n(N, \theta_0)$, although with somewhat bigger variances in this resulting product posterior distribution.

Coming back to the adaptive case when we put a prior on N, instead of posterior $P(\theta|X)$ defined by (2.10), we base the inference on the conditional distribution $P''(\theta|X)$, the rescaled version of $P(\theta|X)$ defined by (2.10) with the scale parameter 2:

$$P''(\theta \in S|X) = \sum_{k\in\mathbb{N}} P''(\theta \in S|X, N = k)P(N = k|X), \tag{2.15}$$

where the conditional distribution $P''(\theta|X, N) = P''_N(\theta|X)$ is defined by (2.14).

Remark 2.5. If one wants to avoid the randomization as in (2.13), then one could use an estimator of the form (2.13) but with a nonrandomized cut-off: $\widehat{\theta} = \widehat{\theta}(\widehat{N}_{\text{MAP}})$, with $\widehat{N}_{\text{MAP}} = \widehat{N}_{\text{MAP}}(X)$ being the so called maximum a posteriori probability (MAP) estimator, defined by

$$\widehat{N}_{\text{MAP}} = \widehat{N}_{\text{MAP}}(X) = \arg\max\{P(N = k|X),\, k \in \mathbb{N}\}, \tag{2.16}$$

i.e. $\widehat{N}_{\text{MAP}}$ is the mode of the posterior distribution $P(N = k|X)$. Since $\sum_{k\in\mathbb{N}} P(N = k|X) = 1$, the selector $\widehat{N}_{\text{MAP}}$ is well defined almost surely.

Remark 2.6. Another candidate for estimator $\widehat{\theta}$ of signal θ is possible: for example, $\widehat{\theta} = 2E(\theta|X)$, with the conditional expectation E taken with respect to the posterior (2.10). Factor 2 is due to the shrinkage phenomenon discussed above. Notice that $\widehat{\theta} = 2E(\theta|X) = E''(\theta|X)$, with the conditional expectation E'' taken with respect to the posterior (2.15). However, this estimator is very much alike the estimator (2.12). This stems from the fact that the posteriors $P'(\theta|X)$ and $P''(\theta|X)$ are very similar, with the only difference between them being a constant factor in the conditional variances $\text{Var}(\theta_i|X, N = k)$ which is $n^{-1}I\{i \le k\}$ for the posterior (2.11) and $2n^{-1}I\{i \le k\}$ for the posterior (2.15). Therefore we will not study the posterior (2.15) and the estimator $\widehat{\theta} = E''(\theta|X)$, as the derivation of their properties is exactly the same as for the posterior (2.11) and the estimator (2.12) respectively, with some modified constants in the proof.

2.2 Main results

This section contains the main results of the chapter. The proofs are deferred to the last section.

2.2.1 Oracle properties of the Bayes estimator and oracle posterior rate

The first theorem claims that the estimator (2.13) mimics the oracle, i.e. satisfies an oracle inequality of the form (1.3). This estimator can therefore be used as

a correcting shifting factor for the posterior distribution $P(\theta|X)$ later on.

Theorem 2.2. *Let $\theta_0 \in \ell_2$ and the prior π_α be defined by (2.9) with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$, and let the estimator $\widehat{\theta}$ be defined either by (2.12) or by (2.13). Then there exist constants $K_1, K_2 > 0$ depending only on α such that*

$$E_{\theta_0}\|\widehat{\theta} - \theta_0\|^2 \le K_1 \mathcal{R}_n(N_o, \theta_0) + \frac{K_2}{n}.$$

Similar oracle inequalities are obtained in Golubev (2004), in Cavalier and Tsybakov (2001) for blockwise Stein estimators and in Birge and Massart (2001) for penalized estimators. The results in those papers are in fact stronger in some respects. The above theorem shows that Bayes estimators with appropriately chosen prior satisfy some projection oracle inequalities as well, which makes it possible to use this estimator as a shifting factor for the posterior distribution $P(\theta|X)$ later on. Besides, the theorem says that the posterior-randomized selector $\widehat{N}$ defined by (2.13) mimics the oracle in the sense that $\widehat{\theta}(\widehat{N})$ satisfies the oracle inequality of the theorem.

The next result, which is the main result in the chapter, establishes that the posterior distributions $P(\theta|X)$ and $P'(\theta|X)$ concentrate about $\theta_0/2$ and θ_0 respectively in P_{θ_0}-probability with the posterior oracle projection rate $r_n(\theta_0)$ defined by (2.4), uniformly over $\theta_0 \in \ell_2$.

Theorem 2.3. *Let the oracle rate $r_n(\theta_0)$ be defined by (2.4), the prior π_α be defined by (2.9) with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$. Then there exist constants $C_1, C_2 > 0$ depending only on α such that, for any $\theta_0 \in \ell_2$ and any $M > 0$,*

$$E_{\theta_0} P\Big\{\big\|\theta - \theta_0/2\big\|^2 \ge M r_n(\theta_0)\,\big|X\Big\} \le \frac{C_1}{M},$$
$$E_{\theta_0} P'\Big\{\big\|\theta - \theta_0\big\|^2 \ge M r_n(\theta_0)\,\big|X\Big\} \le \frac{C_2}{M},$$

where posteriors $P(\theta|X)$ and $P'(\theta|X)$ are defined by (2.10) and (2.11) respectively.

Further, the parts of Theorems 2.2 and 2.3 dealing with the estimator (2.13) and the posterior (2.10) imply that an appropriately shifted posterior distribution $P(\theta|X)$ concentrates about θ_0 in P_{θ_0}-probability with the projection oracle posterior rate $r_n(\theta_0)$ defined by (2.4), uniformly over $\theta_0 \in \ell_2$.

Corollary 2.1. *Let the oracle rate $r_n(\theta_0)$ be defined by (2.4), the prior π_α be defined by (2.9) with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$, and the estimator $\widehat{\theta}$ be defined either by (2.12) or by (2.13). Then there exists a constant $C' > 0$ depending only on α such that for any $\theta_0 \in \ell_2$ and any $M > 0$*

$$E_{\theta_0} P\Big\{\big\|\theta - \theta_0 + \widehat{\theta}/2\big\|^2 \ge M r_n(\theta_0)\,\big|X\Big\} \le \frac{C'}{M}.$$

where posteriors $P(\theta|X)$ is defined by (2.10).

Remark 2.7. Note that the results of the both above theorems are uniform in $\theta_0 \in \ell_2$ and nonasymptotic, i.e. also uniform in $n \in \mathbb{N}$.

Remark 2.8. The results for the estimator $\widehat{\theta} = E''(\theta|X)$ and for the rescaled posterior $P''(\theta|X)$ defined by (2.15) follow in the same way as for the estimator $\widehat{\theta} = E'(\theta|X)$ and for the empirical Bayes posterior $P'(\theta|X)$ defined by (2.11) with slightly different constants.

Remark 2.9. Take any $M_n > 0$ such that $M_n \to \infty$ as $n \to \infty$. Then from Theorem 2.3 and Corollary 2.1 it follows that, under the conditions of Corollary 2.1,

$$P'\Big\{ \big\|\theta - \theta_0\big\|^2 \geq M_n r_n(\theta_0) \,\big|X\Big\} \to 0$$

and

$$P\Big\{ \big\|\theta - \theta_0 + \widehat{\theta}/2\big\|^2 \geq M_n r_n(\theta_0) \,\big|X\Big\} \to 0,$$

as $n \to \infty$ in P_{θ_0}-probability, uniformly over $\theta_0 \in \ell_2$.

Remark 2.10. As is already discussed, with the prior we use, we need either to shift or rescale the resulting posterior distribution to make it concentrate with the oracle rate about θ_0 rather than about $\theta_0/2$. The shifting factor is data driven and good enough to ensure that the concentration rate still remains the same, which is the projection oracle rate. For example, the shifting factor based on the Bayes estimator $\widehat{\theta}$ defined by (2.12) is optimal in the sense of projection oracle properties. So the results with a data-driven optimal shifting or rescaling (in our case the scale parameter 2 is fixed) of the posterior are as good as any other results about the oracle posterior convergence rate when the posterior is not adjusted. Sometimes, adjusting the posterior may be an intrinsic operation. For example, in the classical parametric statistics, the Bernstein-von Mises theorem is about an appropriately shifted posterior distribution. However, manipulating (for example, shifting or rescaling) the posterior distribution after a Bayesian analysis is not in the core of the traditional Bayesian paradigm. The empirical Bayes approach for adjusting the posterior is on the contrary well established and accepted in the Bayesian community. These remarks though are more of a philosophical, methodological nature rather than mathematical.

Remark 2.11. The obtained results establish in some sense the correct behavior of the posterior distributions of θ and N from the P_{θ_0}-perspective, which seems to make it possible to use them in solving another interesting challenging problem – the construction of an adaptive confidence set for θ_0 by using these posteriors. This is a subtle issue to address and will be considered elsewhere.

2.2.2 Bayesian model selection, assessing the false selection probability

The next result concerns the model selection problem. For a $\tau > 0$, $\theta_0 \in \ell_2$ and the oracle $N_o = N_o(\theta_0)$ defined by (2.4), introduce the sets:

$$\begin{aligned} \mathcal{N}(\tau) &= \mathcal{N}(\tau, \theta_0) = \big\{k \in \mathbb{N} : \mathcal{R}_n(k, \theta_0) \le \tau \mathcal{R}_n(N_o, \theta_0)\big\}, \qquad (2.17) \\ \mathcal{N}^-(\tau) &= \mathcal{N}^-(\tau, \theta_0) \\ &= \big\{k \in \mathbb{N} : k < N_o,\, \mathcal{R}_n(k, \theta_0) > \tau \mathcal{R}_n(N_o, \theta_0)\big\}, \qquad (2.18) \\ \mathcal{N}^+(\tau) &= \mathcal{N}^+(\tau, \theta_0) \\ &= \big\{k \in \mathbb{N} : k > N_o,\, \mathcal{R}_n(k, \theta_0) > \tau \mathcal{R}_n(N_o, \theta_0)\big\}. \qquad (2.19) \end{aligned}$$

For any $\tau \ge 1$, these sets form a partition of $\mathbb{N}$, i.e. they do not intersect and $\mathbb{N} = \mathcal{N}^-(\tau) \cup \mathcal{N}(\tau) \cup \mathcal{N}^+(\tau)$. From now on, we deal only with $\tau \ge 1$.

The interpretation of the set $\mathcal{N}(\tau, \theta_0)$ is clear – it specifies the set of acceptable models in the sense that the risks $\mathcal{R}_n(k, \theta_0)$ for all models $k \in \mathcal{N}(\tau, \theta_0)$ are all within the constant factor τ of the oracle risk $\mathcal{R}_n(N_o, \theta_0)$. We call this constant τ tolerance parameter. For a model selector $\tilde{N} = \tilde{N}(X) \in \mathbb{N}$ and a tolerance $\tau \ge 1$, we define our quality measure of model selectors - the false selection probability (FSP):

$$\mathrm{FSP}(\tilde{N}, \tau, \theta_0) = \mathrm{FSP}(\tilde{N}, \tau, \theta_0, n) = P_{\theta_0}\big\{\tilde{N} \notin \mathcal{N}(\tau, \theta_0)\big\}, \qquad (2.20)$$

with $\mathcal{N}(\tau, \theta_0)$ defined by (2.17).

Theorem 2.4. *Let the model selector $\widehat{N}(X, \alpha)$ be defined by (2.13). Then, for any $\alpha \in [\frac{1}{6} - \log\big(\frac{2}{\sqrt{3}}\big), \frac{1}{2}]$ and τ such that $\tau > \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha$ and $\tau \ge \tau_+(\alpha) = \frac{1}{2\alpha}$, there exist constants $B_1, B_2, B_3 > 0$ depending only on α and τ such that for all $\theta_0 \in \ell_2$*

$$\begin{aligned} \mathit{FSP}(\widehat{N}, \tau, \theta_0) &= P_{\theta_0}\big\{\widehat{N} \notin \mathcal{N}(\tau, \theta_0)\big\} \\ &\le \min\Big\{\frac{B_1}{n\mathcal{R}_n(N_o, \theta_0)}, \frac{B_2 n\mathcal{R}_n(N_o, \theta_0)}{e^{B_3 n\mathcal{R}_n(N_o, \theta_0)}}\Big\}. \end{aligned}$$

Remark 2.12. In Theorem 2.4, the constants B_1 and B_2 decrease and B_3 increases as tolerance τ becomes bigger. The exact relations are given in the proof of the theorem.

If the oracle quantity $n\mathcal{R}_n(N_o, \theta_0)$ is large enough, it is easily seen from the above theorem that the false selection probability for the selector $\widehat{N}$ is small, i.e. $\widehat{N}$ selects a "good" model (from $\mathcal{N}(\tau, \theta_0)$) with high probability. The oracle quantity $n\mathcal{R}_n(N_o, \theta_0) \ge K$ if there are at least K coordinates in vector θ_0 which are not less than n^{-1}. If $n\mathcal{R}_n(N_o, \theta_0)$ is small, θ_0 is too close to the zero signal for the selector $\widehat{N}$ to perform well. However, according to the above remark, even for such θ_0's, the method does a good job if we lower our tolerance requirement: making tolerance parameter τ bigger leads to a smaller false selection probability.

Another possibility is the so called maximum a posteriori probability (MAP) model selector $\widehat{N}_{\mathrm{MAP}}(X)$, the mode of the posterior distribution $P(N=k|X)$, defined by (2.16). Recall that it also depends also on the parameter α. It is expected that this selector lives on the set $\mathcal{N}(\tau,\theta_0)$, at least for "good" θ_0's. The following assertion elaborates on this.

Theorem 2.5. *Let the model selector $\widehat{N}_{MAP}(X)$ be defined by (2.16). Then, under the conditions of Theorem 2.4 and for all $\theta_0 \in \ell_2$,*

$$\begin{aligned} FSP(\widehat{N}_{MAP},\tau,\theta_0) &= P_{\theta_0}\big\{\widehat{N}_{MAP} \notin \mathcal{N}(\tau,\theta_0)\big\} \\ &\le (1+|\mathcal{N}(\tau,\theta_0)|^{1/2})^2 P_{\theta_0}\big\{\widehat{N} \notin \mathcal{N}(\tau,\theta_0)\big\} \\ &\le \big(1+\sqrt{\tau n\mathcal{R}_n(N_o,\theta_0)}\,\big)^2 \min\Big\{\frac{B_1}{n\mathcal{R}_n(N_o,\theta_0)}, \frac{B_2 n\mathcal{R}_n(N_o,\theta_0)}{e^{B_3 n\mathcal{R}_n(N_o,\theta_0)}}\Big\}. \end{aligned}$$

Although the above result claims somewhat worse properties for the MAP selector $\widehat{N}_{\mathrm{MAP}}$ as compared to the posterior randomized selector $\hat{N}$, it seems actually that the MAP selector should have at least the same quality as $\hat{N}$, it does perform better in simulations. So, it is more the method of the proof of the properties for the MAP selector, that is not precise enough, rather than the selector itself.

Remark 2.13. In principle, we are interested in the smallest value from the set $\mathcal{N}(\tau)$ in order to select the model with the smallest dimension. So, we can correct our MAP selector slightly to incorporate this requirement. For some positive sequence $\delta = \delta_n$, let

$$\hat{N}_\delta = \min\big\{k \in \mathbb{N}:\ P(N=k|X) \ge \delta_n\big\},$$

with the convention that $\min\{\varnothing\} = \infty$. Then we can take the selector

$$\tilde{N} = \min\{\widehat{N}_{\mathrm{MAP}}, \hat{N}_\delta\}.$$

However, it is not clear what would be an appropriate choice for the sequence δ_n. From the proof of the above theorem, it seems that the sequence δ_n should be of the order $(n\mathcal{R}_n(N_o,\theta_0))^{-1}$. The oracle risk is however unknown, so in practice one should take its reasonable data driven empirical analog, say, $\hat{\delta}_n = c\widehat{N}^{-1}_{\mathrm{MAP}}$.

Remark 2.14. Suppose an estimators family $\widehat{\Theta}(\mathcal{N})$ covers some scale $\{\Theta_\beta,\ \beta \in \mathcal{B}\}$ in the sense we discussed in the general introduction, so that the estimator $\widehat{\theta}(N_\beta)$ is minimax over Θ_β. Then we can relate the above model selection problem to the problem of selecting the smoothness parameter $\beta \in \mathcal{B}$ in case we assume that the signal θ is from the scale $\{\Theta_\beta,\ \beta \in \mathcal{B}\}$. Suppose the above correspondence $\beta \to N_\beta$ can be inverted: $N \to \beta_N$, $N \in \mathcal{N}$. Namely, for a given scale of models $\{\Theta_\beta, \beta \in \mathcal{B}\}$ parameterized by $\mathcal{B}$, we select a model $\widehat{\beta} = \beta_{\widehat{N}}$, with $\widehat{N}$ constructed by using $P(N|X)$. If our selector $\widehat{N}$ satisfies (1.3), the model $\widehat{\beta}$ is then close to the best model in terms of the risk function $R(\widehat{\theta}(N_\beta),\theta)$, i.e. the oracle model $\beta_o = \beta_{N_o}$ for which

$$\inf_{\beta\in\mathcal{B}} R(\widehat{\theta}(N_\beta),\theta) = \inf_{N\in\mathcal{N}} R(\widehat{\theta}(N),\theta) = R(\widehat{\theta}(N_o),\theta).$$

For a example, in case of the Sobolev ellipsoids scale described in the next section $N_\beta = \lfloor cn^{1/(2\beta+1)} \rfloor$, so that $\widehat{\beta} = \frac{\log n}{2\log(\widehat{N}/c)} - \frac{1}{2}$ can be taken to be a smoothness selector.

2.3 Posterior rate: minimax versus oracle

In this section we discuss the relation between the oracle over the family $\widehat{\Theta}(\mathcal{N})$ and minimax over the scale $\{\Theta_\beta,\, \beta \in \mathcal{B}\}$ approaches to the optimality. The two main factors are the estimators family $\widehat{\Theta}(\mathcal{N})$ and the scale $\{\Theta_\beta,\, \beta \in \mathcal{B}\}$. Other important factors are whether there exists an estimator $\widehat{\theta}$ satisfying the oracle inequality (1.3) and how big the set Θ_0 is for which this oracle inequality holds.

As is already mentioned in the introduction, to make the notion of oracle estimator sensible, the family $\widehat{\Theta}(\mathcal{N})$ should not be too poor.This is also discussed in Golubev (2004). Indeed, on the one hand, we want this family to contain some "good" estimators about which we know that they perform well over nonparametric classes Θ_β, $\beta \in \mathcal{B}$. Suppose the family $\widehat{\Theta}(\mathcal{N})$ is sufficiently rich to contain the minimax estimators over all Θ_β, $\beta \in \mathcal{B}$, i.e. for any $\beta \in \mathcal{B}$, there exists an $N_\beta \in \mathcal{N}$ such that $\widehat{\theta}(N_\beta)$ is minimax over the class Θ_β (the family $\widehat{\Theta}(\mathcal{N})$ covers the scale $\{\Theta_\beta,\, \beta \in \mathcal{B}\}$):

$$\sup_{\theta \in \Theta_\beta} R_n(\widehat{\theta}(N_\beta), \theta) \le C'_n r_n(\Theta_\beta),$$

ideally with $C'_n = 1 + o(1)$ as $n \to \infty$, otherwise with some uniform constant $1 \le C'_n = C' < \infty$. This implies that

$$\inf_{N \in \mathcal{N}} \sup_{\theta \in \Theta_\beta} R(\widehat{\theta}(N), \theta) \le C'_n r(\Theta_\beta)$$

for all $\beta \in \mathcal{B}$. Then certainly the oracle approach is stronger as the oracle risk at each point $\theta \in \Theta_\beta$ can only be smaller than a multiple of the minimax risk $r_n(\Theta_\beta)$.

The same reasoning applies to the results about posterior convergence rate. If the family $\widehat{\Theta}(\mathcal{N})$ covers the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$, then the results on oracle posterior convergence rate of the form (1.6) are stronger than the results on adaptive minimax convergence rate of the posterior simply because the oracle rate can only be smaller or the same up to a multiple than the corresponding minimax risk over Θ_β, at least for all $\theta_0 \in \Theta_0 \cap \big(\cup_{\beta \in \mathcal{B}} \Theta_\beta \big)$. Recall that (1.6) holds for $\theta_0 \in \Theta_0$, so it is desirable that $\cup_{\beta \in \mathcal{B}} \Theta_\beta \subseteq \Theta_0$. In our case, $\widehat{\Theta}(\mathcal{N})$ is the class of projection estimators and $\Theta_0 = \ell_2$, so that our results on oracle projection posterior convergence are always stronger than all the adaptation results about the posterior convergence with the minimax rate, simultaneously over all the scales where the corresponding minimax rates are attained by projection estimators. Besides, the uniformity of all the results on adaptive minimax convergence rate of the posterior over all these scales follows immediately from the uniformity of our results over the whole space ℓ_2.

Summarizing, once we establish that the projection estimators family $\widehat{\Theta}(\mathcal{N})$ defined by (2.2) covers the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$, i.e. the minimax rate over Θ_β for each $\beta \in \mathcal{B}$ is attained by a projection estimator from $\widehat{\Theta}(\mathcal{N})$, then, by applying Theorem 2.3, we immediately derive the following result.

Theorem 2.6. *Let the projection estimators class $\widehat{\Theta}(\mathcal{N})$ defined by (2.2) cover the scale $\{\Theta_\beta, \beta \in \mathcal{B}\}$, let $r_n(\Theta_\beta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_\beta} R_n(\widehat{\theta}, \theta)$ be the minimax risk over the class Θ_β and let the prior π_α be defined by (2.9) with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$. Then there exist constants $C_1, C_2 > 0$ depending only on α and Θ_β such that for any $M > 0$,*

$$E_{\theta_0} P\Big\{ \big\|\theta - \theta_0/2\big\|^2 \ge M r_n(\Theta_\beta) \,\big|X\Big\} \le \frac{C_1}{M},$$
$$E_{\theta_0} P'\Big\{ \big\|\theta - \theta_0\big\|^2 \ge M r_n(\Theta_\beta) \,\big|X\Big\} \le \frac{C_2}{M},$$

uniformly over $\theta_0 \in \Theta_\beta$, where the posteriors $P(\theta|X)$ and $P'(\theta|X)$ are defined by (2.10) and (2.11) respectively.

As a consequence we derive that the posterior $P'(\theta|X)$ and the appropriately shifted posterior $P(\theta|X)$ will concentrate about θ_0 uniformly over $\theta_0 \in \Theta_\beta$.

Corollary 2.2. *Under the conditions of Corollary 2.1, for any $M_n > 0$ such that $M_n \to \infty$*

$$P'\Big\{ \big\|\theta - \theta_0\big\|^2 \ge M_n r_n(\Theta_\beta) \,\big|X\Big\} \to 0$$

and

$$P\Big\{ \big\|\theta - \theta_0 + \widehat{\theta}/2\big\|^2 \ge M_n r_n(\Theta_\beta) \,\big|X\Big\} \to 0,$$

as $n \to \infty$ in P_{θ_0}-probability, uniformly over $\theta_0 \in \Theta_\beta$.

Consider a couple of examples of nonparametric scales $\{\Theta_\beta, \beta \in \mathcal{B}\}$ for which the minimax rate is attained by a projection estimator. Denote $\lfloor a \rfloor = \max\{k \in \mathbb{Z},\, k \le a\}$ for $a \in \mathbb{R}$.

Sobolev ellipsoids. Consider the Sobolev ellipsoids

$$\Theta_\beta(Q) = \Big\{\theta : \sum_{i=1}^{\infty} i^{2\beta} \theta_i^2 \le Q\Big\}, \quad \beta > 0.$$

In Belitser and Ghosal (2003) a result on the posterior convergence with the minimax rate for the Sobolev ellipsoid $\Theta_\beta(Q)$ is given. It is well known that the corresponding minimax rate is $r_n(\beta) = n^{-2\beta/(2\beta+1)}$; see, for example, Pinsker (1980) or Belitser and Levit (1995). The uniformity of the main claim in Belitser and Ghosal (2003) was obtained only for sufficiently "small" ellipsoids, although it could have been established for the original ellipsoid $\Theta_\beta(Q)$ as well with somewhat more careful analysis. The prior on θ in Belitser and Ghosal (2003) was different and based on putting a joint prior on the pair (θ, β) rather than

on (θ, N), to model the unknown smoothness parameter β. However, it is easy to see that the projection estimator $\widehat{\theta}(N_\beta)$ defined by (2.2), with $N_\beta = \lfloor cn^{1/(2\beta+1)} \rfloor$, is minimax with respect to the convergence rate:

$$\begin{aligned}\sup_{\theta\in\Theta_\beta(Q)} R(\widehat{\theta}(N_\beta),\theta) &= \sup_{\theta\in\Theta_\beta(Q)} \Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^{\infty} \theta_i^2\Big) \\ &\le \frac{N_\beta}{n} + \sup_{\theta\in\Theta_\beta(Q)} \Big(\sum_{i=N_\beta+1}^{\infty} \frac{\theta_i^2 i^{2\beta}}{N_\beta^{2\beta}}\Big) \\ &\le \frac{N_\beta}{n} + \frac{Q}{N^{2\beta}} = Cn^{-2\beta/(2\beta+1)}.\end{aligned}$$

Then Theorem 2.6 implies that the appropriately shifted posterior distribution $P(\theta|X)$ and the posterior $P'(\theta|X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = n^{-2\beta/(2\beta+1)}$ uniformly over $\theta_0 \in \Theta_\beta(Q)$. Thus our (oracle posterior convergence rate) results are stronger than those (adaptive minimax posterior convergence rate) in Belitser and Ghosal (2003).

Exponential ellipsoids. Suppose our nonparametric class $\Theta_\beta(Q)$ is from the scale of the exponential ellipsoids:

$$\Theta_\beta(Q) = \Big\{\theta : \sum_{k=1}^{\infty} e^{2\beta k}\theta_k^2 \le Q\Big\}, \quad \beta > 0.$$

By the same arguments we derive that the projection estimator $\widehat{\theta}(N)$ defined by (2.2), with $N_\beta = \lfloor \log n/(2\beta) \rfloor$, is minimax with respect to the convergence rate for the exponential ellipsoid $\Theta_\beta(Q)$. Indeed, in this case

$$\begin{aligned}\sup_{\theta\in\Theta_\beta(Q)} R(\widehat{\theta}(N_\beta),\theta) &= \sup_{\theta\in\Theta_\beta(Q)} \Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^{\infty} \theta_i^2\Big) \\ &\le \frac{N_\beta}{n} + \sup_{\theta\in\Theta_\beta(Q)} \Big(\sum_{k=N_\beta+1}^{\infty} \frac{\theta_k^2 e^{2\beta k}}{e^{2\beta N_\beta}}\Big) \\ &\le \frac{\log n}{2\beta n} + \frac{Q}{n} \le \frac{C\log n}{n},\end{aligned}$$

which is of the same order as the minimax rate $r_n(\Theta_\beta(Q)) = \log n/n$ over the exponential ellipsoid $\Theta_\beta(Q)$, cf. Belitser and Levit (1995). Thus, according to Theorem 2.6, the appropriately shifted posterior distribution $P(\theta|X)$ and the posterior $P'(\theta|X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = \log n/n$ uniformly over $\theta_0 \in \Theta_\beta(Q)$.

Hyperrectangles. Consider the so called hyperrectangles in ℓ_2:

$$\Theta_\beta(Q) = \Big\{\theta : |\theta_k| \le \sqrt{Q}k^{-\beta},\, k \in \mathbb{N}\Big\}, \quad \beta > 1/2.$$

It is not difficult to show that the minimax convergence rate over this class is $r_n(\Theta_\beta(Q)) = n^{-(2\beta-1)/(2\beta)}$. We derive that the projection estimator $\widehat{\theta}(N_\beta)$ defined by (2.2), with $N_\beta = \lfloor cn^{1/(2\beta)} \rfloor$, is minimax with respect to the convergence rate for the hyperrectangle $\Theta_\beta(Q)$. Indeed, in this case

$$\begin{aligned}\sup_{\theta\in\Theta_\beta(Q)} R(\widehat{\theta}(N_\beta),\theta) &= \sup_{\theta\in\Theta_\beta(Q)} \Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^{\infty} \theta_i^2\Big) \le \frac{N_\beta}{n} + \sum_{k=N_\beta+1}^{\infty} \frac{Q}{k^{2\beta}} \\ &\le \frac{N_\beta}{n} + \frac{Q}{(2\beta-1)N_\beta^{2\beta-1}} \le Cn^{-(2\beta-1)/(2\beta)},\end{aligned}$$

which is of the same order as the minimax rate over the hyperrectangle $\Theta_\beta(Q)$ $r_n(\beta) = n^{-(2\beta-1)/(2\beta)}$. Theorem 2.6 implies that the appropriately shifted posterior distribution $P(\theta|X)$ and the posterior $P'(\theta|X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = n^{-(2\beta-1)/(2\beta)}$ uniformly over $\theta_0 \in \Theta_\beta(Q)$.

Tail classes. Finally consider the so called tail classes:

$$\Theta_\beta(Q) = \Big\{\theta : \sum_{k=m}^{\infty} \theta_k^2 \le Qm^{-\beta},\, m \in \mathbb{N}\Big\}, \quad \beta > 0.$$

Since hyperrectangle with parameters $\beta' > 1/2$ and Q' can be embedded into a tail class with parameters $\beta' = 2\beta - 1$ and some $Q = Q(\beta, Q')$, we obtain that the minimax risk over the tail class $\Theta_\beta(Q)$ is at least $r_n(\beta) = n^{-\beta/(1+\beta)}$. Now we derive that the projection estimator $\widehat{\theta}(N)$ defined by (2.2), with $N_\beta = \lfloor cn^{1/(1+\beta)} \rfloor$, is minimax with respect to the convergence rate for the tail class $\Theta_\beta(Q)$. Indeed, in this case

$$\begin{aligned}\sup_{\theta\in\Theta_\beta(Q)} R(\widehat{\theta}(N_\beta),\theta) &= \sup_{\theta\in\Theta_\beta(Q)} \Big(\frac{N_\beta}{n} + \sum_{i=N_\beta+1}^{\infty} \theta_i^2\Big) \\ &\le \frac{N_\beta}{n} + \frac{Q}{(N_\beta+1)^\beta} \\ &\le Cn^{-\beta/(1+\beta)},\end{aligned}$$

which is of the same order as the minimax rate $r_n(\beta) = n^{-\beta/(1+\beta)}$ over the tail class $\Theta_\beta(Q)$. Therefore, Theorem 2.6 implies that the appropriately shifted posterior distribution $P(\theta|X)$ and the posterior $P'(\theta|X)$ both concentrate in P_{θ_0}-probability about the true signal θ_0 with at least the minimax rate $r_n(\Theta_\beta(Q)) = n^{-\beta/(1+\beta)}$ uniformly over $\theta_0 \in \Theta_\beta(Q)$.

2.4 Technical results

In this section we provide a couple of technical lemmas used in the proofs of the main results.

