BAYESIAN ASYMPTOTICS
INVERSE PROBLEMS AND IRREGULAR MODELS
VRIJE UNIVERSITEIT

BAYESIAN ASYMPTOTICS
INVERSE PROBLEMS AND IRREGULAR MODELS

ACADEMISCH PROEFSCRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Exacte Wetenschappen
op dinsdag 9 april 2013 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Bartłomiej Tomasz Knapik

geboren te Tarnów, Polen
promotoren: prof.dr. J.H. van Zanten
prof.dr. A.W. van der Vaart
copromotor: dr. B.J.K. Kleijn
Acknowledgments

I am sincerely grateful to many people, without whom this thesis would not exist. Obviously I will thank them personally, but they should be also thanked publicly. Here is my humble attempt to do so.

Certainly, the most important are my supervisors. I still remember the moment at the end of my first year in Amsterdam when Harry van Zanten, who back then supervised my master project, told me I could start my PhD under his supervision. I could always count on him and ask for advice. I would also like to thank him for the patience and effort to understand my vague ideas and ambiguous notes.

I was extremely lucky to have Aad van der Vaart just two doors away from my office. He was always there for me, willing to help and give insightful comments. Sometimes a brief chat by a coffee machine would lead to a long discussion in front of a blackboard. I can only hope that I was not asking the same questions too many times.

Bas Kleijn was the first to introduce me to the world of Bayesian statistics, when I was still a master student. I really enjoyed many interesting discussions we had on various topics. I would like to express my gratitude to him for the constant encouragement to start over every time I ended up going round in circles.

It was really inspiring to work with all of them. Hartelijk bedankt!

I am grateful to the members of the reading committee, including Mathisca de Gunst, Geurt Jongbloed, Andrew Stuart, and Ismaël Castillo, for their time and presence at my defense. I would also like to thank them for many interesting conversations we had during several workshops, lectures, seminars, and conferences in the Netherlands and abroad.

This work has also benefited from a collaboration with my co-author and a fellow PhD student: Botond, let me thank you for your constant optimism, which was especially useful when I had problems noticing obvious things. I have to admit that your persistence in research has always impressed and inspired me. I also wanted to thank another fellow PhD student, with whom I shared a wall at Eurandom in Eindhoven. Kamil, I have always admired your broad mathematical interests. I could always knock on your door just to discuss our research, or solve some first year mathematical problems in a rather sophisticated way.

For a couple of months I had the biggest office at our department only for myself, but I definitely enjoyed more the time when I could share it with my officemates: René, Michał, Markus, Marcin, Alfred, and Tim. Even when they were not around, I could still count on the other members of the statistics group: Marianne, Fetsje, Beata, Suzanne, Jakub, Michael,
Shota, Johannes, Gwen, Wessel, Geert, and Rikkert. Thank you all for the lunch and coffee breaks, and various celebrations, sometimes joined by other colleagues at the department. Special thanks go to Maryke for helping me with many administrative issues, Suzanne for the samenvatting, and Asia for help with PDF issues. I am also grateful to the members of the Bayes Club, who would come to Amsterdam on Fridays to attend the seminar I organized, followed by interesting discussions in The Basket.

I have met many people since I came to Amsterdam in August 2007. I would especially like to mention my housemates: Agnieszka, Marta, Beata and Milosz, Sergiy, Klaudiusz, Kasia, Kinga and Radek, and Katarzyna Izabela. We shared quite a number of apartments, sometimes more than once, sometimes not at the same time. Thank you for everything, you know how to cheer me up when I am not in the mood.

I would also like to thank Maria and Theodoor, Kasia and Paweł, Asia and Wojtek, Ola and Konrad, Ania and Marcin, Agnieszka and Łukasz, Kasia, Gosia, Dorota, Kryśia, Paola, Justyna, Maciek, Michał, Szymon, Marcin, and Andrzej for the time we have spent together on various occasions, mostly starting with taking a bus to the north of Amsterdam.

It was always important to me to visit friends in Poland, and I want to thank them all for their hospitality, especially Agnieszka, and Agnieszka, Masza and Konrad. It was also great to have friends visiting me in Amsterdam. I wish you could have visited me more often, and I hope you liked Amsterdam as much as I do.

Finally, I would like to thank my family for their constant and unconditional love, support, and understanding. These things are never overrated, especially when you are hundreds of kilometers apart. Without them, this thesis would really never exist.

Bartek Knapik
February 2013
# Contents

1 **Introduction**  
1.1 Bayesian statistics ................................................. 3  
1.2 Statistical inverse problems ..................................... 10  
1.3 Regularity and Bernstein–von Mises theorems ..................... 13  
1.4 Overview ....................................................... 15  

I **Inverse problems** ............................................... 19  

2 **Recovery of the full parameter** .................................. 21  
2.1 Introduction ....................................................... 21  
2.2 Prior and posterior distributions .................................. 23  
2.3 Singular value decomposition: Sequence formulation .......... 26  
2.4 Mildly ill-posed problems ....................................... 27  
2.5 Extremely ill-posed problems: heat equation .................. 33  
2.6 Proofs of Theorems 2.3 and 2.5 .................................. 40  
2.7 Proofs of Theorems 2.4 and 2.6 .................................. 41  

3 **Adaptive recovery of the full parameter** .................... 47  
3.1 Introduction ....................................................... 47  
3.2 Description of the empirical and hierarchical Bayes procedures .... 50  
3.3 Adaptation and contraction rates ................................ 52  
3.4 Simulation example: Volterra operator .......................... 55  
3.5 Proof of Lemma 3.1 ............................................... 57  
3.6 Proof of Theorem 3.2 ............................................. 59  
3.7 Proof of Theorem 3.3 ............................................. 67  
3.8 Proof of Theorem 3.6 ............................................. 73  

4 **Recovery of linear functionals of the parameter** ........... 79  
4.1 Introduction ....................................................... 79  
4.2 Marginal posterior distributions. II-measurable linear functionals. .... 81  
4.3 Mildly ill-posed problems ....................................... 83
4.4 Extremely ill-posed problems: heat equation ........................................... 90
4.5 Proofs of Theorems 4.3, 4.4, and 4.7 ......................................................... 95
4.6 Proofs of Theorems 4.5 and 4.8 ................................................................. 97
4.7 Proof of Theorem 4.6 .................................................................................. 100

II Irregular models ......................................................................................... 103

5 Semiparametric posterior limits under LAE ................................................. 105
  5.1 Introduction ............................................................................................... 105
  5.2 Main results ............................................................................................. 108
  5.3 Asymptotic posterior exponentiality .......................................................... 112
  5.4 Estimation of support boundary points .................................................... 118
  5.5 Proofs ...................................................................................................... 125

A Technical lemmas ....................................................................................... 141

Bibliography .................................................................................................. 151

Summary ....................................................................................................... 159

Samenvatting ................................................................................................. 161
The main goal of statistical estimation is to translate observations into conclusions. Quite naturally, the given observations are considered to be uncertain. This might be caused by various factors, from the imprecision of the measurement procedure, to some implicit randomness of the entity represented by the observations. Therefore, we first need to explain how to describe the uncertainty in the observations. The next step in the statistical estimation problem is to define possible sources of observations accounting for the uncertainty. Finally, we describe an estimation procedure that translates observations into conclusions: given the observations selects their source.

We denote the observation by $Y$ and assume it is a random variable with values in a measurable space $\mathcal{Y}$ with $\sigma$-algebra $\mathcal{Y}$. We often refer to $Y$ as data, and consider it to be a single number, a vector, or even an element of an infinite-dimensional space, such as a function space. The space $\mathcal{Y}$ is assumed to be a measurable space to enable the consideration of probability measures on $\mathcal{Y}$, accounting for randomness of the observed data $Y$. This is formalized by considering an underlying probability space $(\Omega, \mathcal{F}, P)$ on which the random variable $Y$ is defined. We then use the random variable $Y$ to “push-forward” the probability measure $P$ on $\Omega$ to a probability measure on $\mathcal{Y}$.

Next, we define a statistical model $\mathcal{P}$ to be a collection of candidate probability measures $P$ on the space $\mathcal{Y}$. We can of course take $\mathcal{P}$ to be a collection of all possible probability measures on $\mathcal{Y}$, but often we consider some meaningful subcollection of probability measures, tailored for the considered problem. In this way we define possible sources of the observation $Y$. We assume that the observation $Y$ is represented by the “true distribution of the data”, denoted by $P_0$. This assumption is essential for the frequentist understanding of statistics. We also assume that the true $P_0$ belongs to the model $\mathcal{P}$.

Since the conclusions that are to be drawn from the observations are often related to certain characteristics of the underlying distribution, rather than to the distribution as a whole, a popular way of representing a statistical model is a description in terms of a parametrization. We choose a parameter space $\Theta$ and map an element $\theta$ of $\Theta$ to an element of $\mathcal{P}$, denoted by $P_\theta$. Throughout this thesis we assume the parametrization to be one-to-one. The “true
distribution of the data” corresponds now to the “true parameter” \( \theta_0 \). A straightforward example is the normal model in which we estimate mean and variance, thus \( \theta = (\mu, \sigma^2) \), and \( P_\theta = N(\mu, \sigma^2) \), using the common notation for a normal distribution with mean \( \mu \) and variance \( \sigma^2 \). The parameter \( \theta \) can be also more complicated, and we note that the model can be parametrized by itself, thus \( \Theta \) can also equal \( \mathcal{P} \).

The term parameter is often used in this thesis, and we refer to both \( \mathcal{P} \) and the parameter space \( \Theta \) as the model. Another term which we use often is nonparametric, even when we consider parametrized models \( \Theta \). This can be clearly misleading, but is so popular that we stick to this terminology. We call the model parametric, when the parameter space \( \Theta \) is finite-dimensional, for instance is a subset of \( \mathbb{R}^k \) for some \( k \geq 1 \). We call the model nonparametric, when the parameter space \( \Theta \) is infinite-dimensional, for instance is a space of square integrable functions on the interval \([0, 1]\), or a space of square summable sequences. Finally, we call the model semiparametric, when it is neither parametric nor nonparametric.

For instance, in this thesis we consider two semiparametric settings that are indexed by an infinite-dimensional parameter, but we estimate only a parametric (thus finite-dimensional) aspect of the parameter.

An estimation procedure is the final ingredient of the statistical estimation problem, as it results in an estimate for \( P_0 \), denoted by \( \hat{P} \) or \( \hat{\theta} \), depending on the chosen perspective. In the parametric setting the frequentist theory of estimation is well established, with procedures like moment estimation, maximum likelihood estimation, to name only a few. Semi- and nonparametric settings are more demanding and involved, but theory of estimation in these settings is also well-established in the literature.

In this thesis we employ Bayesian approach to statistical estimation problems. The reason lies mostly in the conceptual simplicity of the Bayesian paradigm. However, already at this point we want to stress that our approach is pragmatic, and, therefore, hybrid: we consider the Bayesian approach to be an estimation procedure in the frequentist setting. In this chapter we first formally introduce the Bayesian paradigm, and later show how it can be exploited by the frequentist statistician.

An important aspect of statistical estimation is the assessment of the quality of statistical procedures performed on the observed data. To this end, we consider a sequence of observations \( (Y_n) \) indexed by a parameter \( n \) tending to infinity. In this thesis the parameter \( n \) is related to the level of the noise (uncertainty) in the observations: large values of \( n \) indicate the small level of the noise. The true underlying distribution \( P_0 \) from the preceding paragraphs is now denoted by \( P_0^n \). At first sight it is not clear what we want to estimate in this setting, since the underlying distribution varies with \( n \). We therefore turn to the parametrized model and assume that for every \( n \geq 1 \) the distribution \( P^n \) corresponds to some fixed \( \theta \in \Theta \), and as a result, \( P_0^n \) corresponds to some fixed \( \theta_0 \in \Theta \). (More formally, for a given model \( \mathcal{P} \) we consider sequence of models \( \mathcal{P}_n \) and assume that \( P^n \in \mathcal{P}_n \) is a projection of \( P^\infty \in \mathcal{P} \). Here an estimator \( \hat{\theta}_n \) for \( \theta_0 \) is a function of the observation \( Y_n \) thus also a random variable. Under the frequentist assumption that \( Y_n \) is distributed according to the probability measure \( P_0^n \), we might be interested in the asymptotic (i.e., when \( n \) goes to infinity) properties of the estimator \( \hat{\theta}_n \). These include consistency (convergence of \( \hat{\theta}_n \) to \( \theta_0 \)) and rate of convergence (the speed at which \( \hat{\theta}_n \) converges to \( \theta_0 \)). There are other aspects that qualify the statistical
procedure. Together with the two mentioned above, they are elaborated on in the next section on Bayesian procedures.