Lemma 2.1. *Let $\theta_0 \in \ell_2$, $\tau > 0$ and let $\widehat{\theta}(k)$, $k \in \mathbb{N}$, be defined by (2.2), $\widehat{N}$ by (2.13) and the set $\mathcal{N}(\tau)$ by (2.17). Then*

$$E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\}\Big] \le 2\tau \mathcal{R}_n(N_o, \theta_0),$$

$$E_{\theta_0}\Big[\sum_{k \in \mathcal{N}(\tau)} \|\widehat{\theta}(k) - \theta_0\|^2 P(N = k|X)\Big] \le 2\tau \mathcal{R}_n(N_o, \theta_0).$$

Proof. Recall that $\xi_i = \sqrt{n}(X_i - \theta_0) \overset{ind}{\sim} \mathcal{N}(0,1)$, $i \in \mathbb{N}$, under $X \sim P_{\theta_0}$. Write

$$\begin{aligned} & E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\}\Big] \\ &= E_{\theta_0}\Big[\Big(\sum_{i=1}^{\widehat{N}} \frac{\xi_i^2}{n} + \sum_{i=\widehat{N}+1}^{\infty} \theta_{0i}^2\Big) I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\}\Big] \\ &= E_{\theta_0}\Big[\Big(\sum_{i=1}^{\widehat{N}} \frac{\xi_i^2}{n}\Big) I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\}\Big] \\ &\quad + E_{\theta_0}\Big[\Big(\sum_{i=\widehat{N}+1}^{\infty} \theta_{0i}^2\Big) I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\}\Big]. \end{aligned} \tag{2.21}$$

Now we bound the both terms in the right hand side of (2.21) by $\tau \mathcal{R}_n(N_o, \theta_0)$. Since $\mathcal{R}_n(k, \theta_0) = \frac{k}{n} + \sum_{i=k+1}^{\infty} \theta_{0i}^2 \le \tau \mathcal{R}_n(N_o, \theta_0)$ for any $k \in \mathcal{N}(\tau)$, we obtain

$$N_{max} = \max\{\mathcal{N}(\tau)\} \le \tau n \mathcal{R}_n(N_o, \theta_0).$$

This implies the bound for the first term in the right hand side of (2.21):

$$\begin{aligned} E_{\theta_0}\Big[\Big(\sum_{i=1}^{\widehat{N}} \frac{\xi_i^2}{n}\Big) I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\}\Big] &\le E_{\theta_0} \sum_{i=1}^{N_{max}} \frac{\xi_i^2}{n} \\ &= \frac{N_{max}}{n} \le \tau \mathcal{R}_n(N_o, \theta_0). \end{aligned}$$

Finally, we evaluate the second term in the right hand side of (2.21) as follows:

$$\begin{aligned} E_{\theta_0}\Big[\Big(\sum_{i=\widehat{N}+1}^{\infty} \theta_{0i}^2\Big) I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\}\Big] &= E_{\theta_0}\Big[\sum_{k \in \mathcal{N}(\tau)} \Big(\sum_{i=k+1}^{\infty} \theta_{0i}^2\Big) I\{\widehat{N} = k\}\Big] \\ &= \sum_{k \in \mathcal{N}(\tau)} \Big(\sum_{i=k+1}^{\infty} \theta_{0i}^2\Big) E_{\theta_0} I\{\widehat{N} = k\} \\ &\le \sum_{k \in \mathcal{N}(\tau)} \mathcal{R}_n(k, \theta_0) P_{\theta_0}\{\widehat{N} = k\} \\ &\le \tau \mathcal{R}_n(N_o, \theta_0), \end{aligned}$$

since $\mathcal{R}_n(k, \theta_0) \le \tau \mathcal{R}_n(N_o, \theta_0)$ for all $k \in \mathcal{N}(\tau)$. Combining the last two relations with (2.21) completes the proof of the first inequality. The second inequality follows similarly. □

Lemma 2.2. *Let $\widehat{\theta}(k)$, $k \in \mathbb{N}$, be defined by (2.2) and $\widehat{N}$ by (2.13). Then for any $\theta_0 \in \ell_2$ and any $k \in \mathbb{N}$,*

$$E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 I\big\{\widehat{N} = k\big\}\Big] \le \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\big(E_{\theta_0} P(N = k|X)\big)^{1/2},$$

$$E_{\theta_0}\Big[\|\widehat{\theta}(k) - \theta_0\|^2 P(N = k|X)\Big] \le \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\big(E_{\theta_0} P(N = k|X)\big)^{1/2}.$$

Proof. The following fact is well known. Let $Z_1, \dots, Z_k$ be iid $\mathcal{N}(0, \sigma^2)$ random variables, then

$$\begin{aligned}\Big[E\Big(\sum_{i=1}^k Z_i^2\Big)^2\Big]^{\frac{1}{2}} &= \Big[E\Big(\sum_{i=1}^k Z_i^4 + \sum_{i\ne j} Z_i^2 Z_j^2\Big)\Big]^{\frac{1}{2}} \\ &= \big(3k\sigma^4 + (k^2 - k)\sigma^4\big)^{1/2} \\ &\le (k+1)\sigma^2.\end{aligned}$$

Applying this fact and Cauchy-Schwarz inequality, we evaluate

$$\begin{aligned}&E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 I\big\{\widehat{N} = k\big\}\Big] = E_{\theta_0}\Big[\sum_{i=1}^\infty (\widehat{\theta}_i(k) - \theta_{0i})^2 I\big\{\widehat{N} = k\big\}\Big] \\ &= E_{\theta_0}\Big[\Big(\sum_{i=1}^k \frac{\xi_i^2}{n} + \sum_{i=k+1}^\infty \theta_{0i}^2\Big) I\big\{\widehat{N} = k\big\}\Big] \\ &\le \Big[E_{\theta_0}\Big(\sum_{i=1}^k \frac{\xi_i^2}{n}\Big)^2\Big]^{1/2}\Big[E_{\theta_0} I\big\{\widehat{N} = k\big\}\Big]^{1/2} + \Big[\sum_{i=k+1}^\infty \theta_{0i}^2\Big] E_{\theta_0} I\big\{\widehat{N} = k\big\} \\ &\le \frac{k+1}{n}\Big[E_{\theta_0} P(N = k|X)\Big]^{1/2} + \Big[\sum_{i=k+1}^\infty \theta_{0i}^2\Big] E_{\theta_0} P(N = k|X) \qquad (2.22) \\ &\le \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\big(E_{\theta_0} P(N = k|X)\big)^{1/2},\end{aligned}$$

since $\widehat{N}$ is generated from the conditional distribution $P(N|X)$. The second inequality follows similarly. □

Remark 2.15. In fact, we proved a slightly stronger assertion (2.22) as compared to the final statement of the above lemma. Besides, summing up over all values of $\widehat{N} = k$, $k \in \mathbb{N}$, we derive that, for any $\theta_0 \in \ell_2$,

$$E_{\theta_0}\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 \le \sum_{k\in\mathbb{N}} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\big(E_{\theta_0} P(\widehat{N} = k|X)\big)^{1/2}.$$

Recall our Bayesian scheme:

$$X|(\theta, N) \sim P_\theta, \quad \theta|N \sim \pi_N, \quad N \sim \lambda.$$

Two-level hierarchical prior on (θ, N) leads to a joint distribution on (X, θ, N), which in turn gives rise to the conditional marginal $P(X|N)$ and the posterior

distributions $P(\theta|X)$ and $P(N|X)$, at which we can look from the perspective of P_{θ_0}-measure of X. Since $P(X|\theta, N)$, $P(\theta|N)$ are both the products of normals, it is easy to derive due to the conjugacy that the conditional marginal $P(X|N)$ is also the product of normals $\mathcal{N}(0, \tau_i^2(N) + n^{-1})$.

Define

$$a_i(k) = (\tau_i^2(k) + n^{-1})^{-1}, \quad a_i(k, k_0) = a_i(k) - a_i(k_0), \quad k, k_0, i \in \mathbb{N}, \tag{2.23}$$

where $\tau_i^2(k)$ is defined by (2.5). The following Lemma concerns the posterior distribution $P(N|X)$.

Lemma 2.3. *For any $k, k_0 \in \mathbb{N}$ and any $\theta_0 \in \ell^2$,*

$$\begin{aligned} &E_{\theta_0} P(N = k|X) \\ &\quad \leq \frac{\lambda_k}{\lambda_{k_0}} \prod_{i=1}^{\infty} \Big[\frac{a_i(k)}{a_i(k_0)(1 + n^{-1} a_i(k, k_0))} \Big]^{1/2} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{\infty} \frac{a_i(k, k_0)\theta_{0i}^2}{1 + n^{-1} a_i(k, k_0)} \Big\}, \end{aligned}$$

where $a_i(k)$ and $a_i(k, k_0)$ are defined by (2.23).

Proof. Applying the martingale convergence theorem and the dominated convergence theorem in the same manner as in the proof of Lemma 2 of [12], we have that

$$E_{\theta_0} P(N = k|X) = \lim_{m \to \infty} E_{\theta_0} P(N = k|X_1, \ldots, X_m). \tag{2.24}$$

Since the conditional marginal $P(X_1, \ldots, X_m|N = k)$ is the product of normal distributions $\mathcal{N}(0, \tau_i^2(k) + n^{-1})$, $i = 1, \ldots, m$, with densities $f_i(x_i|N = k)$ respectively, it is not difficult to compute the posterior probability

$$\begin{aligned} &P(N = k|X_1, \ldots, X_m) \\ &\quad = \frac{\prod_{i=1}^m f_i(X_i|N = k) P(N = k)}{\sum_{l \in \mathbb{N}} \prod_{i=1}^m f_i(X_i|N = l) P(N = l)} \\ &\quad = \frac{\lambda_k \prod_{i=1}^m \frac{1}{\sqrt{(\tau_i^2(k) + n^{-1})}} \exp\Big\{ -\dfrac{X_i^2}{2(\tau_i^2(k) + n^{-1})} \Big\}}{\sum_{l \in \mathbb{N}} \lambda_l \prod_{i=1}^m \frac{1}{\sqrt{(\tau_i^2(l) + n^{-1})}} \exp\Big\{ -\dfrac{X_i^2}{2(\tau_i^2(l) + n^{-1})} \Big\}}. \end{aligned}$$

Obviously

$$\begin{aligned} &E_{\theta_0} P(N = k|X_1, X_2, \ldots, X_m) \\ &\quad \leq \frac{\lambda_k}{\lambda_{k_0}} \prod_{i=1}^m \sqrt{\frac{a_i(k)}{a_i(k_0)}}\, E_{\theta_0} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^m X_i^2 a_i(k, k_0) \Big\}. \end{aligned}$$

Using this and the elementary identity

$$E\big(\exp\big[-\frac{b}{2} Y^2 \big] \big) = \frac{1}{\sqrt{1 + b\sigma^2}} \exp\big[-\frac{\mu^2 b}{2(1 + b\sigma^2)} \big]$$

for $Y \sim \mathcal{N}(\mu, \sigma^2)$ and $b > -\sigma^{-2}$ (under P_{θ_0}, $X_i \overset{ind}{\sim} \mathcal{N}(\theta_{0i}, n^{-1})$, $i \in \mathbb{N}$), we derive

$$E_{\theta_0} P(N = k | X_1, X_2, \ldots, X_m)$$
$$\leq \frac{\lambda_k}{\lambda_{k_0}} \prod_{i=1}^{m} \Big[\frac{a_i(k)}{a_i(k_0)(1 + n^{-1} a_i(k, k_0))} \Big]^{1/2} \exp \Big\{ -\frac{1}{2} \sum_{i=1}^{m} \frac{a_i(k, k_0) \theta_{0i}^2}{1 + n^{-1} a_i(k, k_0)} \Big\}.$$

Combining the last relation with (2.24) completes the proof of the lemma. □

Note that the above lemma holds for any prior variances $\tau_i^2(k)$'s and any $k_0 \in \mathbb{N}$, including $k_0 = N_o$, the oracle cut-off defined by (2.4). Taking $k_0 = N_o$ and $\tau_i^2(k)$ defined by (2.5), we obtain the following corollary.

Corollary 2.3. *Let $\tau_i^2(k)$, $k \in \mathbb{N}$, be defined by (2.5).*

- *If $k < N_o$, then*

$$E_{\theta_0} \big[P(N = k | X) \big] \leq \frac{\lambda_k}{\lambda_{N_o}} \exp \Big\{ -\frac{n}{6} \Big(\sum_{i=k+1}^{N_o} \theta_{0i}^2 - \frac{6 \log \big(\frac{2}{\sqrt{3}}\big)(N_o - k)}{n} \Big) \Big\}.$$

- *If $k > N_o$, then*

$$E_{\theta_0} \big[P(N = k | X) \big] \leq \frac{\lambda_k}{\lambda_{N_o}} \exp \Big\{ \frac{n}{2} \sum_{i=N_o+1}^{k} \theta_{0i}^2 \Big\}.$$

Proof. Indeed, using relations (2.23) and (2.5), we compute

$$\begin{aligned} a_i(k) &= \frac{n}{2} I\{i \leq k\} + n I\{i > k\}, \\ a_i(k, k_0) &= -\frac{n}{2} I\{k_0 < i \leq k\} + \frac{n}{2} I\{k < i \leq k_0\} \end{aligned}$$

and substitute these values in the right hand side of the inequality of Lemma 2.3. □

Notice that in case $k = N_o$ we obtain a trivial useless bound $E_{\theta_0} \big[P(N = N_o | X) \big] \leq 1$.

Lemma 2.4. *Let the prior π_α be defined by (2.9) with $\alpha \in [\frac{1}{6} - \log\big(\frac{2}{\sqrt{3}}\big), \frac{1}{2}]$, the projection risk $\mathcal{R}_n(k, \theta)$ be defined by (2.3) and the oracle cut-off N_o be defined by (2.4).*

- *For any $\tau_1 \geq 6 \log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha \geq 1$ and any $k < N_o$,*

$$E_{\theta_0} \big[P(N = k | X) \big] \leq \exp \Big\{ -\frac{n}{6} \big(\mathcal{R}_n(k, \theta_0) - \tau_1 \mathcal{R}_n(N_o, \theta_0) \big) \Big\}.$$

- *For any* $\tau_2 \geq \frac{1}{2\alpha} \geq 1$ *and any* $k > N_o$,

$$E_{\theta_0}\big[P(N=k|X)\big] \leq \exp\big\{-n\alpha\big(\mathcal{R}_n(k,\theta_0) - \tau_2\mathcal{R}_n(N_o,\theta_0)\big)\big\}.$$

Proof. To prove the lemma, we use $\frac{\lambda_k}{\lambda_{N_o}} = e^{-\alpha(k-N_o)}$ and Corollary 2.3. For $k < N_o$,

$$\begin{aligned} E_{\theta_0}P(N=k|X) &\leq \exp\Big\{-\frac{n}{6}\Big(\sum_{i=k+1}^{N_o}\theta_{0i}^2 - \frac{(6\log\big(\frac{2}{\sqrt{3}}\big)+6\alpha)(N_o-k)}{n}\Big)\Big\} \\ &\leq \exp\Big\{-\frac{n}{6}\Big(\sum_{i=k+1}^{\infty}\theta_{0i}^2 - \tau_1\sum_{i=N_o+1}^{\infty}\theta_{0i}^2 - \frac{\tau_1 N_o - k}{n}\Big)\Big\} \\ &= \exp\Big\{-\frac{n}{6}\big(\mathcal{R}_n(k,\theta_0) - \tau_1\mathcal{R}_n(N_o,\theta_0)\big)\Big\} \end{aligned}$$

since $\tau_1 \geq 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha \geq 1$. Similarly, for $k > N_o$, we have that

$$\begin{aligned} E_{\theta_0}P(N=k|X) &\leq \exp\Big\{-n\alpha\Big(-\frac{1}{2\alpha}\sum_{i=N_o+1}^{k}\theta_{0i}^2 + \frac{(k-N_o)}{n}\Big)\Big\} \\ &\leq \exp\Big\{-n\alpha\Big(\sum_{i=k+1}^{\infty}\theta_{0i}^2 - \tau_2\sum_{i=N_o+1}^{\infty}\theta_{0i}^2 + \frac{k-\tau_2 N_o}{n}\Big)\Big\} \\ &= \exp\Big\{-n\alpha\big(\mathcal{R}_n(k,\theta_0) - \tau_2\mathcal{R}_n(N_o,\theta_0)\big)\Big\} \end{aligned}$$

because $\tau_2 \geq \frac{1}{2\alpha} \geq 1$. □

Remark 2.16. The above exponential inequalities describe essentially a "correct" frequentist behavior of the posterior distribution $P(N|X)$. The bigger the difference is between the risk $\mathcal{R}_n(k,\theta_0)$ at point k and the oracle risk $\mathcal{R}_n(N_o,\theta_0)$, the smaller the exponential bound for $E_{\theta_0}P(N=k|X)$ becomes.

For brevity sake, denote

$$\begin{aligned} &R_-(\theta_0,\beta,\tau,\tau_1,n) \\ &\quad = \sum_{k\in\mathcal{N}^-(\tau)} n\mathcal{R}_n(k,\theta_0)\exp\big\{-\beta n\big(\mathcal{R}_n(k,\theta_0) - \tau_1\mathcal{R}_n(N_o,\theta_0)\big)\big\}, \quad (2.25) \end{aligned}$$

$$\begin{aligned} &R_+(\theta_0,\beta,\tau,\tau_2,n) \\ &\quad = \sum_{k\in\mathcal{N}^+(\tau)} n\mathcal{R}_n(k,\theta_0)\exp\big\{-\beta n\big(\mathcal{R}_n(k,\theta_0) - \tau_2\mathcal{R}_n(N_o,\theta_0)\big)\big\}, \quad (2.26) \end{aligned}$$

where the projection risk $\mathcal{R}_n(k,\theta)$ is defined by (2.3), the oracle cut-off $N_o = N_o(\theta_0)$ is defined by (2.4), the sets $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ are defined by (2.18) and (2.19) respectively.

Lemma 2.5. *Let the quantities $R_-(\theta_0,\beta,\tau,\tau_1,n)$ and $R_+(\theta_0,\beta,\tau,\tau_2,n)$ be defined by (2.25) and (2.26) respectively.*

(i) For any $\tau>\tau_1>0$, $\beta>0$, the following inequality holds:

$$R_-(\theta_0,\beta,\tau,\tau_1,n)\le \min\Big\{\frac{c_1}{B^2},\frac{\big(n\mathcal{R}_n(N_o,\theta_0)\big)^2}{e^{Bn\mathcal{R}_n(N_o,\theta_0)}}\Big\},$$

where $B=B(\tau,\tau_1,\beta)=\beta(\tau-\tau_1)/\tau$, $c_1=4e^{-2}$.

(ii) For any $\tau>\tau_2>0$, $\beta>0$, the following inequality holds:

$$\begin{aligned}R_+(\theta_0,\beta,\tau,\tau_2,n) &\le \min\Big\{g(D),\frac{c_3\big(n\mathcal{R}_n(N_o,\theta_0)\big)^2}{e^{Dn\mathcal{R}_n(N_o,\theta_0)}}\Big\}\\ &\le \min\Big\{\frac{c_2}{D^2},\frac{c_3\big(n\mathcal{R}_n(N_o,\theta_0)\big)^2}{e^{Dn\mathcal{R}_n(N_o,\theta_0)}}\Big\},\end{aligned}$$

where $g(u)=e^{-1}u^{-2}+e^{-u}(e^u-1)^{-2}$, $D=D(\tau,\tau_2,\beta)=\beta(\tau-\tau_2)/\tau$, $c_2=e^{-1}+1$ and $c_3=(e+e^2)/4$.

Proof. Since $\tau>\tau_1$, we obtain for all $k\in\mathcal{N}^-(\tau)$ that

$$\begin{aligned}&\mathcal{R}_n(k,\theta_0)-\tau_1\mathcal{R}_n(N_o,\theta_0)\\ &\quad=\frac{\tau_1}{\tau}(\mathcal{R}_n(k,\theta_0)-\tau\mathcal{R}_n(N_o,\theta_0))+\Big(1-\frac{\tau_1}{\tau}\Big)\mathcal{R}_n(k,\theta_0)\\ &\quad\ge\Big(1-\frac{\tau_1}{\tau}\Big)\mathcal{R}_n(k,\theta_0).\end{aligned}\tag{2.27}$$

Recall the notation $B=B(\tau,\tau_1,\beta)=\frac{\beta(\tau-\tau_1)}{\tau}>0$ and let $a_k=a_k(\theta_0,n)=n\mathcal{R}_n(k,\theta_0)$. Let b be any constant such that $0<b\le B$. Certainly $a_k\ge a_{N_o}=n\mathcal{R}_n(N_o,\theta_0)$ by the definition of oracle. Then, using (2.27),

$$\begin{aligned}R_-(\theta_0,\beta,\tau,\tau_1,n) &\le \sum_{k\in\mathcal{N}^-(\tau)} n\mathcal{R}_n(k,\theta_0)\exp\{-Bn\mathcal{R}_n(k,\theta_0)\}\\ &= \sum_{k\in\mathcal{N}^-(\tau)} a_k\exp\{-Ba_k\}\\ &\le e^{-(B-b)a_{N_o}}\sum_{k\in\mathcal{N}^-(\tau)} a_k\exp\{-ba_k\}.\end{aligned}\tag{2.28}$$

Consider separately two cases: $1\le N_o\le b^{-1}$ and $N_o>b^{-1}$. First assume that $1\le N_o\le b^{-1}$, then obviously

$$\begin{aligned}\sum_{k\in\mathcal{N}^-(\tau)} a_k\exp\{-ba_k\} &\le |\mathcal{N}^-(\tau)|\max_{x\ge0}\{xe^{-bx}\}\\ &\le N_o(be)^{-1}=b^{-2}e^{-1}.\end{aligned}\tag{2.29}$$

Now suppose that $N_o>b^{-1}$. Let us look at the term $\sum_{k\in\mathcal{N}^-(\tau)}a_k\exp\{-ba_k\}$. First note that $a_k=n\mathcal{R}_n(k,\theta_0)\ge n\mathcal{R}_n(N_o,\theta_0)\ge N_o>b^{-1}$ for all $k\in\mathcal{N}^-(\tau)$.

The function xe^{-bx}, $x \geq 0$, is increasing on the interval $[0, b^{-1}]$ and decreasing afterwards. This implies that $a_k \exp\{-ba_k\} \leq N_o \exp\{-bN_o\}$. Recall also that $|\mathcal{N}^-(\tau)| \leq N_o$. Using these relations, we derive the following bound:

$$\begin{aligned} \sum_{k \in \mathcal{N}^-(\tau)} a_k \exp\{-ba_k\} &\leq |\mathcal{N}^-(\tau)| N_o \exp\{-bN_o\} \\ &\leq N_o^2 \exp\{-bN_o\} \\ &\leq \max_{x \geq 0}\{x^2 e^{-bx}\} = 4(be)^{-2}. \end{aligned} \tag{2.30}$$

Combining (2.29) and (2.30) leads to

$$\sum_{k \in \mathcal{N}^-(\tau)} a_k \exp\{-ba_k\} \leq \max\left\{b^{-2}e^{-1}, 4(be)^{-2}\right\} = 4(be)^{-2},$$

which, together with (2.28), implies that for any $b \in (0, B]$

$$R_-(\theta_0, \beta, \tau, \tau_1, n) \leq 4e^{-2} e^{-(B-b)a_{N_o}} b^{-2}. \tag{2.31}$$

We have a family of upper bounds for different values of $b \in (0, B]$. Now we minimize the upper bound over $b \in (0, B]$. The upper bound in (2.31) is of the form $Ce^{a_{N_o} b} b^{-2}$, with constants $C = 4e^{-2} e^{-Ba_{N_o}}$, $a_{N_o} = n\mathcal{R}_n(N_o, \theta_0)$. The minimum of function $Ce^{a_{N_o} b} b^{-2}$ over $b \in (0, B]$ is attained in $\min\{B, 2/a_{N_o}\}$:

$$\min_{b \in (0,B]} Ce^{a_{N_o} b} b^{-2} = \min\Big\{\frac{Ce^{a_{N_o} B}}{B^2}, \frac{Ce^2 a_{N_o}^2}{4}\Big\}. \tag{2.32}$$

Thus the corresponding sharpest upper bound becomes

$$\min_{b \in (0,B]} \left\{4e^{-2} e^{-(B-b)a_{N_o}} b^{-2}\right\} = \min\left\{4e^{-2}B^{-2}, \big(n\mathcal{R}_n(N_o, \theta_0)\big)^2 e^{-Bn\mathcal{R}_n(N_o, \theta_0)}\right\},$$

which, combined with (2.31), establishes part (i) of the lemma.

Let us prove part (ii) of the lemma. We derive in the same manner as in (2.27) that, for all $k \in \mathcal{N}^+(\tau)$,

$$\beta n\big(\mathcal{R}_n(k, \theta_0) - \tau_2 \mathcal{R}_n(N_o, \theta_0)\big) \geq \beta n\Big(1 - \frac{\tau_2}{\tau}\Big)\mathcal{R}_n(k, \theta_0) = Da_k, \tag{2.33}$$

where $D = D(\tau, \tau_2, \beta) = \frac{(\tau - \tau_2)\beta}{\tau} > 0$ and $a_k = n\mathcal{R}_n(k, \theta_0)$ as before. Using (2.33), we obtain that for any $0 < d \leq D$

$$\begin{aligned} R_+(\theta_0, \beta, \tau, \tau_2, n) &\leq \sum_{k \in \mathcal{N}^+(\tau)} a_k \exp\{-Da_k\} \\ &\leq e^{-(D-d)a_{N_0}} \sum_{k \in \mathcal{N}^+(\tau)} a_k \exp\{-da_k\}. \end{aligned} \tag{2.34}$$

Next, introduce the sets

$$S_1(d) = S_1(d, \tau, \theta_0) = \{k \in \mathcal{N}^+(\tau, \theta_0) : k \leq d^{-1}\},$$

$$S_2(d) = S_2(d, \tau, \theta_0) = \{k \in \mathcal{N}^+(\tau, \theta_0) : k > d^{-1}\},$$

so that $\mathcal{N}^+(\tau) = S_1(d) \cup S_2(d)$. Obviously, $|S_1(d)| \le d^{-1}$. Besides, recall that $k \le a_k$ for all $k \in \mathbb{N}$ and the function xe^{-dx}, $d \ge 0$, is decreasing on $[d^{-1}, +\infty)$. Thus, as $a_k \ge k > d^{-1}$,

$$a_k \exp\{-da_k\} \le k \exp\{-dk\}$$

for all $k \in S_2(d)$. Using this, we obtain that

$$\begin{aligned}
\sum_{k \in \mathcal{N}^+(\tau)} a_k \exp\{-da_k\} &\le \sum_{k \in S_1(d)} a_k \exp\{-da_k\} + \sum_{k \in S_2(d)} a_k \exp\{-da_k\} \\
&\le |S_1(d)| \max_{x \ge 0}\{xe^{-dx}\} + \sum_{k \in S_2(d)} k \exp\{-dk\} \\
&\le d^{-2}e^{-1} + \sum_{k=N_o+1}^{\infty} k \exp\{-dk\} \\
&= d^{-2}e^{-1} + \frac{e^{-d(N_o+1)}e^{-d}}{(1-e^{-d})^2} \\
&\le d^{-2}e^{-1} + \frac{e^{-3d}}{(1-e^{-d})^2} \\
&= \frac{e^{-1}}{d^2} + \frac{e^{-d}}{(e^d-1)^2} = g(d) \\
&\le \frac{e^{-1}}{d^2} + \frac{e^{-d}}{d^2} \le \frac{e^{-1}+1}{d^2}, \qquad (2.35)
\end{aligned}$$

which, combined with (2.34), implies that for any $0 < d \le D$

$$R_+(\theta_0, \beta, \tau, \tau_2, n) \le (e^{-1}+1)e^{-(D-d)a_{N_0}} d^{-2}. \qquad (2.36)$$

Again we have a family of upper bounds for different values of $d \in (0, D]$. Using (2.32) (this time with $C = (e^{-1}+1)e^{-Da_{N_o}}$), we minimize the upper bound over $d \in (0, D]$ in the same way as before:

$$\begin{aligned}
&\min_{d \in (0,D]} \left\{(e^{-1}+1)e^{-(D-d)a_{N_0}} d^{-2}\right\} \\
&\quad = \min\left\{\frac{e^{-1}+1}{D^2}, \frac{e+e^2}{4}\big(n\mathcal{R}_n(N_o, \theta_0)\big)^2 e^{-Dn\mathcal{R}_n(N_o,\theta_0)}\right\}.
\end{aligned}$$

Besides, by taking $d = D = \frac{\beta(\tau-\tau_2)}{\tau}$ in (2.34) and (2.35), we also have a bound

$$R_+(\theta_0, \beta, \tau, \tau_2, n) \le \frac{e^{-1}}{D^2} + \frac{e^{-D}}{(e^D-1)^2} = g(D) \le \frac{e^{-1}+1}{D^2}.$$

Finally combine (2.36) with the last two relations to establishes part (ii) of the lemma. □

Remark 2.17. The minimum in the right hand sides of the inequalities of the above lemma is attained by the second term if $\min\{B, 2/a_{N_o}\} = 2/a_{N_o}$ and $\min\{D, 2/a_{N_o}\} = 2/a_{N_o}$, i.e. the oracle risk is sufficiently large $a_{N_o} = n\mathcal{R}_n(N_o, \theta_0) \geq \max\{2/B, 2/D\}$ or $\mathcal{R}_n(N_o, \theta_0) \geq 2\max\{B^{-1}, D^{-1}\}n^{-1}$. Thus, the oracle risk should be bigger than the parametric rate with sufficiently large constant. Clearly, the upper bounds in the right hand sides of the inequalities of the above lemma will therefore be small if $n\mathcal{R}_n(N_o, \theta_0)$ is large, which typically the case for the so called "nonparametric" θ_0's: $n\mathcal{R}_n(N_o, \theta_0) \to \infty$ as $n \to \infty$, i.e. the oracle risk is of a bigger order than the parametric rate n^{-1} as $n \to \infty$.

For brevity sake, denote

$$P_-(\theta_0, \beta, \tau, \tau_1, n) = \sum_{k \in \mathcal{N}^-(\tau)} \exp\left\{-\beta n\big(\mathcal{R}_n(k, \theta_0) - \tau_1 \mathcal{R}_n(N_o, \theta_0)\big)\right\}, \quad (2.37)$$

$$P_+(\theta_0, \beta, \tau, \tau_2, n) = \sum_{k \in \mathcal{N}^+(\tau)} \exp\left\{-\beta n\big(\mathcal{R}_n(k, \theta_0) - \tau_2 \mathcal{R}_n(N_o, \theta_0)\big)\right\}. \quad (2.38)$$

where the projection risk $\mathcal{R}_n(k, \theta)$ is defined by (2.3), the oracle cut-off $N_o = N_o(\theta_0)$ is defined by (2.4), the sets $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ are defined by (2.18) and (2.19) respectively.