In this chapter we provide a brief introduction to Bayesian statistics, and several key aspects of the Bayesian estimation procedure when used in the frequentist context. The main aim of Section 1.1 is to familiarize the reader with important concepts that we later study in two particular statistical problems. In Section 1.2 we introduce nonparametric inverse problems, viewed from a statistical perspective. The presentation is rather brief, and will be extended in the first part of this work. In both statistical settings considered in this thesis we present semiparametric versions of the Bernstein–von Mises theorem, the ultimate link between Bayesian and frequentist statistics. We discuss this important result in the regular, parametric setting in Section 1.3, together with an overview of the existing literature on Bernstein–von Mises-type results in other semi- and nonparametric models. The chapter is concluded with an overview section, in which we describe both parts of the thesis, provide brief introductions to each of the chapters and indicate the contribution of this work to the theory of Bayesian asymptotics.

1.1 Bayesian statistics

In this section we formally describe the Bayesian paradigm and compare it with the frequentist paradigm. We first consider an example of a very abstract setting, leading to the derivation of an expression for the posterior distribution. Later, when we take the frequentist perspective, we use this expression as a definition of an estimation procedure.

We also introduce three definitions of Bayesian asymptotics: consistency, posterior rate of contraction, and credibility. All these concepts have their equivalents in the frequentist notions of estimation and confidence sets. We assume the reader is well acquainted with these frequentist concepts, and we mention them rather briefly.

For a more detailed treatment on both Bayesian and frequentist concepts discussed in this section we refer the reader to, e.g., [56, 69, 71, 99].

Prior and posterior distributions

Let $\Theta$ be the parameter space of a model $\mathcal{P}$ (and later also referred to as the model). Consider the observation $Y$ with values in a measurable Polish space $\mathcal{Y}$ with $\sigma$-algebra $\mathcal{Y}$, and suppose that $\Theta$ is a measurable Polish space with $\sigma$-algebra $\mathcal{T}$. The model parameter takes values $\theta \in \Theta$ but is a random variable in this context. We denote this random variable by $\vartheta$, but later drop this notation, since in most situations it is clear whether $\theta$ denotes a fixed value, or the random variable $\vartheta$. We assume that on the product space $\mathcal{Y} \times \Theta$ with product $\sigma$-algebra $\mathcal{Y} \times \mathcal{T}$ we have a probability measure

$$
\Pi : \mathcal{Y} \times \mathcal{T} \to [0, 1],
$$

which is not a product measure. The probability measure $\Pi$ provides a joint probability distribution for $(Y, \vartheta)$ (later we write simply $(Y, \theta)$), where $Y$ is the data and $\vartheta$ is (the random variable associated with) the parameter of the model.
Implicitly the choice for the measure $\Pi$ defines the model in Bayesian context, by the possibility to condition on $\vartheta = \theta$ for some $\theta \in \Theta$. The conditional distribution $\Pi_{Y|\vartheta}: \mathcal{Y} \times \Theta \to [0, 1]$ describes the distribution of the observations $Y$ given the parameter $\vartheta$. It follows from classical probability theory that the conditional probability $\Pi_{Y|\vartheta}$ can be taken to be a so-called regular conditional distribution since $\mathcal{Y}$ is a Polish space, under the additional assumption that the $\sigma$-algebra $\mathcal{Y}$ equals the Borel $\sigma$-algebra. As such it defines the elements $P_\theta$ of the (parametrized) model $P = \{P_\theta: \theta \in \Theta\}$, although the role they play in Bayesian context is slightly different from that in a frequentist setting. The measures $\Pi_{Y|\vartheta} (\cdot | \vartheta = \theta)$ form a ($\Pi$-almost sure) version of the elements $P_\theta$ of the model $P$:

$$P_\theta = \Pi_{Y|\vartheta} (\cdot | \vartheta = \theta): \mathcal{Y} \to [0, 1].$$

The probability measure $\Pi$ gives rise to two distributions that we call prior and posterior distributions.

**Definition 1.1.** (Prior and posterior distributions). The marginal probability $\Pi(\mathcal{Y} \times \cdot)$ on $\mathcal{T}$ is called the prior distribution. The conditional distribution

$$\Pi_{\vartheta|Y}: \mathcal{T} \times \mathcal{Y} \to [0, 1],$$

is called a posterior distribution.

We often write $\Pi$ for the prior, slightly abusing the notation: for $A \in \mathcal{T}$ let $\Pi(A)$ denote $\Pi(\mathcal{Y} \times A)$. We again note that the conditional distribution in the above definition can be taken to be a regular distribution, under the additional assumption that the $\sigma$-algebra $\mathcal{T}$ equals the Borel $\sigma$-algebra.

Assuming that the model $\mathcal{P} = \{P_\theta: \theta \in \Theta\}$ is dominated by a $\sigma$-finite measure $\mathcal{M}$ on $\mathcal{Y}$, the posterior distribution can be also expressed in terms of $\mathcal{M}$-densities $p_\theta = dP_\theta/d\mathcal{M}: \mathcal{Y} \to \mathbb{R}$. The Bayes rule then yields the following expression for the posterior distribution:

$$\Pi(\vartheta \in A | Y) = \int_A p_\theta(Y) d\Pi(\theta) \left/ \int_\Theta p_\theta(Y) d\Pi(\theta), \right. 
\tag{1.1}$$

where $A \in \mathcal{T}$ is a measurable subset of the model $\Theta$.

Before we introduce the use of the Bayesian perspective in the frequentist context, we discuss one more definition related to the notion of the model in Bayesian statistics. Suppose we have a model $\Theta$ and its subset $\Theta_1 \subset \Theta$. If the prior assigns mass zero to the submodel $\Theta_1$, then it does not play a role, since omission of $\Theta_1$ from $\Theta$ does not influence the posterior (cf. (1.1)). We therefore make the following definition.

**Definition 1.2.** (Prior support). In addition to $(\Theta, \mathcal{T}, \Pi)$ being a probability space, let $(\Theta, \tau)$ be a Polish space. Assume that $\mathcal{T}$ equals the Borel $\sigma$-algebra corresponding to the topology $\tau$. The support $\text{supp}(\Pi)$ of the prior $\Pi$ is defined as:

$$\text{supp}(\Pi) = \bigcap\{A \in \mathcal{T}: A \text{ closed}, \Pi(A) = 1\}.$$
1.1. Bayesian statistics

**Proposition 1.3.** Let $\Pi$ be a probability measure on $(\Theta, T)$, and let $T$ be the Borel $\sigma$-algebra corresponding to the topology $\tau$ on $\Theta$. Then $\text{supp}(\Pi)$ is closed. If in addition $(\Theta, \tau)$ is a Polish space, then $\Pi(\text{supp}(\Pi)) = 1$.

**Proof.** First note that $\text{supp}(\Pi)$ is closed as an intersection of closed sets. Next, note that $\Theta \setminus \text{supp}(\Pi) = \bigcup \{U \in \tau : \Pi(U) = 0\}$. Since $\Theta$ is Polish, there exists a countable collection $\mathcal{U} \subset \tau$ such that $U = \bigcup \{O \in \mathcal{U} : O \subset U\}$ for each $U \in \tau$. But then there exists a countable collection $\mathcal{V}$ of open sets in $\Theta$ such that $\Theta \setminus \text{supp}(\Pi) = \mathcal{V}$, and $\Pi(V) = 0$ for each $V \in \mathcal{V}$.\]

Up to this point we did not use an “underlying distribution of the data”, the key notion of the frequentist statistics. Suppose that we take the frequentist perspective, so we choose a model $P$ and look for an estimation procedure based on the Bayesian paradigm.

The expression in (1.1) indicates that we can choose the parameter space $\Theta$ (and thus the model $P$) together with a prior $\Pi$, and then calculate the posterior distribution. The prior distribution is introduced as a measure on the model, rather than a marginal of a measure implicitly defining the model, and confining the parameter to the space of observations. Thus we might keep the observations and the model separated, and then define the posterior distribution. The Bayesian methodology serves then as a starting point for an estimation procedure. The prior distribution might restrict the model, and the posterior distribution gives rise to many possible estimators.

We consider a sequence of observations $(Y_n)$ and introduce three notions of Bayesian asymptotics, expressing the quality of a posterior distribution. Recall that the observation at the $n$th stage is denoted by $Y_n$ and distributed according to $P^n_0$. We assume that $P^n$ is parametrized by $\theta$ for every $n$. We denote the prior distribution on $\Theta$ by $\Pi$, and the resulting posterior distribution by $\Pi_n(\cdot|Y_n)$.

**Consistency**

We start with the most basic concept regarding the quality of the posterior distribution, which is called consistency. In frequentist statistics an estimator for $\theta_0$ is called consistent if it converges under the true distribution of the data to $\theta_0$. In Bayesian context we speak of consistency of the posterior. We again assume that the parameter space is a Polish space. Let $d$ denote a metric such that $(\Theta, d)$ is separable.

**Definition 1.4.** (Posterior consistency). The posterior distribution $\Pi_n(\cdot|Y_n)$ is said to be consistent at $\theta_0$ if for every neighborhood $U$ of $\theta_0$,

$$\Pi_n(U^c|Y_n) \to 0,$$

in $P^n_0$-probability, as $n \to \infty$.

Consistency can be summarized as saying that the posterior distribution concentrates at $\theta_0$. A prior or posterior distribution stands for the statistician’s knowledge (or belief) about unknown parameters. A perfect knowledge implies a degenerate prior, which later results in a degenerate posterior. Therefore, the concentration of the posterior at $\theta_0$ is highly desired. In other words “the observation overrides prior beliefs asymptotically”.
Two remarks are in order concerning the above definition. First of all, we note that if \( U' \subset U'' \) are neighborhoods of \( \theta_0 \), then if the convergence in the definition of consistency holds for \( U' \), then it also holds for \( U'' \). As a result, we may consider only small neighborhoods of \( \theta_0 \). Moreover, we may consider neighborhoods of the form \( U_\delta = \{ \theta \in \Theta : d(\theta, \theta_0) \leq \delta \} \) for all \( \delta \) small enough, where \( d \) is the metric on \( \Theta \).

In the present work we do not focus our attention on consistency but rather finer properties of the posterior distribution. Therefore, we only briefly mention two theorems on posterior consistency that apply to the nonparametric setting.

Doob’s theorem is the first well-known result on posterior consistency. It basically says that for any fixed prior the posterior distribution is consistent at every \( \theta \) except those in a “bad set” that is “small” when seen from the prior point of view (i.e., it has a prior mass zero). We note, however, that the \( \Pi \)-null sets might be very large, especially when the parameter space is infinite-dimensional.

Schwartz’ theorem is a more informative result on posterior consistency for nonparametric dominated models. The posterior distribution is consistent under two conditions on the prior and the model. In contrast to Doob’s theorem, Schwartz’ theorem answers the question whether the prior yields posterior consistency at the given element of the model.

**Posterior rate of contraction**

The definition of consistency says that the posterior distribution concentrates on balls of radius \( \delta \), for all \( \delta > 0 \). This suggests that the radii \( \delta \) can also depend on \( n \), and results in the following definition:

**Definition 1.5.** (Posterior rate of contraction). The posterior distribution \( \Pi_n(\cdot | Y_n) \) is said to contract at \( \theta_0 \) at rate \( \varepsilon_n \), \( \varepsilon_n \downarrow 0 \), if

\[
\Pi_n(\theta: d(\theta, \theta_0) \geq M_n \varepsilon_n | Y_n) \to 0
\]

in \( P_0^n \)-probability, for every \( M_n \to \infty \) as \( n \to \infty \).

We first note that the rate of contraction is related to the distance \( d \), and implies consistency (since \( \varepsilon_n \downarrow 0 \)). Moreover, if \( \varepsilon_n \) is a rate of contraction, then every sequence that tends to zero at a slower rate is also a contraction rate, according to the definition. Usually there is the fastest rate of contraction, depending again on the metric \( d \) and the model \( \mathcal{P} \).

Since the choice of the metric \( d \) is rather natural, we now mainly focus on the model dependence of the rate. More specifically, let us assume that we consider a collection of submodels \( \mathcal{P}_\beta \subset \mathcal{P} \) (or, equivalently, \( \Theta_\beta \subset \Theta \)), indexed by some index \( \beta \). In this thesis we mainly consider submodels consisting of parameters of certain “regularity”. For example, when the parameter space \( \Theta \) is a function space, then \( \Theta_\beta \) is a subset of \( \Theta \) consisting of \( \beta \)-smooth functions (in Hölder or Sobolev sense) for \( \beta > 0 \). Our interest lies in the fastest contraction rate, given the true \( \theta_0 \) is in \( \Theta_\beta \) for a given “regularity” \( \beta \). It is rather intuitive that the fastest rate of contraction depends on \( \beta \).