Corollary 2.4. *Under the conditions of Lemma 2.5, the following bounds hold:*

$$\begin{aligned} P_-(\theta_0, \beta, \tau, \tau_1, n) &\leq \frac{1}{\tau} \min\left\{\frac{c_1}{B^2 n\mathcal{R}_n(N_o, \theta_0)}, \frac{n\mathcal{R}_n(N_o, \theta_0)}{e^{Bn\mathcal{R}_n(N_o, \theta_0)}}\right\} \\ &\leq \frac{\min\left\{4(Be)^{-1}, 1\right\}}{\tau Be}, \end{aligned}$$

$$\begin{aligned} P_+(\theta_0, \beta, \tau, \tau_2, n) &\leq \min\left\{\frac{g(D)}{\tau n\mathcal{R}_n(N_o, \theta_0)}, \frac{c_3 n\mathcal{R}_n(N_o, \theta_0)}{\tau e^{Dn\mathcal{R}_n(N_o, \theta_0)}}, \frac{e^{-DN_o}}{e^D - 1}\right\} \\ &\leq \min\left\{\frac{g(D)}{\tau}, \frac{1+e}{4D\tau}, \frac{e^{-D}}{e^D - 1}\right\}. \end{aligned}$$

Proof. Indeed, by using Lemma 2.5 and the facts that $n\mathcal{R}_n(N_o, \theta_0) \geq 1$, $\max_{x \geq 0}\{xe^{-Bx}\} = (Be)^{-1}$ and

$$\frac{1}{\mathcal{R}_n(k, \theta_0)} \leq \frac{1}{\tau \mathcal{R}_n(N_o, \theta_0)} \quad \text{for all } k \notin \mathcal{N}(\tau, \theta_0),$$

we obtain

$$\begin{aligned} P_-(\theta_0, \beta, \tau, \tau_1, n) &\leq \frac{R_-(\theta_0, \beta, \tau, \tau_1, n)}{\tau n\mathcal{R}_n(N_o, \theta_0)} \\ &\leq \tau^{-1} \min\left\{\frac{c_1}{B^2 n\mathcal{R}_n(N_o, \theta_0)}, \frac{n\mathcal{R}_n(N_o, \theta_0)}{e^{Bn\mathcal{R}_n(N_o, \theta_0)}}\right\} \\ &\leq \tau^{-1} \min\left\{\frac{4}{(Be)^2}, \frac{1}{Be}\right\} = (\tau Be)^{-1} \min\left\{4(Be)^{-1}, 1\right\}. \end{aligned}$$

The inequality

$$\begin{aligned} P_+(\theta_0,\beta,\tau,\tau_2,n) &\le \frac{1}{\tau}\min\Big\{\frac{g(D)}{n\mathcal{R}_n(N_o,\theta_0)}, \frac{c_3 n\mathcal{R}_n(N_o,\theta_0)}{e^{Dn\mathcal{R}_n(N_o,\theta_0)}}\Big\} \\ &\le \min\Big\{\frac{g(D)}{\tau}, \frac{1+e}{4D\tau}\Big\}. \end{aligned}$$

follows similarly from Lemma 2.5. Besides, due to (2.38) and (2.33) and the fact that $a_k \ge k$ for all $k \in \mathbb{N}$, we derive the bound

$$\begin{aligned} P_+(\theta_0,\beta,\tau,\tau_2,n) &\le \sum_{k\in\mathcal{N}^+(\tau)} \exp\{-Da_k\} \\ &\le \sum_{k=N_o+1}^{\infty} \exp\{-Dk\} \\ &\le \frac{e^{-D(N_o+1)}}{1-e^{-D}} \le \frac{e^{-DN_o}}{e^D-1}. \end{aligned}$$

Combining the last two relations completes the proof. □

2.5 Proofs of the theorems

Proof of Theorem 2.1. Using the conditional Chebyshev's inequality, we have

$$P_N\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\|^2 \ge M\mathcal{R}_n(N,\theta_0)\,\Big|X\Big\} \le \frac{E_N\big(\big\|\theta-\frac{\theta_0}{2}\big\|^2\,\big|X\big)}{M\mathcal{R}_n(N,\theta_0)}, \tag{2.39}$$

Analogously, for the posterior $P'_N(\theta|\boldsymbol{y})$

$$P'_N\big\{\|\theta-\theta_0\|^2 \ge M\mathcal{R}_n(N,\theta_0)\,|X\big\} \le \frac{E'_N\big(\|\theta-\theta_0\|^2\,\big|X\big)}{M\mathcal{R}_n(N,\theta_0)}, \tag{2.40}$$

In the next step we compute $E_N\big(\big\|\theta-\theta_0\big\|^2\,\big|X\big)$, where the conditional expectation is taken with respect to $P_N(\theta|X)$,

$$\begin{aligned} E_N\Big[\big\|\theta-\frac{\theta_0}{2}\big\|^2\,\big|X\Big] &= \sum_{i=1}^{\infty}\mathrm{Var}_N(\theta_i|X) + \sum_{i=1}^{\infty}\Big(E_N(\theta_i|X)-\frac{\theta_0}{2}\Big)^2 \\ &= \frac{N}{2n} + \frac{1}{4n}\sum_{i=1}^{N}\xi_i^2 + \sum_{i=N+1}^{\infty}\frac{\theta_{0_i}^2}{4}. \end{aligned} \tag{2.41}$$

The similar formula holds for $E'_N\big(\big\|\theta-\theta_0\big\|^2\,\big|X\big)$, regarding the distribution $P'_N(\theta|X)$

$$E'_N\Big[\big\|\theta-\theta_0\big\|^2\,\big|X\Big] = \frac{N}{2n} + \frac{1}{n}\sum_{i=1}^{N}\xi_i^2 + \sum_{i=N+1}^{\infty}\theta_{0_i}^2. \tag{2.42}$$

Combining (2.39), (2.41) and (2.3) we obtain

$$\begin{aligned} & E_{\theta_0} P_N\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \geq M\mathcal{R}_n(N,\theta_0)\,\Big|X\Big\} \\ \leq\ & \frac{1}{M\mathcal{R}_n(N,\theta_0)} E_{\theta_0}\Big(\frac{N}{2n} + \frac{1}{4}\sum_{i=1}^{N}\frac{\xi_i^2}{n} + \sum_{i=N+1}^{\infty}\frac{\theta_{0_i}^2}{4}\Big) \\ \leq\ & \frac{(1/2)\mathcal{R}_n(N,\theta_0) + (1/4)\mathcal{R}_n(N,\theta_0)}{M\mathcal{R}_n(N,\theta_0)}. \end{aligned} \tag{2.43}$$

Substitution of equation (2.42) into (2.40), combined with (2.3), yields

$$E_{\theta_0} P'_N\big\{\|\theta-\theta_0\|^2 \geq M\mathcal{R}_n(N,\theta_0)\,\big|X\big\} \leq \frac{(1/2)\mathcal{R}_n(N,\theta_0) + \mathcal{R}_n(N,\theta_0)}{M\mathcal{R}_n(N,\theta_0)}. \tag{2.44}$$

Thus, it follows from (2.43) and (2.44) that

$$\begin{aligned} E_{\theta_0} P_N\Big\{\Big\|\theta - \frac{\theta_0}{2}\Big\|^2 \geq M\mathcal{R}_n(N,\theta_0)\,\Big|X\Big\} &\leq \frac{c_1}{M} \\ E_{\theta_0} P'_N\big\{\|\theta-\theta_0\|^2 \geq M\mathcal{R}_n(N,\theta_0)\,\big|X\big\} &\leq \frac{c_2}{M}, \end{aligned}$$

with $c_1 = 3/4$ and $c_2 = 3/2$, which proves the theorem. □

Proof of Theorem 2.2. First recall that, according to (2.12),

$$\widehat{\theta} = E'(\theta|X) = \sum_{k\in\mathbb{N}} E'(\theta|X, N=k)P(N=k|X).$$

That is

$$\widehat{\theta}_i = E'(\theta_i|X) = \sum_{k\in\mathbb{N}} E'(\theta_i|X,N=k)P(N=k|X) = \sum_{k\in\mathbb{N}}\widehat{\theta}_i(k)P(N=k|X),$$

with $\widehat{\theta}_i(k) = X_i I\{i \leq k\}$. Now, by Fubini's theorem and Cauchy-Schwarz inequality,

$$\begin{aligned} E_{\theta_0}\|\widehat{\theta}-\theta_0\|^2 &= E_{\theta_0}\sum_{i\in\mathbb{N}}\Big(\sum_{k\in\mathbb{N}}\widehat{\theta}_i(k)P(N=k|X) - \theta_{0i}\Big)^2 \\ &\leq E_{\theta_0}\sum_{i\in\mathbb{N}}\sum_{k\in\mathbb{N}}\big(\widehat{\theta}_i(k)-\theta_{0i}\big)^2 P(N=k|X) \\ &= E_{\theta_0}\sum_{k\in\mathbb{N}}\|\widehat{\theta}(k)-\theta_0\|^2 P(N=k|X). \end{aligned}$$

Fix a $\tau \geq 1$ which is to be chosen later and let the sets $\mathcal{N}(\tau)$, $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ be defined by (2.17), (2.18) and (2.19) respectively. Split the last sum in three sums over the sets $\mathcal{N}(\tau)$, $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ and apply Lemmas 2.1 and 2.2 to derive the bound for the estimator (2.12):

$$E_{\theta_0}\|\widehat{\theta}-\theta_0\|^2 \leq 2\tau\mathcal{R}_n(N_o,\theta_0) + T_1 + T_2, \tag{2.45}$$

for any $\theta_0 \in \ell_2$, where

$$T_1 = \sum_{k \in \mathcal{N}^-(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\Big(E_{\theta_0} P(N=k|X)\Big)^{1/2},$$

$$T_2 = \sum_{k \in \mathcal{N}^+(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big)\Big(E_{\theta_0} P(N=k|X)\Big)^{1/2}.$$

Similarly, applying Lemmas 2.1 and 2.2 for the estimator (2.13), we obtain the same bound for the estimator (2.13):

$$\begin{aligned} & E_{\theta_0}\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 \\ &= E_{\theta_0}\Big[\|\widehat{\theta}(\widehat{N}) - \theta_0\|^2 \Big(I\big\{\widehat{N} \in \mathcal{N}(\tau)\big\} + I\big\{\widehat{N} \in \mathcal{N}^-(\tau)\big\} + I\big\{\widehat{N} \in \mathcal{N}^+(\tau)\big\}\Big)\Big] \\ &\le 2\tau \mathcal{R}_n(N_o,\theta_0) + T_1 + T_2 \end{aligned} \tag{2.46}$$

for any $\theta_0 \in \ell_2$, with T_1 and T_2 defined above.

Now we apply Lemma 2.4 to evaluate both terms T_1 and T_2. Take some $\tau_1 \ge \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha \ge 1$ which is to be chosen later. The inequality $\tau_-(\alpha) \ge 1$ follows from the condition on α. Then, by using Lemma 2.4 and the definitions (2.25) and (2.37), we obtain

$$\begin{aligned} T_1 &\le \sum_{k \in \mathcal{N}^-(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big) \exp\Big\{-\frac{n}{12}\big(\mathcal{R}_n(k,\theta_0) - \tau_1 \mathcal{R}_n(N_o,\theta_0)\big)\Big\} \\ &\le n^{-1}\big(R_-(\theta_0, 1/12, \tau, \tau_1, n) + P_-(\theta_0, 1/12, \tau, \tau_1, n)\big). \end{aligned}$$

Assume now that τ and τ_1 are chosen in such a way that $\tau > \tau_1 \ge \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha$. Then we can apply part (i) of Lemma 2.5 and Corollary 2.4 to obtain

$$T_1 \le \frac{C_1}{n}, \qquad C_1 = C_1(\tau,\tau_1) = \frac{c_1}{B^2} + \frac{1}{\tau B e}, \tag{2.47}$$

with

$$c_1 = 4e^{-2} \quad \text{and} \quad B = B(\tau,\tau_1) = \frac{\tau - \tau_1}{12\tau}.$$

To bound the term T_2, we apply Lemmas 2.4 and 2.5 again. Assume that τ and τ_2 are chosen in such a way that $\tau > \tau_2 \ge \tau_+(\alpha) = \frac{1}{2\alpha} \ge 1$. The inequality $\tau_+(\alpha) \ge 1$ follows from the condition on α. By consequent applying first Lemma 2.4 and then Lemma 2.5 with Corollary 2.4, we obtain that

$$\begin{aligned} T_2 &\le \sum_{k \in \mathcal{N}^+(\tau)} \Big(\mathcal{R}_n(k,\theta_0) + \frac{1}{n}\Big) \exp\Big\{-\frac{n\alpha}{2}\big(\mathcal{R}_n(k,\theta_0) - \tau_2 \mathcal{R}_n(N^o,\theta_0)\big)\Big\} \\ &\le n^{-1}\big(R_+(\theta_0, \alpha/2, \tau, \tau_2, n) + P_+(\theta_0, \alpha/2, \tau, \tau_2, n)\big) \le \frac{C_2}{n}, \end{aligned} \tag{2.48}$$

$$C_2 = C_2(\tau,\tau_2,\alpha) = g(D) + \min\Big\{\frac{g(D)}{\tau}, \frac{e^{-D}}{e^D - 1}\Big\},$$

with

$$D = D(\tau, \tau_2, \alpha) = \frac{\alpha(\tau - \tau_2)}{2\tau} \quad \text{and} \quad g(u) = e^{-1}u^{-2} + e^{-u}(e^u - 1)^{-2}.$$

It remains to choose τ, τ_1, τ_2 so that $\tau > \tau_1 \geq \tau_-(\alpha) = 6\log\left(\frac{2}{\sqrt{3}}\right) + 6\alpha$ and $\tau \geq \tau_2 \geq \tau_+(\alpha) = \frac{1}{2\alpha}$. Take, for example, $\tau_1 = \tau_-(\alpha)$, $\tau_2 = \tau_+(\alpha)$ and $\tau = \max\{\tau_1, \tau_2\} + 1$. Finally, we combine the relations (2.45), (2.46), (2.47) and (2.48) to establish the statement of the theorem with $K_1 = 2\tau$ and $K_2 = C_1 + C_2$. The theorem is proved. □

Remark 2.18. For example, we take $\tau_1 = \frac{1}{2\alpha}$, $\tau_2 = (6\log\left(\frac{2}{\sqrt{3}}\right) + 6\alpha)$ with $\alpha \in [\frac{1}{6} - \log\left(\frac{2}{\sqrt{3}}\right), \frac{1}{2}]$ such that $\tau_1 = \tau_2$ which leads to

$$\alpha = \frac{-6\log\left(\frac{2}{\sqrt{3}}\right) + \sqrt{36\log^2\left(\frac{2}{\sqrt{3}}\right) + 12}}{12} \approx 0.2255789,$$

and $\tau_1 = \tau_2 = \frac{1}{2\alpha} \approx 2.21652$. Take further $\tau = \tau_1 + 7$. Then we compute the constants $K_1 \approx 18.433$ and $K_2 \approx 310.904$, or the following (non asymptotic) inequality holds true for all $\theta_0 \in \ell_2$:

$$E_{\theta_0}\|\widehat{\theta} - \theta_0\|^2 \leq 19\mathcal{R}_n(N_o, \theta_0) + \frac{311}{n}.$$

Of course, this is not the optimal choice of the involved parameters $\alpha, \tau_1, \tau_2, \tau$.

Remark 2.19. We could also try to improve the constants K_1 and K_2 by refining the bounds in the proof. However, if we try to make K_1 as small as possible, the resulting K_2 becomes big. On the other hand, we can make K_2 smaller by sacrificing the constant K_1. Actually, the constant K_2 can be strongly improved if n is a reasonable number and we assume a nonparametric behavior of the oracle risk $\mathcal{R}_n(N_o, \theta_0)$, which means that $n\mathcal{R}_n(N_o, \theta_0)$ is large. This follows from the proof of the theorem, where we used the following uniform trivial bounds: $n\mathcal{R}_n(N_o, \theta_0) \geq 1$,

$$n\mathcal{R}_n(N_o, \theta_0)\exp\{-cn\mathcal{R}_n(N_o, \theta_0)\} \leq \max_{x \geq 0}\{xe^{-cx}\} = (ce)^{-1},$$

$$\left(n\mathcal{R}_n(N_o, \theta_0)\right)^2\exp\{-cn\mathcal{R}_n(N_o, \theta_0)\} \leq \max_{x \geq 0}\{x^2e^{-cx}\} = 4(ce)^{-2}.$$

Another possibility to improve the constants is by introducing a factor c in the expression for the prior variance $\tau_i^2(N) = cn^{-1}I\{i \leq N\}$ in the definition of the prior (2.5). We also used a somewhat rough estimate $P^2(N = k|X) \leq P(N = k|X)$ in some places of the proof.

Proof of Theorem 2.3. First consider the posterior (2.10). By the conditional

Chebyshev's inequality,

$$
\begin{aligned}
&P\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\|^2\ge Mr_n(\theta_0)\,\Big|X\Big\}\\
&\quad=\sum_{k\in\mathbb{N}}P\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\|^2\ge Mr_n(\theta_0)\,\Big|X,N=k\Big\}P(N=k|X)\\
&\quad\le\sum_{k\in\mathbb{N}}\frac{E\big(\big\|\theta-\frac{\theta_0}{2}\big\|^2\big|X,N=k\big)}{Mr_n(\theta_0)}P(N=k|X),
\end{aligned}
\tag{2.49}
$$

where the conditional distribution $P(\theta|X,N=k)$ is defined by (2.6). Analogously, for the posterior (2.11) we have that

$$
\begin{aligned}
&P'\big\{\|\theta-\theta_0\|^2\ge Mr_n(\theta_0)\,|X\big\}\\
&\quad\le\quad\sum_{k\in\mathbb{N}}\frac{E'\big(\|\theta-\theta_0\|^2|X,N=k\big)}{Mr_n(\theta_0)}P(N=k|X),
\end{aligned}
\tag{2.50}
$$

where the conditional distribution $P'(\theta|X,N=k)$ is defined by (2.8).

Using these, we compute

$$
\begin{aligned}
&E\Big[\big\|\theta-\frac{\theta_0}{2}\big\|^2\big|X,N=k\Big]\\
&\quad=\sum_{i=1}^{\infty}\mathrm{Var}(\theta_i|X,N=k)+\sum_{i=1}^{\infty}\Big(E(\theta_i|X,N=k)-\frac{\theta_0}{2}\Big)^2\\
&\quad=\frac{k}{2n}+\frac{1}{4n}\sum_{i=1}^{k}\xi_i^2+\sum_{i=k+1}^{\infty}\frac{\theta_{0_i}^2}{4},\\
&E'\Big[\big\|\theta-\theta_0\big\|^2\big|X,N=k\Big]=\frac{k}{2n}+\frac{1}{n}\sum_{i=1}^{k}\xi_i^2+\sum_{i=k+1}^{\infty}\theta_{0_i}^2.
\end{aligned}
\tag{2.51}
$$

The relations (2.49) and (2.51) imply that

$$
\begin{aligned}
&E_{\theta_0}P\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\|^2\ge Mr_n(\theta_0)\,\Big|X\Big\}\\
&\le\quad\frac{1}{Mr_n(\theta_0)}E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\frac{k}{2n}+\frac{1}{4}\sum_{i=1}^{k}\frac{\xi_i^2}{n}+\sum_{i=k+1}^{\infty}\frac{\theta_{0_i}^2}{4}\Big)P(N=k|X)\Big]\\
&\le\quad\frac{\sum_{k\in\mathbb{N}}\mathcal{R}_n(k,\theta_0)E_{\theta_0}P(N=k|X)}{2Mr_n(\theta_0)}\\
&\qquad+\frac{E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k|X)\Big]}{4Mr_n(\theta_0)}.
\end{aligned}
\tag{2.52}
$$

Similarly, from the relations (2.50) and (2.51) it follows that

$$\begin{aligned} & E_{\theta_0} P'\big\{\|\theta-\theta_0\|^2 \geq M r_n(\theta_0)\,\big|X\big\} \\ & \quad \leq \frac{\sum_{k\in\mathbb{N}} \mathcal{R}_n(k,\theta_0) E_{\theta_0} P(N=k|X)}{M r_n(\theta_0)} \\ & \qquad + \frac{E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k|X)\Big]}{M r_n(\theta_0)}. \end{aligned} \tag{2.53}$$

Fix a $\tau > 0$ which is to be chosen later and let the sets $\mathcal{N}(\tau)$, $\mathcal{N}^-(\tau)$ and $\mathcal{N}^+(\tau)$ be defined by (2.17), (2.18) and (2.19) respectively. Suppose τ, τ_1, τ_2 are chosen in such a way that $\tau > \tau_1 \geq \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha$ and $\tau \geq \tau_2 \geq \tau_+(\alpha) = \frac{1}{2\alpha}$. Clearly, $\mathcal{R}_n(k,\theta_0) \leq \tau \mathcal{R}_n(N_o,\theta_0)$ for all $k \in \mathcal{N}(\tau)$, so that

$$\begin{aligned} \sum_{k\in\mathbb{N}} \mathcal{R}_n(k,\theta_0) E_{\theta_0} P(N=k|X) &\leq \sum_{k\in\mathcal{N}(\tau)} \mathcal{R}_n(k,\theta_0) E_{\theta_0} P(N=k|X) \\ &\quad + \sum_{k\in\mathcal{N}^-(\tau)} \mathcal{R}_n(k,\theta_0) E_{\theta_0} P(N=k|X) \\ &\quad + \sum_{k\in\mathcal{N}^+(\tau)} \mathcal{R}_n(k,\theta_0) E_{\theta_0} P(N=k|X) \\ &\leq \tau \mathcal{R}_n(N_o,\theta_0) E_{\theta_0} P(N\in\mathcal{N}(\tau)|X) + T_1 + T_2 \\ &\leq \tau \mathcal{R}_n(N_o,\theta_0) + \frac{C_1+C_2}{n}, \end{aligned} \tag{2.54}$$

where the terms T_1 and T_2 are defined in the proof of Theorem 2.2, constants $C_1 = C_1(\tau,\tau_1)$ and $C_2 = C_2(\tau,\tau_2,\alpha)$ are defined by (2.47) and (2.48) respectively. The last inequality follows from the bounds $T_1 \leq C_1/n$ and $T_2 \leq C_2/n$ established in the proof of Theorem 2.2 (relations (2.47) and (2.48)).

Since

$$\mathcal{R}_n(k,\theta_0) = \frac{k}{n} + \sum_{i=k+1}^{\infty} \theta_{0i}^2 \leq \tau \mathcal{R}_n(N_o,\theta_0)$$

for any $k \in \mathcal{N}(\tau)$, we obtain

$$N_{max} = \max\{\mathcal{N}(\tau)\} \leq \tau n \mathcal{R}_n(N_o,\theta_0).$$

This implies that

$$
\begin{aligned}
&E_{\theta_0}\Big[\sum_{k\in\mathbb{N}}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k|X)\Big]\\
&\quad= E_{\theta_0}\Big[\sum_{k\in\mathcal{N}(\tau)}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k|X)\Big]\\
&\qquad+E_{\theta_0}\Big[\sum_{k\in\mathcal{N}^-(\tau)\cup\mathcal{N}^+(\tau)}\Big(\sum_{i=1}^{k}\frac{\xi_i^2}{n}\Big)P(N=k|X)\Big]\\
&\quad\le E_{\theta_0}\sum_{i=1}^{N_{max}}\frac{\xi_i^2}{n}\sum_{k\in\mathcal{N}(\tau)}P(N=k|X)\\
&\qquad+\sum_{k\in\mathcal{N}^-(\tau)\cup\mathcal{N}^+(\tau)}\frac{k+1}{n}\Big(E_{\theta_0}P^2(N=k|X)\Big)^{1/2}\\
&\quad\le E_{\theta_0}\sum_{i=1}^{N_{max}}\frac{\xi_i^2}{n}+\sum_{k\in\mathcal{N}^-(\tau)\cup\mathcal{N}^+(\tau)}\Big(\mathcal{R}_n(k,\theta_0)+\frac{1}{n}\Big)\Big(E_{\theta_0}P(N=k|X)\Big)^{1/2}\\
&\quad\le \frac{N_{max}}{n}+T_1+T_2\le\tau\mathcal{R}_n(N_o,\theta_0)+\frac{C_1+C_2}{n},
\end{aligned}
\tag{2.55}
$$

where we again used the relations (2.47) and (2.48) to bound the terms T_1 and T_2.

It remains to choose τ,τ_1,τ_2 so that $\tau>\tau_1\ge\tau_-(\alpha)=6\log\big(\frac{2}{\sqrt{3}}\big)+6\alpha$ and $\tau>\tau_2\ge\tau_+(\alpha)=\frac{1}{2\alpha}$. Take, for example, $\tau_1=\tau_-(\alpha)$, $\tau_2=\tau_+(\alpha)$ and $\tau=\max\{\tau_1,\tau_2\}+1$. Thus the constants τ, C_1 and C_2 depend only on α.

Combining (2.52), (2.54) and (2.55) and taking into account that $r_n(\theta_0)=\mathcal{R}_n(N_o,\theta_0)\ge\frac{1}{n}$, we obtain that for any $\theta_0\in\ell_2$

$$
\begin{aligned}
E_{\theta_0}P\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\|^2\ge Mr_n(\theta_0)\,\Big|X\Big\}&\le\frac{3\tau r_n(\theta_0)+3(C_1+C_2)n^{-1}}{4Mr_n(\theta_0)}\\
&\le\frac{3(\tau+C_1+C_2)}{4M}.
\end{aligned}
$$

Finally, combining (2.53), (2.54) and (2.55) in a similar manner, we obtain that for any $\theta_0\in\ell_2$

$$
\begin{aligned}
E_{\theta_0}P'\big\{\|\theta-\theta_0\|^2\ge Mr_n(\theta_0)\,\big|X\big\}&\le\frac{2\tau r_n(\theta_0)+2(C_1+C_2)n^{-1}}{Mr_n(\theta_0)}\\
&\le\frac{2(\tau+C_1+C_2)}{M}.
\end{aligned}
$$

The theorem follows. □

Proof of Corollary 2.1. Write

$$\begin{aligned}
&P\Big\{\big\|\theta-\theta_0+\widehat{\theta}/2\big\|^2 \geq Mr_n(\theta_0)\,\big|X\Big\}\\
&\quad= P\Big\{\big\|\theta-\theta_0+\widehat{\theta}/2\big\| \geq \sqrt{Mr_n(\theta_0)}\,\big|X\Big\}\\
&\quad\leq P\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\| \geq \frac{\sqrt{Mr_n(\theta_0)}}{2}\,\Big|X\Big\} + P\Big\{\Big\|\frac{\widehat{\theta}}{2}-\frac{\theta_0}{2}\Big\| \geq \frac{\sqrt{Mr_n(\theta_0)}}{2}\,\Big|X\Big\}\\
&\quad= P\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\|^2 \geq \frac{Mr_n(\theta_0)}{4}\,\Big|X\Big\} + P\Big\{\big\|\widehat{\theta}-\theta_0\big\|^2 \geq Mr_n(\theta_0)\,\big|X\Big\}.
\end{aligned}$$

Using the conditional Chebyshev inequality, Theorem 2.2 and the fact that always $r_n(\theta_0) = \mathcal{R}_n(N_o,\theta_0) \geq 1/n$, we bound the expectation of the second term in the right hand side of the above inequality: for any $\theta_0 \in \ell_2$,

$$E_{\theta_0} P\Big\{\big\|\widehat{\theta}-\theta_0\big\|^2 \geq Mr_n(\theta_0)\,\big|X\Big\} \leq \frac{E_{\theta_0}\|\widehat{\theta}-\theta_0\|^2}{Mr_n(\theta_0)} \leq \frac{K_1+K_2}{M}.$$

The corollary follows since, by Theorem 2.3, we have that, for any $\theta_0 \in \ell_2$,

$$E_{\theta_0} P\Big\{\Big\|\theta-\frac{\theta_0}{2}\Big\|^2 \geq \frac{Mr_n(\theta_0)}{4}\,\Big|X\Big\} \leq \frac{K_3}{M} \tag{2.56}$$

for some constant $K_3 > 0$ depending only on α. □

Proof of Theorem 2.4. Recall the definitions (2.18), (2.19) and the fact that the random variable $\widehat{N}$ is generated according to $P(N|X)$. We apply Lemma 2.4 and Corollary 2.4. For a given $\alpha \in [\frac{1}{6} - \log\big(\frac{2}{\sqrt{3}}\big), \frac{1}{2}]$, take some values τ, τ_1 and τ_2 such that $\tau > \tau_1 \geq \tau_- = \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha \geq 1$, $\tau > \tau_2 \geq \tau_+(\alpha) = \frac{1}{2\alpha} \geq 1$. Then, according to Lemma 2.4, we obtain

$$\begin{aligned}
P_{\theta_0}\big\{\widehat{N} \notin \mathcal{N}(\tau)\big\} &= E_{\theta_0} P\big\{\widehat{N} \notin \mathcal{N}(\tau)|X\big\}\\
&= \sum_{k\in N_-(\tau)\cup N_+(\tau)} E_{\theta_0} P\big\{\widehat{N}=k|X\big\}\\
&\leq P_-(\theta_0, 1/6, \tau, \tau_1, n) + P_+(\theta_0, \alpha, \tau, \tau_1, n).
\end{aligned}$$

Now fix values $\tau_1 = \tau_-(\alpha)$ and $\tau_2 = \tau_+(\alpha)$. Finally apply Corollary 2.4 to the right hand side of the last inequality and use the trivial relation $\min\{a_1,b_1\} + \min\{a_2,b_2\} \leq \min\{a_1+a_2, b_1+b_2\}$ to get the statement of the theorem:

$$P_{\theta_0}\big\{\widehat{N} \notin \mathcal{N}(\tau)\big\} \leq \min\Big\{\frac{B_1}{n\mathcal{R}_n(N_o,\theta_0)}, \frac{B_2 n\mathcal{R}_n(N_o,\theta_0)}{e^{B_3 n\mathcal{R}_n(N_o,\theta_0)}}\Big\},$$

where

$$\begin{aligned}
B_1 &= B_1(\alpha,\tau) = \frac{c_1 B^{-2} + g(D)}{\tau},\\
B_2 &= B_2(\alpha,\tau) = \frac{1+c_3}{\tau},\\
B_3 &= B_3(\alpha,\tau) = \min\{B, D\},
\end{aligned}$$

$c_1 = 4e^{-2}$, $c_3 = (e+e^2)/4$, $B = B(\tau, \tau_-(\alpha), 1/6) = \frac{\tau - \tau_-(\alpha)}{6\tau}$, $D = D(\tau, \tau_+(\alpha), \alpha) = \frac{\alpha(\tau - \tau_+(\alpha))}{\tau}$ and $g(D) = e^{-1}D^{-2} + e^{-D}(e^D - 1)^{-2}$. □

Remark 2.20. To get uniform constants, fix some $\alpha \in [\frac{1}{6} - \log\big(\frac{2}{\sqrt{3}}\big), \frac{1}{2}]$, take some value τ such that $\tau > \tau_-(\alpha) = 6\log\big(\frac{2}{\sqrt{3}}\big) + 6\alpha$ and $\tau > \tau_+(\alpha) = \frac{1}{2\alpha}$. There are many choices possible, for example, take α such that $\tau_-(\alpha) = \tau_+(\alpha) = \frac{1}{2\alpha}$, which is $\alpha \approx 0.226$ and $\tau_-(\alpha) = \tau_+(\alpha) \approx 2.217$. Next, take $\tau = \tau_-(\alpha) + 2 \approx 4.217$, then $B_1 \approx 44.87$, $B_2 \approx 0.84$, $B_3 \approx 0.08$.

Proof of Theorem 2.5. By the definition of the MAP selector $\widehat{N}_{\text{MAP}}$,

$$\big\{P(N = \widehat{N}_{\text{MAP}}|X) \le \delta\big\} \subseteq \big\{P(N \notin \mathcal{N}(\tau, \theta_0)|X) \ge 1 - |\mathcal{N}(\tau, \theta_0)|\delta\big\},$$

$$\big\{\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau, \theta_0), P(N = \widehat{N}_{\text{MAP}}|X) > \delta\big\} \subseteq \big\{P(N \notin \mathcal{N}(\tau, \theta_0)|X) \ge \delta\big\}.$$

Notice that $|\mathcal{N}(\tau, \theta_0)|$ is finite for any $\theta_0 \in \ell_2$. Next, apply the above relation and the Markov inequality:

$$\begin{aligned} P_{\theta_0}\big(\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau, \theta_0)\big) &= P_{\theta_0}\big(\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau, \theta_0), P(N = \widehat{N}_{\text{MAP}}|X) \le \delta\big) \\ &\quad + P_{\theta_0}\big(\widehat{N}_{\text{MAP}} \notin \mathcal{N}(\tau, \theta_0), P(N = \widehat{N}_{\text{MAP}}|X) > \delta\big) \\ &\le \frac{E_{\theta_0} P(N \notin \mathcal{N}(\tau, \theta_0)|X)}{1 - |\mathcal{N}(\tau, \theta_0)|\delta} + \frac{E_{\theta_0} P(N \notin \mathcal{N}(\tau, \theta_0)|X)}{\delta}. \end{aligned}$$

Take

$$\delta = (|\mathcal{N}(\tau, \theta_0)|^{1/2} + |\mathcal{N}(\tau, \theta_0)|)^{-1},$$

apply Theorem 2.4 and recall that $|\mathcal{N}(\tau, \theta_0)| \le \tau n \mathcal{R}(N_o, \theta_0)$ to complete the proof. □

Chapter 3

Lower bound for the posterior convergence rate

In the previous chapter, a new notion for the oracle posterior concentration rate is proposed. The program of oracle estimation and Bayes oracle posterior optimality is fully implemented in the above chapter for the Gaussian white noise model and the projection estimators family.

Let us recall the model from Chapter 2. We observe $X = (X_i)_{i\in\mathbb{N}}$:

$$X_i \overset{ind}{\sim} N(\theta_i, n^{-1}), \quad i \in \mathbb{N},$$

where $\mathbb{N} = \{1, 2, \ldots\}$, $\theta = (\theta_i)_{i\in\mathbb{N}}$ is an unknown parameter of interest from the space ℓ_2 with the usual norm $\|\cdot\|$, the noise variables ξ_i are independent, identically distributed $N(0, 1)$ random variables, parameter n^{-1} is the noise intensity. The model is known to be a sequence version of the Gaussian white noise model; the details can be found in Section 1.5. The goal is to make an inference on the unknown parameter θ by using a Bayes approach in the asymptotic setup as $n \to \infty$. In Chapter 2, we put a simple conjugate prior on the signal θ and investigated the performance of the resulting Bayesian procedure, namely the convergence rate of the resulting posterior. A new benchmark for the posterior concentration rate was proposed, the so called oracle (quadratic) risk rate,

$$r_n(\theta_0) = \min_{N\in\mathbb{N}} \mathcal{R}_n(N, \theta_0) = \inf_{\widehat{\theta}\in\widehat{\Theta}(\mathcal{N})} E_\theta \|\widehat{\theta} - \theta_0\|^2,$$

the smallest possible risk over the class of projection estimators $\widehat{\Theta}$, or equivalently, over the family of the posterior rates corresponding to the family of priors $\{\pi_N,\ N \in \mathbb{N}\}$. In Chapter 2, the relation of type (1.5) was derived for the oracle projection posterior rate $r_n(\theta_0)$ (as compared to (1.5), here we deal with the squared distance function). It means that the resulting posterior convergence rate is at least $r_n(\theta_0)$, i.e. we have established an upper bound result.