We connect the theory of posterior rates of contraction to the (frequentist) theory of “optimal” rates of estimation, typically defined by the minimax criterion: given \( \theta_0 \in \Theta_\beta \) for
some $\beta$, we look at all estimators for $\theta_0$ and call the\ minimax risk\ the maximal risk of an estimator with minimal risk among all estimators. The following proposition shows that the posterior cannot give better estimates than the best frequentist estimation procedure, since the posterior yields a point estimator that converges to the true $\theta_0$ at the same rate as the rate of contraction.

**Proposition 1.6.** Suppose that the posterior distribution $\Pi_n(\cdot|Y_n)$ contracts at rate $\varepsilon_n$ at $\theta_0$ relative to the metric $d$ on $\Theta$. Let $\hat{\theta}_n$ be defined as the center of a (nearly) smallest ball that contains posterior mass at least $1/2$. Then $\varepsilon_n^{-1}d(\hat{\theta}_n, \theta_0)$ is bounded in $P^n$-probability.

**Proof.** For $B(\theta, r) = \{s \in \Theta: d(s, \theta) \leq r\}$ the closed ball of radius $r$ around $\theta \in \Theta$, let $\hat{\theta}_n(\theta) = \inf\{r: \Pi_n(B(\theta, r)|Y_n) \geq 1/2\}$, where the infimum over the empty set is $\infty$. Taking the balls closed ensures that $\Pi_n(B(\theta, \hat{\theta}_n(\theta))|y_n) \geq 1/2$, for every $\theta$. Let $\hat{\theta}_n$ be a near minimizer of $\theta \mapsto \hat{\theta}_n(\theta)$ in the sense that $\hat{\theta}_n(\theta) \leq \inf_{\theta} \hat{\theta}_n(\theta) + \varepsilon_n$.

Since the posterior contracts at rate $\varepsilon_n$, $\Pi_n(B(\theta_0, M_n\varepsilon_n)|Y_n) \to 0$ in probability, for every $M_n \to \infty$. As a first consequence $\hat{\theta}_n(\theta_0) \leq M_n\varepsilon_n$ with probability tending to one, and hence $\hat{\theta}_n(\theta_0) \leq M_n\varepsilon_n + \varepsilon_n$ with probability tending to one. As a second consequence the balls $B(\theta_0, M_n\varepsilon_n)$ and $B(\hat{\theta}_n, \hat{\theta}_n(\theta_0))$ cannot be disjoint, as their union would contain mass nearly $1 + 1/2$. This shows that $d(\theta_0, \hat{\theta}_n) \leq 2M_n\varepsilon_n + \varepsilon_n$ or, equivalently, $\varepsilon_n^{-1}d(\theta_0, \hat{\theta}_n) \leq 2M_n + 1$ with probability tending to one. This being true for every $M_n \to \infty$ implies the assertion. \hfill $\square$

Clearly the posterior distribution depends on the prior distribution, and in many situations this dependence is inherited by the rate of contraction. Consider the following example: suppose that the draws from some prior are $\alpha$-smooth in the Sobolev sense, the true function is $\beta$-smooth on interval $[0, 1]$, and we observe the function in $n$ equidistant points, perturbed by independent Gaussian errors. The posterior rate of contraction is then of the order $n^{-(\alpha \wedge \beta)/(1+2\alpha)}$, where $\alpha \wedge \beta$ denotes the minimum of $\alpha$ and $\beta$. It clearly depends on the (smoothness of the) prior and is not optimal unless $\alpha = \beta$.

Since the true $\beta$ is assumed to be unknown, prior distributions should not depend on the regularity of the true parameter. If for such a prior $\Pi$, the posterior contraction rate is optimal for every possible value of $\beta$, the prior $\Pi$ is called rate-adaptive: it adapts to the optimal rate of contraction for every $\beta$, without the knowledge of the “regularity” of the true $\theta_0$.

It is clear that such priors are highly desirable. It should be noted that finding rate-adaptive priors is not an easy task, and often requires subtle reasoning. It has been also noted in the literature that some priors can even lead to inconsistency. In this thesis we will not see examples of such priors, but we will show how the choice of parameters of the prior affects the rate of contraction, and how to choose these parameters in order to achieve the optimal rate of contraction.

One of the first general treatments of theory posterior contraction has been introduced by Ghosal et al. in [39]. The authors present several theorems, starting with a general theorem with an in-probability statement, later improved to an almost sure assertion. They also provide a generalization yielding the right result in the finite-dimensional situation. It should be noted that the most general result of the paper is formulated in terms of tests of the truth versus the complement of a shrinking ball around it. The existence of tests with certain bounds...
on error probabilities is automatically guaranteed for the Hellinger distance, provided the alternative is convex. Then next step is to combine these tests into a test for the complement of the shrinking ball around the truth, and under the conditions of the theorems in [39] such a construction is possible. Therefore, the results in [39] are stated in terms of the Hellinger distance, with consequences for other equivalent distances. The general theory of posterior contraction has been developed and applied in other statistical settings, see, for instance, [40–42, 63, 64, 86, 101, 103], but never with other norms.

We now turn to the white noise model considered in [40]. We consider an equivalent formulation, the so-called infinite-dimensional normal mean model, that will be studied in the first part of this thesis. We present a slight generalization of the model and the corresponding result on the posterior contraction.

Suppose we observe an infinite-dimensional random vector \( Y_n = (X_1, X_2, \ldots) \), where \( X_i \)'s are independent, \( X_i \) has distribution \( N(\kappa_i \theta, n^{-1}) \), for \( i \geq 1 \), \( \theta = (\theta_1, \theta_2, \ldots) \in \ell_2 \), and \( \kappa = (\kappa_1, \kappa_2, \ldots) \in \ell_2 \) is given. The parameter \( \theta \) is unknown and the goal is to make inference about \( \theta \). Let \( \| \theta \| \) denote the \( \ell_2 \)-norm of \( \theta \), and let \( P^n_\theta \) denote the distribution of the vector \( Y_n \).

For \( \varepsilon > 0 \) let \( N(\varepsilon, B, \| \cdot \|) \) be the \( \varepsilon \)-covering number, which is the minimal number of balls of radius \( \varepsilon \) needed to cover \( B \). The following theorem is a version of Theorem 6 in [40].

**Theorem 1.7.** For any \( \theta \in \ell_2 \), and \( \kappa \in \ell_2 \) let \( K \theta = (\kappa_1 \theta_1, \kappa_2 \theta_0, \ldots) \). Suppose that for \( \varepsilon_n \to 0 \), \( n \varepsilon_n^2 \) bounded away from 0, and \( \Theta \in \ell_2 \), the following conditions are satisfied:

\[
\sup_{\varepsilon > \varepsilon_n} \log N(\varepsilon/8, \{ \theta \in \Theta : \| K \theta - K \theta_0 \| < \varepsilon \}, \| \cdot \|) \leq n \varepsilon_n^2,
\]

for every \( j \in \mathbb{N} \)

\[
\frac{\Pi_n(\theta \in \Theta : \| K \theta - K \theta_0 \| \leq j \varepsilon_n)}{\Pi_n(\theta \in \Theta : \| K \theta - K \theta_0 \| \leq \varepsilon_n)} \leq e^{n \varepsilon_n^2 j^2/64}.
\]

Then

\[
\Pi_n(\theta \in \Theta : \| K \theta - K \theta_0 \| \geq M_n \varepsilon_n | Y_n) \to 0
\]

for every \( M_n \to \infty \) in \( P^n_\theta \)-probability.

The above theorem gives the posterior rate of contraction with respect to the \( "K-\ell_2" \)-norm, which is an \( \ell_2 \)-type norm weighed by the sequence \( \kappa \). The \( "K-\ell_2" \)-norm above is equivalent to the standard \( \ell_2 \)-norm (and thus the contraction rate corresponds to the estimation of \( \theta_0 \) rather than \( K \theta_0 \)) if and only if the sequence \( \kappa \) is bounded away from zero and infinity, i.e., there exists a constant \( c \geq 1 \) such that \( c^{-1} \leq \kappa_i \leq c \), for \( i \geq 1 \).

**Credibility**

Bayesian credible sets are counterparts of frequentist confidence regions. We again consider a parametrized model \( \Theta \) and let \( Y_n \) be distributed according to \( P^n_\theta = P^n_{\theta_0} \). We choose a confidence level \( 1 - \gamma \in (0, 1) \). We then say that a subset of \( \Theta \) dependent only on the observations \( Y_n \) denoted by \( C_n(Y_n) \) is a confidence region for \( \theta \) of confidence level \( 1 - \gamma \) if

\[
P^n_\theta(\theta \in C_n(Y_n)) \geq 1 - \gamma,
\]
for all $\theta \in \Theta$ and $n \geq 1$. If the above inequality holds in the limit, rather than for every $n \geq 1$, we say that $C_n(Y_n)$ is an asymptotic confidence region for $\theta$ of confidence level $1 - \gamma$. Even in parametric frequentist statistics, we usually encounter asymptotic confidence regions, rather than exact confidence regions.

Consider the same model $\Theta$, with prior $\Pi$, and choose a level $1 - \gamma \in (0, 1)$. Let $D_n(Y_n)$ be a subset of $\Theta$. We say that $D_n(Y_n)$ is a credible set for $\theta$ of credibility level $1 - \gamma$ if

$$\Pi_n(\theta \in D_n(Y_n)|Y_n) \geq 1 - \gamma.$$ 

There are plenty of ways to choose the set $D_n(Y_n)$. In practice one tends to consider sets that capture the most posterior mass, being of reasonable size at the same time. For instance, in the case the posterior is unimodal, one could consider $D_n(Y_n)$ to be a ball (or ellipsoid) centered at the posterior mode.

The performance of credible sets is certainly of high practical importance. In many situations it might be hard to obtain frequentist confidence regions. On the other hand, numerical methods for simulation from the posterior distribution are well-developed and often used in practice. Therefore, one simulates the posterior distribution based on the prior and the observations, to derive a subset $D_n(Y_n)$ such that it captures posterior mass $1 - \gamma$. The natural question arises: can credible sets be used as frequentist confidence regions? To answer this question we consider the frequentist confidence or frequentist coverage of the credible set $D_n(Y_n)$:

$$P_\theta^n(\theta \in D_n(Y_n)),$$

and our interest lies mostly in its asymptotic behavior, that is as $n \to \infty$. It is not immediately clear how the above frequentist coverage is related to the credibility level. As a matter of fact, this important aspect of Bayesian asymptotics seems to be underrepresented in the literature.

In regular, parametric models for independently and identically distributed data there is a close, asymptotic link between Bayesian credible sets and frequentist confidence sets centered on the maximum likelihood estimator (or any other “optimal” estimator). This follows from the Bernstein–von Mises theorem (discussed in more detail in Section 1.3), which states that under the aforementioned regularity condition on the model, and a mild condition on the prior we have

$$\sup_A \left| \Pi_n(\theta \in A|Y_n) - N(\hat{\theta}_n, n^{-1}I_{\hat{\theta}_0}^{-1})(A) \right| \to 0$$

in $P_\theta^n$-probability, where $\hat{\theta}_n$ is the maximum likelihood estimator for $\theta_0$, $I_{\hat{\theta}_0}$ is the Fisher information matrix, and the supremum is taken over all measurable subsets of $\Theta$. We denote by $N(\hat{\theta}_n, n^{-1}I_{\hat{\theta}_0}^{-1})(A)$ the probability of the event $A$ under $N(\hat{\theta}_n, n^{-1}I_{\hat{\theta}_0}^{-1})$. Let $C_n(Y_n)$ be the confidence region of confidence level $1 - \gamma$ centered at the maximum likelihood estimator $\hat{\theta}_n$. Then $N(\hat{\theta}_n, n^{-1}I_{\hat{\theta}_0}^{-1})(C_n(Y_n))$ is asymptotically equal to $1 - \gamma$, by the normality (and optimality) of the maximum likelihood estimator under the condition on the model. By the above assertion, it is also an asymptotic credible set. Analogously, if $D_n(Y_n)$ is a credible set of credibility level $1 - \gamma$, then it is also an asymptotic confidence region.

In nonparametric (and also semiparametric) models little is known about the frequentist coverage of Bayesian credible sets. For a long time the Bernstein–von Mises phenomenon
was studied only in the parametric setting, only to conclude that the limiting behavior of the posterior is a very delicate matter in nonparametric models (see also Section 1.3). For instance, Cox in [26] considered the frequentist performance of Bayesian credible sets in the nonparametric fixed design regression setting and came to the alarming conclusion that Bayesian credible sets have frequentist coverage zero. More precisely, for almost every parameter \( \theta_0 \) from the prior the coverage of a credible set (of any level) is 0.