In this chapter we complement the upper bound results of Chapter 2 on posterior concentration rate by a lower bound result, namely that the concentration rate of the posterior distribution around the true signal value θ_0 can not

be faster than the oracle projection rate:

$$E_{\theta_0} P\big\{r_n^{-1}(\theta_0)\|\theta-\theta_0\|^2 \le \delta | X\big\} \le \psi(\delta),$$

for some function ψ monotonically decreasing to zero as $\delta \to 0$ and all $\theta_0 \in \ell_2$. This means that the oracle posterior rate $r_n(\theta_0)$ can not be improved, which together with the upper bound result implies that the posterior mass is concentrated in the annulus

$$\{\theta \in \ell_2 : \delta r_n(\theta_0) \le \|\theta-\theta_0\|^2 \le M r_n(\theta_0)\}$$

for a sufficiently small $\delta > 0$ and a sufficiently large $M > 0$.

An asymptotic lower bound result for the posterior rate was first obtained by Castillo (2008) in case of a Gaussian prior. The method was essentially based on the so called concentration function of the Gaussian process and the small ball probabilities for a Gaussian process; see Lifshits (1995), van der Vaart and van Zanten (2008) and further references therein. As compared to Castillo (2008), our result is somewhat more specific, as it concerns a particular prior (a mixture of some normals), but also more refined in that it is nonasymptotic and uniform over ℓ_2. Our result is also local in the sense that it concerns the oracle projection rate $r_\epsilon(\theta_0)$, which is a local quantity, since it depends on the "true" values θ_0.

3.1 Lower bound

Consider first a general situation. We adopt the notations and notions from the Introduction. Assume we have a model $X \sim P_\theta^{(\epsilon)}$ and a family of priors $\Pi = \{\pi_\gamma,\, \gamma \in \Gamma\}$ with a countable index set Γ, which leads to the corresponding family of posteriors $\{P_\gamma(\theta|X),\, \gamma \in \Gamma\}$. Let $\{r_\epsilon(\pi_\gamma, \theta),\, \gamma \in \Gamma\}$ be a family of posterior rates and the oracle posterior rate is then

$$r_\epsilon(\Pi, \theta) = \inf_{\gamma \in \Gamma} r_\epsilon(\pi_\gamma, \theta).$$

A good candidate for an adaptive prior is a mixture over the priors from the family Π. If the adaptive prior is a mixture, then the resulting adaptive posterior is a (random) mixture of posteriors. Let $\pi = \sum_{\gamma\in\Gamma} \lambda_\gamma \pi_\gamma$ be a mixture prior for some distribution λ on Γ and let $P(\theta|X)$ be the corresponding posterior.

A way to establish a lower bound with the oracle rate for this adaptive posterior is described by the proposition below. It expresses the following simple intuitive idea: the smallest lower bound among the members of the mixture will clearly be a lower bound for the resulting adaptive posterior.

Proposition 3.1. *Assume that for a $\delta > 0$, some function ψ and all $\theta_0 \in \Theta_0$ the following lower bound relation holds for each $\gamma \in \Gamma$:*

$$P_\gamma\{r_\epsilon^{-1}(\pi_\gamma, \theta_0)\, d(\theta, \theta_0) \le \delta | X\} \le \psi(\delta)$$

with P_{θ_0}-probability 1. Then, for all $\theta_0 \in \Theta_0$,

$$P\{r_\epsilon^{-1}(\Pi, \theta_0)\, d(\theta, \theta_0) \le \delta | X\} \le \psi(\delta)$$

with P_{θ_0}-probability 1.

Proof. The proof is straightforward. Denote by $P(G = \gamma|X)$, $\gamma \in \Gamma$, the posterior distribution of the index parameter γ given X. Then, as $P\{\theta|X, G = \gamma\} = P_\gamma\{\theta|X\}$,

$$\begin{aligned} & P\{r_\epsilon^{-1}(\Pi, \theta_0)\, d(\theta, \theta_0) \le \delta | X\} \\ &= \sum_{\gamma \in \Gamma} P_\gamma\{d(\theta, \theta_0) \le \delta r_\epsilon(\Pi, \theta_0)|X\} P(G = \gamma|X) \\ &\le \sum_{\gamma \in \Gamma} P_\gamma\{d(\theta, \theta_0) \le \delta r_\epsilon(\pi_\gamma, \theta_0)|X\} P(G = \gamma|X) \le \psi(\delta) \end{aligned}$$

with P_{θ_0}-probability 1. □

In what follows we employ the notations and notions about the oracle projection posterior rate developed in Chapter 2. Below is a lower bound result for the family of posteriors $\{P'_N(\theta|X),\, N \in \mathbb{N}\}$ defined by (2.8): the posterior $P'_N(\theta|X)$ concentrates around θ_0 with the rate $\mathcal{R}_n(N, \theta_0)$, and not faster.

Theorem 3.1. *Let the oracle rate $\mathcal{R}_n(N, \theta_0)$ be defined by (2.3) and $P'_N(\theta|X)$ be defined by (2.8). Then there exists a constant $C > 0$ such that for any $\theta_0 \in \ell_2$ and all $N \in \mathbb{N}$*

$$P'_N\big\{\|\theta - \theta_0\|^2 \le \delta \mathcal{R}_n(N, \theta_0)\,\big|X\big\} \le C \frac{(2e\delta)^{N/2}}{\sqrt{N}}$$

for any $0 < \delta \le 1$ with P_{θ_0}-probability 1.

The proof of this theorem is given in the next section. Notice that trivially for all $0 < \delta \le e^{-1}/2$ and all $N \in \mathbb{N}$

$$\frac{(2e\delta)^{N/2}}{\sqrt{N}} \le C' \sqrt{\delta}$$

for some constant C'. Combining the above theorem and Proposition 3.1, we derive the following corollary.

Corollary 3.1. *Let the oracle rate $r_n(\theta_0)$ be defined by (2.4) and $P'(\theta|X)$ be defined by (2.11) with any $\alpha > 0$. Then there exists a constant $C > 0$ such that for any $\theta_0 \in \ell_2$*

$$P'\big\{\|\theta - \theta_0\|^2 \le \delta r_n(\theta_0)\,\big|X\big\} \le C\sqrt{\delta}$$

for any $0 < \delta \le e^{-1}/2$ with P_{θ_0}-probability 1.

Note that the above result is uniform over n (i.e. nonasymptotic) and $\theta_0 \in \ell_2$.

The following corollary states essentially that the posterior rate $r_n(\theta_0)$ is sharp.

Corollary 3.2. *Under the conditions of Corollary 3.1 and Theorem 2.3 from Chapter 2, there exist constants $C_1, C_2 > 0$ such that for any $\theta_0 \in \ell_2$*

$$E_{\theta_0} P'\big\{\delta r_n(\theta_0) \le \|\theta - \theta_0\|^2 \le M r_n(\theta_0) \,\big|X\big\} \ge 1 - \frac{C_1}{M} - C_2\sqrt{\delta}.$$

The above corollary essentially means that almost all of the posterior mass is concentrated in the annulus $\{\theta \in \ell_2 : \ \delta r_n(\theta_0) \le \|\theta - \theta_0\|^2 \le M r_n(\theta_0)\}$ for sufficiently small δ and sufficiently large M.

Remark 3.1. It is a challenging problem to derive this kind of results for a wide class of priors instead of specific ones. It should be a nontrivial class of priors for which some reasonable upper bound result hold, because lower bound results are only interesting when related to "good" upper bound results. Indeed, if one takes some trivial degenerate prior, then the corresponding lower bound will be "very good", while the upper bound will certainly be "very poor". A lower bound makes sense only together with a reasonable corresponding upper bound, and only in cases when the posterior rates in these two bounds coincide (up to a constant factor or maybe some log factor).

3.2 Proof of the theorem

Denote the Euclidean ball of radius r in space $\mathbb{R}^k$ by $B_k(r) = \{x \in \mathbb{R}^k : \|x\|_k^2 \le r^2\}$, here $\|\cdot\|_k$ is the usual Euclidean norm in space $\mathbb{R}^k$. Let $\Lambda(S)$ be the Lebesgue measure (or volume) of a set $S \subset \mathbb{R}^k$. It is known that the Lebesgue measure of the ball $B_k(r)$ is

$$\Lambda(B_k(r)) = r^k \Lambda(B_k(1)) = \frac{r^k \pi^{k/2}}{\Gamma(k/2+1)},$$

where $\Gamma(u)$ is the Gamma function. Using Stirling's approximation for the Gamma function

$$\Gamma(x) = \sqrt{2\pi}\, x^{x-1/2} e^{-x+\kappa/(12x)}$$

for all $x \ge 1$ and $0 \le \kappa \le C$, we have a bound

$$\begin{aligned}
\Gamma(k/2+1) &= \sqrt{2\pi}\Big(\frac{k}{2}+1\Big)^{(k+1)/2} e^{-k/2-1+\kappa/(6k+12)} \\
&= \frac{(1+2/k)^{(k+1)/2}\sqrt{\pi}}{e^{1-\kappa/(6k+12)}} k^{(k+1)/2} (2e)^{-k/2} \\
&= c_k k^{(k+1)/2} (2e)^{-k/2} \ge c k^{(k+1)/2} (2e)^{-k/2},
\end{aligned}$$

with

$$c_k = \frac{(1+2/k)^{(k+1)/2}\sqrt{\pi}}{e^{1-\kappa/(6k+12)}} > \frac{\sqrt{\pi}}{e} = c.$$

Of course, $c_k \to \sqrt{\pi}$ as $k \to \infty$ and more accurate estimates $c_k > c$ are possible. Therefore

$$\Lambda(B_k(r)) \le c^{-1} r^k \pi^{k/2} k^{-(k+1)/2} (2e)^{k/2}. \tag{3.1}$$

Let $Z_1, \dots, Z_k$ be independent $N(0,1)$ random variables and $\xi = (\xi_i)_{i\in\mathbb{N}} = n^{1/2}(X_i - \theta_{i0})$ are iid standard normal random variables under $X \sim P_{\theta_0}$. Further recall the conditional distribution $P'_k(\theta|X) = P'(\theta|X, N = k)$ given by (2.8):

$$\theta_i|(X, N = k) \overset{ind}{\sim} N\Big(X_i I\{i \le k\}, \frac{n^{-1}}{2} I\{i \le k\}\Big), \quad i \in \mathbb{N}.$$

Using this and Anderson's inequality, we obtain that for $\delta \le 1$, with P_{θ_0}-probability 1,

$$\begin{aligned}
&P'_k\big\{\|\theta - \theta_0\|^2 \le \delta\mathcal{R}_n(k,\theta_0)\,\big|X\big\} \\
&= P'_k\Big\{\sum_{i=1}^{k}(\theta_i - \theta_{i0})^2 + \sum_{i=k+1}^{\infty}\theta_{i0}^2 \le \delta\mathcal{R}_n(k,\theta_0)\,\big|X\Big\} \\
&= P\Big\{n^{-1}\sum_{i=1}^{k}\Big(\frac{Z_i}{\sqrt{2}} + \xi_i\Big)^2 + \sum_{i=k+1}^{\infty}\theta_{i0}^2 \le \delta\mathcal{R}_n(k,\theta_0)\,\big|\xi\Big\} \\
&\le P\Big\{n^{-1}\sum_{i=1}^{k}\frac{Z_i^2}{2} + \sum_{i=k+1}^{\infty}\theta_{i0}^2 \le \delta\mathcal{R}_n(k,\theta_0)\Big\}.
\end{aligned}$$

Now, taking into account the trivial inequality

$$P\big\{\sum_{i=1}^{k} Z_i^2 \le R^2\big\} \le (2\pi)^{-k/2}\Lambda\big(B_k(R)\big)$$

and (3.1), we proceed

$$\begin{aligned}
&P\Big\{n^{-1}\sum_{i=1}^{k}\frac{Z_i^2}{2} + \sum_{i=k+1}^{\infty}\theta_{i0}^2 \le \delta\mathcal{R}_n(k,\theta_0) = \frac{\delta k}{n} + \delta\sum_{i=k+1}^{\infty}\theta_{i0}^2\Big\} \\
&\le P\Big\{\sum_{i=1}^{k} Z_i^2 \le 2k\delta\Big\} \le (2\pi)^{-k/2}\Lambda\big(B_k(\sqrt{2k\delta})\big) \\
&\le (2\pi)^{-k/2}c^{-1}(2k\delta)^{k/2}\pi^{k/2}k^{-(k+1)/2}(2e)^{k/2} = \frac{1}{c}\frac{(2e\delta)^{k/2}}{\sqrt{k}}.
\end{aligned}$$

Combining the last two inequalities completes the proof of the theorem.

Chapter 4

Posterior studies by numerical simulations

In this chapter we apply the Bayesian approach to the problem of recovering a signal corrupted by a Gaussian noise. Utilizing the Bayesian approach allows us to study the performance of posterior distributions, which are the main objects within the Bayesian framework. We put a two level hierarchical prior on the multidimensional normal mean and on the dimension itself, which leads to the posterior distribution of the signal and the dimension parameter N, also called a cut-off parameter. Using the posterior distribution of the cut-off N, we construct a Bayesian selection method for the dimension parameter, which mimics the oracle cut-off.

We measure the performance of our Bayesian procedure by comparing the risk for the Bayesian cut-off with the oracle risk over projection estimators, i.e. the risk of the best projection estimator, which is attained by the oracle cut-off. Our goal is to "mimic" the oracle cut-off by a Bayesian selector.

In this chapter, we choose the Bayesian cut-off parameter as a median or a maximum of the posterior probability (MAP), or as the smallest natural number larger than the half of the MAP. In many cases it turns out that the Bayesian cut-off parameters do not coincide with the oracle cut-off. However, the ratio of the risk of the projection estimator, constructed with the Bayesian cut-off, and the oracle risk is close to one. Firstly, we conduct the simulation studies for each of the above described Bayesian cut-off parameters on test sequences. Secondly, we look at the posterior distribution of signal given data. As it has been shown in Chapter 2, the posterior distribution corresponding to the Gaussian priors has to be adjusted to achieve good asymptotic properties. One of the methods to adjust the posterior, proposed in the above chapter, is the empirical Bayesian approach (other approaches are also considered in Chapter 2). The numerical studies show that the adjusted priors indeed concentrate in some annulus around the true value of the signal, while the posterior mass of unadjusted prior does not. Moreover, the simulations show that the annulus shrinks towards the value of the true signal as the noise level decreases, which corresponds to an increasing

sample size in the asymptotically equivalent nonparametric discrete regression model. In addition, we also study the performance of our Bayesian procedure for the asymptotically equivalent nonparametric discrete regression model. We construct an $(1-\alpha)$-credible band for the true signal f_0.

To conduct the numerical studies the test sequences of signal are chosen in the form of Step, Damped-Sine, Outliers, and as the Fourier coefficients of Bumps function introduced by Donoho and Johnstone (1995).

4.1 Model and preliminaries

Suppose that we observe the data $(Y_1, \dots, Y_n)$ such that

$$Y_i = f\Big(\frac{i}{n}\Big) + Z_i, \quad i = 1, 2, \dots, n, \tag{4.1}$$

where Z_i's are independent $\mathcal{N}(0,1)$ random variables and f is an unknown function from $L_2[0,1]$ at the equidistant points i/n. Let $\{\varphi_j\}_{j=1}^{\infty}$ be an orthonormal basis in $L_2[0,1]$ then $f(t) = \sum\limits_{j=1}^{\infty} \widetilde{\theta}_j \varphi_j(t)$, where $\widetilde{\theta}_j = \int f(t)\varphi_j(t)dt$. Since the function $f(t)$ is from $L_2[0,1]$, the series $\sum\limits_{j=1}^{\infty} \widetilde{\theta}_j^2$ converges, which implies vanishing $\widetilde{\theta}_j$ as j goes to infinity.

Using the discrete Fourier transformation of equation (4.1) we convert the problem of estimating the function $f(\frac{i}{n})$, $i = 1, \dots, n$ to the problem of estimating n values of its Fourier coefficients. Indeed, this model can be represented as the equivalent Gaussian sequence model,

$$y_j = \theta_j + \sigma\xi_j, \quad j = 1, 2, \dots, n, \tag{4.2}$$

where $y_j = \frac{1}{n}\sum\limits_{i=1}^{n} Y_i \varphi_j\big(\frac{i}{n}\big)$, $\theta_j = \frac{1}{n}\sum\limits_{i=1}^{n} f\big(\frac{i}{n}\big)\varphi_j\big(\frac{i}{n}\big) \approx \widetilde{\theta}_j$, ξ_j's are i.i.d. random variables distributed according to $\mathcal{N}(0,1)$ and $\sigma = 1/\sqrt{n}$. In general, for an arbitrary basis this model is asymptotically, as n goes to infinity, equivalent to the Gaussian sequence model (4.2) as was shown in Chapter 1. In what follows we will focus on the equivalent to (4.1) model (4.2) and all the inference on f can be in principle derived from the inference on θ.

Notice that the approximate relation $\theta_j \approx \widetilde{\theta}_j$ becomes asymptotically exact as n goes to infinity and thus, θ_j also vanishes as j goes to infinity. Therefore, it makes sense to consider the first N most relevant components and neglect the others with significant noise contribution. This gives rise to projection estimation. Actually, neglecting noisy Fourier coefficients y_j with $j = N+1, \dots, n$ leads to the same projection estimation of a function as Fourier decomposition in the truncated basis, φ_j, $j = 1, \dots, N$. The projection estimator of a function in the truncated basis has been considered in details by Tsybakov (2008), Efromovich (1999). The projection estimation in general was studied in different scopes by many authors, e.g. Birgé Massart (2001), Cavalier and Tsybakov (2001), Tsybakov (2008), Efromovich (1999), Kneip (1994), etc.

Let us recall the setting. The class of projection estimators $\{\widehat{\boldsymbol{\theta}}(N),\, N \in \mathbb{N}\}$ is parameterized by the cut-off parameter N, with $\widehat{\boldsymbol{\theta}}(N)$ defined by the rule

$$\widehat{\boldsymbol{\theta}}(N) = \big(\widehat{\theta}_1(N), \dots, \widehat{\theta}_n(N)\big), \quad \widehat{\theta}_j(N) = y_j I\{j \le N\}, \quad j = 1, 2, \dots, n \tag{4.3}$$

The estimation quality of the estimator $\widehat{\boldsymbol{\theta}}(N)$ is measured by the risk in ℓ_2 norm

$$\mathcal{R}(N, \boldsymbol{\theta}) = \mathcal{R}(\widehat{\boldsymbol{\theta}}(N), \boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}(N) - \boldsymbol{\theta}\|_2^2 = E_\theta \sum_{j=1}^{n} ((\widehat{\theta}(N))_j - \theta_j)^2, \tag{4.4}$$

which in the case of fixed (known) parameter N is equal to

$$\mathcal{R}(N, \boldsymbol{\theta}) = E_\theta \Big[\sum_{j=1}^{N} \sigma^2 \xi_j^2 + \sum_{j=N+1}^{n} \theta_j^2 \Big] = N\sigma^2 + \sum_{j=N+1}^{n} \theta_j^2.$$

For a fixed $\boldsymbol{\theta} \in \ell_2$, we define the oracle $N_o = N_o(\boldsymbol{\theta}) \in \mathbb{N}$ and the *oracle projection risk* $r_\sigma(\boldsymbol{\theta}) = \mathcal{R}(N_o, \boldsymbol{\theta})$ by the following relation:

$$r_\sigma(\boldsymbol{\theta}) = \mathcal{R}(N_o, \boldsymbol{\theta}) = N_o \sigma^2 + \sum_{j=N_o+1}^{n} \theta_j^2 = \min_{N \in \mathbb{N}} \mathcal{R}(N, \boldsymbol{\theta}). \tag{4.5}$$

Further we apply Bayesian approach to construct adaptive projection estimators (or adaptive cut-offs) and study the performance of the Bayesian procedures by numerical simulations on the test examples.

4.2 Bayesian analysis

In this section we recall the Bayesian approach to the problem of adaptive projection estimation.

First we have to design a Bayesian model, which means to impose a prior distribution on a probability measure on the set of parameters, i.e. devise a hierarchical prior on $(\boldsymbol{y}, \boldsymbol{\theta}, N)$:

$$\begin{aligned} y_j | \theta_j, N &\sim \mathcal{N}[\theta_j, \sigma^2], \\ \theta_j | N &\sim \mathcal{N}[0, \tau_j(N)], \\ N &\sim \lambda, \end{aligned} \tag{4.6}$$

where $\tau_j(N)$ is defined by

$$\tau_j(N) = \sigma^2 I\{j \le N\}, \quad j \in \{1, \dots, n\}. \tag{4.7}$$

The prior λ is determined as follows:

$$P(N = k) = \lambda_k = c(\alpha) e^{-\alpha k}, \quad k \in \mathbb{N},$$

where $c(\alpha)$ is a normalizing constant and the constant $\alpha = 0.226$. Other choices of α are possible, see the discussion on this issue in Chapter 2 or Babenko and Belitser (2010).

By convention we assume that if $\theta_j \sim \mathcal{N}(0,0)$ then $P(\theta_j = 0) = 1$. Denote the probability density function (pdf) of Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ by $\mathrm{p}(\mathrm{x}; \mu, \sigma^2)$. Then the joint posterior probability density function $p(\boldsymbol{y}, \boldsymbol{\theta}, N)$, corresponding to the imposed priors (4.6), is readily obtained as

$$p(\boldsymbol{y}, \boldsymbol{\theta}, N) = \lambda_k \prod_{i=1}^{n} \mathrm{p}\big(y_i; \theta_i, \sigma^2\big)\mathrm{p}\big(\theta_i; 0, \tau_i(k)\big),$$

the posterior pdf of N given $\boldsymbol{y}$ is

$$p(N = k|\boldsymbol{y}) = \frac{\prod_{i=1}^{n} \mathrm{p}(y_i; 0, \sigma^2 + \tau_i(k))\lambda_k}{\sum_s \big(\lambda_s \prod_{i=1}^{n} \mathrm{p}(y_i; 0, \sigma^2 + \tau_i^2(s))\big)}, \tag{4.8}$$

and the posterior pdf of $\boldsymbol{\theta}$ given $\boldsymbol{y}$ is

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\sum_k \lambda_k \prod_{i=1}^{n} \mathrm{p}(y_i; \theta_i, \sigma^2)\mathrm{p}(\theta_i; 0, \tau_i(k))}{\sum_s \big(\lambda_s \prod_{i=1}^{n} \mathrm{p}(y_i; 0, \sigma^2 + \tau_i^2(s))\big)}. \tag{4.9}$$

The product $\mathrm{p}(y_i; \theta_i, \sigma^2)\mathrm{p}(\theta_i; 0, \tau_i(k))$ in the right hand side of (4.9) can be rewritten as

$$\mathrm{p}(y_i; 0, \sigma^2 + \tau_i(k))\mathrm{p}\Big(\theta_i; \frac{\tau_i(k)y_i}{\sigma^2 + \tau_i(k)}, \frac{\sigma^2\tau_i(k)}{\sigma^2 + \tau_i(k)}\Big). \tag{4.10}$$

Thus, from (4.9) using (4.10) we obtain

$$\begin{aligned} & p(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \frac{1}{S}\sum_k \lambda_k \prod_{i=1}^{n} \mathrm{p}(y_i; 0, \sigma^2 + \tau_i(k))\mathrm{p}\Big(\theta_i; \frac{\tau_i(k)y_i}{\sigma^2 + \tau_i(k)}, \frac{\sigma^2\tau_i(k)}{\sigma^2 + \tau_i(k)}\Big), \end{aligned} \tag{4.11}$$

with $S = \sum_s \big(\lambda_s \prod_{i=1}^{n} \mathrm{p}(y_i; 0, \sigma^2 + \tau_i^2(s))\big)$.

The posterior pdf (4.8) and (4.11) represent the resulting posterior densities of the Bayesian procedure with respect to imposed priors (4.6). Next, we construct the adaptive projection estimator $\widehat{\boldsymbol{\theta}}(\widehat{N})$. Define the Bayesian cut-offs as

$$N_{MAP} = \mathrm{argmax}\big\{p\big(N = k|\boldsymbol{y}\big),\ k \in \mathbb{N}\big\},$$

the point with the maximum posterior pdf,

$$N_{MED} = \mathrm{median}\big\{p\big(N = k|\boldsymbol{y}\big),\ k \in \mathbb{N}\big\},$$

the median of the posterior pdf, and

$$\widetilde{N}(\delta) = \min\big\{k \in \mathbb{N} :\ p\big(N = k|\boldsymbol{y}\big) \geq \delta\big\},$$

with $\delta = \frac{1}{2}\max\big\{p\big(N = k|\boldsymbol{y}\big),\ k \in \mathbb{N}\big\}$.

Accomplishing the construction of the adaptive projection estimators $\widehat{\boldsymbol{\theta}}(\widehat{N})$, we take $\widehat{N}$ as N_{MAP}, N_{MED}, $\widetilde{N}(\delta)$ and then set

$$\widehat{\boldsymbol{\theta}}(\widehat{N}) = \big(\widehat{\theta}_1(\widehat{N}), \dots, \widehat{\theta}_n(\widehat{N})\big), \quad \widehat{\theta}_j(\widehat{N}) = y_j I\{j \le \widehat{N}\}, \quad j = 1, 2, \dots, n. \tag{4.12}$$

Thus, we get $\widehat{\boldsymbol{\theta}}(N_{MAP})$, $\widehat{\boldsymbol{\theta}}(N_{MED})$ and $\widehat{\boldsymbol{\theta}}(\widetilde{N}(\delta))$.

Notice that the Bayes estimator in this case is $\widehat{\boldsymbol{\theta}}^B = E(\boldsymbol{\theta}|\boldsymbol{y})$, where the conditional expectation $E(\boldsymbol{\theta}|\boldsymbol{y})$ is taken with respect to (4.11). However, it turns out that the Bayes estimator converges to $\boldsymbol{\theta}/2$ instead of $\boldsymbol{\theta}$. To obtain a consistent estimator, we modify the original priors by introducing new shift parameters $\{\mu_j\}_{j=1}^n$, which leads to an empirical Bayes estimator. As it was shown in Babenko and Belitser (2010) (see also the previous chapter), by choosing the shift parameters as maximum likelihood estimates of the marginal distribution of $\boldsymbol{y}$, one obtains a consistent empirical Bayes estimator.

Now we adjust the posterior (4.11) and accomplish the construction of the empirical Bayes estimator. Therefore, instead of the posterior pdf $p(\boldsymbol{\theta}|\boldsymbol{y})$, defined by (4.11), we can use the posterior pdf defined as $p'(\boldsymbol{\theta}|\boldsymbol{y})$:

$$\begin{aligned} & p'(\boldsymbol{\theta}|\boldsymbol{y}) \\ & \quad = \quad \frac{1}{S}\sum_k \lambda_k \prod_{i=1}^n \mathrm{p}(y_i; 0, \sigma^2 + \tau_i(k))\mathrm{p}\Big(\theta_i; \frac{2\tau_i(k) y_i}{\sigma^2 + \tau_i(k)}, \frac{\sigma^2 \tau_i(k)}{\sigma^2 + \tau_i(k)}\Big). \end{aligned} \tag{4.13}$$

By using the posterior pdf $p'(\boldsymbol{\theta}|\boldsymbol{y})$ we construct the empirical Bayes estimator $\widehat{\boldsymbol{\theta}}^{EB} = E'(\boldsymbol{\theta}|\boldsymbol{y}) = \int \boldsymbol{\theta} p'(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$, i.e.

$$\widehat{\boldsymbol{\theta}}^{EB} = (\widehat{\theta}_1, \dots \widehat{\theta}_n), \quad \widehat{\theta}_j = E'(\theta_j|\boldsymbol{y}) = \int \theta_j p'(\boldsymbol{\theta}|\boldsymbol{y}) \prod_{i=1}^n d\theta_i, \quad j = 1, \dots, n,$$

where the conditional expectation $E'(\theta_j|\boldsymbol{y})$ taken with respect to the posterior (4.13) is

$$E'(\theta_j|\boldsymbol{y}) = \sum_k p(N = k|\boldsymbol{y}) y_j I\{j \le k\}.$$

The exact calculation of the conditional expectation $E'(\theta_j|\boldsymbol{y})$ is given in appendix. The estimators and their efficiency is further studied numerically in the next section of the chapter.

4.3 Numerical simulations

In order to compare the performance of the estimators, we conducted a simulation study on Fourier coefficients $\{\theta_j\}_{j=1}^n$ of the following form: Step, Damped-Sine, and Outliers. In addition, we also considered the Fourier coefficients of Bumps function introduced by Donoho and Johnstone (1995). The functions are defined in Table 4.1. The Bumps function is defined explicitly, while the

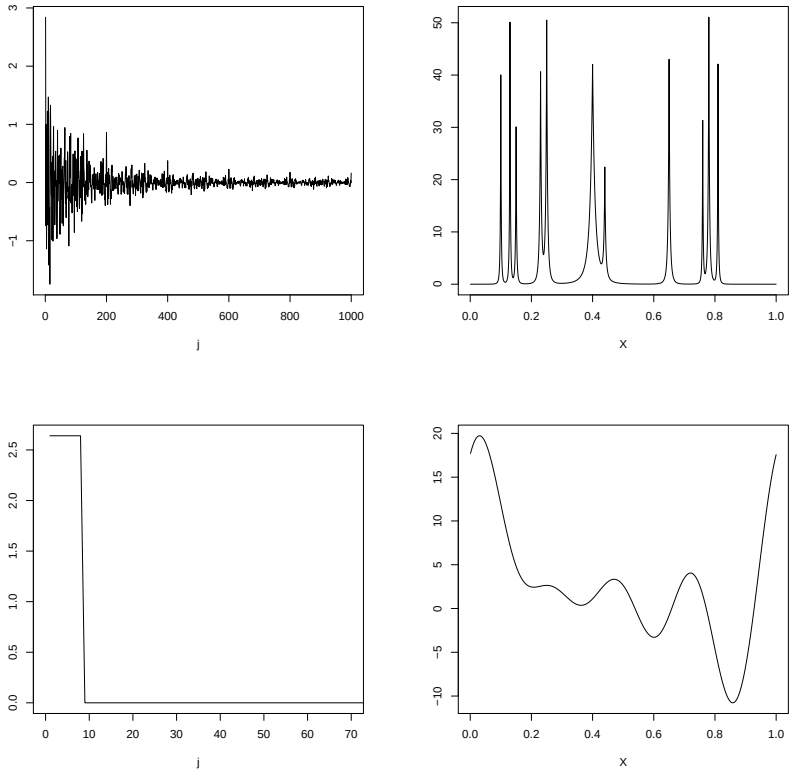


Figure 4.1: The plot of Bumps function (on the top right panel) and its Fourier coefficients (top left). The Step form of the Fourier coefficients (bottom left) and the reconstructed function (bottom right).

others are defined through their Fourier coefficients. In order to obtain the Fourier coefficients of Bumps function we use a discrete Fourier transformation of the function with respect to the orthonormal trigonometric basis defined by

$$\begin{aligned} \varphi_1(x) &\equiv 1 \\ \varphi_{2k}(x) &= \sqrt{2}\sin(2\pi kx) \\ \varphi_{2k+1}(x) &= \sqrt{2}\cos(2\pi kx), \quad k = 1, 2, \ldots \end{aligned}$$

The plots of the Fourier coefficients and the Bumps function itself are shown on Figure 4.1 (top left and top right correspondingly). The Step Fourier coefficients and the corresponding function, reconstructed via the inverse discrete Fourier transformation, are depicted as well on Figure 4.1 (bottom left and bottom right correspondingly).

The Fourier coefficients of Damped-Sine and the Outliers are presented on

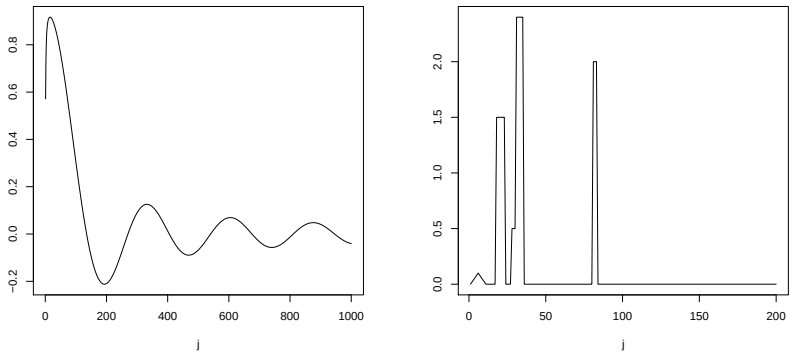


Figure 4.2: The test Fourier coefficients: Damped-Sine (on the left) and Outliers (on the right).