In the first part of this thesis we study the infinite-dimensional normal mean model, equivalent to that of [26]. Again, as in the posterior contraction, the coverage depends on the true parameter \( \theta_0 \) and the prior together. Moreover, in the setting of this thesis the coverage can be understood in terms of a bias-variance trade-off, much as the coverage of frequentist nonparametric procedures. Our results expand the story about the frequentist coverage of Bayesian credible sets.

### 1.2 Statistical inverse problems

One possible definition of a statistical inverse problem is that instead of observing a noisy version of the parameter of interest \( \mu \) directly, we observe its transformed version. (To comply with the notation in the rest of the thesis, in the inverse problem setting we use the Greek letter \( \mu \), instead of \( \theta \), to denote the parameter of interest.) We assume that the transformation, which we denote by \( K \), does not induce any noise in the observation. We note, however, that it might have a strong influence on the estimation procedure.

Nonparametric statistical inverse problems arise in many fields of applied science, including quantum physics (quantum homodyne tomography), geophysics (soil contamination, seismic activity), econometrics (instrumental variables), financial mathematics (model calibration of the volatility), genomics (gene expressions), medical image analysis (X-ray tomography) and astronomy (blurred images of the Hubble Space Telescope), to mention but a few. Some of these examples clearly indicate indirectness of the problem, e.g., in the problem of X-ray tomography one has to reconstruct the internal structure of a human body, by use of external observations. Similarly in the problem of seismic activity, only the observation of top layers of the lithosphere is available, whereas much more is needed to study the seismic activity of Earth.

In this thesis we consider a particular setting of nonparametric inverse problems, related to the infinite-dimensional normal mean model introduced in the previous section. This is an idealized version of many models encountered often in practical problems, e.g., fixed design regression, or density estimation. The problems which we consider are ill-posed, meaning that even if the transformation of the parameter of interest can be inverted, the noise present in the observations destroys the naive estimator significantly. For that reason some form of regularization is required.

In practice this is often achieved by employing the Bayesian paradigm. One possible explanation of the increasing popularity of Bayesian methods is the fact that assigning a prior distribution to an unknown functional parameter is a natural way of specifying a degree of regularization. Probably at least as important is the fact that various computational meth-
1.2. Statistical inverse problems

Methods exist to carry out the inference in practice, including MCMC methods and approximate methods like expectation propagation, Laplace approximations and approximate Bayesian computation. A third important aspect that appeals to users of Bayes methods is that an implementation of a Bayesian procedure typically produces not only an estimate of the unknown quantity of interest (usually a posterior mean or mode), but also a large number of samples from the whole posterior distribution. These can then be used to report a credible set that serves as a quantification of the uncertainty in the estimate.

Some examples of papers using Bayesian methods in nonparametric inverse problems in various applied settings include, e.g., [2, 38, 65, 81, 82]. The paper [94] provides a nice overview and many additional references, for instance several papers discussing the Bayesian approach to nonparametric inverse problems, e.g., [34, 36, 75]. These papers, however, do not provide general answers regarding asymptotic performance of such Bayesian methods.

In the remainder of this section we briefly explain why the existing literature on posterior contraction cannot be used in a wide range of nonparametric inverse problems. Recall the infinite-dimensional normal mean model introduced in the previous section: we observe an infinite-dimensional random vector $Y_n = (X_1, X_2, \ldots)$, where $X_i$’s are independent, $X_i$ has distribution $N(\kappa_i \mu_i, n^{-1})$, for $i \geq 1$, $\mu = (\mu_1, \mu_2, \ldots) \in \ell_2$, and $\kappa = (\kappa_1, \kappa_2, \ldots) \in \ell_2$ is given. The parameter $\mu$ is unknown and the goal is to make inference about $\mu$.

Many nonparametric inverse problems are equivalent to the above formulation: if the $\kappa_i$’s are not equal to 1, the recovery of $\mu$ is an inverse problem. If, additionally, the $\kappa_i$’s are such that $\kappa_i \to 0$ as $i \to \infty$, the problem is ill-posed. The $\kappa_i$’s can decay to zero polynomially, exponentially, or even faster (e.g., in a sub-Gaussian manner). The faster the decay, the more ill-posed the problem.

In the previous section we discussed the posterior contraction result for this model. However, we also noted that as long as the $\kappa_i$’s are not bounded away from zero or infinity, the general theory of posterior contraction yields rates in the “$K-\ell_2$”-norm rather than usual $\ell_2$-norm. These results are not interesting, since they do not deal with the inverse nature of this statistical problem. However, many important inverse problems are ill-posed, and, therefore, nothing can be said about posterior contraction with respect to the natural norm for the parameter of interest based on the existing literature.

To our best knowledge the research contained in this thesis is a first study of posterior contraction and frequentist coverage of Bayesian credible sets in the nonparametric inverse problem setting. Another papers in this area include [1] and [35]. Both papers study the same or similar problem. Agapiou et al. in [1] prove posterior contraction for a smaller range of priors and impose an additional constraint on the true parameter. In [35], Florens and Simoni restrict the covariance structure of the noise. Both papers do not consider Bayesian credible sets and their asymptotic frequentist performance, or linear functionals of the parameter of interest.

Finally, we end this section by giving two examples of nonparametric inverse problems, discussed in more details in Chapter 2 of the present thesis. Both examples are formulated in the generalized infinite-dimensional normal mean model introduced earlier in this chapter. In the first example the $\kappa_i$’s decay polynomially, whereas in the latter example the decay is sub-Gaussian.
**Volterra operator**

Suppose we observe a noisy version of the primitive of a function $\mu$. More formally, it can be written in “signal in white noise” form as follows: observe the process $(Y_t: t \in [0, 1])$ where

$$Y_t = \int_0^t \int_0^s \mu(u) \, du \, ds + \frac{1}{\sqrt{n}} W_t, \quad t \in [0, 1],$$

for a Brownian motion $W$.

Spectral decomposition of the classical Volterra operator yields the following equivalent sequence formulation: we observe the sequence $Y = (Y_1, Y_2, \ldots)$ of noisy coefficients of the primitive of $\mu$ satisfying

$$Y_i = \kappa_i \mu_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,$$

where $(\mu_i)$ are the coefficients of the function $\mu$ in the eigenbasis of $K^T K$, and $K^T$ denotes the adjoint of $K$, $(\kappa_i)$ satisfy $\kappa_i = (i - 1/2)^{-1} \pi^{-1}$, and $Z_1, Z_2, \ldots$ are independent, standard normal random variables. We therefore have an example of the infinite-dimensional normal mean model with the sequence $(\kappa_i)$ tending to zero.

**Heat equation**

Consider the Dirichlet problem for the heat equation

$$\frac{\partial}{\partial t} u(x, t) = \frac{\partial^2}{\partial x^2} u(x, t), \quad u(x, 0) = \mu(x), \quad u(0, t) = u(1, t) = 0,$$

where $u$ is defined on $[0, 1] \times [0, T]$ and the function $\mu \in L^2[0, 1]$ satisfies $\mu(0) = \mu(1) = 0$. One obtains an ordinary differential equation in the Fourier domain, which yields the solution to the above problem given by

$$u(x, t) = \sqrt{2} \sum_{i=1}^{\infty} \mu_i e^{-i^2 \pi^2 t} \sin(i \pi x),$$

where $(\mu_i)$ are the coordinates of $\mu$ in the basis $e_i = \sqrt{2} \sin(i \pi x)$, for $i \geq 1$.

We assume we observe the solution $u(\cdot, T)$ in white noise of intensity $1/n$. By expanding in the basis $(e_i)$ this is equivalent to observing the sequence of noisy, transformed Fourier coefficients $Y = (Y_1, Y_2, \ldots)$ satisfying

$$Y_i = \kappa_i \mu_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,$$

for $(\mu_i)$ as above, $(\kappa_i)$ satisfying $\kappa_i = \exp(-i^2 \pi^2 T)$, and $Z_1, Z_2, \ldots$ independent, standard normal random variables. We therefore have an example of the infinite-dimensional normal mean model with the sequence $(\kappa_i)$ tending to zero.
1.3 Regularity and Bernstein–von Mises theorems

In this section we again discuss an asymptotic aspect of Bayesian procedures. As noted before, the quality of Bayesian procedure might depend on the choice of the prior measure, e.g., the regularity of the prior appears in the posterior contraction rate. Bernstein–von Mises theorems, among other things, indicate that in certain settings, the choice of \( \Pi \) matters little, when the size of the sample is large. In this section we consider a specific type of observation \( Y_n = (X_1, \ldots, X_n) \), where \( X_1, \ldots, X_n \) are independent and identically distributed (i.i.d.) according to an unknown \( P_{\theta_0} \). Therefore, \( P_{\theta_0}^n \) is the \( n \)-fold product of \( P_{\theta_0} \).

We consider regular parametric models, and the regularity of the model \( P \) parametrized by \( \Theta \subset \mathbb{R}^k \) for some \( k \geq 1 \) is defined as follows: we say that \( \theta \mapsto P_{\theta} \) is locally asymptotically normal (LAN) if for every sequence \((h_n) \subset \mathbb{R}^k, h_n \rightarrow h, \text{as } n \rightarrow \infty\)

\[
\prod_{i=1}^{n} \frac{{P_{\theta_0 + n^{-1/2}h_n}}(X_i)}{p_{\theta_0}} = \exp \left( h^T \Delta_{n,\theta_0} - \frac{1}{2} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1) \right),
\]

where

\[
\Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i),
\]

\( \dot{\ell}_{\theta_0} \) is the score function of the model, and \( I_{\theta_0} \) is the Fisher information for \( \theta_0 \). This expansion is of quadratic form, and suggests the limiting shape of the (rescaled and shifted) posterior.

Before we state the parametric version of the Bernstein–von Mises theorem, we note, following historical remarks in [71] and [99], that the Bernstein–von Mises theorem was noted, in the i.i.d. case, by Laplace in 1810. It was further studied by Bernstein in 1917 and von Mises in 1931, and then by Le Cam in 1953.

The following statement is adapted from [99].

**Theorem 1.8.** (Bernstein–von Mises theorem, parametric). Let \( X_1, \ldots, X_n \) be distributed i.i.d. \( P_{\theta_0} \in \mathcal{P} \). The model \( \mathcal{P} \) is parametrized by \( \theta \in \Theta \), open in \( \mathbb{R}^k \) for some \( k \geq 1 \). Suppose that \( \theta \mapsto P_{\theta} \) is locally asymptotically normal at \( \theta_0 \) with nonsingular Fisher information matrix \( I_{\theta_0} \), and suppose that for every \( \varepsilon > 0 \) there exists a sequence of test \( \phi_n \) such that

\[
P_{\theta_0}^n(\phi_n) \rightarrow 0, \quad \sup_{\|\theta - \theta_0\| \geq \varepsilon} P_{\theta}^n(1 - \phi_n) \rightarrow 0.
\]

Furthermore, let the prior measure be absolutely continuous in a neighborhood of \( \theta_0 \) with a continuous positive density at \( \theta_0 \). Then the corresponding posterior distributions satisfy

\[
\sup_A \left| \Pi_n(\sqrt{n}(\theta - \theta_0) \in A|X_1, \ldots, X_n) - N(I_{\theta_0}^{-1}\Delta_{n,\theta_0}, I_{\theta_0}^{-1})(A) \right| \rightarrow 0,
\]

in \( P_{\theta_0}^n \)-probability, where the supremum is taken over all measurable subsets of (localized) \( \Theta \).

We localize the parameter space \( \Theta \) by taking the union of \( \sqrt{n}(\Theta - \theta_0) \) over \( n \geq 1 \). Using the theory of efficiency of estimators, the assertion of the above theorem can be written
in another form, already mentioned in the section on Bayesian procedures. The sequence 
$I_{\theta_0}^{-1}\Delta_{n,\theta_0}$ can be shown to be asymptotically equivalent in probability to any sequence of 
standardized asymptotically efficient estimators $\sqrt{n}(\hat{\theta}_n - \theta_0)$. We may rewrite the assertion 
to arrive at 
$$ \sup_A \left| \Pi_n(\theta \in A | X_1, \ldots, X_n) - N\left(\hat{\theta}_n, n^{-1}I_{\theta_0}^{-1}\right) (A) \right| \to 0, $$
in $P_{\theta_0}$-probability, where $\hat{\theta}_n$ is any asymptotically efficient estimator.

The Bernstein–von Mises theorem is a remarkable result, since the interpretations of ran-
domness in the two situations considered in the theorem are quite different and the two quan-
tities involved in the process are obtained from very different principles: Bayesian and purely 
frequentist.

As already mentioned, the importance of the above theorem (viewed from an application 
point of view) lies in its ability to construct frequentist confidence sets using Bayesian meth-
ods. This is very useful especially in complex problems where the sampling distribution of 
a frequentist estimator might be hard to compute. On the other hand, simulation of a large 
sample from the posterior distribution has become comparatively straightforward, with recent 
developments in MCMC techniques.