Table 4.1: Formulas of test sequences

(a) $Bumps function$
$f(t) = 10 \sum h_j K((t - t_j)/w_j) \quad K(t) = (1 + \lvert t \rvert)^{-4}$
$(t_j) = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81)$
$(h_j) = (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2)$
$(w_j) = (.005, .005, .006, .01, .01, .03, .01, .01, .005, .008, .005)$
(b) $Step$
$\theta_j = 2.64u(j) - 2.64u(j - 9) \quad u(t) = I\{t \geq 0\}$
(c) $Damped - Sine$
$\theta_j = (295/(7j + 5)) \sin(j/43)$
(d) $Outliers$
$\theta_j = (jh - h)u(j - 2) - (2jh - 12h)u(j - 7) + (jh - 11h)u(j - 12)$
$+1.5u(j - 18) - 1.5u(j - 24) + 0.5u(j - 28) + 1.9u(j - 31) - 2.4u(j - 36)$
$+2u(j - 81) - 2u(j - 84),$
$u(t) = I\{t \geq 0\}$ *is a unit step function,* $\quad h = 1/50$

Figure 4.2. Notice that the Fourier coefficients in the form of Step and Outliers are sparse, i.e. the Fourier coefficients with relatively few non zero elements.

As it has been mentioned in Section 4.1, we observe the noisy Fourier coefficients $\boldsymbol{y}$, which are displayed in Figure 4.3 and Figure 4.4. After that we apply the estimators, described in Section 4.2, to the observations. Thus, firstly, we consider projection estimators to choose the best (on average) Bayesian cut-off and, secondly, we compare the projection estimators with the empirical Bayes

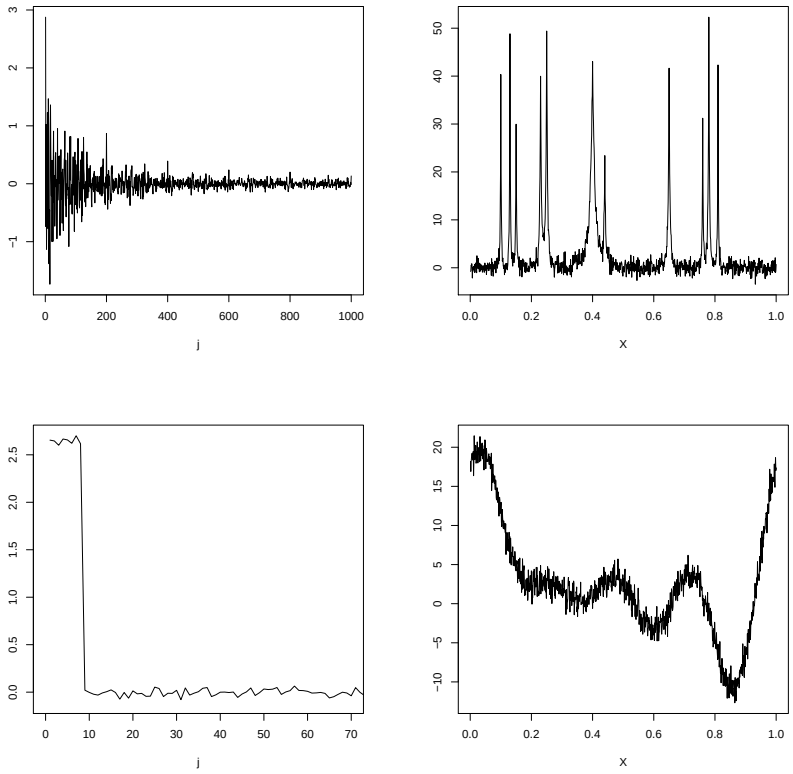


Figure 4.3: The noisy version of Bumps Fourier coefficients (top left) and Step Fourier coefficients (bottom left) within the Gaussian sequence model, signal-to-noise ratio $SD(\boldsymbol{\theta})/\sigma = 7.4$, $n = 1000$, $\sigma^2 = n^{-1}$. In the right column the function recoveries are displayed.

estimator.

4.3.1 Projection estimators

We move on to illustrate the performance of the projection estimators (4.12) with the Bayesian cut-offs N_{MAP}, N_{MED} and $\widetilde{N}$, for the purpose of which we use the statistical environment R 2.6.0. The estimators are applied to the above described test sequences, having 1000 points in the domain, i.e j ranges from 1 to 1000. This means that the vectors $\boldsymbol{y}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$ have dimension 1000.

In our simulations we use 600 realizations of $\boldsymbol{\xi}$. Each particular realization of the noise $\boldsymbol{\xi}$ yields a realization of the Fourier coefficients $\boldsymbol{y}$, which leads to a posterior pdf $p(N|\boldsymbol{y})$ defined by (4.8). For each pdf we compute N_{MAP}, N_{MED},

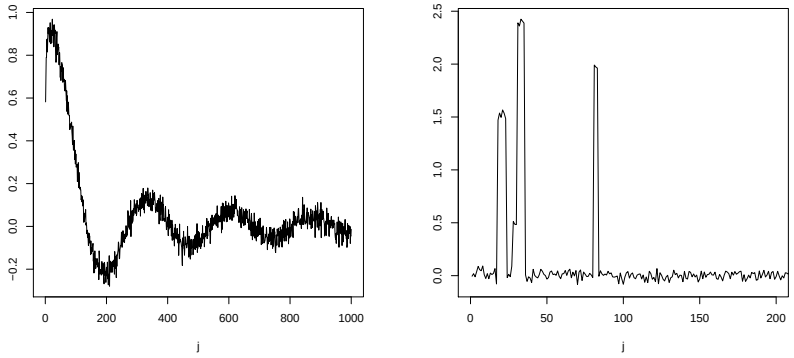


Figure 4.4: The noisy Fourier coefficients of Damped-Sine and Outliers within the Gaussian sequence model. Signal-to-noise ratio $SD(\boldsymbol{\theta})/\sigma = 7.4$, $n = 1000$, $\sigma^2 = n^{-1}$.

and $\widetilde{N}$.

The average values of the cut-off N_{MAP}, N_{MED}, $\widetilde{N}$ with n equal 1000 and the ratio of the corresponding risks (4.4) to the oracle risk (4.5) versus the corresponding Fourier coefficients $\boldsymbol{\theta}_{bumps}$, $\boldsymbol{\theta}_{step}$, $\boldsymbol{\theta}_{d-sine}$, $\boldsymbol{\theta}_{outliers}$ are shown in Table 4.2. The average is taken over all 600 realizations.

Table 4.2: The average value of the cut-off parameters and the ratio of the risks to the oracle risk with $n = 1000$, $\sigma^2 = n^{-1}$.

$\boldsymbol{\theta}$	N_o	$\overline{N}_{map}$	$\overline{N}_{med}$	$\overline{\widetilde{N}}$	$\frac{\mathcal{R}(N_{map},\boldsymbol{\theta})}{\mathcal{R}(N_o,\boldsymbol{\theta})}$	$\frac{\mathcal{R}(N_{med},\boldsymbol{\theta})}{\mathcal{R}(N_o,\boldsymbol{\theta})}$	$\frac{\mathcal{R}(\widetilde{N},\boldsymbol{\theta})}{\mathcal{R}(N_o,\boldsymbol{\theta})}$
$\boldsymbol{\theta}_{bumps}$	1000	935	926	777	1.034	1.046	1.1
$\boldsymbol{\theta}_{step}$	8	8.48	9.5	8.1	1.2	1.27	1
$\boldsymbol{\theta}_{d-sine}$	913	810.46	810.63	761.1	1.035	1.033	1.045
$\boldsymbol{\theta}_{outliers}$	83	83.48	84.54	83.1	1.01	1.02	1

Table 4.2 reveals that for Step and Outliers Fourier coefficients the adaptive projection estimator with the cut-off parameter $\widetilde{N}$ yields the same risk as the oracle risk. The slightly worse results in those cases are obtained by N_{MAP} and N_{MED}.

In the Bumps case we obtain the oracle cut-off N_o equal 1000, which means that the oracle does not do a good job in this case, it says that we have to keep all the observations $\boldsymbol{y}$ to get the oracle projection risk. However, the considered cut-offs N_{MAP}, N_{MED}, and $\widetilde{N}$, whose average values are smaller than 1000, give the adaptive projection estimators with risks close to the projection oracle risk, their ratios are around one.

So, as we see from Table 4.2, all the considered cut-offs N_{MAP}, N_{MED}, and

$\widetilde{N}$ illustrate "good" results, in the sense that the ratios of the risks are around one, while the cut-off parameters N_{MAP}, N_{MED}, and $\widetilde{N}$ might be different from the value of the oracle cut-off N_o.

We present more information about the whole distributions of N_{MAP}, N_{MED}, $\widetilde{N}$ in Table 4.4, which contains histograms describing frequency of appearing of a cut-off with a certain value. In particular, we can observe that there are certain values of the cut-off parameters, which exactly coincide with the oracle N_o (the small point on the bottom marks the oracle cut-off parameter N_o, indicated in Table 4.2).

Now we see, for example, for Step and Outliers Fourier coefficients, that the maximum of the frequency of the cut-off parameters N_{MAP} and $\widetilde{N}$ concentrates around the oracle N_o. However, the situation is different with N_{MED} parameter, where even the largest peak does not concentrate around the oracle cutoff. The cutoff parameters N_{MAP} and N_{MED} for Bumps Fourier coefficients are most frequently appearing around N_o, while the most frequently appearing value of $\widetilde{N}$ is not equal to N_o. The histograms of N_{MAP} and N_{MED} cutoff parameters for Damped-Sine Fourier coefficients are similar and have values concentrate around N_o, while the histogram for $\widetilde{N}$ is concentrated rather far from N_o.

4.3.2 Risk evaluation

At this point we compare the different estimators by measuring their quality by the risk

$$\mathcal{R}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 = \mathrm{Ave} \sum_{j=1}^{n} (\widehat{\theta}_j - \theta_j)^2, \tag{4.14}$$

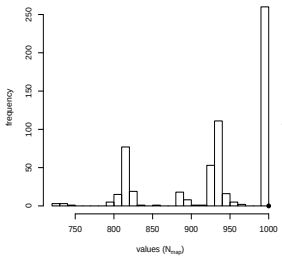
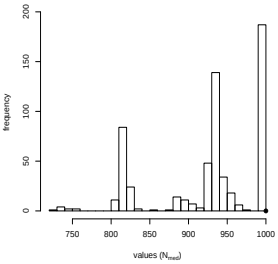
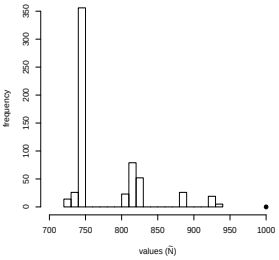
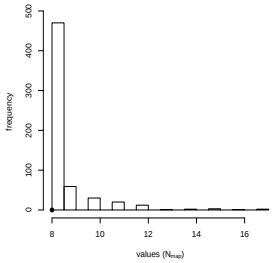
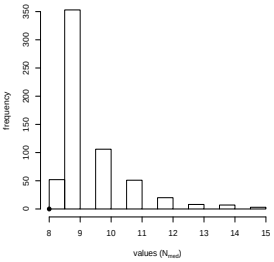
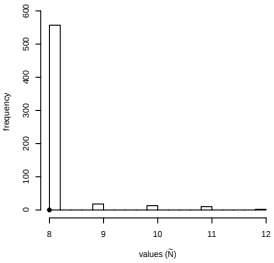
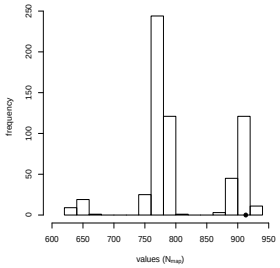
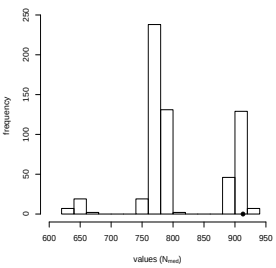
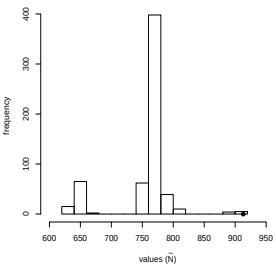
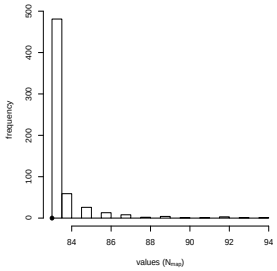
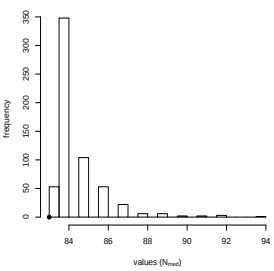
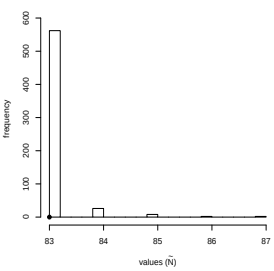
where the average is taken over 600 realizations. Thus, we compare efficiency of the adaptive projection estimators and the empirical Bayes estimator by (4.14), and summarize the results in Table 4.3.

The realizations of risk (4.14) for projection estimators with the cut-off N_{MAP}, N_{MED}, and $\widetilde{N}$ are placed in the first three columns. In the last column we put the realizations of risk (4.14) for the empirical Bayes estimator. The realizations of the estimators showed that the adaptive projection estimators can have a better risk than the empirical Bayes estimator, typically for small values of N_o.

Table 4.3: An average operator of norm loss for different estimators

$\boldsymbol{\theta}$	Projec.(N_{map})	Projec.(N_{med})	Projec.($\widetilde{N}$)	Emp. Bayes
$\boldsymbol{\theta}_{bumps}$	1.034	1.046	1.1	1.013
$\boldsymbol{\theta}_{step}$	0.009	0.010	0.008	0.010
$\boldsymbol{\theta}_{d-sine}$	0.998	0.995	1.006	0.969
$\boldsymbol{\theta}_{outliers}$	0.084	0.0845	0.083	0.084

Table 4.4: Histograms of frequency of appearing of the cut-off parameters. The small point on the bottom marks the oracle cut-off parameter N_o, indicated in Table 4.2.

$\boldsymbol{\theta}_0$	N_{MAP}	N_{MED}	$\widetilde{N}$
$\boldsymbol{\theta}_b$			
$\boldsymbol{\theta}_{st}$			
$\boldsymbol{\theta}_{d-s}$			
$\boldsymbol{\theta}_{out}$			

4.4 Posterior studies

The study of posterior distribution provides an insight into the validity of the imposed priors. In this section we illustrate the appropriate adjustment of the posterior pdf $p(\boldsymbol{\theta}|\boldsymbol{y})$, which was performed in the process of construction of the empirical Bayes estimator. Notice that the empirical Bayes estimator is based on posterior distribution $P'(\boldsymbol{\theta}|\boldsymbol{y})$ defined by pdf (4.13), while for the adaptive projection estimation (4.12) the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{y})$ is defined by pdf (4.11).

From now on we denote by $\boldsymbol{\theta}_0$ the vector of Fourier coefficients of some function in $L_2[0,1]$. Recall, that within the model (4.2), we assume that $\boldsymbol{y} \sim P_{\boldsymbol{\theta}_0} = P^{(\sigma)}_{\boldsymbol{\theta}_0}$ for some 'true' $\boldsymbol{\theta}_0 = (\theta_{0j})_{j=1}^n$ and (at least for a good Bayesian procedure) we want the corresponding posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{y})$ to concentrate around $\boldsymbol{\theta}_0$ as $n \to \infty$. To characterize the quality of the Bayesian procedure, we look at the rate at which $P(\boldsymbol{\theta}|\boldsymbol{y})$ concentrates around $\boldsymbol{\theta}_0$. Recall that a positive sequence r_n is called the posterior rate if for any $M_n \to \infty$,

$$P\Big\{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \geq M_n r_\sigma |\boldsymbol{y}\Big\} \to 0 \quad \text{as} \quad n \to \infty \tag{4.15}$$

in P_{θ_0}-probability. See Ghosal, Ghosh and Van der Vaart (2000), Shen and Wasserman (2001) for details. The idea of the posterior rate was extended later on in Babenko and Belitser (2010) to cover the posterior oracle rate. It was shown that for r_σ being the sequence of the oracle risks $r_\sigma(\boldsymbol{\theta}_0)$ the equation (4.15) is also satisfied. The relation between the posterior oracle rate and posterior minimax rate is discussed in more details in Chapter 2.

Below we remind the reader the theoretical result from Chapter 2, which characterizes the quality of the imposed priors (4.6) and justifies the changes of the posterior pdf (4.11).

- In case the oracle $r_\sigma(\boldsymbol{\theta}_0)$ is defined by (4.5) and the prior on $(\boldsymbol{y}, \boldsymbol{\theta}, N)$ is defined by (4.6) with $\sigma^2 = n^{-1}$, then there exists constant $C_1 > 0$ such that for any $\boldsymbol{\theta}_0 \in \ell_2$ and any $M > 0$

$$E_{\boldsymbol{\theta}_0} P'\Big\{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \geq M r_\sigma(\boldsymbol{\theta}_0)\,|\boldsymbol{y}\Big\} \leq \frac{C_1}{M}, \tag{4.16}$$

  where the posterior $P'(\boldsymbol{\theta}|\boldsymbol{y})$ is defined by (4.13).

It immediately follows from the above that the right hand side of inequality (4.16) converges to zero as $M \to \infty$, providing the posterior oracle rate. In other words, it says that the posterior distributions $P'(\boldsymbol{\theta}|\boldsymbol{y})$ concentrates around $\boldsymbol{\theta}_0$ in $P_{\boldsymbol{\theta}_0}$-probability with the posterior projection oracle rate $r_\sigma(\boldsymbol{\theta}_0)$, which implies the validity of Bayesian procedure.

Our numerical studies, represented by Figure 4.5, indeed show that $E_{\boldsymbol{\theta}_0} P'\Big\{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \geq M r_\sigma(\boldsymbol{\theta}_0)|\boldsymbol{y}\Big\}$ goes to zero, as required by (4.16), while $E_{\boldsymbol{\theta}_0} P\Big\{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \geq M r_\sigma(\boldsymbol{\theta}_0)|\boldsymbol{y}\Big\}$ stays close to one for the whole range of considered M. Thus, the posterior $P(\boldsymbol{\theta}|\boldsymbol{y})$ has been adjusted by the empirical

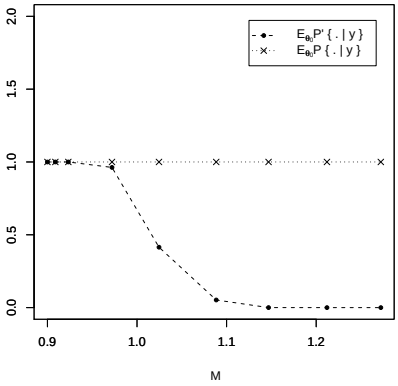


Figure 4.5: The plot of $E_{\boldsymbol{\theta}_0} P'\Big\{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \geq M r_\sigma(\boldsymbol{\theta}_0)\,|\boldsymbol{y}\Big\}$ and $E_{\boldsymbol{\theta}_0} P\Big\{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \geq M r_\sigma(\boldsymbol{\theta}_0)\,|\boldsymbol{y}\Big\}$ with $\boldsymbol{\theta}_0$ as a vector of Fourier coefficients of Bumps function, $n \in [1000, 1800]$.

Bayes approach in a way described in Section 4.2, so that the resulting posterior $P'(\boldsymbol{\theta}|\boldsymbol{y})$ concentrates around $\boldsymbol{\theta}_0$.

To demonstrate the rate of posterior concentration for different $\boldsymbol{\theta}_0$, we perform the simulations of (4.16) for two types of $\boldsymbol{\theta}_0$. In Figure 4.6 we plot the expectation of the conditional probability with $\sigma^2 = (1500)^{-1}$ for Bumps Fourier coefficients, whose oracle risk is 0.8269, and for Step Fourier coefficients with the value of the oracle risk 0.005. Indeed, for Step Fourier coefficients the posterior contracts much faster than for Bumps Fourier coefficients.

Recall a more refined theoretical result about the posterior $P'(\boldsymbol{\theta}|\boldsymbol{y})$:

- If the oracle $r_\sigma(\boldsymbol{\theta}_0)$ is defined by (4.5) and the prior on $(\boldsymbol{y}, \boldsymbol{\theta}, N)$ is defined by (4.6) with $\sigma^2 = n^{-1}$, then there exist constants $C_1, C_2 > 0$ and $0 < \delta < e^{-1}/2$ such that for any $\boldsymbol{\theta}_0 \in \ell_2$ and any $M > 0$

$$E_{\boldsymbol{\theta}_0} P'\Big\{\delta r_\sigma(\boldsymbol{\theta}_0) \leq \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \leq M r_\sigma(\boldsymbol{\theta}_0)\,|\boldsymbol{y}\Big\} \geq 1 - \frac{C_1}{M} - C_2\sqrt{\delta},$$

  where the posterior $P'(\boldsymbol{\theta}|\boldsymbol{y})$ is defined by (4.13).

This essentially says that almost all of the posterior mass is concentrated in the annulus $\{\boldsymbol{\theta} : \delta r_\sigma(\boldsymbol{\theta}_0) \leq \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \leq M r_\sigma(\boldsymbol{\theta}_0)\,|\boldsymbol{y}\}$, which we are going to study numerically in the next part of the section. Let us comment on Figure 4.6(a). The upper bound of expectation of conditional probability (4.16) tells that the probability for a point $\boldsymbol{\theta}$ to be outside the n-dimensional sphere centered at $\boldsymbol{\theta}_0$ with radius $\sqrt{M r_\sigma(\boldsymbol{\theta}_0)}$ is not larger than some constant divided by M. As it can be clearly seen from Figure 4.6(a), if $M r_\sigma(\boldsymbol{\theta}_0)$ is smaller than 0.75

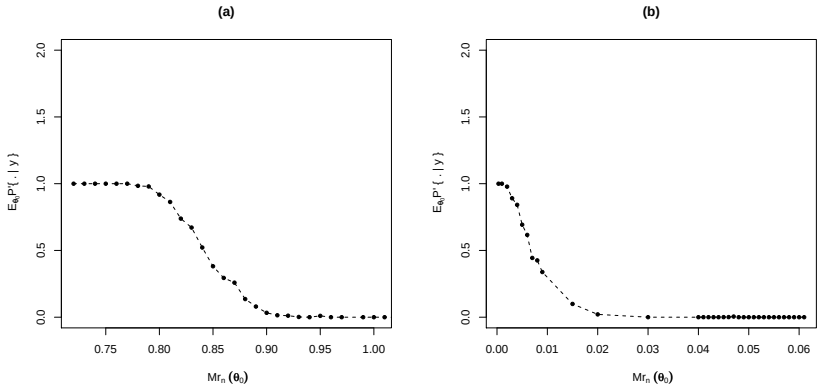


Figure 4.6: The plot of $E_{\boldsymbol{\theta}_0}P'\Big\{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \geq Mr_\sigma(\boldsymbol{\theta}_0)|\boldsymbol{y}\Big\}$, $\sigma^2=(1500)^{-1}$, with $\boldsymbol{\theta}_0$ being a vector of: a) Fourier coefficients of Bumps function, b) Step Fourier coefficients.

the expectation of conditional probability is very close to 1 within 10^{-15}, which tells that the points are inside the sphere with probability less than 10^{-15}. On the other hand, for $Mr_\sigma(\boldsymbol{\theta}_0)$ larger than 0.96 the expectation of the conditional probability is smaller than 10^{-15}, which means that the points are outside the sphere with radius $Mr_\sigma(\boldsymbol{\theta}_0)=\sqrt{0.96}$ with probability less than 10^{-15}. Thus, the random draws from the posterior pdf $p'(\boldsymbol{\theta}|\boldsymbol{y})$, in average, for a fixed n concentrates around $\boldsymbol{\theta}_0$ in some annulus. We place the values of $Mr_\sigma(\boldsymbol{\theta}_0)$, characterizing the size of the annulus for the other Fourier coefficients, in Table 4.5 for $\sigma^2=(1500)^{-1}$, in Table 4.6 for $\sigma^2=(1000)^{-1}$, and in Table 4.7 for $\sigma^2=(500)^{-1}$. The three dimensional plot of the annulus is presented in Figure 4.7(a). For the purpose of illustration, the density plot is also added in Figure 4.7(b).

Table 4.5: The values of $Mr_\sigma(\boldsymbol{\theta}_0)$ and $\delta r_\sigma(\boldsymbol{\theta}_0)$ characterizing the size of the annulus for $\sigma^2=(1500)^{-1}$, $\rho=1-10^{-15}$.

$\boldsymbol{\theta}_0$	$r_\sigma(\boldsymbol{\theta}_0)$	δ	M	The size of the annulus
$\boldsymbol{\theta}_b$	0.8269	0.907	1.161	$E_{\boldsymbol{\theta}_0}P'\Big\{0.75<\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2<0.96\,\|\boldsymbol{y}\Big\}\geq\rho$
$\boldsymbol{\theta}_{st}$	0.0053	0.183	10	$E_{\boldsymbol{\theta}_0}P'\Big\{0.00097<\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2<0.053\,\|\boldsymbol{y}\Big\}\geq\rho$
$\boldsymbol{\theta}_{d-s}$	0.9282	0.927	1.153	$E_{\boldsymbol{\theta}_0}P'\Big\{0.86<\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2<1.07\,\|\boldsymbol{y}\Big\}\geq\rho$
$\boldsymbol{\theta}_{out}$	0.0553	0.542	1.989	$E_{\boldsymbol{\theta}_0}P'\Big\{0.03<\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2<0.11\,\|\boldsymbol{y}\Big\}\geq\rho$

The Tables 4.5-4.7 show that the size of each annulus is shrinking as σ decreasing. Such a behavior is expected, since the oracle risk $r_\sigma(\boldsymbol{\theta}_0)$ tends to

Table 4.6: The values of $Mr_\sigma(\boldsymbol{\theta}_0)$ and $\delta r_\sigma(\boldsymbol{\theta}_0)$ characterizing the size of the annulus for $\sigma^2 = (1000)^{-1}$, $\rho = 1 - 10^{-15}$.

$\boldsymbol{\theta}_0$	$r_\sigma(\boldsymbol{\theta}_0)$	δ	M	The size of the annulus
$\boldsymbol{\theta}_b$	1	0.89	1.25	$E_{\boldsymbol{\theta}_0} P'\left\{0.89 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 1.25 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$
$\boldsymbol{\theta}_{st}$	0.008	0.125	10.625	$E_{\boldsymbol{\theta}_0} P'\left\{0.001 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 0.085 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$
$\boldsymbol{\theta}_{d-s}$	0.963	0.872	1.215	$E_{\boldsymbol{\theta}_0} P'\left\{0.84 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 1.17 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$
$\boldsymbol{\theta}_{out}$	0.083	0.602	1.807	$E_{\boldsymbol{\theta}_0} P'\left\{0.05 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 0.15 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$

Table 4.7: The values of $Mr_\sigma(\boldsymbol{\theta}_0)$ and $\delta r_\sigma(\boldsymbol{\theta}_0)$ characterizing the size of the annulus for $\sigma^2 = (500)^{-1}$, $\rho = 1 - 10^{-15}$.

$\boldsymbol{\theta}_0$	$r_\sigma(\boldsymbol{\theta}_0)$	δ	M	The size of the annulus
$\boldsymbol{\theta}_b$	1	0.83	1.24	$E_{\boldsymbol{\theta}_0} P'\left\{0.83 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 1.24 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$
$\boldsymbol{\theta}_{st}$	0.016	0.19	9.38	$E_{\boldsymbol{\theta}_0} P'\left\{0.003 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 0.15 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$
$\boldsymbol{\theta}_{d-s}$	1	0.84	1.31	$E_{\boldsymbol{\theta}_0} P'\left\{0.84 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 1.31 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$
$\boldsymbol{\theta}_{out}$	0.166	0.66	1.87	$E_{\boldsymbol{\theta}_0} P'\left\{0.11 < \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 < 0.31 \,\middle|\, \boldsymbol{y}\right\} \geq \rho$

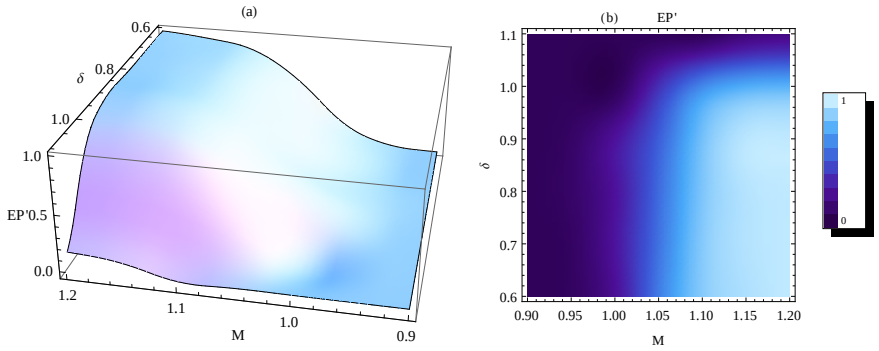


Figure 4.7: (a) The plot of $E_{\boldsymbol{\theta}_0} P'\left\{\delta r_n(\boldsymbol{\theta}_0) \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2 \leq M r_n(\boldsymbol{\theta}_0) | \boldsymbol{y}\right\}$, $n = 1000$. The vector $\boldsymbol{\theta}_0$ is a vector of Fourier coefficients of Bumps function. (b) The density plot of the same quantity.

zero in this case. The dependence of the size of the annuluses on the oracle risk can also be observed in Tables 4.5-4.7 and Figure 4.6.

Next we consider the posterior distributions $p(N = N_{MAP}|\boldsymbol{y})$, $p(N = N_{MED}|\boldsymbol{y})$, and $p(N = \widetilde{N}|\boldsymbol{y})$. Since the analytical formulas for the posteriors are

difficult to derive, we restrict ourselves to numerical simulations. Notice that the histograms in Table 4.4, which we already discussed in Section 4.3, approach the posterior distributions $p(N = N_{MAP}|\boldsymbol{y})$, $p(N = N_{MED}|\boldsymbol{y})$, $p(N = \bar{N}|\boldsymbol{y})$ as the number of realizations goes to infinity.

4.5 Approximate credible band

Let us now return to the equivalent discrete regression model (4.1), where the function can be approximated by $f(t) \approx \sum_{j=1}^{n} \theta_j \varphi_j(t)$. According to the discrete Fourier transformation the above correspondence is exact at the points $t_i = \frac{i}{n}$, $i = 1, \ldots, n$. Using this correspondence, we construct credible bands for the regression function f on the basis of the posterior distribution $P'(\boldsymbol{\theta}|\boldsymbol{y})$. It is convenient to denote by $f_0(t_i) = \sum_{j=1}^{n} \theta_{0j} \varphi_j(t_i)$ the function with fixed Fourier coefficients $\boldsymbol{\theta}_0$.

In this chapter we use $\boldsymbol{\Theta} = \{\Theta_j\}_{j=1}^{n}$ to denote random draws from the posterior distribution $P'(\boldsymbol{\theta}|\boldsymbol{y})$. Let us define

$$F(t) = F_y(t) = \sum_{j=1}^{n} \Theta_j \varphi_j(t), \quad t \in [0,1], \tag{4.17}$$

with $\Theta_j = \Theta_j(\boldsymbol{y})$, $j = 1, \ldots, n$, independent random variables from posterior $P'(\boldsymbol{\theta}|\boldsymbol{y})$. Define $P_{f_0(t)} = P^{(n)}_{f_0(t)}$, $t \in [0,1]$, with $f_0(t) = \sum_{j=1}^{n} \theta_{0j} \varphi_j(t)$, the probability measure of $F(t) = \sum_{j=1}^{n} \Theta_j \varphi_j(t)$, where $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_n) \sim P'(\boldsymbol{\theta}|\boldsymbol{y})$ with $\boldsymbol{y} \sim P_{\boldsymbol{\theta}_0}$.