The story becomes far more complicated when it comes to nonparametric problems. 
Since the parametric Bernstein–von Mises theorem serves as a link between Bayesian infer-
ence and frequentist efficiency, we should note that neither the frequentist theory on asymp-
totic optimality nor the above theorem generalize fully to nonparametric problems. Several 
negative examples have been studied by Cox [26], and Freedman [37].

Considering semiparametric problems might in principle be a viable intermediate step. 
Recall that like in the parametric case, the semiparametric problems focuses on the estimation 
of a finite-dimensional aspect of the model. Instances of the semiparametric Bernstein–von 
Mises theorem (with several specific examples of nonparametric results) have been studied in 
various statistical models. Kim and Lee studied models from survival analysis. In [54], they 
considered the infinite-dimensional posterior for the cumulative hazard function under right-
censoring for a class of neutral-to-the-right process priors. The posterior for the baseline 
cumulative hazard function and regression coefficients in Cox’s proportional hazard model 
are considered with similar priors. In [53], the author shows that the nonparametric posterior 
for the survival function based on doubly censored data satisfies a nonparametric Bernstein–
von Mises theorem. Several aspects of Bernstein–von Mises phenomenon in Gaussian white 
noise model have been presented by Leahu in [72], and more recently, by Castillo and Nickl 
in [19].

Most general works on the semiparametric Bernstein–von Mises theorem include the fol-
lowing papers. First consider a semiparametric problem in which the model parametrization 
can be decomposed as a pair, with a parameter of interest and a nuisance parameter. A general 
approach has been given by Shen [91], but his conditions are implicit and may prove hard 
to verify in practice. Bickel and Kleijn in [7] consider such a decomposed semiparametric 
problem in which the model is parametrized by a pair. The authors provide a general set 
of conditions, comparable to the standard conditions encountered in the theory of posterior 
contraction and illustrate their result with the study of the partial linear regression model.
1.4. Overview

Similar setting is considered by Castillo [18], where the author considers Gaussian process priors, with the application to the estimation of the center of symmetry of a symmetric function in Gaussian white noise model, a time-discrete functional data analysis and Cox’s proportional hazard model. Another general semiparametric setting is considered by Rivoirard and Rousseau [85], where the authors consider the asymptotic posterior distribution of linear functionals of the density. They present a general result and also an application to a class of exponential families, where they obtain the asymptotic posterior distribution that can be either Gaussian or a mixture of Gaussian distributions with different centering points.

We also note more (model- and/or prior-) specific derivations of the Bernstein–von Mises limit. In the paper [27], De Blasi and Hjort consider semiparametric competing risk models with beta-process priors. De Jonge and van Zanten in [29] study the asymptotic behavior of the marginal posterior for the error standard deviation in a nonparametric, fixed design regression model with Gaussian errors. Semiparametric estimation of the long-memory parameter in a Gaussian time series setting is considered by Kruijer and Rousseau in [62]. Boucheron and Gassiat in [13] study the Bernstein–von Mises result for families of discrete distributions.

Several aspects of limiting posteriors in the Gaussian sequence model have been studied by Johnstone in [51].

Finally we note that there is no reason why a Bernstein–von Mises theorem should be restricted to the $n^{-1/2}$ rate of contraction.

1.4 Overview

This thesis consist of two parts on Bayesian approach to nonparametric inverse problems and Bayesian asymptotics in irregular models, followed by an appendix gathering some technical results used in the proofs in the main text.

1.4.1 Part I

This part of the thesis is focused on several aspects of nonparametric inverse problems. Chapters 2–4 are based on the three papers [59–61].

Chapter 2

In Chapter 2 we first describe a nonparametric inverse problem in the context of the canonical signal-in-white noise model with the operator acting between two Hilbert spaces, and show its equivalence to the infinite-dimensional normal mean model. We briefly recall some essential facts on Gaussian distributions on Hilbert spaces, and next show that a Gaussian prior on the parameter of interest leads to a posterior distribution that is also Gaussian.

The main contribution of this chapter is the study of the asymptotic properties of the posterior in two settings of inverse problems: mildly and extremely ill-posed. The former setting covers, among others, estimation of a derivative of a function, and the latter is presented by a study of the Dirichlet problem for the heat equation. We investigate how different choices of priors affect the rate of contraction and frequentist properties of Bayesian credible balls.
We compare the Bayesian methodology with the existing frequentist approaches to nonparametric inverse problems. We consider a family of prior distributions \((\Pi_{\alpha,\tau}: \alpha > 0, \tau > 0)\) indexed by two fixed, deterministic parameters (and later we let \(\tau\) to depend on the sample size) and show that the rate of contraction depends on the parameters of the prior, characteristics of the inverse problem, and the regularity of the true parameter of interest, assumed to belong to a submodel \(\mathcal{P}_\beta \subset \mathcal{P}\) for some fixed, but unknown regularity parameter \(\beta > 0\). The existing literature on theory of posterior contraction does not cover ill-posed inverse problems, and the study of frequentist coverage of Bayesian credible sets is still underdeveloped, as discussed in Section 1.1.

The results on contraction and credibility are illustrated by simulation examples in both inverse problem settings. In the mildly ill-posed setting, we consider the problem of recovering a function from observation of a noisy version of its primitive. The extremely ill-posed setting is illustrated with the underlying problem of recovery of the initial condition for the heat equation.

Chapter 3

Adaptive Bayesian approaches to the problem of estimating the parameter of interest in mildly ill-posed inverse problems are the topic of Chapter 3. We extend the results on posterior contraction obtained in Chapter 2, by considering two data-driven methods for choosing the regularity of the prior, as introduced in Chapter 2.

Adaptation properties of Bayes procedures for mildly ill-posed nonparametric inverse problems have until now not been studied in the literature. This is again related to the use of general theory of posterior contraction when such a study is performed. We consider a family \((\Pi_{\alpha}: \alpha > 0)\) of Gaussian priors for the parameter of interest. These priors are indexed by a parameter \(\alpha\) quantifying the “regularity” of the prior. In Chapter 2 we considered \(\alpha\) fixed, and in this chapter we select this parameter using the data.

A first approach is fully Bayesian: we endow the parameter \(\alpha\) with a prior distribution itself. Such a hierarchical procedure is typically preferred by Bayesian statisticians. A second approach we study mixes the Bayesian and the frequentist paradigm already on the level of estimation procedure: we first “estimate” \(\alpha\) from the data in a frequentist manner, and then substitute the estimator \(\hat{\alpha}_n\) for \(\alpha\) in the posterior distribution obtained in Chapter 2. This so-called empirical Bayes procedure is not really Bayesian in the strict sense of the word. However, such methods are widely use in practice, yet their theoretical performance is rarely studied. We show that both methods lead to adaptation and rate-optimality (up to lower order factors) over two families of submodels containing the true parameter of interest, and describing its regularity.

With this chapter we contribute to the literature on empirical Bayes procedures. Moreover, and more importantly, we present the first theoretical study of adaptive Bayesian procedures for nonparametric inverse problems.

We illustrate both methods by the simulation example introduced in Chapter 2 in the mildly ill-posed inverse problem setting.
Chapter 4

We also consider a semiparametric aspect of inverse problems: recovery of linear functionals of the parameter of interest. Similar to Chapter 2, we first formally obtain the marginal posterior of a linear functional. We consider not only continuous, but also certain discontinuous functionals, belonging to a wider class of prior-measurable linear functionals.

The contribution of this chapter is similar to the one of Chapter 2: we study posterior contraction that is not covered by the existing literature on the subject, and we investigate the frequentist coverage of Bayesian credible intervals. The regularity of the linear functional plays an important role in the asymptotic behavior of Bayesian procedures. We show that certain continuous linear functionals cancel the inverse nature of the problem, and put the problem in the regular regime. In this chapter we obtain a first Bernstein–von Mises theorem, not only with a typical $n^{-1/2}$ rate, but also with a rate slowed down by a slowly varying factor.

The results of this chapter are illustrated by the same simulation examples as in Chapter 2.

1.4.2 Part II

In this part of the thesis we continue the study of asymptotic properties of the Bayesian approach to statistics, and consider semiparametric posterior limits under local asymptotic exponentiality, an irregular counterpart of local asymptotic normality. An irregular semiparametric version of the Bernstein–von Mises theorem is presented, with the $n^{-1}$ rate of contraction, much faster than the usual $n^{-1/2}$ rate. Chapter 5 is based on the paper [57].

We start Chapter 5 by presenting a simple irregular model, consisting of shifted exponential distributions with scale 1. We then state a parametric version of irregular Bernstein–von Mises theorem, in which the posterior distribution converges to a negative exponential limit, and introduce local asymptotic exponentiality (LAE). This is further adapted to the semiparametric setting in which we decompose the parameter as a pair $(\theta, \eta)$, where the parameter of interest $\theta$ lies in an open subset of the real line, and the nuisance parameter $\eta$ is an element of an infinite-dimensional space. Since we study the marginal posterior for $\theta$ under the frequentist assumption, we assume the existence of the true $\theta_0$ and $\eta_0$. Subsequently, the LAE property is assumed to hold for the nuisance parameters in a neighborhood of the true nuisance $\eta_0$.

We next state the main theorem of the chapter and discuss its conditions. We also present a simplified version of the theorem, if there is no need to control the specific rate of contraction of the marginal posterior for the nuisance parameter. The proof of the main theorem consists of three steps, following the idea of [7]. We first show that taking a sequence of perturbed models $\mathcal{P}_n$ that approximate the nonparametric nuisance model at the true $\theta_0$ we obtain posterior convergence. This result allows us to show that the LAE property is also obtained for the likelihoods, integrated with respect to the nuisance parameter. We then show the main assertion, following ideas of Le Cam and Yang in the parametric regular case. A separate section of the chapter is dedicated to the most demanding condition of the main theorem, namely marginal consistency at $n^{-1}$ rate. Some discussion is provided, followed by a lemma verifying the condition based on a condition on the likelihood ratio.
Two semiparametric models exhibiting the LAE property are presented. Both problems are related to the problem of estimation of the boundary point of a distribution. The first is a generalization of the shifted exponential model. More specifically, we consider the model consisting of probability distributions $P_{\theta,\eta}$ with corresponding probability density functions $p_{\theta,\eta}$, where for $\theta$ in an open subset of $\mathbb{R}$, and $\eta$ a monotone decreasing, continuously differentiable density function supported on the half-line $[0, \infty)$ we have

$$p_{\theta,\eta}: [\theta, \infty) \ni x \mapsto \eta(x - \theta) 1_{[\theta, \infty)}(x) \in [0, \infty).$$

The other one generalizes the uniform distribution on the interval $[0, \theta]$. In this case we take $\eta$ to be a monotone increasing, continuously differentiable density function supported on the interval $[0, 1]$, and define model densities as

$$p_{\theta,\eta}: [0, \theta] \ni x \mapsto \frac{1}{\theta} \eta\left(\frac{x}{\theta}\right) 1_{[0, \theta]}(x) \in [0, \infty).$$

In both settings, based on an i.i.d. sample distributed according to an unknown, but fixed element $P_{\theta_0, \eta_0}$, we obtain exponential limits for the marginal posterior distributions.
Part I

Inverse problems
Chapter 2

Recovery of the full parameter

2.1 Introduction

In this chapter we introduce a Bayesian approach to nonparametric statistical inverse problems and study it in its nonparametric aspect.

We describe the inverse problem in the context of the canonical signal-in-white-noise model, or, equivalently, the infinite-dimensional normal mean model. We consider estimating an unknown parameter of interest $\mu$ from indirect noisy observations $Y$ following the model

$$Y = K\mu + \frac{1}{\sqrt{n}}Z.$$  \hspace{1cm} (2.1)

The unknown parameter $\mu$ is an element of a separable Hilbert space $H_1$, and is mapped into another Hilbert space $H_2$ by a known, injective, linear operator $K: H_1 \to H_2$. The image $K\mu$ is perturbed by unobserved, scaled Gaussian white noise $Z$. There are many special examples of this infinite-dimensional regression model, which can be also viewed as an idealized version of other statistical models, including density estimation. Inverse problems of this type are often ill-posed in the sense that the operator $K$ does not have a well-behaved, continuous inverse. This means that some form of regularization is necessary to solve the inverse problem and to deal with the noise, otherwise even a small error in the observation may produce a large variation in the estimated parameter.

The noise process $Z$ in (2.1) is the standard normal or iso-Gaussian process for the Hilbert space $H_2$. Because this is not realizable as a random element in $H_2$, the model (2.1) is interpreted in process form (as in [11]). The iso-Gaussian process is the zero-mean Gaussian process $Z = (Z_h; h \in H_2)$ with covariance function $\text{E}Z_hZ_{h'} = \langle h, h' \rangle_2$, and the measurement equation (2.1) is interpreted in that we observe a Gaussian process $Y = (Y_h; h \in H_2)$ with mean and covariance functions

$$\text{E}Y_h = \langle K\mu, h \rangle_2, \quad \text{cov}(Y_h, Y_{h'}) = \frac{1}{n} \langle h, h' \rangle_2.$$  \hspace{1cm} (2.2)
Sufficiency considerations show that it is statistically equivalent to observe the subprocess \((Y_h, : i \in \mathbb{N})\), for any orthonormal basis \(h_1, h_2, \ldots\) of \(H_2\). This is further discussed in Section 2.3, and used in the rest of the thesis.

The Bayesian approach to (2.1) consists of putting a prior on the parameter \(\mu\), and computing the posterior distribution. We study Gaussian priors, which are conjugate to the model, so that the posterior distribution is also Gaussian and easy to derive. Our interest is in studying the properties of this posterior distribution, under the frequentist assumption that the data \(Y\) has been generated according to the model (2.1) with a given “true” parameter \(\mu_0\). We investigate whether and at what rate the posterior distributions contract to \(\mu_0\) as \(n \to \infty\) (as in [39]). Note that general theorems on contraction rates for posterior distributions (as in [39] or [40], see also Chapter 1) are not suitable to deal with ill-posed inverse problems considered in this chapter. The reason is that if these general theorems are applied in the inverse case, we only obtain convergence rates relative to the operator norm induced by the operator \(K\), which is not very interesting (this is related to a direct estimation of \(K\mu\)). We also focus our interest on the performance of credible sets for measuring the uncertainty about the parameter.

Work on the fundamental properties of Bayes procedures for nonparametric inverse problems has only started to appear recently. The few papers in this area include [1] and [35], besides the papers forming the basis for this thesis. This is in sharp contrast with the work on frequentist methodology, which is quite well developed. See, for instance, the overviews given by Cavalier [20, 21], and also [11, 20, 30, 88, 94, 105].

The posterior distribution is shown to contract to the true parameter at a rate that depends on the smoothness of the true parameter \(\mu_0\), and the smoothness and scale of the prior. This dependence is also determined by the type of inverse problem. It is rather intuitive, since the operator \(K\), usually assumed to be compact, acts on the parameter \(\mu\). Therefore, the “degree of compactness” of the operator \(K\) affects the smoothness of the parameter \(\mu\) that can be “observed” in the data \(Y\). In this thesis, the “degree of compactness” is measured by the decay of eigenvalues of the operator \(K\). The rates of convergence we encounter are powers of \(1/n\) and \(1/\log n\). Our results show that for a wide choice of the parameters of the prior, the posterior distribution recovers the truth at the optimal rate. In some cases priors suffice to be smooth enough and, therefore, we are able to devise a rate-adaptive method that achieves the optimal rate of recovery. In the other cases, however, priors recovering the truth at the optimal rate require the knowledge of the regularity of the parameter \(\mu\), which is in most situations not available beforehand.

The frequentist coverage of credible sets is also shown to depend on the combination of the prior and true parameter, and also depends on the type of the inverse problem. It can be understood in terms of a bias-variance trade-off, much as the coverage of frequentist nonparametric procedures. A nonparametric procedure that oversmoothes the truth (too big a bandwidth in a frequentist procedure, or a prior that puts too much weight on “smooth” parameters) will be biased, and a confidence or credible region based on such a procedure will be both too concentrated and wrongly located, giving zero coverage. On the other hand, undersmoothing does work (to a certain extent), also in the Bayesian setup, as we show below. In the undersmoothed case credible regions are conservative in general, with coverage tending to 1. The good news is that typically they are of the correct order of magnitude, so that they
do give a reasonable idea of the uncertainty in the estimate. However, we also consider an example in which credible sets, if only correctly located, are an order of magnitude bigger than the frequentist confidence sets. Moreover, we see that the rate-adaptive procedure results in credible sets that are very bad confidence sets, due to the oversmoothing effect of rate-adaptive priors.

Of course, whether a prior under- or oversmoothes depends on the regularity of the true parameter. In practice, we may not want to consider this known, and adapt the prior smoothness to the data. In this chapter we do consider the effect of changing the “length scale” of a prior, but do not study data-dependent length scales. We show that knowing the regularity of the truth we can scale a possibly wrong prior and obtain satisfying coverage and an optimal contraction rate. In Chapter 3, we study another data-dependent approach to selecting a correct regularity of the prior distribution and its influence on the contraction rates.

In the next section we give a more precise statement of the problem, and describe the priors that we consider and derive the corresponding posterior distributions. Section 2.3 gives an equivalent description of the problem (used also in the remainder of the thesis), that enables us to derive main results of this chapter. In Sections 2.4 and 2.5, we introduce two types of inverse problems. We obtain the rate of contraction of the posterior distribution in both settings, in its dependence on parameters of the prior. Furthermore, we study the frequentist coverage of credible regions for $\mu$ in both settings. We also illustrate the results by simulations and pictures. Proofs of the main results of this chapter are placed in Sections 2.6, and 2.7.

**Notation**

Throughout the chapter $\langle \cdot, \cdot \rangle_1$ and $\| \cdot \|_1$, and $\langle \cdot, \cdot \rangle_2$ and $\| \cdot \|_2$ denote the inner products and norms of the Hilbert spaces $H_1$ and $H_2$. The adjoint of an operator $A$ between two Hilbert spaces is denoted by $A^T$.

For $\beta \geq 0$, the Sobolev norm $\|\mu\|_\beta$ and the $\ell_2$-norm $\|\mu\|$ of an element $\mu \in \ell_2$ are defined in a usual way by

$$\|\mu\|_\beta^2 = \sum_{i=1}^{\infty} i^{2\beta} \mu_i^2, \quad \|\mu\|^2 = \sum_{i=1}^{\infty} \mu_i^2,$$

and the corresponding Sobolev space by $S^\beta = \{ \mu \in \ell_2 : \|\mu\|_\beta < \infty \}$.

For two sequences $(a_n)$ and $(b_n)$ of numbers, $a_n \asymp b_n$ means that $|a_n/b_n|$ is bounded away from zero and infinity as $n \to \infty$, $a_n \lesssim b_n$ means that $a_n/b_n$ is bounded, $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$, and $a_n \ll b_n$ means that $a_n/b_n \to 0$ as $n \to \infty$. For two real numbers $a$ and $b$, we denote by $a \vee b$ their maximum, and by $a \wedge b$ their minimum.

### 2.2 Prior and posterior distributions

We assume a mean-zero Gaussian prior for the parameter $\mu$. In the next three paragraphs we recall some essential facts on Gaussian distributions on Hilbert spaces.

A Gaussian distribution $N(\nu, \Lambda)$ on the Borel sets of the Hilbert space $H_1$ is characterized by a mean $\nu$, which can be any element of $H_1$, and a covariance operator $\Lambda : H_1 \to H_1$,
which is a nonnegative-definite, self-adjoint, linear operator of trace class: a compact operator with eigenvalues \((\lambda_i)\) that are summable \(\sum_{i=1}^{\infty} \lambda_i < \infty\) (see, e.g., [92], pages 18–20). A random element \(G\) in \(H_1\) is \(N(\nu, \Lambda)\)-distributed if and only if the stochastic process \((\langle G, h \rangle_1 : h \in H_1)\) is a Gaussian process with mean and covariance functions

\[
E\langle G, h \rangle_1 = \langle \nu, h \rangle_1, \quad \text{cov}(\langle G, h \rangle_1, \langle G, h' \rangle_1) = \langle h, \Lambda h' \rangle_1.
\] (2.3)

The coefficients \(G_i = \langle G, \varphi_i \rangle_1\) of \(G\) relative to an orthonormal eigenbasis \((\varphi_i)\) of \(\Lambda\) are independent, univariate Gaussians with means the coordinates \(\nu_i = \langle \nu, \varphi_i \rangle_1\) of the mean vector \(\nu\) and variances the eigenvalues \(\lambda_i\).

The iso-Gaussian process \(Z\) in (2.1) may be thought of as a \(N(0, I)\)-distributed Gaussian element, for \(I\) the identity operator (on \(H_2\)), but as \(I\) is not of trace class, this distribution is not realizable as a proper random element in \(H_2\). Similarly, the data \(Y\) in (2.1) can be described as having a \(N(K\mu, n^{-1}I)\)-distribution.

For a stochastic process \(W = (W_h : h \in H_2)\) and a continuous, linear operator \(A : H_2 \to H_1\), we define the transformation \(AW\) as the stochastic process with coordinates \(\langle AW, h \rangle_1 = W_{A^\text{op}h}, h \in H_1\). If the process \(W\) arises as \(W_h = \langle W, h \rangle_2\) from a random element \(W\) in the Hilbert space \(H_2\), then this definition is consistent with identifying the random element \(AW\) in \(H_1\) with the process \((\langle AW, h \rangle_1 : h \in H_1)\), as in (2.3) with \(G = AW\). Furthermore, if \(A\) is a Hilbert–Schmidt operator (i.e., \(AA^T\) is of trace class), and \(W = Z\) is the iso-Gaussian process, then the process \(AW\) can be realized as a random variable in \(H_1\) with a \(N(0, AA^T)\)-distribution.

In the Bayesian setup the prior, which we take \(N(0, \Lambda)\), is the marginal distribution of \(\mu\), and the noise \(Z\) in (2.1) is considered independent of \(\mu\). The joint distribution of \((Y, \mu)\) is then also Gaussian, and so is the conditional distribution of \(\mu\) given \(Y\), the posterior distribution of \(\mu\). In general, one must be a bit careful with manipulating possibly “improper” Gaussian distributions (see [75]), but in our situation the posterior is a proper Gaussian conditional distribution on \(H_1\).

**Proposition 2.1.** (Full posterior). If \(\mu\) is \(N(0, \Lambda)\)-distributed and \(Y\) given \(\mu\) is \(N(K\mu, n^{-1}I)\)-distributed, then the conditional distribution of \(\mu\) given \(Y\) is Gaussian \(N(AY, S_n)\) on \(H_1\), where

\[
S_n = \Lambda - A_n(n^{-1}I + K\Lambda K^T)A_n^T,
\] (2.4)

and \(A_n : H_2 \to H_2\) is the continuous linear operator

\[
A_n = \Lambda^{1/2} \left( \frac{1}{n} I + \Lambda^{1/2} K^T K \Lambda^{1/2} \right)^{-1} \Lambda^{1/2} K^T = \Lambda K^T \left( \frac{1}{n} I + K\Lambda K^T \right)^{-1}.
\] (2.5)

The posterior distribution is proper (i.e., \(S_n\) has finite trace) and equivalent (in the sense of absolute continuity) to the prior.

**Proof.** Identity (2.5) is a special case of the identity \((I + BB^T)^{-1}B = B(I + B^T B)^{-1}\), which is valid for any compact, linear operator \(B : H_1 \to H_2\). Indeed, the polar representation gives that \(B = U(B^T B)^{1/2}\) for a partial isometry \(U\), and then \(BB^T = U(B^T B)U^T\). If \((\varphi_i), (\zeta_i)\) are the eigenvectors and eigenvalues of \(B^T B\), then \((U \varphi_i)\) and \((\zeta_i)\) are the eigenvectors...
and eigenvalues of $BB^T$, since $BB^T U \varphi_i = U(B^T B) U^T U \varphi_i = U(B^T B) \varphi_i = U \zeta_i \varphi_i$. We now obtain that $(I + BB^T)^{-1} B \varphi_i = (I + BB^T)^{-1} U \sqrt{\zeta_i} \varphi_i = (1 + \zeta_i)^{-1} \sqrt{\zeta_i} U \varphi_i$ and $B(I + B^T B)^{-1} \varphi_i = B(1 + \zeta_i)^{-1} \varphi_i = U \sqrt{\zeta_i} (1 + \zeta_i)^{-1} \varphi_i$, for any $i$. That $S_n$ is of trace class is a consequence of the fact that it is bounded above by $\Lambda$ (i.e., $\Lambda - S_n$ is nonnegative definite), which is of trace class by assumption.

The operator $\Lambda^{1/2} K^T K \Lambda^{1/2}: H_1 \to H_1$ has trace bounded by $\|K^T K\| \text{tr}(\Lambda)$ and hence is of trace class. It follows that the variable $\Lambda^{1/2} K^T Z$ can be defined as a random element in the Hilbert space $H_1$, and so can $A_n Y$, for $A_n$ given by the first expression in (2.5). The joint distribution of $(Y, \mu)$ is Gaussian with zero mean and covariance operator

$$
\begin{pmatrix}
  n^{-1} I + K \Lambda K^T & K \Lambda \\
  \Lambda K^T & \Lambda
\end{pmatrix}.
$$

Using this with the second form of $A_n$ in (2.5), we can check that the cross covariance operator of the variables $\mu - A_n Y$ and $Y$ (the latter viewed as a Gaussian stochastic process in $\mathbb{R}^{H_2}$) vanishes and, hence, these variables are independent. Thus, the two terms in the decomposition $\mu = (\mu - A_n Y) + A_n Y$ are conditionally independent and degenerate given $Y$, respectively. The distribution of $\mu - A_n Y$ is zero-mean Gaussian with covariance operator $\text{Cov}(\mu - A_n Y) = \text{Cov}(\mu) - \text{Cov}(A_n Y)$, by the independence of $\mu - AY$ and $AY$. This gives the form of the posterior distribution.