From the equation (4.17) it follows that the family $\{F(t),\, t \in [0,1]\}$ is a Gaussian random process such that for each fixed t, $F(t)$ is a random variable with distribution $\mathcal{N}(\widetilde{\mu}_n(t), \widetilde{\sigma}_n^2(t))$ for some $\widetilde{\mu}_n(t) = \widetilde{\mu}_n(t, \boldsymbol{y})$ and $\widetilde{\sigma}_n^2(t) = \widetilde{\sigma}_n^2(t, \boldsymbol{y})$, since the posterior distribution (generating mechanism) depends on the observation $\boldsymbol{y}$. According the theoretical results from Chapter 3 as $n = \sigma^{-2}$ ($\sigma \to 0$) goes to infinity, the sample $\boldsymbol{\Theta}$ from the posterior distribution $P'(\boldsymbol{\theta}|\boldsymbol{y})$ concentrates around $\boldsymbol{\theta}_0$ in some annulus. This should lead to the concentration of the posterior distribution of $F(t)$ around $f_0(t)$. In particular $\widetilde{\mu}_n(t) \to f_0(t)$ and $\widetilde{\sigma}_n^2(t) \to 0$ in $P_{f_0(t)}$-probability.

In principle one can try to calculate the quantities $\widetilde{\mu}_n(t)$ and $\widetilde{\sigma}_n^2(t)$. Instead, for each $t \in [0,1]$ we can generate a sample $X_1, \ldots, X_M$ from the Gaussian distribution $\mathcal{N}(\widetilde{\mu}_n(t), \widetilde{\sigma}_n^2(t))$ and then approximate $\widetilde{\mu}_n(t)$ and $\widetilde{\sigma}_n^2(t)$ by using the empirical counterparts $\overline{X}_M = \frac{1}{M} \sum_{i=1}^{M} X_i$ for $\widetilde{\mu}_n(t)$ and $S_X^2 = \frac{1}{M-1} \sum_{i=1}^{M} (X_i - \overline{X}_M)^2$ for $\widetilde{\sigma}_n^2(t)$. We do it for a sufficiently fine grid $t_1, \ldots, t_L$ and sufficiently large M, which is limited only by computer power. In the simulations, we always take $L = n$, $t_i = \frac{i}{n}$, $i = 1, \ldots, n$. It is well known that the $(1-\alpha)$-highest posterior density (HPD) region for a Gaussian distribution $\mathcal{N}(\omega, \nu^2)$ is $\left[\omega - \widetilde{\xi}_{1-\alpha/2}\nu,\, \omega + \widetilde{\xi}_{1-\alpha/2}\nu\right]$, where $\widetilde{\xi}_{1-\alpha/2}$ is the quantile of normal distribution of level $1-\alpha/2$. Thus, for sufficiently large M we take the following approximate

$(1-\alpha)$-credible interval:

$$\Big[\overline{X}_M - \widetilde{\xi}_{1-\alpha/2} S_X,\, \overline{X}_M + \widetilde{\xi}_{1-\alpha/2} S_X\Big]. \tag{4.18}$$

We simulate the values of $F(t)$ defined by (4.17), $M = 400$ times and then apply formula (4.18) to calculate 95% credible intervals for each $F(t_i)$, $i = 1, \ldots, n$. The credible intervals at points t_i, $i = 1, \ldots, n$ form together a 95% pointwise credible band.

On Figure 4.8 we present the simulation results. We consider the test example $f_0(t)$ corresponding to $\boldsymbol{\theta}_{step}$ and $\boldsymbol{\theta}_{bumps}$. As we see from the plots, the credible band contracts towards the function f_0 (solid line) as σ decreases.

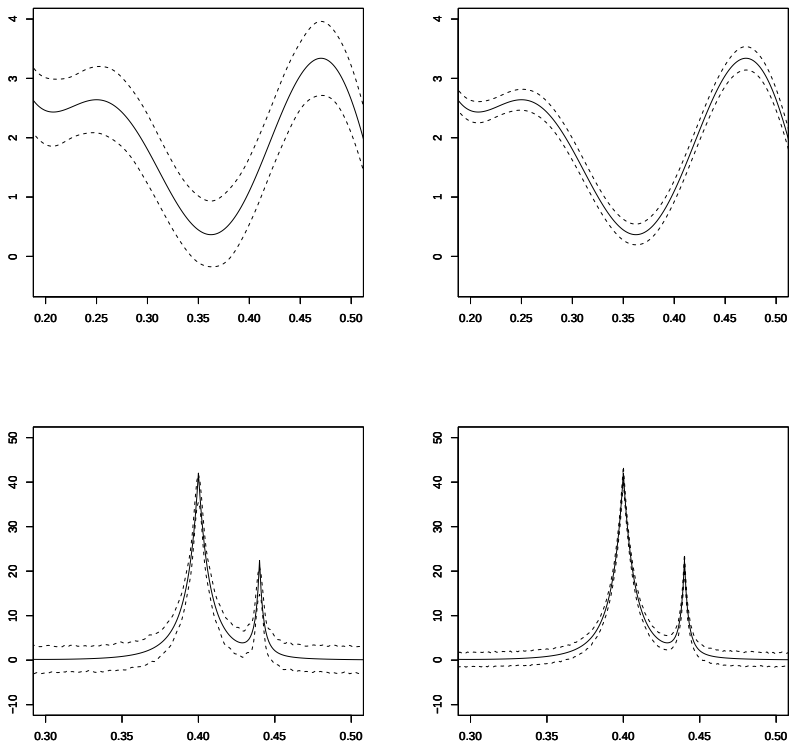


Figure 4.8: The plots of the credible bands for F with Θ_j's from posterior distributions $P'(\boldsymbol{\theta}|\boldsymbol{y})$ with the observation $\boldsymbol{y}$ obtained on the basis of Fourier coefficients $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{step}$ and $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{bumps}$. Left column: $\sigma^2 = (100)^{-1}$, right column: $\sigma^2 = (1000)^{-1}$. The solid lines are the corresponding functions f_0.

As we already saw in the previous section, the behavior of the posterior is better for Step signal than for Bumps signal, see Figure 4.6. The same

phenomenon occurs for the behavior of the posterior distributions of the corresponding Step and Bumps functions.

4.6 Conclusions

In our numerical simulations, performed for the four types of Fourier coefficients (Step, Damped-Sine, Outliers, and for Fourier coefficients of Bumps function) the empirical Bayes estimator shows the best performance in the case of large oracle cut-off N_o (Bumps and Damped-Sine), while the adaptive projection estimators are more efficient in the case of small values N_o (Step and Outliers). Moreover, the adaptive projection estimation and the empirical Bayes estimator allow to study their corresponding posterior distributions. Our theoretical studies formulated in Chapter 2 show that the empirical Bayesian approach, used for adjusting the posterior, implies the posterior distribution, which concentrates around the "true" θ_0 as $M \to \infty$ and $n \to \infty$.

The numerical studies of $E_{\boldsymbol{\theta}_0} P'\Big\{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \geq M_n r_\sigma(\boldsymbol{\theta}_0)|\boldsymbol{y}\Big\}$ and $E_{\boldsymbol{\theta}_0} P\Big\{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \geq M_n r_\sigma(\boldsymbol{\theta}_0)|\boldsymbol{y}\Big\}$ indeed confirm the theoretical result. Moreover, we studied the dependance of annulus size on σ. Our simulations show that the size of each annulus is shrinking to the "true" value of $\boldsymbol{\theta}_0$ as σ goes to zero. In addition we also illustrate the construction of an $(1-\alpha)$-credible band on the basis of the adjusted posterior $P'(\boldsymbol{\theta}|\boldsymbol{y})$, which width tends to zero as $\sigma \to 0$.

4.7 Appendix

Now we describe the exact calculation of the conditional expectation $E'(\theta_j|\boldsymbol{y})$. The expectation $E'(\theta_j|\boldsymbol{y})$ taken with respect to the posterior (4.13) can be written as

$$\begin{aligned} E'(\theta_j|\boldsymbol{y}) &= \frac{1}{S}\sum_k \lambda_k \prod_{i=1}^n \mathrm{p}(y_i; 0, \sigma^2+\tau_i(k)) \\ &\quad \times \int \theta_j \prod_{i=1}^n \mathrm{p}\Big(\theta_i; \frac{2\tau_i(k)y_i}{\sigma^2+\tau_i(k)}, \frac{\sigma^2\tau_i(k)}{\sigma^2+\tau_i(k)}\Big) \prod_{i=1}^n d\theta_i. \end{aligned} \quad (4.19)$$

Using the elementary identity

$$\int \theta_j \prod_{i=1}^n \mathrm{p}(\theta_i; \mu_i, \sigma_i^2) \prod_{i=1}^n d\theta_i = \mu_j$$

from (4.19) we obtain

$$E'(\theta_j|\boldsymbol{y}) = \frac{1}{S}\sum_k \lambda_k \prod_{i=1}^n \mathrm{p}(y_i; 0, \sigma^2+\tau_i(k)) \frac{2\tau_j(k)y_j}{\sigma^2+\tau_j(k)},$$

which is in view of (4.8) equivalent to

$$E'(\theta_j|\boldsymbol{y}) = \sum_k \frac{2\tau_j(k)y_j}{\sigma^2 + \tau_j(k)} p(N = k|\boldsymbol{y}).$$

The following substitution of $\tau_j(k)$ by (4.7) leads to

$$E'(\theta_j|\boldsymbol{y}) = \sum_k p(N = k|\boldsymbol{y}) y_j I\{j \leq k\},$$

which completes the calculations. □

Chapter 5

Posterior convergence rate and adaptive filtration of a linear functional

Suppose we observe $X = X^{(n)} = (X_1, X_2, \ldots)$:

$$X_i = \theta_i + n^{-1/2}\xi_i, \quad i = 1, 2, \ldots, \tag{5.1}$$

where ξ_i are independent $\mathcal{N}(0,1)$ noises, $\theta = (\theta_1, \theta_2 \ldots)$ is an unknown infinite dimensional parameter, which we will call the signal, parameter $n \to \infty$ reflects the increase of information. The goal is to make an inference on the linear functional of the signal θ in the asymptotic setup as $n \to \infty$. For a given fixed $\theta \in \ell_2$, denote the corresponding probability measure of X by $P_\theta = P_\theta^{(n)}$. If θ is a stochastic element, P_θ denotes a conditional probability measure of X given θ. Many quantities will depend on the information parameter n, but for the sake of notational simplicity we will often skip this dependence.

The interest in model (5.1) is mostly motivated by its well known equivalence to the white noise model (see the relevant section in the Introduction of this thesis) :

$$dX_\epsilon(t) = f(t)dt + \epsilon\, dW(t), \quad 0 \le t \le 1, \tag{5.2}$$

where $X_\epsilon(\cdot)$ is an observation process, $f(\cdot) \in L_2[0,1]$ is an unknown signal, $W(t)$ is a standard Wiener process, ϵ is the noise intensity. The problem is studied in the asymptotic setup as $\epsilon \to 0$.

Recall that, given an orthonormal basis $\{\phi_i,\, i \in \mathbb{N}\}$ in $L_2[0,1]$, the model (5.2) can be translated into an equivalent sequence model (5.1) with observations $X_i = \int_0^1 \phi_i(t)dX_\epsilon(t)$, the unknown Fourier coefficients $\theta_i = \int_0^1 \phi_i(t)f(t)dt$, independent Gaussian noises $\xi_i = \int_0^1 \phi_i(t)dW(t)$ and $n = \epsilon^{-2}$. Since there is an isometric (with respect to the scalar products in ℓ_2 and L_2) correspondence between θ and $f(\cdot)$, we call both θ and $f(\cdot)$ signals. As is already discussed in the Introduction, while interesting in communication theory in its own right,

model (5.2) also provides a good approximation to a variety of curve estimation problems.

Suppose we are interested in estimating the signal $f(t)$ from the model (5.2) at a fixed point $t \in [0,1]$, the so called pointwise estimation problem. Then, in terms of parameter θ, this problem corresponds to the estimation of the linear functional $\Phi = \Phi(\theta) = \sum_{i=1}^{\infty} b_i \theta_i$ in the model (5.1), where $b_i = \phi_i(t)$. Assume throughout without loss of generality that $|b_i| \le 1$ for all $i \in \mathbb{N}$. The above two problems are equivalent if the pointwise convergence $f(t) = \sum_{i=1}^{\infty} \theta_i \phi_i(t) = \Phi(\theta)$ holds, which can be provided by some conditions on $f(\cdot)$. For example, for the standard trigonometrical basis, continuity and the Dini condition at point t would do.

In the first part of this chapter, we continue the line of investigation started by Li and Zhao (2002). We assume that $\theta \in \Theta_\beta$, a Sobolev ellipsoid of a smoothness $\beta > 1/2$. Li and Zhao (2002) studied this problem and showed that a certain choice of normal prior leads to the Bayes estimator which attains the minimax rate over the Sobolev ball. However, as Li and Zhao (2002) demonstrate, this prior (and posterior) assigns zero mass to the underlying Sobolev space. In fact, they show that there does not exist a Gaussian prior on Sobolev space such that the corresponding Bayes estimator attains the optimal minimax rate. Related results about the Bayesian estimation of the entire signal θ in ℓ_2-norm can be found in Zhao (2000) and in Belitser and Ghosal (2003).

We propose a family of conjugate priors $\{\pi_{\beta,\delta},\ \delta < 2\beta - 1\}$ and establish that any prior from this family leads to the minimax (over Θ_β) rate Bayes estimator $\tilde{\Phi} = \tilde{\Phi}_\beta = \tilde{\Phi}_{\beta,\delta}$ of the functional $\Phi(\theta)$. The prior from Li and Zhao (2002) is the only member from this family (corresponding to the choice $\delta = 0$) which does not depend on the information parameter n. Next we show that the Bayes risk is of the same order as the minimax risk. In a way, this means that any prior from the above family adequately models the deterministic condition $\theta \in \Theta_\beta$. However, as in Li and Zhao (2002), the prior and the corresponding posterior mass "escape" the parameter space of interest. Despite this undesirable property, the posterior still does the "right thing" as regards the estimation of the linear functional $\Phi(\theta)$: using the results on the frequentist and Bayesian performance of our Bayes estimator, we finally establish that the posterior distribution of the functional $\Phi(\theta)$ concentrates around the true value of the functional with the minimax rate, uniformly over the Sobolev class.

Our constructions in the first part of this chapter are based on the knowledge of the smoothness parameter β. If the parameter β in unknown, then the problem of adaptive estimation of the functional $\Phi(\theta)$ arises. In the minimax setup, Lepski (1990, 1991, 1992) showed that a penalty log factor in the minimax risk is unavoidable for the problem of adaptive pointwise estimation and proposed an adaptive estimator attaining this adaptive (i.e. degraded by the log factor) minimax rate. As to the Bayesian approach, a natural candidate for an adaptive prior π would be a mixture of π_β over some distribution λ for the parameter β, cf. with Belitser and Ghosal (2003) for estimation of the signal θ itself. However, it turned out to be a very difficult problem to establish the adaptive minimax rate for the resulting posterior convergence and for the

corresponding Bayes estimator.

Instead, in the second part of this chapter, Sections 5.4 and 5.5, we consider a Bayesian version of the adaptive estimation of the functional $\Phi(\theta)$, which is in fact an adaptive filtering problem. We assume that the signal θ is distributed according a prior π_β with unknown "true smoothness" β. Clearly, the Bayes estimator $\tilde{\Phi}_\beta$ can not be used, because β is unknown. We regard $\tilde{\Phi}_\beta$ as a Bayesian oracle and its risk becomes our benchmark, which we call the oracle Bayes risk. The main goal is then to mimic the Bayesian oracle, that is to find such a procedure $\hat{\Phi}$ whose Bayes risk with respect to the prior π_β is within the constant factor of the oracle Bayes risk. We apply the empirical Bayes approach, which is due to Robbins (1956). We construct a marginal likelihood (an empirical Bayes) estimator $\hat{\beta}$ for the smoothness parameter β which leads to the plug-in procedure $\tilde{\Phi}_{\hat{\beta}}$ for the functional Φ and show that it mimics the Bayesian oracle. Actually we show that our adaptive estimator $\tilde{\Phi}_{\hat{\beta}}$ is asymptotically sharp, i.e. its risk coincides asymptotically with the oracle Bayes risk. The accompanying problem of estimating the smoothness parameter β by the empirical Bayes procedure $\hat{\beta}$ can be seen as a Bayesian counterpart of the inference problem on the smoothness parameter β. In the minimax frequentist setting, it is impossible to estimate the smoothness in any meaningful sense, while it is a well defined problem from the Bayesian point of view as the problem of estimating a parameter β of the prior distribution π_β.

5.1 Preliminaries

In this chapter, we study the problem of estimation of the linear functional $\Phi(\theta)$ in the model (5.1) by using a Bayesian approach. The prior knowledge about the signal θ can basically be modelled in two ways: either the signal θ is assumed to be deterministic and to belong to a given subset Θ of ℓ_2; or the signal θ is assumed to be random according to a certain prior distribution π. In this chapter we combine these two approaches to the problem of estimating the linear functional $\Phi(\theta)$ and study their interplay.

In the first approach, we assume that the signal θ belongs to a Sobolev class Θ_β of smoothness β (to be defined later). We put a simple conjugate prior π_β on the signal θ and investigate the performance of the resulting Bayesian procedure.

In the second approach, when the signal is intrinsically random ($\theta \sim \pi_\beta$), the estimation problem becomes a filtering problem, often studied within the framework of stationary random processes. The quality of estimation is typically measured by the Bayesian risk and the optimal estimator with respect to the mean square error risk is the posterior mean $\hat{\Phi} = E(\Phi(\theta)|X)$. The two approaches are actually closely related. Indeed, the minimax estimator over Θ_β is Bayesian with respect to the so called less informative prior π_β. On the other hand, a Bayesian estimator $\hat{\Phi}$ for nonrandom $\Phi(\theta)$ is optimal if the quality of estimation is measured by the Bayesian risk.

For a smoothness parameter $\beta > 0$, introduce a Sobolev space $\Theta_\beta = \{\theta :$

$\sum_{i=1}^{\infty} i^{2\beta}\theta_i^2 < \infty\Big\}$ and an ℓ_2-ellipsoid which we call the Sobolev ellipsoid (or class) of smoothness β and size Q:

$$\Theta_\beta(Q) = \Big\{\theta : \sum_{i=1}^{\infty} i^{2\beta}\theta_i^2 \le Q\Big\}.$$

Denote by $\hat\Phi = \hat\Phi(X)$ an estimator of linear functional $\Phi = \Phi(\theta) = \sum_{i=1}^{\infty} b_i\theta_i$, a measurable function of observation X from model (5.1).

Suppose that $\theta \in \Theta_\beta(Q)$, $\beta > 1/2$ and recall that $|b_i| \le 1$, $i \in \mathbb{N}$. The functional $\Phi(\theta)$ is then well defined. Indeed, by the Cauchy-Schwartz inequality

$$\big(\Phi(\theta)\big)^2 \le \Big(\sum_{i=1}^{\infty} |b_i\theta_i|\Big)^2 \le \sum_{i=1}^{\infty} b_i^2 i^{-2\beta} \sum_{i=1}^{\infty} \theta_i^2 i^{2\beta} < \infty.$$

For an estimator $\hat\Phi$, the estimation quality is measured by the risk function

$$R(\hat\Phi) = R(\hat\Phi, \Phi(\theta)) = E_\theta\big(\hat\Phi - \Phi(\theta)\big)^2.$$

It is known that, under appropriate assumptions on $b = (b_1, b_2, \ldots)$, the minimax risk

$$r_n(\Theta_\beta(Q)) = \inf_{\hat\Phi} \sup_{\theta\in\Theta_\beta(Q)} R(\hat\Phi, \Phi(\theta))$$

is of order $n^{-(2\beta-1)/2\beta}$ and it is sharp (see Donoho and Low (1992)), i.e. there exist $0 < C_1 \le C_2 < \infty$, such that

$$C_1 \le \liminf_{n\to\infty} n^{(2\beta-1)/2\beta} r_n(\Theta_\beta(Q)) \le \limsup_{n\to\infty} n^{(2\beta-1)/2\beta} r_n(\Theta_\beta(Q)) \le C_2.$$

The lower bound in the above relations holds only for the so called "nonparametric" $b = (b_1, b_2, \ldots)$; a simple example is $|b_i| \ge \kappa > 0$, $i \in \mathbb{N}$.

Remark 5.1. This lower bound holds also for the minimax risk over the ellipsoid $\widetilde\Theta_\beta(Q) = \big\{\theta : \sum_{k=1}^{\infty} k^{2\beta}(\theta_{2k}^2 + \theta_{2k+1}^2) \le Q\big\}$ and sequence $b = (b_1, b_2, \ldots)$ such that $b_{2k}^2 + b_{2k+1}^2 = 2$, $k \in \mathbb{N}$. The reason for looking at this case is that the estimation of functional $\Phi(\theta)$ over class $\widetilde\Theta_\beta(Q)$ for $\beta \in \mathbb{N}$ corresponds to the pointwise estimation of signal $f(t)$ in the model (5.2), where signal f belongs to a periodic Sobolev functional class of smoothness β (which is a subset of the usual Sobolev functional class, so the same lower bound holds for the Sobolev class); see Efromovich (1999) and Tsybakov (2004). In this case one takes the basis $\{\phi_k,\ k \in \mathbb{N}\}$ to be standard trigonometric: $\phi_1(t) = 1$, $\phi_{2k}(t) = \sqrt{2}\cos(2\pi kt)$, $\phi_{2k+1}(t) = \sqrt{2}\sin(2\pi kt)$, $k \in \mathbb{N}$, so that the above relations on $b_k = \phi_k(t)$, $k \in \mathbb{N}$, are indeed fulfilled.

Here we use a Bayesian approach. Namely, we put a prior π on signal θ and study the performance of the resulting Bayes procedure. The prior should be chosen in such a way that it adequately models the deterministic condition $\theta \in \Theta_\beta(Q)$. We have however to specify exactly what we mean by "adequately models". Our prime statistical inference problem under study is the estimation

of a linear functional Φ of signal θ and our main performance criterion is the convergence rate of the risk for this estimation problem. We propose a family of priors and look at the performance of the resulting Bayes procedures from the two different perspectives: frequentist (maximal risk) and Bayesian (Bayes risk). The priors from this family are shown to be adequate in the sense that for each prior from this family the corresponding Bayes estimator $\hat{\Phi}$ of the functional $\Phi(\theta)$ turns out to have the minimax convergence rate, simultaneously under both Bayesian and frequentist formulations. Finally, we show that these results lead to adequate behavior of the posterior: the posterior distribution of the functional $\Phi(\theta)$ concentrates around the true value of the functional with the minimax rate, uniformly over the Sobolev class.

To be precise, let $\theta_0 \in \Theta_\beta(Q)$ be the true value of unknown deterministic parameter θ. Now we look at the posterior distribution of $r_n^{-1/2}|\Phi(\theta) - \Phi(\theta_0)|$, given X, for some appropriate positive normalizing sequence $r_n \to 0$ as $n \to \infty$, from the point of view of the "true" distribution $X \sim P_{\theta_0}$. We say that the posterior of the functional Φ converges with rate r_n if for any $M_n \to \infty$ as $n \to \infty$,

$$\pi(r_n^{-1/2}|\Phi(\theta) - \Phi(\theta_0)| \geq M_n|X) \to 0$$

in P_{θ_0}-probability as $n \to \infty$.

If we put a prior π on signal θ, we can then regard the measure P_θ as the conditional distribution of X given θ, so that we get a joint distribution on (X, θ). Denote by E_θ the expectation with respect to the conditional distribution of X given θ, by E_π the expectation with respect to the prior distribution of θ and by E the expectation with respect to the joint distribution of (X, θ). Introduce the Bayesian mean square risk of an estimator $\hat{\Phi}$

$$R_\pi(\hat{\Phi}) = E_\pi E_\theta|\hat{\Phi} - \Phi(\theta)|^2 = E\big(\hat{\Phi} - \Phi(\theta)\big)^2.$$

5.2 Frequentist and Bayesian risks

For a given fixed smoothness parameter $\beta > 1/2$, introduce the family of priors $\{\pi_{\beta,\delta},\ \delta < 2\beta - 1\}$, with $\pi_{\beta,\delta} = \pi_{\beta,\delta,n}$ such that if $\theta \sim \pi_{\beta,\delta}$, then

$$\theta_i \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2(\beta, \delta)), \tag{5.3}$$

where

$$\tau_i^2(\beta, \delta) = \tau_i^2(\beta, \delta, n) = i^{-2\beta+\delta} n^{-\delta/(2\beta)}, \quad i \in \mathbb{N}.$$

Remark 5.2. In terms of the continuous white noise model (5.2), the prior on signal $f(\cdot)$ represents a Gaussian random process $\eta(t) = \sum_{i=1}^\infty \theta_i \phi_i(t)$, $t \in [0, 1]$, with zero mean and covariance $\mathrm{Cov}(\eta(t), \eta(s)) = \sum_{i=1}^\infty \tau_i^2(\beta, \delta)\phi_i(t)\phi_i(s)$. If $\{\phi_k,\ k \in \mathbb{Z}\}$ is the standard Fourier basis, i.e. $\phi_k(t) = e^{2\pi k t i}$, this random process is stationary.

Remark 5.3. Notice that there is only one prior from the family which does not depend on the information parameter n: $\pi_{\beta,0}$ (i.e. when $\delta = 0$), the one

considered in Li and Zhao (2002). According to the (pure) Bayesian tradition, one should use a prior which does not depend on n. In order to allow more modelling flexibility, we, however, do not impose this requirement. After all, using such priors does not affect formal constructions.

Remark 5.4. As discussed by Li and Zhao (2002), the prior $\pi_{\beta,0}$ assigns zero mass to the Sobolev space Θ_β since the series $\sum_{i=1}^\infty i^{2\beta}\tau_i^2(\beta,0)$ diverges. Of course, we can take prior $\pi_{\beta,\delta}$ with $\delta < -1$ to get a convergent series $\sum_{i=1}^\infty i^{2\beta}\tau_i^2(\beta,\delta) = n^{-\delta/(2\beta)}\sum_{i=1}^\infty i^\delta = C(\delta)n^{-\delta/(2\beta)}$. The above undesirable property of the prior essentially remains though: the prior mass 'escapes' Θ_β as $n \to \infty$.

Recall the following simple fact: if $Z|Y \sim \mathcal{N}(Y,\sigma^2)$ and $Y \sim \mathcal{N}(\mu,\tau^2)$, then

$$Y|Z \sim \mathcal{N}\Big(\frac{Z\tau^2+\mu\sigma^2}{\tau^2+\sigma^2}, \frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\Big).$$

Thus, the prior (5.3) leads to the following posterior distribution $\pi_{\beta,\delta}(\theta|X)$:

$$\theta_i|X \overset{ind}{\sim} \mathcal{N}\Big(\frac{\tau_i^2(\beta,\delta)X_i}{\tau_i^2(\beta,\delta)+n^{-1}}, \frac{\tau_i^2(\beta,\delta)n^{-1}}{\tau_i^2(\beta,\delta)+n^{-1}}\Big) \tag{5.4}$$

Therefore the classical Bayes estimator of the functional $\Phi(\theta)$ is

$$\tilde\Phi = \tilde\Phi_\beta = \tilde\Phi_\beta(\delta,X) = E(\Phi|X) = \sum_{i=1}^\infty b_i E(\theta_i|X) = \sum_{i=1}^\infty b_i\tilde\theta_i, \tag{5.5}$$

where

$$\tilde\theta_i = \tilde\theta_i(\beta,\delta,X) = E(\theta_i|X) = E(\theta_i|X_i) = \frac{\tau_i^2(\beta,\delta)X_i}{\tau_i^2(\beta,\delta)+n^{-1}}, \quad i \in \mathbb{N}.$$

For $p,q>0$, $r>-1$, $pq>r+1$, introduce the function

$$B(p,q,r) = \int_0^\infty \frac{u^r}{(1+u^p)^q}\,du = p^{-1}\mathrm{Beta}\Big(q-\frac{r+1}{p}, \frac{r+1}{p}\Big), \tag{5.6}$$

where $\mathrm{Beta}(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du$ is the beta function. From the properties of the beta function it is known that

$$B(p,1,r) = p^{-1}\mathrm{Beta}\Big(1-\frac{r+1}{p}, \frac{r+1}{p}\Big) = \frac{\pi}{p^2\sin(\pi(r+1)/p)}. \tag{5.7}$$

We will need the following technical lemma which is a slightly more general version of the corresponding result from Belitser and Enikeeva (2008).

Lemma 5.1. *Suppose $0<p<\infty$, $0<q<\infty$, $-1<r<pq-1$. Let $B(p,q,r)$ be defined by (5.6). For $r \ge 0$ define $D(p,q,r) = r^{r/p}(pq-r)^{q-(r/p)}(pq)^{-q} =$*

$\big(1-\frac{r}{pq}\big)^q\big(\frac{pq}{r}-1\big)^{-r/p}$, *with the convention* $0^0=1$. *Let us have positive* $\gamma_n\to\infty$ *as* $n\to\infty$. *Then*

$$\sum_{i=1}^{\infty}\frac{i^r}{(\gamma_n+i^p)^q}=B(p,q,r)\gamma_n^{\frac{r+1}{p}-q}+\phi_n,$$

where $\phi_n=\phi_n(p,q,r)$ *such that* $|\phi_n|\le D(p,q,r)\gamma_n^{-q+r/p}$ *if* $r\ge 0$ *and* $|\phi_n|\le\gamma_n^{-q}$ *if* $r<0$.

Proof. Define $g(u)=\frac{u^r}{(\gamma_n+u^p)^q}$, $u\in\mathbb{R}_+=\{u:\,u>0\}$. If $r\ge 0$, the function $g(u)$ is increasing on $u\in[0,u_{max}]$ and decreasing on $[u_{max},\infty)$ with $u_{max}=\big(r\gamma_n/(pq-r)\big)^{1/p}$. If $-1<r<0$, the function $g(u)$ is decreasing on $\mathbb{R}_+$ and in this case we let $u_{max}=1$. Now we have that

$$\int_0^{\infty}\frac{u^r\,du}{(\gamma_n+u^p)^q}-g(u_{max})\le\sum_{i=1}^{\infty}\frac{i^r}{(\gamma_n+i^p)^q}\le\int_0^{\infty}\frac{u^r\,du}{(\gamma_n+u^p)^q}+g(u_{max}),$$

with $g(u_{max})=D(p,q,r)\gamma_n^{(r/p)-q}$ if $r\ge 0$ and $|g(u_{max})|\le\gamma_n^{-q}$ if $r<0$, which establishes the lemma. □

Remark 5.5. If $r\le -1$, then for some $C_1=C_1(p,q,r)$ and $C_2=C_2(p,q,r)$ we can get a trivial asymptotic relation

$$C_1\gamma_n^{-q}\le\sum_{i=1}^{\infty}\frac{i^r}{(\gamma_n+i^p)^q}\le C_2\gamma_n^{-q},$$

which is not really informative.

The following theorem illustrates that the chosen prior adequately reflects the requirement $\theta\in\Theta_\beta(Q)$: both frequentist and Bayes risks of the Bayes estimator $\tilde\Phi$ have minimax convergence rate uniformly over Sobolev ellipsoid $\Theta_\beta(Q)$.

Theorem 5.1. *Let* $\pi=\pi_{\beta,\delta}$ *be a prior from the family (5.3) with* $\delta<\min\{2\beta-1,\beta+1/2\}$, *let* $\tilde\Phi=\tilde\Phi_\beta(\delta,X)$ *be the corresponding Bayes estimator (of the functional* Φ*) defined by (5.5), function* $B(p,q,r)$ *be defined by (5.6) and* $|b_i|\le 1$, $i\in\mathbb{N}$. *Then*

$$\limsup_{n\to\infty}n^{\frac{2\beta-1}{2\beta}}\sup_{\theta\in\Theta_\beta(Q)}R(\tilde\Phi,\Phi(\theta))\le B(2\beta-\delta,2,0)+QB(2\beta-\delta,2,2\beta-2\delta)$$

and

$$\limsup_{n\to\infty}n^{\frac{2\beta-1}{2\beta}}R_\pi(\tilde\Phi)\le\frac{\pi}{(2\beta-\delta)^2\sin(\pi/(2\beta-\delta))}.$$

Proof. Split the risk $R(\tilde\Phi,\Phi(\theta))$ into the sum of the bias and variance terms:

$$\begin{aligned}R(\tilde\Phi,\Phi(\theta))&=E_\theta\big(\tilde\Phi-\Phi(\theta)\big)^2=E_\theta\Big(\sum_{i=1}^{\infty}\frac{b_i\tau_i^2(\beta,\delta)X_i}{\tau_i^2(\beta,\delta)+n^{-1}}-b_i\theta_i\Big)^2\\&=\Big(\sum_{i=1}^{\infty}\frac{n^{-1}b_i\theta_i}{\tau_i^2(\beta,\delta)+n^{-1}}\Big)^2+\sum_{i=1}^{\infty}\frac{n^{-1}b_i^2\tau_i^4(\beta,\delta)}{(\tau_i^2(\beta,\delta)+n^{-1})^2}.\end{aligned}$$

First evaluate the variance term. Since $\delta < 2\beta - 1$, by Lemma 5.1 we have

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{n^{-1} b_i^2 \tau_i^4(\beta,\delta)}{(\tau_i^2(\beta,\delta)+n^{-1})^2} &\le \sum_{i=1}^{\infty} \frac{n^{1-\delta/\beta}}{\left(n^{1-\delta/(2\beta)} + i^{2\beta-\delta}\right)^2} \\ &= n^{-\frac{2\beta-1}{2\beta}} B(2\beta-\delta, 2, 0)(1+o(1)).\end{aligned}$$

By the Cauchy-Schwartz inequality and Lemma 5.1, we obtain the following upper bound for the bias term: uniformly over $\theta \in \Theta_\beta(Q)$,

$$\begin{aligned}\Big(\sum_{i=1}^{\infty} \frac{n^{-1} b_i \theta_i}{\tau_i^2(\beta,\delta)+n^{-1}}\Big)^2 &\le \sum_{i=1}^{\infty} \theta_i^2 i^{2\beta} \sum_{i=1}^{\infty} \frac{n^{-2} b_i^2}{i^{2\beta}(\tau_i^2(\beta,\delta)+n^{-1})^2} \\ &\le \sum_{i=1}^{\infty} \frac{Q i^{2\beta-2\delta}}{(n^{1-\delta/(2\beta)} + i^{2\beta-\delta})^2} \qquad (5.8) \\ &= n^{-\frac{2\beta-1}{2\beta}} Q B(2\beta-\delta, 2, 2\beta-2\delta)(1+o(1)),\end{aligned}$$

since we assumed also $\delta < \beta + 1/2$; see the remark on this condition below. The first relation is proved.

Using (5.7), Lemma 5.1 and the fact that $\delta < 2\beta - 1$, we derive the asymptotic expression for the Bayes risk

$$\begin{aligned}R_\pi(\tilde{\Phi}) &= \sum_{i=1}^{\infty} \frac{b_i^2 \tau_i^2(\beta,\delta) n^{-1}}{\tau_i^2(\beta,\delta)+n^{-1}} \le \sum_{i=1}^{\infty} \frac{n^{-\delta/(2\beta)}}{n^{1-\delta/(2\beta)} + i^{2\beta-\delta}} \\ &= n^{-\frac{2\beta-1}{2\beta}} B(2\beta-\delta, 1, 0)(1+o(1)) \\ &= \frac{n^{-\frac{2\beta-1}{2\beta}} \pi}{(2\beta-\delta)^2 \sin(\pi/(2\beta-\delta))}(1+o(1)),\end{aligned}$$

as $n \to \infty$. □

Remark 5.6. Clearly the condition $\delta < 2\beta - 1$ is needed to ensure the convergence of the variance term of the frequentist risk and the Bayes risk. Let us explain why we need the condition $\delta < \beta + 1/2$. If we had $\delta > \beta + 1/2$, then the upper bound for the bias term (see (5.8) and Remark 5.5) would be as follows:

$$C_2 n^{-2+\delta/\beta} \ge \sum_{i=1}^{\infty} \frac{Q i^{2\beta-2\delta}}{(n^{1-\delta/(2\beta)} + i^{2\beta-\delta})^2} \ge C_1 n^{-2+\delta/\beta} \gg n^{-\frac{2\beta-1}{2\beta}}$$

as $n \to \infty$, i.e. of a worse order as compared to the optimal rate $n^{-\frac{2\beta-1}{2\beta}}$.

Notice that if $\delta = \beta + 1/2$, the bias term is of the correct order $n^{-\frac{2\beta-1}{2\beta}}$, but with some other constant. So, the theorem holds also under milder condition $\delta \le \beta + 1/2$ with an adjusted constant for the case $\delta = \beta + 1/2$.

5.3 Posterior rate

Using the results from the previous section, it is not difficult to derive the result on the posterior convergence rate of the linear functional of the signal. The theorem below says essentially that the posterior distribution of $\Phi(\theta)$ behaves correctly in the sense that it concentrates around the true value $\Phi(\theta_0)$ with the correct rate $r_n = r_n(\beta) = n^{-(2\beta-1)/(2\beta)}$ from the perspective of P_{θ_0}-measure.

Theorem 5.2. *Let $\beta > 1/2$, $\delta < \min\{2\beta - 1, \beta + 1/2\}$ and $r_n = r_n(\beta) = n^{-\frac{2\beta-1}{2\beta}}$. Then for any $M_n \to \infty$,*

$$\pi(r_n^{-1/2}|\Phi(\theta) - \Phi(\theta_0)| \geq M_n|X) \to 0$$

in P_{θ_0}-probability as $n \to \infty$, uniformly over $\theta_0 \in \Theta_\beta(Q)$.

Proof. Using (5.4), it is not difficult to derive the posterior distribution of $\Phi(\theta)$ given X:

$$\Phi(\theta)|X \sim \mathcal{N}\big(\tilde{\Phi}, \mathrm{Var}(\Phi(\theta)|X)\big), \quad \mathrm{Var}(\Phi(\theta)|X) = \sum_{i=1}^{\infty} \frac{b_i^2 \tau_i^2(\beta,\delta) n^{-1}}{\tau_i^2(\beta,\delta) + n^{-1}},$$

where $\tilde{\Phi}$ is the Bayes estimator of Φ, defined by (5.5). By the conditional Chebyshev inequality, we have

$$\begin{aligned} \pi(r_n^{-1/2}|\Phi(\theta) - \Phi(\theta_0)| \geq M_n|X) &\leq \frac{r_n^{-1} E\big((\Phi(\theta) - \Phi(\theta_0))^2|X\big)}{M_n^2} \\ &\leq \frac{r_n^{-1}\mathrm{Var}(\Phi(\theta)|X)}{M_n^2} + \frac{r_n^{-1}(\tilde{\Phi} - \Phi(\theta_0))^2}{M_n^2}. \end{aligned}$$

It is enough to show that the expectation of the right hand side of the last relation with respect to distribution P_{θ_0} tends to zero as $n \to \infty$. Since $r_n^{-1} = n^{\frac{2\beta-1}{2\beta}}$ and

$$E_{\theta_0}\big[\mathrm{Var}(\Phi(\theta)|X) + (\tilde{\Phi} - \Phi(\theta_0))^2\big] = R_\pi(\tilde{\Phi}) + R(\tilde{\Phi}, \Phi(\theta)),$$

the result follows from Theorem 5.1. □

Remark 5.7. The above theorem guarantees good frequentist properties of credible intervals for the functional $\Phi(\theta)$, constructed on the basis of posterior distribution.

5.4 Adaptive filtering by empirical Bayes

In what follows, we consider a Bayesian version of the adaptive estimation of the functional $\Phi(\theta)$, which is in fact an adaptive filtering problem. Namely, we assume that $\theta \sim \pi_\beta$, i.e. θ is a random element distributed according a prior π_β with unknown "true smoothness" β.

We restrict ourselves to the simplest (and the most natural from the Bayesian point of view) prior $\pi_\beta = \pi_{\beta,0}$ from the family of priors $\{\pi_{\beta,\delta},\, \delta < 2\beta - 1\}$, i.e. we take $\delta = 0$ in the expression of the variances $\tau_i^2(\beta, \delta)$. Slightly abusing the notations, we denote

$$\tau_i^2(\beta) = \tau_i^2(\beta, 0) = i^{-2\beta}, \quad i \in \mathbb{N},$$

in what follows. It is possible in principle to consider a general case of $\delta < 2\beta - 1$ as well, but the mathematical treatment becomes more involved.

From now on we denote by $\beta_0 > 1/2$ the true value of the unknown parameter β. Thus, the Bayes estimator $\tilde{\Phi}_{\beta_0}$ given by (5.5) cannot be used, because β_0 is unknown. The Bayes estimator $\tilde{\Phi}_{\beta_0}$ plays now a role of the *Bayesian oracle* and we call its risk $R_\pi(\tilde{\Phi}_{\beta_0})$ the *oracle Bayes risk*.

Recall that now we have the following marginal distribution of X: the X_i's are independent and $X_i \sim N\big(0, \tau_i^2(\beta_0) + n^{-1}\big)$, $i \in \mathbb{N}$. Let $L_n(\beta) = L_n(\beta, X)$ be the marginal likelihood of the data $X = (X_i)_{i\in\mathbb{N}}$:

$$L_n(\beta) = \prod_{i=1}^{\infty} \frac{1}{\sqrt{2\pi\big(\tau_i^2(\beta) + n^{-1}\big)}} \exp\Big\{ - \frac{X_i^2}{2\big(\tau_i^2(\beta) + n^{-1}\big)} \Big\}.$$

Maximizing the function $L_n(\beta)$ is equivalent to minimizing

$$Z_n(\beta) = -2 \log L_n(\beta).$$

To avoid complications in defining the minimum of $Z_n(\beta)$ under the events $\{Z_n(\beta) = \pm\infty\}$, for some fixed reference value $\bar{\beta} > 0$ it is convenient to introduce $\bar{Z}_n(\beta) = Z_n(\beta, \bar{\beta}) = -2 \log \frac{L_n(\beta)}{L_n(\bar{\beta})}$, which is finite almost surely. For any set $S_n \subseteq (0, +\infty)$, define the marginal likelihood estimator of β restricted to the set S_n:

$$\hat{\beta} = \hat{\beta}(S_n) = \hat{\beta}(S_n, X, n) = \arg\min_{\beta \in S_n} \bar{Z}_n(\beta). \tag{5.9}$$

This means that $Z_n(\hat{\beta}(S_n)) \le Z_n(\beta')$ for all $\beta' \in S_n$, or equivalently $Z_n(\hat{\beta}(S_n), \beta') \le 0$ for all $\beta' \in S_n$.

Remark 5.8. Certainly, this is not the only possible way to estimate the smoothness parameter. The estimator is easy to implement in practice and it has an appealing feature that it is based on the fundamental principle in statistics: maximization of the likelihood. However, the analytic treatment of this approach is somewhat involved, even in our case for conjugate pair of normal model and normal prior. One can in principle try to find another estimator for β (for example, by using the method of moments), which should be good enough to plug in $\hat{\Phi}_\beta$ and is easier to treat.

Denote for brevity

$$\begin{aligned} a_i = a_i(\beta, \beta') &= \frac{1}{\tau_i^2(\beta) + n^{-1}} - \frac{1}{\tau_i^2(\beta') + n^{-1}} \\ &= \frac{\tau_i^2(\beta') - \tau_i^2(\beta)}{(\tau_i^2(\beta') + n^{-1})(\tau_i^2(\beta) + n^{-1})}\,, \end{aligned} \tag{5.10}$$

$$b_i = b_i(\beta, \beta') = \frac{\tau_i^2(\beta) + n^{-1}}{\tau_i^2(\beta') + n^{-1}}\,. \tag{5.11}$$

Then

$$Z_n(\beta, \beta') = \sum_{i=1}^{\infty} a_i(\beta, \beta') X_i^2 + \sum_{i=1}^{\infty} \log b_i(\beta, \beta'),$$

and for all $\beta' \in S_n$, according to (5.9), we have

$$\sum_{i=1}^{\infty} a_i(\hat\beta(S), \beta') X_i^2 \le \sum_{i=1}^{\infty} \log \big[b_i(\hat\beta(S), \beta')\big]^{-1}. \tag{5.12}$$

From now on we define the set S_n to be as follows:

$$S_n = \{1/2 + \kappa_n + k\epsilon_n,\ k = 0, 1, \dots, M_n - 1\}, \tag{5.13}$$

where the positive sequences $\kappa_n \to 0$, $\epsilon_n \to 0$ and $M_n \in \mathbb{N}$, $M_n \to \infty$ so that $M_n\epsilon_n \to \infty$, as $n \to \infty$.

Next, denote $\hat\Phi(\beta) = \tilde\Phi_\beta$ and introduce the empirical Bayesian plug-in estimator for the functional Φ:

$$\hat\Phi = \hat\Phi(\hat\beta) = \tilde\Phi_{\hat\beta} = \sum_{i=1}^{\infty} \frac{b_i \tau_i^2(\hat\beta) X_i}{\tau_i^2(\hat\beta) + n^{-1}}, \tag{5.14}$$

with $\tilde\Phi_\beta$ defined by (5.5), $\hat\beta = \hat\beta(S_n)$ defined by (5.9) and S_n defined by (5.13).

To avoid uninteresting cases (when the value of the functional Φ is close to zero), we assume that $R_\pi(\tilde\Phi_{\beta_0}) \ge cn^{-\alpha}$ for some $c, \alpha > 0$. This requirement is not restrictive since it will be fulfilled if there exists an $i \in \{1, 2, \dots, n\}$ such that $b_i^2 \ge 2c > 0$. Indeed, then

$$R_\pi(\tilde\Phi_{\beta_0}) = \sum_{i=1}^{\infty} \frac{b_i^2}{i^{2\beta_0} + n} \ge cn^{-2\beta_0}.$$

The next theorem claims that, under the above condition and very mild conditions on the choice of the set S_n, the adaptive empirical Bayes estimator $\hat\Phi$ mimics the Bayesian oracle, i.e. its Bayes risk is asymptotically not worse than the Bayesian oracle risk.

Theorem 5.3. *Suppose $R_\pi(\tilde\Phi_{\beta_0}) \ge cn^{-\alpha}$ for some $c, \alpha > 0$. Let $\hat\Phi(\hat\beta)$ be defined by (5.14) and the sequences κ_n, ϵ_n and M_n from the definition (5.13) of the set S_n be such that $\epsilon_n = o(1/\log n)$ and for some $C < (4\beta_0)^{-2}/2$*

$$\kappa_n^{-1} M_n \exp\big\{ - C\epsilon_n^2 (\log n)^2 n^{1/(2\beta_0)} \big\} = o(n^{-\alpha})$$

as $n \to \infty$. Then

$$R_\pi(\hat\Phi(\hat\beta)) \le R_\pi(\tilde\Phi_{\beta_0})(1 + o(1)) \le \frac{\pi n^{-(2\beta_0 - 1)/(2\beta_0)}}{(2\beta_0)^2 \sin(\pi/(2\beta_0))}(1 + o(1)) \quad \textit{as} \quad n \to \infty.$$

Remark 5.9. Many choices of sequences ϵ_n, κ_n and M_n satisfying the conditions of the theorem are possible. For example, $\epsilon_n = 1/(\log n)^2$, $\kappa_n = 1/\log n$ and $M_n = (\log n)^3$ will do. In fact, there is no need to take a sequence κ_n converging to zero faster than $1/\log n$ since already for $\beta_0 = \kappa_n = 1/\log n$ the risk will not converge to zero. Neither does it make sense to take the sequence M_n converging to infinity faster than a sequence for which $\log n = o(M_n\epsilon_n)$, since already for $\beta_0 = \log n$ we will get the unimprovable parametric convergence rate n^{-1}.

Proof. Write

$$\begin{aligned} R_\pi(\hat\Phi(\hat\beta)) &= E(\hat\Phi(\hat\beta)-\Phi)^2 = E\Big[(\hat\Phi(\hat\beta)-\Phi)^2 I\{|\hat\beta-\beta_0|\ge 2\epsilon_n\}\Big] \\ &\quad + E\Big[(\hat\Phi(\hat\beta)-\Phi)^2 I\{|\hat\beta-\beta_0|< 2\epsilon_n\})\Big] = T_1+T_2. \end{aligned}$$

First notice that

$$T_2 = E\Big[(\hat\Phi(\hat\beta)-\Phi)^2 I\{|\hat\beta-\beta_0|< 2\epsilon_n\}\Big] \le \max_{\beta:\,|\beta-\beta_0|<2\epsilon_n} E(\hat\Phi(\beta)-\Phi)^2. \quad (5.15)$$

Now, recall that $E(X_i-\theta_i)^2 = n^{-1}$ and $E\theta_i^2 = i^{-2\beta_0}$. Therefore,

$$\begin{aligned} E(\hat\Phi(\beta)-\Phi)^2 &= E_\pi E_\theta(\hat\Phi(\beta)-\Phi)^2 \\ &= \sum_{i=1}^\infty \frac{n^{-1}b_i^2\tau_i^4(\beta)}{(\tau_i^2(\beta)+n^{-1})^2} + E_\pi\Big(\sum_{i=1}^\infty \frac{n^{-1}b_i\theta_i}{\tau_i^2(\beta)+n^{-1}}\Big)^2 \\ &= \sum_{i=1}^\infty \frac{b_i^2 n}{(i^{2\beta}+n)^2} + \sum_{i=1}^\infty \frac{b_i^2 i^{4\beta-2\beta_0}}{(i^{2\beta}+n)^2} \\ &= \sum_{i=1}^\infty \frac{b_i^2}{i^{2\beta}+n} + \sum_{i=1}^\infty \frac{b_i^2(i^{4\beta-2\beta_0}-i^{2\beta})}{(i^{2\beta}+n)^2}. \end{aligned} \quad (5.16)$$

Next, let $K_n = \lfloor n^{\alpha/(2\beta_0-1)+\delta}\rfloor$ for some fixed $\delta>0$. Then, as $n\to\infty$,

$$\sum_{i=K_n+1}^\infty i^{-2\beta_0} = o(n^{-\alpha}) = o\big(R_\pi(\tilde\Phi_{\beta_0})\big)$$

due to the condition $R_\pi(\tilde\Phi_{\beta_0}) \ge cn^{-\alpha}$. Using this relation and the elementary inequality $a-1\ge 1-a^{-1}$ for any $a>0$, we obtain that, uniformly over

$|\beta - \beta_0| < 2\epsilon_n$ with $\epsilon_n = o(1/\log n)$,

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{b_i^2 |i^{4\beta-2\beta_0} - i^{2\beta}|}{(i^{2\beta}+n)^2} &= \sum_{i=1}^{\infty} \frac{b_i^2 i^{4\beta} |i^{-2\beta_0} - i^{-2\beta}|}{(i^{2\beta}+n)^2} \\
&\le \sum_{i=1}^{\infty} \frac{b_i^2 i^{4\beta_0+8\epsilon_n} |i^{-2\beta_0} - i^{-2\beta}|}{(i^{2\beta_0+4\epsilon_n}+n)^2} \\
&= \sum_{i=1}^{\infty} \frac{b_i^2 i^{2\beta_0+8\epsilon_n} |1 - i^{-2(\beta-\beta_0)}|}{(i^{2\beta_0+4\epsilon_n}+n)^2} \\
&\le \sum_{i=1}^{K_n} \frac{b_i^2 (i^{4\epsilon_n}-1)}{i^{2\beta_0}+n} + \sum_{i=K_n+1}^{\infty} \frac{1}{i^{2\beta_0}} \\
&= o(1) \sum_{i=1}^{K_n} \frac{b_i^2}{i^{2\beta_0}+n} + \sum_{i=K_n+1}^{\infty} \frac{1}{i^{2\beta_0}} \\
&= o\big(R_\pi(\tilde{\Phi}_{\beta_0})\big) \qquad (5.17)
\end{aligned}$$

as $n \to \infty$.

Since $R_\pi(\tilde{\Phi}_{\beta_0}) \ge cn^{-\alpha}$ and $\epsilon_n = o(1/\log n)$ as $n \to \infty$, it is not difficult to establish, similarly to (5.17), that

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{b_i^2}{i^{2\beta}+n} - R_\pi(\tilde{\Phi}_{\beta_0}) &= \sum_{i=1}^{\infty} \frac{b_i^2}{i^{2\beta}+n} - \sum_{i=1}^{\infty} \frac{b_i^2}{i^{2\beta_0}+n} \\
&= o\big(R_\pi(\tilde{\Phi}_{\beta_0})\big),
\end{aligned}$$

uniformly over $|\beta - \beta_0| < 2\epsilon_n$. By combining (5.15), (5.16), (5.17) and the last relation, we derive that, as $n \to \infty$,

$$T_2 \le \max_{\beta:\, |\beta-\beta_0|<2\epsilon_n} E\big(\hat{\Phi}(\beta) - \Phi(\theta)\big)^2 = R_\pi(\tilde{\Phi}_{\beta_0})(1 + o(1)).$$

To finish the proof of the theorem, it remains to show that, as $n \to \infty$,

$$T_1 = o(n^{-\alpha}) = o\big(R_\pi(\tilde{\Phi}_{\beta_0})\big).$$

Recall the elementary c_r-inequality $|a+b|^r \le c_r(|a|^r + |b|^r)$ for $r > 0$ and $c_r = 1$ if $r \le 1$ and $c_r = 2^{r-1}$ if $c_r > 1$. Using this and the Cauchy-Schwartz inequality, we obtain that

$$\begin{aligned}
&\big(\hat{\Phi}(\hat{\beta}) - \Phi\big)^4 \\
&= \Big[\sum_{i=1}^{\infty} b_i \Big(\frac{\tau_i^2(\hat{\beta})(X_i - \theta_i)}{\tau_i^2(\hat{\beta}) + n^{-1}} - \frac{n^{-1}\theta_i}{\tau_i^2(\hat{\beta}) + n^{-1}}\Big)\Big]^4 \\
&\le 8\Big[\sum_{i=1}^{\infty} \frac{b_i \tau_i^2(\hat{\beta}) \xi_i n^{-1/2}}{\tau_i^2(\hat{\beta}) + n^{-1}}\Big]^4 + 8\Big[\sum_{i=1}^{\infty} \frac{b_i n^{-1}\theta_i}{\tau_i^2(\hat{\beta}) + n^{-1}}\Big]^4 \\
&\le 8 \sum_{\beta_k \in S_n} \Big[\sum_{i=1}^{\infty} \frac{b_i \tau_i^2(\beta_k) \xi_i n^{-1/2}}{\tau_i^2(\beta_k) + n^{-1}}\Big]^4 + 8 \sum_{\beta_k \in S_n} \Big[\sum_{i=1}^{\infty} \frac{b_i n^{-1}\theta_i}{\tau_i^2(\beta_k) + n^{-1}}\Big]^4.
\end{aligned}$$

Recall the following fact. Let $Z_1, Z_2, \dots$ be independent, $Z_i \sim \mathcal{N}(0, \sigma_i^2)$, with $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$, then

$$E\Big(\sum_{i=1}^{\infty} Z_i\Big)^4 \le 3\Big(\sum_{i=1}^{\infty} \sigma_i^2\Big)^2.$$

Apply this relation and again the c_r-inequality (for $r = 1/2$) to get that

$$\begin{aligned}
&\Big[E\big(\hat\Phi(\hat\beta) - \Phi\big)^4\Big]^{1/2} \\
&\quad\le\ 2\sqrt{6M_n}\sum_{i=1}^{\infty}\frac{b_i^2 n}{(i^{1+2\kappa_n}+n)^2} + 2\sqrt{6M_n}\sum_{i=1}^{\infty}\frac{b_i^2 n^{-2} i^{-2\beta_0}}{(\tau_i^2(\beta_{M_n})+n^{-1})^2} \\
&\quad\le\ 2\sqrt{6M_n}\big(1 + (2\kappa_n)^{-1} + 1 + (2\beta_0 - 1)^{-1}\big) \\
&\quad\le\ c_1\kappa_n^{-1}\sqrt{M_n}
\end{aligned}$$

for sufficiently large n. Using the above estimate and the Cauchy-Schwartz inequality,

$$\begin{aligned}
T_1 &= E\Big[\big(\hat\Phi(\hat\beta) - \Phi\big)^2 I\{|\hat\beta - \beta_0| \ge 2\epsilon_n\}\Big] \\
&\le \Big[E\big(\hat\Phi(\hat\beta) - \Phi\big)^4\Big]^{1/2}\Big[P\{|\hat\beta - \beta_0| \ge 2\epsilon_n\}\Big]^{1/2} \\
&\le c_1\kappa_n^{-1}\sqrt{M_n}\Big[P\{|\hat\beta - \beta_0| \ge 2\epsilon_n\}\Big]^{1/2}.
\end{aligned}$$

Since $\epsilon_n = o(1/\log n)$, by Lemma 5.3 (see also Remark 5.12 after this lemma) we have that for any $c < (4\beta_0)^{-2}$ and all $n \ge N(c, \beta_0)$

$$P\{\hat\beta = \beta\} \le \exp\big\{-c\epsilon_n^2(\log n)^2 n^{1/(2\beta_0)}\big\}$$

uniformly over all β such that $|\beta - \beta_0| \ge 2\epsilon_n$. Therefore,

$$\begin{aligned}
P\big\{|\hat\beta - \beta_0| \ge 2\epsilon_n\big\} &= \sum_{\beta:\,|\beta-\beta_0|\ge 2\epsilon_n} P\big\{\hat\beta = \beta\big\} \\
&\le M_n \exp\big\{-c\epsilon_n^2(\log n)^2 n^{1/(2\beta_0)}\big\}
\end{aligned}$$

for all $n \ge N(c, \beta_0)$. Combining the last relations with the condition of the theorem, we obtain that

$$\begin{aligned}
T_1 &\le c_1\kappa_n^{-1} M_n \exp\big\{-c\epsilon_n^2(\log n)^2 n^{1/(2\beta_0)}/2\big\} \\
&= o(n^{-\alpha}) = o\big(R_\pi(\tilde\Phi_{\beta_0})\big)
\end{aligned}$$

as $n \to \infty$, which completes the proof. □

Remark 5.10. In case the basis $\{\phi_i,\, i \in \mathbb{N}\}$ is trigonometric, the signal f in the equivalent Gaussian white noise model (5.2) is a stationary Gaussian process and its Bayesian estimator is the Wiener filter in the classical problem of filtering of a stationary Gaussian process with Gaussian white noise. In this light, the problem we consider can be interpreted as adaptive pointwise filtration problem of a Gaussian process of an unknown smoothness β.

Remark 5.11. Note that the obtained Bayesian results can also be looked at from the frequentist point of view in the sense that we estimated the deterministic functional $\Phi(\theta)$ in the best way if we measure the performance of our estimator by the Bayesian risk with the prior π_β where β in unknown.

5.5 Estimating the smoothness parameter

In our plug-in estimator $\hat{\Phi}_{\hat{\beta}}$ for the functional $\Phi(\theta)$, we use a smoothness estimator $\hat{\beta}$. Recall that from the Bayesian perspective,

$$X_i \overset{ind}{\sim} \mathcal{N}(0, n^{-1} + \tau_i^2(\beta_0)), \quad i \in \mathbb{N}, \tag{5.18}$$

where β_0 denotes the true value of the unknown smoothness parameter. The problem of estimating the smoothness parameter β_0 is an auxiliary problem, but it is of interest on its own right. This a peculiar problem of parametric estimation with infinitely many non-identically distributed normal observations and a peculiar asymptotics: in our case $n \to \infty$ some information parameter involved in the variances of the observations, and not the traditional size of the observation sample. In fact, we use a version of the maximum likelihood method. Similar approach was previously considered by Belitser and Enikeeva (2008) for another estimation problem, the signal estimation in the ℓ_2-norm.

The next lemma is essentially from Belitser and Enikeeva (2008), we reproduce it here for the completeness sake.

Lemma 5.2. *For any $\beta, \beta' \in S_n$ and $0 < \lambda \le 1/2$, where either $\beta' < \beta$ or $\beta < \beta' < \beta_0$,*

$$\begin{aligned} P\{\hat{\beta} = \beta\} \quad &\le \quad \prod_{i=1}^{\infty} \Big(\frac{\tau_i^2(\beta') + n^{-1}}{\tau_i^2(\beta) + n^{-1}} \Big)^{\lambda} \\ &\quad \times \Big(1 + 2\lambda \frac{(\tau_i^2(\beta') - \tau_i^2(\beta))(\tau_i(\beta_0) + n^{-1})}{(\tau_i^2(\beta') + n^{-1})(\tau_i^2(\beta) + n^{-1})} \Big)^{-1/2}. \end{aligned} \tag{5.19}$$

Proof. In this proof we denote $a_i = a_i(\beta, \beta')$, $b_i = b_i(\beta, \beta')$. The proof follows from the Markov inequality. Indeed, since $\beta' \in S_n$, from the definition of $\hat{\beta}$ we have

$$\begin{aligned} P\{\hat{\beta} = \beta\} \quad &= \quad P\big\{ Z_n(\beta, \beta'') \le 0 \quad \forall\, \beta'' \in S_n \big\} \\ &\le \quad P\big\{ Z_n(\beta, \beta') \le 0 \big\} \\ &= \quad P\Big\{ -\sum_{i=1}^{\infty} a_i X_i^2 \ge \sum_{i=1}^{\infty} \log b_i \Big\} \\ &\le \quad E \exp\Big\{ -\lambda \sum_{i=1}^{\infty} a_i X_i^2 \Big\} \exp\Big\{ \sum_{i=1}^{\infty} \log(b_i^{-\lambda}) \Big\}. \end{aligned}$$

Using the simple formula for a Gaussian random variable $\eta \sim \mathcal{N}(\mu, \sigma^2)$,

$$E \exp\{\kappa \eta^2\} = (1 - 2\kappa\sigma^2)^{-1/2} \exp\Big\{ \frac{\kappa \mu^2}{1 - 2\kappa\sigma^2} \Big\} \qquad \text{for} \quad \kappa < \frac{1}{2\sigma^2}$$

and the fact that $X_i \sim \mathcal{N}(0, n^{-1} + \tau_i^2(\beta_0))$, we obtain that

$$E\exp\Big\{-\lambda\sum_{i=1}^{\infty} a_i X_i^2\Big\} = \prod_{i=1}^{\infty}\big(1+2\lambda a_i(\beta,\beta')(\tau_i^2(\beta_0)+n^{-1})\big)^{-1/2}$$
$$= \prod_{i=1}^{\infty}\Big(1+2\lambda\frac{(\tau_i^2(\beta')-\tau_i^2(\beta))(\tau_i(\beta_0)+n^{-1})}{(\tau_i^2(\beta')+n^{-1})(\tau_i^2(\beta)+n^{-1})}\Big)^{-1/2}$$

as $-\lambda a_i < 1/(2(n^{-1}+\tau_i(\beta_0)))$. □

Lemma 5.3. *Let $\beta \in S_n$, $|\beta-\beta_0| \geq 2\epsilon_n$ and for any $C < 1/\beta_0$*

$$\gamma_n = \min\Big\{C\max\{\epsilon_n, |\beta-\beta_0|-2\epsilon_n\}\log n, 0.77\Big\}, \quad \gamma_n' = \min\Big\{C\epsilon_n\log n, 0.77\Big\}.$$

Then there exists $N = N(C,\beta_0)$ such that for all $n \geq N$ the following inequality holds:

$$P\{\hat\beta = \beta\} \leq \exp\Big\{-\frac{\gamma_n\gamma_n' n^{1/(2\beta_0+2\epsilon_n)}}{16}\Big\}.$$

Remark 5.12. Note that if $\epsilon_n = o(1/\log n)$ as $n\to\infty$, then for any $\beta \in S_n$ such that $|\beta-\beta_0| \geq 2\epsilon_n$ and any $c < (4\beta_0)^{-2}$ there exists $N = N(c,\beta_0)$ such that for all $n \geq N$

$$P\{\hat\beta = \beta\} \leq \exp\big\{-c\epsilon_n^2(\log n)^2 n^{1/(2\beta_0)}\big\}.$$

Proof. For $\lambda = 1/2$ the relation (5.19) becomes: $P\{\hat\beta=\beta\} \leq \prod_{i=1}^{\infty}\big(1+b_i\big)^{1/2}$ with

$$\begin{aligned} b_i &= b_{i,n}(\beta,\beta_0,\beta') \\ &= \frac{(\tau_i^2(\beta_0)-\tau_i^2(\beta'))(\tau_i^2(\beta)-\tau_i^2(\beta'))}{\tau_i^2(\beta)\tau_i^2(\beta')+\tau_i^2(\beta')\tau_i^2(\beta_0)-\tau_i^2(\beta)\tau_i^2(\beta_0)+2n^{-1}\tau_i^2(\beta')+n^{-2}} \\ &= -1+\frac{2n^{-1}\tau_i^2(\beta')+n^{-2}+\tau_i^4(\beta')}{\tau_i^2(\beta)\tau_i^2(\beta')+\tau_i^2(\beta')\tau_i^2(\beta_0)-\tau_i^2(\beta)\tau_i^2(\beta_0)+2n^{-1}\tau_i^2(\beta')+n^{-2}}. \end{aligned}$$

If $\beta' \in S_n$ is between β_0 and β, then, as one can easily see from the above formulas, $i\in\mathbb{N}$, $-1 \leq b_i \leq 0$ for all $i\in\mathbb{N}$. Therefore, for any $\beta'\in S_n$ between β_0 and β (i.e. $\beta_0\leq\beta'\leq\beta$ or $\beta\leq\beta'\leq\beta_0$), any $N_1\in\mathbb{N}$, $N_2\in\mathbb{N}\cup\{+\infty\}$, such that $N_1\leq N_2$, we have

$$P\{\hat\beta=\beta\} \leq \prod_{i=1}^{\infty}\big(1+b_i\big)^{1/2} \leq \exp\Big\{\frac12\sum_{i=1}^{\infty} b_i\Big\} \leq \exp\Big\{\frac12\sum_{i=N_1}^{N_2} b_i\Big\}. \tag{5.20}$$

Write

$$b_i = \frac{n^2(i^{2\beta'}-i^{2\beta})(i^{2\beta'}-i^{2\beta_0})}{n^2 i^{2\beta_0+2\beta'}+n^2 i^{2\beta'+2\beta}-n^2 i^{4\beta'}+2ni^{2\beta_0+2\beta'+2\beta}+i^{2\beta_0+4\beta'+2\beta}} = \frac{a_i}{d_i}.$$

First consider the case $\beta - \beta_0 \geq 2\epsilon_n$. Denote $\delta_n = 1 - N_1^{2(\beta'-\beta)}$. Then for all $i \geq N_1$

$$i^{2\beta} - i^{2\beta'} = i^{2\beta}(1 - i^{2(\beta'-\beta)}) \geq \delta_n i^{2\beta}$$

because $i^{2(\beta'-\beta)} \leq N_1^{2(\beta'-\beta)} = 1 - \delta_n$ for all $i \geq N_1$.