The final assertion may be proved by explicitly comparing the Gaussian prior and posterior. Easier is to note that it suffices to show that the model consisting of all $N(K\mu, n^{-1} I)$-distributions is dominated. In that case the posterior can be obtained using Bayes’ rule, which reveals the normalized likelihood as a density relative to the (in fact, any) prior. To prove domination, we may consider equivalently the distributions $\bigotimes_{i=1}^\infty N(\kappa_i \mu_i, n^{-1})$ on $\mathbb{R}^{\infty}$ of the sufficient statistic $(Y_i)$ defined as the coordinates of $Y$ relative to the conjugate spectral basis (see the next subsection). These distributions, for $(\mu_i) \in \ell_2$, are equivalent to the distribution $\bigotimes_{i=1}^\infty N(0, n^{-1})$, as can be seen with the help of Kakutani’s theorem, the affinity being $\exp(-\sum_i \kappa_i^2 \mu_i^2 / 8) > 0$. \hfill \square

In the remainder of this chapter we study the asymptotic behavior of the posterior distribution, under the assumption that $Y = K \mu_0 + n^{-1/2} Z$ for a fixed $\mu_0 \in H_1$. The posterior is characterized by its center $AY$, the posterior mean, and its spread, the posterior covariance operator $S_n$. The first depends on the data, but the second is deterministic. From a frequentist-Bayes perspective both are important: one would like the posterior mean to give a good estimate for $\mu_0$, and the spread to give a good indication of the uncertainty in this estimate.

The posterior mean is a regularization, of the Tikhonov type (called Tikhonov method with a different prior), of the naive estimator $K^{-1} Y$. It can also be characterized as a penalized least squares estimator (see [21, 79, 96]): it minimizes the functional

$$
\mu \mapsto \|Y - K \mu\|^2_2 + \frac{1}{n} \|\Lambda^{-1/2} \mu\|^2_1,
$$

(2.6)

The penalty $\|\Lambda^{-1/2} \mu\|_1$ is interpreted as $\infty$ if $\mu$ is not in the range of $\Lambda^{1/2}$. Because this range is precisely the reproducing kernel Hilbert space (RKHS) of the prior (cf. [102]), with
\[ \|\Lambda^{-1/2}\mu\|_1 \] as the RKHS-norm of \( \mu \), the posterior mean also fits into the general regularization framework using RKHS-norms (see [80]). In the inverse problem literature \( 1/n \) in the above expression is often denoted by \( \gamma \) and called a regularization parameter.

Scaling the prior corresponds to the tuning of the regularization parameter of the Tikhonov method. Indeed, if we take the prior covariance to be \( n \) dependent \( \Lambda_n = \tau_n^2 \Lambda \), for some sequence \( \tau_n \) such that \( n\tau_n^2 \to \infty \), then the functional in (2.6) takes the form

\[ \mu \mapsto \|Y - K\mu\|^2_2 + \frac{1}{n\tau_n^2} \|\Lambda^{-1/2}\mu\|^2_1. \]

and we have \((n\tau_n^2)^{-1} = \gamma \) (see also [21]).

If we impose regularity conditions on the true \( \mu \), the regularization method we use might not be able to recover the parameter optimally. The qualification of a method is the largest regularity of the true parameter for which the bias of the method converges with the optimal rate. It can be shown that the qualification of the Tikhonov method with a different prior is finite (see [21]), which suggest that the Bayesian approach will also inherit this limitation, and it will be impossible to recover the truth optimally, unless the prior is smooth enough.

In any case the posterior mean is a well-studied point estimator in the literature on inverse problems. In this chapter we add a Bayesian interpretation to it, and are (more) concerned with the full posterior distribution.

### 2.3 Singular value decomposition: Sequence formulation

In the next two sections we study the full posterior distribution \( N(AY, S_n) \) in two particular classes of inverse problems: mildly and extremely ill-posed. To facilitate the study of the posterior distribution, we choose an appropriate form of the prior distribution on \( H_1 \) based on the singular value decomposition of the operator \( K \). We also provide an equivalent description of the model (2.1) used throughout the rest of this thesis.

If the operator \( K \) is compact, then the spectral decomposition of the self-adjoint operator \( K^TK: H_1 \to H_1 \) provides a convenient basis. In the compact case the operator \( K^TK \) possesses countably many positive eigenvalues \( \kappa_i^2 \) and there is a corresponding orthonormal basis \( (e_i) \) of \( H_1 \) of eigenfunctions (hence, \( K^TKe_i = \kappa_i^2 e_i \) for \( i \in \mathbb{N} \); see, e.g., [87]). The sequence \( (f_i) \) defined by \( Ke_i = \kappa_i f_i \) forms an orthonormal “conjugate” basis of the range of \( K \) in \( H_2 \). An element \( \mu \in H_1 \) can be identified with its sequence \( (\mu_i) \) of coordinates relative to the eigenbasis \( (e_i) \), and its image \( K\mu = \sum_i \mu_i Ke_i = \sum_i \mu_i \kappa_i f_i \) can be identified with its coordinates \( (\mu_i \kappa_i) \) relative to the conjugate basis \( (f_i) \). Since we can identify the element \( \mu \in H_1 \) with its sequence of coordinates \( (\mu_i) \in \ell_2 \) relative to the (orthonormal) eigenbasis \( (e_i) \), we write \( \mu \in \ell_2 \) interchangeably with \( \mu \in H_1 \). This identification applies also to the corresponding norms, and other elements of \( H_1 \) and \( H_2 \).

The model (2.1) is equivalent to the following sequence model. If we write \( Y_i \) for \( Y_{f_i} \), then (2.2) shows that \( Y_1, Y_2, \ldots \) are independent Gaussian variables with means \( EY_i = \mu_i \kappa_i \) and variance \( 1/n \). Therefore, a concrete equivalent description of the statistical problem is to
recover the sequence \((\mu_i) \in \ell_2\) from independent observations \(Y_1, Y_2, \ldots\) satisfying
\[
Y_i = \kappa_i \mu_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,
\] (2.7)
for \((\kappa_i)\) as above, and \(Z_1, Z_2, \ldots\) independent, standard normal random variables.

In the following we do not require \(K\) to be compact, but we do assume the existence of an orthonormal basis of eigenfunctions of \(K^T K\). The main additional example we then cover is the white noise model, in which \(K\) is the identity operator. The description of the problem remains the same.

The singular value decomposition of the operator \(K\) (so the spectral decomposition of the operator \(K^T K\)) allows us to define the operator \(\Lambda\) for a given sequence \((\lambda_i)\), where \(\lambda_i \to 0\), by \(\Lambda e_i = \lambda_i e_i\) for \(i \geq 1\). In other words, we put product priors \(\Pi\) on \(\ell_2\) given by
\[
\Pi = \bigotimes_{i=1}^{\infty} N(0, \lambda_i),
\] (2.8)
and study the corresponding sequence of posterior distribution. It is straightforward to verify (either by countably many posterior computations in conjugate normal models, or by Proposition 2.1) that in this case the posterior distribution \(N(A_n Y, S_n)\), denoted by \(\Pi_n(\cdot \mid Y)\), is given by
\[
\Pi_n(\cdot \mid Y) = \bigotimes_{i=1}^{\infty} N\left(\frac{n \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Y_i, \frac{\lambda_i}{1 + n \lambda_i \kappa_i^2}\right),
\] (2.9)
and we will often refer to the posterior mean, denoted by \(\hat{\mu} = (\hat{\mu}_i)\), where
\[
\hat{\mu}_i = \frac{n \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Y_i.
\] (2.10)

In the remainder of this part of the thesis we investigate how different choices of the sequences \((\kappa_i)\) and \((\lambda_i)\) together with regularity conditions on the parameter \(\mu\) influence the performance of the posterior distribution.

If \(\kappa_i \to 0\), this problem is ill-posed, and the recovery of \(\mu\) from \(Y\) an inverse problem. The ill-posedness can be quantified by the speed of decay \(\kappa_i \downarrow 0\), and we consider polynomial and sub-Gaussian decay. In general the estimation of \(\mu\) is harder if the decay \(\kappa_i \downarrow 0\) is faster. The difficulty of estimation may be measured by the minimax risks over the scale of Sobolev spaces relative to the orthonormal basis \((e_i)\) of eigenfunctions of \(K^T K\). For \(\beta > 0\) let \(S^\beta = \{\mu \in \ell_2 : \|\mu\|_{\beta} < \infty\}\). The minimax rates of estimation over the unit ball of this space relative to the loss \(\|t - \mu\|\) of an estimate \(t\) for \(\mu\) are known for various types of inverse problems (see, e.g., [20]). These rates are attained by various “regularization” methods, such as generalized Tikhonov and Moore–Penrose regularization, or spectral cut-off \([6, 11, 12, 20, 45, 46, 73, 74]\). The Bayesian approach is closely connected to these methods: in (2.6) the posterior mean is shown to be a regularized estimator.

### 2.4 Mildly ill-posed problems

In this section we consider the mildly ill-posed problem (in the terminology of [20]) and assume that the decay of the eigenvalues \((\kappa_i^2)\) of the operator \(K^T K\) is polynomial. We
investigate the influence of the prior on the performance of the posterior distributions for various true parameters \( \mu_0 \). We study this in the following setting.

**Assumption 2.2.** (Mildly ill-posed problem). The sequences \((\kappa_i)\) and \((\lambda_i)\) in (2.7) and (2.8) satisfy
\[
\lambda_i = \tau_n^2 i^{-1-2\alpha}, \quad C^{-1} i^{-p} \leq \kappa_i \leq C i^{-p}
\]
for some \( \alpha > 0 \), \( p \geq 0 \), \( C \geq 1 \) and \( \tau_n > 0 \) such that \( n \tau_n^2 \to \infty \). Furthermore, the true parameter \( \mu_0 \) belongs to \( S^\beta \) for some \( \beta > 0 \): that is, it satisfies
\[
\sum_{i=1}^{\infty} \mu_{0,i}^2 \iota_{-1-2\alpha} < \infty.
\]
We refer to the parameter \( \beta \) as the “regularity” of the true parameter \( \mu_0 \). In the special case that \( \mu_0 \) is a sequence of coordinates of an element of a function space relative to its Fourier basis, this parameter gives smoothness of \( \mu_0 \) in the classical Sobolev sense. Because the coefficients \((\mu_i)\) of the prior parameter \( \mu \) are normally \( N(0, \lambda_i) \)-distributed, under Assumption 2.2 we have
\[
\mathbb{E} \sum_{i=1}^{\infty} i^{2\alpha'} \mu_i^2 = \tau_n^2 \sum_{i=1}^{\infty} i^{2\alpha'} i^{-1-2\alpha} < \infty \quad \text{if and only if} \quad \alpha' < \alpha.
\]
Thus, \( \alpha \) is “almost” the smoothness of the parameters generated by the prior. This smoothness is modified by the scaling factor \( \tau_n \). Although this leaves the relative sizes of the coefficients \( \mu_i \), and hence the qualitative smoothness of the prior, invariant, we shall see that scaling can completely alter the performance of the Bayesian procedure. Intuitively, rates \( \tau_n \downarrow 0 \) increase, and rates \( \tau_n \uparrow \infty \) decrease the “regularity” of the prior.

### 2.4.1 Main results

We first study the contraction of the posterior distribution \( \Pi_n(\cdot | Y) \). Our first theorem, proved in Section 2.6, shows that it contracts as \( n \to \infty \) to the true parameter at a rate \( \varepsilon_n \) that depends on all four parameters \( \alpha, \beta, \tau_n, p \) of the (Bayesian) inverse problem.

**Theorem 2.3.** (Contraction). If \( \mu_0, (\lambda_i), (\kappa_i) \) and \( (\tau_n) \) are as in Assumption 2.2, then
\[
\sup_{\|\mu - \mu_0\|_\beta \leq R} \mathbb{E}_{\mu_0} \Pi_n(\mu: \|\mu - \mu_0\| \geq M_n \varepsilon_n | Y) \to 0,
\]
for every \( R > 0 \) and \( M_n \to \infty \), where
\[
\varepsilon_n = (n \tau_n^2)^{-\alpha/1+2\alpha+2p} + \tau_n (n \tau_n^2)^{-\beta/1+2\alpha+2p}.
\]
In particular:

(i) If \( \tau_n \equiv 1 \), then \( \varepsilon_n = n^{-\frac{\alpha}{1+2\alpha+2p}} \).