Denote $\delta_n' = 1 - N_1^{2(\beta_0-\beta')}$, then in the same way we derive that for all $i \geq N_1$

$$i^{2\beta'} - i^{2\beta_0} = i^{2\beta'}(1 - i^{2(\beta_0-\beta')}) \geq \delta_n' i^{2\beta'}.$$

These relations imply that for all $i \geq N_1$

$$a_i = n^2(i^{2\beta'} - i^{2\beta})(i^{2\beta'} - i^{2\beta_0}) \leq -\delta_n\delta_n' n^2 i^{2\beta'+2\beta}.$$

Now, note that for $i \leq n^{1/(\beta_0+\beta')}$, we have that

$$ni^{2\beta_0+2\beta'+2\beta} \leq n^2 i^{2\beta'+2\beta} \quad \text{and} \quad i^{2\beta_0+4\beta'+2\beta} \leq n^2 i^{2\beta'+2\beta}.$$

Then for all $i \leq n^{1/(\beta_0+\beta')}$,

$$\begin{aligned} d_i &= n^2 i^{2\beta_0+2\beta'} + n^2 i^{2\beta'+2\beta} - n^2 i^{4\beta'} + 2ni^{2\beta_0+2\beta'+2\beta} + i^{2\beta_0+4\beta'+2\beta} \\ &\leq n^2 i^{2\beta'+2\beta} + 2ni^{2\beta_0+2\beta'+2\beta} + i^{2\beta_0+4\beta'+2\beta} \leq 4n^2 i^{2\beta'+2\beta}. \end{aligned}$$

Combining the last relations for a_i and d_i, we obtain that for all $N_1 \leq i \leq N_2 \leq n^{1/(\beta_0+\beta')}$,

$$\sum_{i=N_1}^{N_2} b_i = \sum_{i=N_1}^{N_2} \frac{a_i}{d_i} \leq -\sum_{N_1}^{N_2} \frac{\delta_n\delta_n' n^2 i^{2\beta'+2\beta}}{4n^2 i^{2\beta'+2\beta}} = -\sum_{N_1}^{N_2} \frac{\delta_n\delta_n'}{4} = -\frac{(N_2 - N_1 + 1)\delta_n\delta_n'}{4}.$$

Thus, using (5.20) and the last inequality, for any $N_1 \leq N_2 \leq n^{1/(\beta_0+\beta')}$ we have

$$P\{\hat{\beta} = \beta\} \leq \exp\Big\{\frac{1}{2}\sum_{i=N_1}^{N_2} b_i\Big\} \leq \exp\Big\{-\frac{(N_2 - N_1 + 1)\delta_n\delta_n'}{8}\Big\}. \tag{5.21}$$

If $\beta \in S_n$ and $\beta - \beta_0 \geq 2\epsilon_n$ then there exists $\beta' \in S_n$ between β_0 and β such that $\beta_0 + \epsilon_n \leq \beta' \leq \min\{\beta_0 + 2\epsilon_n, \beta - \epsilon_n\}$ for sufficiently large n. Indeed, $\beta_0 > \frac{1}{2} + \kappa_n = \min\{S_n\}$ for sufficiently large n as $\kappa_n \to 0$ and $\beta' = \min\{\beta \in S_n : \beta \geq \beta_0 + \epsilon_n\}$ satisfies the above inequalities.

Take $N_2 = \lfloor n^{1/(\beta_0+\beta')} \rfloor$ and $N_1 = \lfloor N_2/2 \rfloor + 1$. Recall that $\delta_n = 1 - N_1^{2(\beta'-\beta)}$. Since for any $C < C(\beta_0) = 1/\beta_0$ there exists $M_1 = M_1(C, \beta_0)$ such that for all $n \geq M_1$, $2(\log 2 - \frac{\log n}{2\beta_0+2\epsilon_n}) \leq -C \log n$, we have that, for a constant $C <$

$C(\beta_0) = 1/\beta_0$ and all $n \geq M_1(C, \beta_0)$,

$$\begin{aligned}
\delta_n &= 1 - N_1^{2(\beta'-\beta)} \\
&\geq 1 - \exp\Big\{2(\beta-\beta')\Big(\log 2 - \frac{\log n}{\beta_0+\beta'}\Big)\Big\} \\
&\geq 1 - \exp\Big\{2(\beta-\beta')\Big(\log 2 - \frac{\log n}{2\beta_0+2\epsilon_n}\Big)\Big\} \\
&\geq 1 - \exp\Big\{-C(\beta-\beta')\log n\Big\} \\
&\geq \min\Big\{C(\beta-\beta')\log n, 1-e^{-3/2}\Big\} \\
&\geq \min\Big\{C(\beta-\beta')\log n, 0.77\Big\} \geq \gamma_n.
\end{aligned}$$

We also used that $1-e^{-x} \geq \frac{x}{2}$ for all $0 \leq x \leq \frac{3}{2}$ and, by construction, $\beta-\beta' \geq \epsilon_n$ and $\beta-\beta' = \beta-\beta_0-(\beta'-\beta_0) \geq \beta-\beta_0-2\epsilon_n$, so that $\beta-\beta' \geq \max\{\epsilon_n, \beta-\beta_0-2\epsilon_n\}$. We derive a similar bound for $\delta_n' = 1-N_1^{2(\beta_0-\beta')}$: for a constant $C < C(\beta_0) = 1/\beta_0$ and for all $n \geq M_1(C,\beta_0)$,

$$\delta_n' \geq \min\Big\{C(\beta'-\beta_0)\log n, 0.77\Big\} \geq \min\Big\{C\epsilon_n \log n, 0.77\Big\} = \gamma_n'$$

since by construction $\beta'-\beta_0 \geq \epsilon_n$.

Finally, using (5.21) and the last three relations, we obtain that for a constant $C < C(\beta_0) = 1/\beta_0$ and for all $n \geq M_1(C,\beta_0)$,

$$\begin{aligned}
P\{\hat\beta = \beta\} &\leq \exp\Big\{-\frac{(N_2-N_1+1)\delta_n\delta_n'}{8}\Big\} \\
&\leq \exp\Big\{-\frac{\gamma_n\gamma_n' n^{1/(\beta_0+\beta')}}{16}\Big\} \\
&\leq \exp\Big\{-\frac{\gamma_n\gamma_n' n^{1/(2\beta_0+2\epsilon_n)}}{16}\Big\},
\end{aligned}$$

which completes the proof for the case $\beta-\beta_0 \geq 2\epsilon_n$.

Now consider the case $\beta-\beta_0 \leq -2\epsilon_n$. The same reasoning applies with a couple of small modifications. In fact, all estimates are somewhat easier in this case. We take

$$\beta' = \max\{\beta \in S_n : \beta \leq \beta_0 - \epsilon_n\}, \quad N_2 = \lfloor n^{1/(\beta+\beta')}\rfloor, \quad N_1 = \lfloor N_2/2\rfloor + 1.$$

Now the main term in a_i and d_i is $n^2 i^{2\beta'+2\beta_0}$. Precisely, with $\delta_n = 1-N_1^{2(\beta-\beta')}$ and $\delta_n' = 1-N_1^{2(\beta'-\beta_0)}$, we have that

$$a_i = n^2(i^{2\beta'} - i^{2\beta})(i^{2\beta'} - i^{2\beta_0}) \leq -\delta_n\delta_n' n^2 i^{2\beta_0+2\beta'},$$

for all $i \geq N_1$ and

$$d_i \leq 4n^2 i^{2\beta_0+2\beta'}$$

for all $i \le N_2$. We obtain slightly different estimates for δ_n and δ'_n: for a constant $C < C(\beta_0) = 1/\beta_0$ and for all $n \ge M_2(C, \beta_0)$,

$$\delta_n \ge \min\Big\{C(\beta' - \beta)\log n, 0.77\Big\} \ge \gamma_n,$$

$$\delta'_n \ge \min\Big\{C(\beta_0 - \beta')\log n, 0.77\Big\} \ge \gamma'_n.$$

Now, combining all the above relations with (5.20) in the same way as for the case $\beta - \beta_0 \ge 2\epsilon_n$, we conclude that, for a constant $C < C(\beta_0) = 1/\beta_0$ and for all $n \ge M_2(C, \beta_0)$ (in fact, as one can show, $M_2(C, \beta_0) \le M_1(C, \beta_0)$),

$$\begin{aligned} P\{\hat\beta = \beta\} &\le \exp\Big\{-\frac{(N_2 - N_1 + 1)\delta_n\delta'_n}{8}\Big\} \\ &\le \exp\Big\{-\frac{\gamma_n\gamma'_n n^{1/(\beta+\beta')}}{16}\Big\} \\ &\le \exp\Big\{-\frac{\gamma_n\gamma'_n n^{1/(2\beta_0)}}{16}\Big\}. \end{aligned}$$

□

Remark 5.13. As we already mentioned, although the problem of estimating the smoothness parameter β is not our prime goal, it is interesting on its own. From the frequentist perspective (when θ is nonrandom and $\theta \in \Theta_\beta$ with smoothness β) in the minimax setup, the problem of smoothness estimation is not well defined, since it is not clear what the smoothness of nonrandom θ is: $\theta \in \Theta_\beta(Q)$ may also belong to a smoother Sobolev ellipsoid $\Theta_{\beta'}(Q')$, i.e. with $\beta' > \beta$, but with a bigger size $Q' > Q$. On the other hand, the problem of estimating the smoothness parameter from the Bayesian perspective is certainly sensible: one simply needs to estimate the unknown parameter β_0 in the model (5.18).

Remark 5.14. Interestingly, as one can see from the above proof, the bound for the probability $P(\hat\beta = \beta)$ appears to be slightly better for the case $\beta \le \beta_0 - 2\epsilon$ than for the case $\beta \ge \beta_0 + 2\epsilon$. The claimed bound for β from the set $\{\beta \in S_n : |\beta - \beta_0| \ge 2\epsilon\}$ is in fact the worst of these two bounds. This asymmetric behavior is discussed in detail in Belitser and Levit (2003) for some other related problem where this phenomenon is even more prominent. The heuristics is that the method deals a little better with undersmoothing (smoothness parameters which are less than the true smoothness) than with oversmoothing (smoothness parameters which are bigger than the true smoothness).

Remark 5.15. One can formally calculate the Fisher information $I(\beta_0)$ about β_0 in the parametric model (5.18) which is

$$I(\beta_0) = 2\sum_{i=1}^{\infty} \frac{(n\log i)^2}{(n + i^{2\beta_0})^2}.$$

It is not difficult to evaluate that the Fisher information $I(\beta_0)$ is of the order

$O((\log n)^2 n^{1/(2\beta_0)})$. On the other hand, according to Remark 5.12, the choice for ϵ_n should be such that $\epsilon_n^2 \gg C(\log n)^2 n^{-1/(2\beta_0)}$ in order to guarantee the consistency of our estimator. Relating these facts, the rate $n^{-1/\beta_0}(\log n)^{-1}$ seems to be the best possible in estimating β_0 (in an appropriate sense, e.g. minimax). However, such a preciseness in estimating β_0 is not needed in the proof of the adaptive filtration result.

Bibliography

[1] Babenko, A. (2010). *Posterior studies for the Gaussian white noise model.* Submitted.

[2] Babenko, A. and Belitser, E. (2009). *On posterior pointwise convergence rate of a Gaussian signal under a conjugate prior.* Statist. Prob. Lett. **79**, 670–675.

[3] Babenko, A. and Belitser, E. (2010). *Oracle convergence rate of posterior under projection prior and Bayesian model selection.* Math. Methods Statist. **19**, 219–245.

[4] Babenko, A. and Belitser, E. (2010). *Lower bound for the oracle projection posterior convergence rate.* Tentatively accepted.

[5] Baraud, Y. (2004). *Confidence balls in Gaussian regression.* Ann. Statist. **32**, 528–551.

[6] Barron, A., Birgé, L. and Massart, P. (1999). *Risk bounds for model selection.* Probab. Theory Related Fields **113**, 301–413.

[7] Barron, A., Schervish, M.J. and Wasserman, L. (1999). *The consistency of posterior distributions in nonparametric problems.* Ann. Statist. **27**, 536–561.

[8] Bayes, T. (1763). *An Essay towards solving a Problem in the doctrine of chances.* Phil. Trans. Roy. Soc. **53**, 370–418.

[9] Belitser, E. (2000). *Minimax estimation in regression and random censorship models.* CWI Tract 127, CWI, Amsterdam.

[10] Belitser, E. and Enikeeva, F. (2008). *Adaptive filtering of a random signal in Gaussian white noise.* Problemy Peredachi Informatsii **44**, 39–51; translation in Probl. Inf. Transm. **44**, 321–332.

[11] Belitser, E. and Enikeeva, F. (2008). *Empirical Bayesian test for the smoothness.* Math. Methods Statist. **17**, 1–18.

[12] Belitser, E. and Ghosal, S. (2003). *Adaptive Bayesian inference on the mean of an infinite dimensional normal distribution.* Ann. Statist. **31**, 536–559.

[13] Belitser, E. and Levit, B. (1995). *On minimax filtering over ellipsoids.* Math. Methods Statist. **3**, 259–273.

[14] Belitser, E. and Levit, B. (2003). *On the empirical Bayes approach to adaptive filtering.* Math. Meth. Statist. **12**, 131–154.

[15] Beran, R. and Dümbgen, L. (1998). *Modulation of estimators and confidence sets.* Ann. Statist. **26**, 1826–1856.

[16] Berger, J. (1985). *Statistical decision theory and Bayesian analysis.* Springer, New York.

[17] Birgé, L. and Massart, P. (2001). *Gaussian model selection.* J. Eur. Math. Soc. **3**, 203–268.

[18] Brown, L.D. and Low, M.G. (1995). *Asymptotic equivalence of nonparametric regression and white noise.* Ann. Statist. **24**, 2384–2398.

[19] Cai, T.T. and Low, M.G. (2004). *An adaptation theory for nonparametric confidence intervals.* Ann. Statist. **32**, 1805–1840.

[20] Cai, T.T. and Low, M.G. (2006). *Adaptive confidence balls.* Ann. Statist. **34**, 202–228.

[21] Claeskens, G. and Hjort, N. (2004). *Goodness-of-fit via nonparametric likelihood ratios.* Scandinavian Journal of Statistics **31**, 487–513.

[22] Castillo, I. (2008). *Lower bounds for posterior rates for Gaussian process priors.* Electr. Journal Statist. **2**, 1281-1299.

[23] Cavalier, L., Golubev, G.K., Picard, D. and Tsybakov, A.B. (2002). *Oracle inequalities for inverse problems.* Ann. Statist. **30**, 843–874.

[24] Cavalier, L. and Tsybakov, A. (2001). *Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation.* Math. Methods Statist. **10**, 247–282.

[25] Cox, D.D. (1993). *An analysis of Bayesian inference for nonparametric regression.* Ann. Statist. **21**, 903–923.

[26] Diaconis, P. and Freedman, D. (1986). *On the consistency of Bayes estimates.* Ann. Statist. **14**, 1–26.

[27] Diaconis, P. and Freedman, D. (1986). *On inconsistent Bayes estimates of location.* Ann. Statist. **14**, 68–87.

[28] Donoho, D.L. and Johnstone, I.M. (1994). *Ideal spatial adaptation by wavelet shrinkage.* Biometrika **81**, 425–455.

[29] Donoho, D.L. and Johnstone, I.M. (1995). *Adapting to unknown smoothness via wavelet shrinkage.* J. Amer. Statist. Assoc. **90**, 1200–1224.

[30] Donoho, D.L. and Low, M.G. (1992). *Renormalization exponents and optimal pointwise rates of convergence.* Ann. Statist. **20**, 944–970.

[31] Doob, J. (1948). *Applications of the theory of martingales.* Le calcul des Probabilités et ses Applications, Colloques Internationales du CNRS, Paris, 22–28.

[32] Efromovich, S. (1999). *Nonparametric Curve Estimation.* Springer, New York.

[33] Efromovich, S. (2010). *Oracle inequality for conditional density estimation and an actuarial example.* Annals of the Institute of Statistical Mathematics **62**, 249–275.

[34] Efromovich, S. and Pinsker, M. (1984). *A learning algorithm for nonparametric filtering.* Automat. Remote Control. **24**, 1434–1440 (1984).

[35] Freedman, D. (1963). *On the asymptotic behaviour of Bayes estimates in the discrete case.* Ann. Math. Statist. **34**, 1386–1403.

[36] Freedman, D. (1999). *On the Bernstein-von Mises theorem with infinite dimensional parameters.* Ann. Statist. **27**, 1119–1140.

[37] Freedman, D. and Diaconis, P. (1983). *On inconsistent Bayes estimates in the discrete case.* Ann. Statist. **11**, 1109–1118.

[38] Genovese, C. and Wasserman, L. (2005). *Nonparametric confidence sets for wavelet regression.* Ann. Statist. **33**, 698–729.

[39] Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000). *Convergence rates of posterior distributions.* Ann. Statist. **28**, 500–531.

[40] Ghosal, S., Lember, J. and van der Vaart, A.W. (2008). *Nonparametric Bayesian model selection and averaging.* Electronic Journal of Statistics **2**, 63–89.

[41] Ghosal, S. and van der Vaart, A.W. (2001). *Maximum likelihood estimation for mixture of normal densities.* Ann. Statist. **29**, 1233–1263.

[42] Ghosal, S. and van der Vaart, A.W. (2007). *Convergence rates of posterior distributions for non-iid observations.* Ann. Statist. **35**, 192–225.

[43] Ghosal, S. and Van der Vaart, A.W. (2010). Draft of the book *Theory of Nonparametric Bayesian Inference.*

[44] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian nonparametrics.* Springer-Verlag, New York.

[45] Gill, R.D. and Levit, B.Y. (2002). *Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound.* Bernoulli **1**, 59–79.

[46] Golubev, G.K. (2004). *On a method for minimizing empirical risk.* Problems Inform. Transmission. **40**, 202–211.

[47] Golubev, Y. and Levit, B. (2004). *An oracle approach to adaptive estimation of linear functionals in a Gaussian model.* Math. Methods Statist. **13**, 392–408.

[48] Hjort, N.L. (2003). *Topics in non-parametric Bayesian statistics. With part A by Aad van der Vaart and part B by Sonia Petrone.* Oxford Statist. Sci. Ser. **27**, 455–487.

[49] Hoffmann, M. and Lepski, O. (2002). *Random rates in anisotropic regression.* Ann. Statist. **28**, 325–396.

[50] Ibragimov, I.A. and Khasminski, R.Z. (1981). *Statistical estimation: asymptotic theory.* Springer, New York.

[51] Ingster, Yu.I. and Suslina, I.A. (2003). *Nonparametric goodness-of-fit testing under Gaussian models.* Springer, New York.

[52] Johnstone, I. (2004). *Function estimation and Gaussian sequence models.* Monograph draft, http://www-stat.stanford.edu/∼imj/

[53] Johnstone, M. and Silverman, B.W. (2004). *Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences.* Ann. Statist. **32**, 1594–1649.

[54] Juditsky, A. and Lambert-Lacroix, S. (2003). *Nonparametric confidence set estimation.* Math. Meth. Statist. **12**, 410–428.

[55] Kerkyacharian, G. and Picard, D. (2002). *Minimax or maxisets?* Bernoulli **8**, 219–253.

[56] Kneip, A. (1994). *Ordered linear smoothers.* Ann. Statist. **22**, 835–866.

[57] Kotelnikov, V.A. (1959). *The theory of optimum noise immunity.* McGraw Hill, New York.

[58] Laurent, B. and Massart, P. (2000). *Adaptive estimation of a quadratic functional by model selection.* Ann. Statist. **30**, 325–396.

[59] Lepski, O.V. (1990). *One problem of adaptive estimation in Gaussian white noise.* Theory Probab. Appl. **35**, 459–470.

[60] Lepski, O.V. (1991). *Asymptotic minimax adaptive estimation. 1. Upper bounds.* Theory Probab. Appl. **36**, 645–659.

[61] Lepski, O.V. (1992). *Asymptotic minimax adaptive estimation. 2. Statistical model without optimal adaptation. Adaptive estimators.* Theory Probab. Appl. **37**, 468–481.

[62] Lepski, O.V. (1999). *How to improve the accuracy of estimation.* Math. Meth. Statist. **8**, 441–486.

[63] Lepski, O.V. and Spokoiny, V.G. (1997). *Optimal pointwise adaptive methods in nonparametric estimation.* Ann. Statist. **25**, 2512–2546.

[64] Li, K.-C. (1989). *Honest confidence regions for nonparametric regression.* Ann. Statist. **17**, 1001–1008.

[65] Li, X. and Zhao, L.H. (2002). *Bayesian nonparametric point estimation under a conjugate prior.* Statist. Prob. Lett. **58**, 23–30.

[66] Lifshits, M.A. (1995). *Gaussian random functions.* Kluwer Academic Publishers.

[67] Low, M.G. (1997). *On nonparametric confidence intervals.* Ann. Statist. **25**, 2547–2554.

[68] Nussbaum, M. (1996). *Asymptotic equivalence of density estimation and Gaussian white noise.* Ann. Statist. **24**, 2399–2430.

[69] Picard, D. and Tribouley, K. (2000). *Adaptive confidence interval for pointwise curve estimation.* Ann. Statist. **28**, 298–335.

[70] Pinsker, M. (1980). *Optimal filtration of square-integrable signal in Gaussian white noise.* Problems Inform. Transmission. **16**, 120–133.

[71] Robbins, H. (1955). *An empirical Bayes approach to statistics.* In: Proc. 3rd Berkeley Symp. on Math. Statist. and Prob. **1**, Berkeley, Univ. of California Press, 157–164.

[72] Robins, J. and van der Vaart, A.W. (2006). *Adaptive nonparametric confidence sets.* Ann. Statist. **34**, 229–253.

[73] Rosenblatt, M. (1956). *Remarks on some nonparametric estimates of a density function.* Ann. Statist. **27**, 832–837.

[74] Schwartz, L. (1965). *On Bayes procedures.* Z. Wahrscheinlichkeitstheorie und Verw. Gebiete **4**, 10–26.

[75] Shen, X. and Wasserman, L. (2001). *Rates of convergence of posterior distribution.* Ann. Statist. **29**, 687–714.

[76] Stone, C.J. (1980). *Optimal rates of convergence for nonparametric estimators.* Ann. Statist. **8**, 1348–1360.

[77] Tsybakov, A.B. (2004). *Introduction à l'estimation non-paramétrique.* Springer, Berlin.

[78] Tsybakov, A. (2008). *Introduction to nonparametric estimation.* Springer.

[79] VAN DER VAART, A.W. (1998). *Asymptotic statistics.* Cambridge University press.

[80] VAN DER VAART, A.W. and VAN ZANTEN, H. (2007). *Bayesian inference with rescaled Gaussian process priors.* Electr. Journal Statist. **1**, 433–448.

[81] VAN DER VAART, A.W. and VAN ZANTEN, H. (2008). *Rates of contraction of posterior distributions based on Gaussian process priors.* Ann. Statist. **36**, 1435–1463.

[82] WALKER, S. (2004). *New approaches to Bayesian consistency.* Ann. Statist. **32**, 2028–2043.

[83] WALKER, S. and HJORT, N.L. (2001). *On Bayesian consistency.* J. R. Stat. Soc. Ser. B Stat. Methodol. **63**, 811–821.

[84] WASSERMAN, L. (2006). *All of nonparametric statistics.* Springer.

[85] ZHAO, L.H. (2000). *Bayesian aspects of some nonparametric problems.* Ann. Statist. **28**, 532–552.

Samenvatting

We passen de bayesiaanse methode toe op het probleem van 'statistical inference' bij een onbekende oneindigdimensionale parameter in het model van gaussische rijen. De belangrijkste grootheid in de bayesiaanse analyse is de resulterende a-posterioriverdeling (posterior) van alle onbekende (eindig- of oneindigdimensionale) parameters die betrokken zijn bij de modellering. Stel dat we de prestaties van een bayesiaanse methode evalueren vanuit het standpunt van een frequentist. We doen dit door te kijken naar de snelheid waarmee de posterior zich concentreert rond de 'echte' waarden van de betrokken parameters met betrekking tot de kansverdeling van data onder die echte parameterwaarden. Traditioneel gezien neemt men aan dat de relevante onbekende grootheid behoort tot een niet-parametrische klasse die beschreven wordt door gladheidscondities. Alle tot nu toe verkregen resultaten over posterior snelheden zijn daarom gerelateerd aan de optimale minimaxsnelheden voor het schattingsprobleem over de corresponderende niet-parametrische gladheidsklassen, en derhalve globaal van aard. Intussen is er echter een nieuwe locale manier om optimaliteit te bereiken ontwikkeld binnen het kader van het statistische schatten, namelijk de orakelaanpak. Het hoofddoel van dit proefschrift is het ontwikkelen en bestuderen van de orakelbenadering tot de prestaties van posteriorconvergentie in het model van gaussische rijen. We stellen daarom een nieuw ijkpunt voor de snelheid waarmee de posterior convergeert, de zogenaamde posterior orakelsnelheid. Dit is de kleinst mogelijke snelheid over een familie van posteriorsnelheden corresponderend met een toepasselijk gekozen familie a-prioriverdelingen.

We introduceren het begrip van 'orakel projectieconvergentiesnelheid' ofwel de beste convergentiesnelheid over de familie van de projectieschatters. Verder stellen we een toepasselijke hiërarchische a-prioriverdeling (prior) voor, hetgeen ons in staat stelt om het probleem van bayesiaanse modelselectie aan te pakken, en vast te stellen dat de posterior zich concentreert rond de ware parameter met de orakel projectieconvergentiesnelheid, wat sterker is dan de posterior convergentie met de minimaxsnelheid over de niet-parametrische klasse onder de voorwaarde dat onze familie projectieschatters een minimaxschatter over die klasse bevat. We construeren ook een bayesiaanse schatter gebaseerd op de posterior en tonen aan dat deze aan een orakelongelijkheid voldoet. Daarnaast vullen we de bovengrens op de snelheid van convergentie van de posterior aan met een ondergrens voor de orakelsnelheid. De snelheden voor de boven- en ondergrens blijken samen te vallen met de orakel projectieconvergentiesnelheid.

Dit impliceert dat de orakel posteriorsnelheid scherp is en dat alle posterior kansmassa zich concentreert in een annulus rond de ware parameterwaarde. Al deze resultaten zijn niet-asymptotisch en uniform over l_2.

Verder bestuderen we de a-posterioriverdelingen in het 'many normal means model' met behulp van numerieke simulaties en vergelijken we de resultaten hiervan met de theoretisch verkregen resultaten uit het model van gaussische rijen. Om onze bevindingen te illustreren, construeren we geloofwaardigheidsbanden (credible bands) voor de signaalfunctie in het equivalente witteruismodel.

In dit proefschrift beschouwen we ook het probleem van bayesiaanse schatting van een lineaire functionaal, onder aanname dat het onbekende signaal afkomstig is van een Sobolev gladheidsklasse. De geconstrueerde bayesiaanse schatter van de functionaal blijkt de minimaxsnelheid over de Sobolev gladheidsklasse te bereiken, vanuit zowel het bayesiaanse perspectief als vanuit dat van de frequentist. Met behulp hiervan bewijzen we een resultaat over de convergentiesnelheid van de a-posterioriverdeling van de lineaire functionaal, die in de minimaxzin optimaal blijkt te zijn over de Sobolev-klasse. Onder de minst informatieve (ten opzichte van de Sobolev gladheidsklasse) prior beschouwen we een bayesiaanse versie van het adaptieve schatten van de functionaal, hetgeen in feite een probleem is van adaptieve filtering. We stellen het bayesiaans orakel voor, alsmede het orakel bayesiaans risico, en we laten zien dat onze adaptieve schatter van de functionaal asymptotisch scherp is, dat wil zeggen dat zijn risico asymptotisch samenvalt met het orakel bayesiaans risico. In andere woorden, we laten zien dat de resulterende adaptieve schatter van de functionaal de rol speelt van het bayesiaans orakel. Het bijkomend probleem de gladheidsparameter te schatten met behulp van een empirische bayesiaanse methode is ook bekeken.

Acknowledgements

The four years I have spent in Utrecht University during my PhD have been certainly a great time, for this I am grateful to all the nice people I have met. Though it is quite difficult to make this section complete, I would like to acknowledge at least some people.

First of all, I would like to express my gratitude to my co-promotor Eduard Belitser from whom I have learnt a lot during this years. It is due to his brilliant ideas and patient guidance that I was able to complete this thesis. Thank you for the freedom you gave me and for providing interesting and arduous problems, especially, for the attractive and difficult problem of estimating a functional with an unknown smoothness parameter, which turned out to be more complicated than we expected at the beginning. I very much appreciate your help in preparing the thesis.

Further, I would like to thank my promotor Roberto Fernandez for his responsiveness and the inspiring discussions we had.

Special gratitude is addressed to my reading committee: professors Subhashis Ghoshal, Richard Gill, Aad van der Vaart, and Harry van Zanten for their time invested in reading the manuscript.

I am indebted to Pieter Trapman, Bas Janssens, Sergey Anisov, Gil Cavalcanti, and Shinji Nakaoka for their useful suggestions and fruitful discussions.

I am also thankful to Charlene Kalle, Martin Bootsma, and specially to Pieter Trapman for their help in organizing the Stochastics Colloquium.

Special thanks to my colleagues Bas (thanks for the Dutch translation of the summary and teaching me some Dutch), Jeroen (thanks for explaining me some tricks in LaTex), Tanya, Janne, Jaap, Jan-Willem, Sebastiaan, Albert Jan, Job, Lee, Sebastian, Esther, Wouter, Bart, Jantien, Arjen, Bas, Jan Jitse, David, Dana, Ionut, Marti, Maria, Vincent, Charlene, Arthur (thanks for the beautiful LaTex template), Tammo Jan, Alex, Rogier, Jakub, Oliver, Alex and Slavik for the great time that we spent together.

A big thank to Shinji Nakaoka for his help in providing computational resources.

Lastly, I would like especially thank Dima for his support and help.

Curriculum Vitae

Aleksandra Babenko was born on November 14^{th}, 1983 in Rudny, Kazakhstan, which was a part of the Soviet Union at that time. In 2001 she completed her study at the Special Physics and Mathematics School associated with Novosibirsk State University and was enrolled in the Faculty of Mathematics and Mechanics of the same university. On the 3rd year of her university education she has followed the courses in Probability Theory and Statistics, the beauty of which made a very strong impression on her and determined the future research area. In 2006 she has successfully graduated (cum laude) from the faculty. In the same year she came to Utrecht University to start her PhD in theory of probability and statistics with Dr. Eduard Belitser. During her PhD she participated and gave talks at numerous conferences and workshops in the Netherlands and abroad. Moreover, she assisted six different bachelor courses and organized Stochastic Colloquium.