(ii) If \( \beta \leq 1 + 2\alpha + 2p \) and \( \tau_n \gg n^{-\frac{\alpha-\beta}{1+2\beta+2p}} \), then \( \varepsilon_n = n^{-\frac{\beta}{1+2\beta+2p}} \).

(iii) If \( \beta > 1 + 2\alpha + 2p \), then \( \varepsilon_n \gg n^{-\frac{\alpha-\beta}{1+2\beta+2p}} \), for every scaling \( \tau_n \).
2.4. Mildly ill-posed problems

It is known that the minimax rate of convergence over a Sobolev ball in $S^β$ is of the order $n^{-β/(1+2β+2p)}$ (see [20]). By (i) of the theorem the posterior contraction rate is the same if the regularity of the prior is chosen to match the regularity of the truth ($α = β$) and the scale $τ_n$ is fixed. Alternatively, the optimal rate is also attained by appropriately scaling ($τ_n \propto n^{(α−β)/(1+2β+2p)}$, determined by balancing the two terms in $ε_n$) a prior that is regular enough ($β ≤ 1 + 2α + 2p$). In all other cases (no scaling and $α ≠ β$, or any scaling combined with a rough prior $β > 1 + 2α + 2p$), the contraction rate is slower than the minimax rate.

That “correct” specification of the prior gives the optimal rate is presumably comforting to the statistician using Bayesian methodology. Perhaps the main message of the theorem is that even if the prior mismatches the truth, it may be scalable to give the optimal rate. Here, similar as found by [100] in a different setting, a smooth prior can be scaled to make it “rougher” to any degree, but a rough prior can be “smoothed” relatively little (namely, from $α$ to any $β ≤ 1 + 2α + 2p$). As mentioned already, any regularization method can be associated with its qualification: the largest regularity of the true parameter for which the bias converges with the optimal rate, hence the optimal recovery is possible. Here, we see that the qualification of the Bayesian method equals $1 + 2α + 2p$. In Chapter 3 we consider empirical and hierarchical Bayes approaches to the same problem that recover the truth at the nearly optimal rate, regardless the regularity of the truth.

Bayesian inference takes the spread in the posterior distribution as an expression of uncertainty. This practice is not validated by (fast) contraction of the posterior. Instead we consider the frequentist coverage of credible sets. As the posterior distribution is Gaussian, it is natural to center a credible region around the posterior mean. Different shapes of such a set could be considered. The natural counterpart of the preceding theorem is to consider balls. Alternatively, one might consider ellipsoids, depending on geometry of the support of the posterior. This, however, adds complications and leads to similar conclusions.

Because the posterior spread $S_n$ (or the corresponding sequence of variances in (2.9), also referred to as the posterior spread) is deterministic, the radius is the only degree of freedom when we choose a ball, and we fix it by the desired “credibility level” $1 − γ \in (0, 1)$. A credible ball centered at the posterior mean $\hat{μ}$ (see (2.10)) takes the form, where $B(r)$ denotes an $ℓ_2$-ball of radius $r$ around 0,

$$\hat{μ} + B(r_{n,γ}) := \{μ ∈ ℓ_2: ∥μ − \hat{μ}∥ < r_{n,γ}\},$$  \hspace{1cm} (2.14)

where the radius $r_{n,γ}$ is determined so that

$$Π_n(\hat{μ} + B(r_{n,γ}) | Y) = 1 − γ.$$  \hspace{1cm} (2.15)

Because the posterior spread $S_n$ is not dependent on the data, neither is the radius $r_{n,γ}$. The frequentist coverage or confidence of the set (2.14) is

$$P_μ(μ_0 ∈ \hat{μ} + B(r_{n,γ})),$$  \hspace{1cm} (2.16)

where under the probability measure $P_μ$, the variable $Y$ follows (2.1) with $μ = μ_0$. We shall consider the coverage as $n → ∞$ for fixed $μ_0$, uniformly in Sobolev balls, and also along sequences $μ^n_0$ that change with $n$. 
The following theorem, proved in Section 2.7, shows that the relation of the coverage to the credibility level $1 - \gamma$ is mediated by all parameters of the problem. For further insight, the credible region is also compared to the “correct” frequentist confidence ball $\hat{\mu} + B(\tilde{r}_{n, \gamma})$, which has radius $\tilde{r}_{n, \gamma}$ chosen so that the probability in (2.16) with $r_{n, \gamma}$ replaced by $\tilde{r}_{n, \gamma}$ is equal to $1 - \gamma$.

**Theorem 2.4.** (Credibility). Let $\mu_0$, $(\lambda_i)$, $(\kappa_i)$, and $\tau_n$ be as in Assumption 2.2, and set $\tilde{\beta} = \beta \wedge (1 + 2\alpha + 2p)$, and $\tilde{\tau}_n = n^{(\alpha-\beta)/(1+2\tilde{\beta}+2p)}$. The asymptotic coverage of the credible region (2.14) is:

(i) $1$, uniformly in $\mu_0$ with $\|\mu_0\|_\beta \leq 1$, if $\tau_n \gg \tilde{\tau}_n$; in this case $\tilde{r}_{n, \gamma} \approx r_{n, \gamma}$.

(ii) $1$, for every fixed $\mu_0 \in S^\beta$, if $\beta < 1 + 2\alpha + 2p$ and $\tau_n \asymp \tilde{\tau}_n$; $c$, along some $\mu^n_0$ with $\sup_n \|\mu^n_0\|_\beta < \infty$, if $\tau_n \asymp \tilde{\tau}_n$ (any $c \in [0, 1]$).

(iii) $0$, along some $\mu^n_0$ with $\sup_n \|\mu^n_0\|_\beta < \infty$, if $\tau_n \ll \tilde{\tau}_n$.

If $\tau_n \equiv 1$, then the cases (i), (ii) and (iii) arise if $\alpha < \beta$, $\alpha = \beta$ and $\alpha > \beta$, respectively. In case (iii) the sequence $\mu^n_0$ can then be chosen a fixed element $\mu_0$.

The theorem is easiest to interpret in the situation without scaling ($\tau_n \equiv 1$). Then oversmoothing the prior (case (iii): $\alpha > \beta$) has disastrous consequences for the coverage of the credible sets, whereas undersmoothing (case (i): $\alpha < \beta$) leads to conservative confidence sets. Choosing a prior of correct regularity (case (ii): $\alpha = \beta$) gives mixed results.

Inspection of the proofs shows that the lack of coverage in case of oversmoothing arises from a bias in the positioning of the posterior mean combined with a posterior spread that is smaller even than in the optimal case. In other words, the posterior is off mark, but believes it is very right. The message is that (too) smooth priors should be avoided; they lead to overconfident posteriors, which reflect the prior information rather than the data, even if the amount of information in the data increases indefinitely.

Under- and correct smoothing give very conservative confidence regions (coverage equal to 1). However, (i) and (ii) also show that the credible ball has the same order of magnitude as a correct confidence ball $(1 \geq \tilde{r}_{n, \gamma}/r_{n, \gamma} \gg 0)$, so that the spread in the posterior does give the correct order of uncertainty. This at first sight surprising phenomenon is caused by the fact that the posterior distribution concentrates near the boundary of a ball around its mean, and is not spread over the inside of the ball. The coverage is 1, because this sphere is larger than the corresponding sphere of the frequentist distribution of the posterior mean $\hat{\mu}$, even though the two radii are of the same order.

By Theorem 2.3 the optimal contraction rate is obtained (only) by a prior of the correct smoothness. Combining the two theorems leads to the conclusion that priors that slightly undersmooth the truth might be preferable. They attain a nearly optimal rate of contraction and the spread of their posterior gives a reasonable sense of uncertainty.

Scaling of the prior modifies these conclusions. The optimal scaling found in Theorem 2.3 is covered in case (ii). This rescaling leads to a balancing of square bias, variance and spread, and to credible regions of the correct order of magnitude, although the precise (uniform)
coverage can be any number in [0, 1). Alternatively, bigger rescaling rates are covered in case (i) and lead to coverage 1. The optimal or slightly bigger rescaling rate seems the most sensible. In Chapter 3, we consider a data-dependent choice of the prior and study its influence on the posterior contraction. It is an interesting and non-trivial problem to extend this study to the frequentist coverage of credible balls.

2.4.2 Simulation example: Volterra operator

The classical Volterra operator \( K: L^2[0, 1] \to L^2[0, 1] \) and its adjoint \( K^T \) are given by

\[
K \mu(x) = \int_0^x \mu(s) \, ds, \quad K^T \mu(x) = \int_x^1 \mu(s) \, ds.
\]

The resulting problem (2.1) can also be written in “signal in white noise” form as follows: observe the process \((Y_t: t \in [0, 1])\) given by

\[
Y_t = \int_0^t \int_0^s \mu(u) \, du \, ds + \frac{1}{\sqrt{n}} W_t,
\]

for a Brownian motion \( W \).

The eigenvalues, eigenfunctions of \( K^T K \) and conjugate basis are given by (see [49]), for \( i = 1, 2, \ldots \),

\[
\kappa_i^2 = \frac{1}{(i - 1/2)^2 \pi^2}, \quad e_i(x) = \sqrt{2} \cos\left((i - 1/2)\pi x\right),
\]

\[
f_i(x) = \sqrt{2} \sin\left((i - 1/2)\pi x\right).
\]

The \((f_i)\) are the eigenfunctions of \( KK^T \), relative to the same eigenvalues, and \( K e_i = \kappa_i f_i \) and \( K^T f_i = \kappa_i e_i \), for every \( i \in \mathbb{N} \). If, slightly abusing notation, we define \( Y_i = \int_0^1 e_i(t) \, dY_t \), for \( e_i \) as above, then it is easily verified that the observations \( Y_i \) satisfy (2.7).

To illustrate our results with simulated data, we start by choosing a true function \( \mu_0 \), which we expand as \( \mu_0 = \sum_{i=1}^{\infty} \mu_{0,i} e_i \) on the basis \((e_i)\). The simulated data are the noisy and transformed coefficients

\[
Y_i = \kappa_i \mu_{0,i} + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,
\]

where \( Z_1, Z_2, \ldots \) are independent standard normal random variables. The posterior distribution of \( \mu \) is Gaussian, and can be described coordinate-wise (cf. (2.9)) by

\[
\mu_i | Y \sim N \left( \frac{n \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Y_i, \frac{\lambda_i}{1 + n \lambda_i \kappa_i^2} \right).
\]

We obtained posterior credible balls by drawing 100 realizations from the posterior distribution, and plotting 95 realizations with the smallest \( \ell_2 \)-distance from the posterior mean. Overlapping draws from the posterior result in a darker color in the plot.

Figure 2.1 illustrates these balls for \( n = 1000 \). In every one of the 10 panels in the figure the black curve represents the function \( \mu_0 \), defined by the coefficients \( i^{-3/2} \sin(i) \) relative
Figure 2.1: Realizations of the posterior mean (red) and 95 draws from the posterior (gray curves). In all ten panels $n = 10^3$ and $\beta = 1$. Left 5 panels: $\alpha = 1$; right 5 panels: $\alpha = 5$.

Figure 2.2: Realizations of the posterior mean (red) and 95 draws from the posterior (gray curves). In all ten panels $\beta = 1$ and $\alpha = 1/2, 1, 2, 3, 5$ (top to bottom). Left 5 panels: $n = 10^3$; right 5 panels: $n = 10^8$. 
to $e_i$ (thus $\mu_0 \in S^\beta$ for every $\beta < 1$). The 10 panels represent 10 independent realizations of the data, yielding 10 different realizations of the posterior mean (the red curves) and the posterior credible balls. In the left five panels the prior is given by $\lambda_i = i^{-2\alpha-1}$ with $\alpha = 1$, whereas in the right panels the prior corresponds to $\alpha = 5$.

Clearly, the posterior mean is not estimating the true curve very well, even for $n = 1000$. This is mostly caused by the intrinsic difficulty of the inverse problem: better estimation requires bigger sample size. A comparison of the left and right panels shows that the rough prior ($\alpha = 1$) is aware of the difficulty: it produces credible balls that in (almost) all cases contain the true curve. On the other hand, the smooth prior ($\alpha = 5$) is overconfident; the spread of the posterior distribution poorly reflects the imprecision of estimation.

Specifying a prior that is too smooth relative to the true curve yields a posterior distribution which gives both a bad reconstruction and a misguided sense of uncertainty. Our theoretical results show that the inaccurate quantification of estimation error remains even as $n \to \infty$.

The reconstruction, by the posterior mean or any other posterior quantiles, will eventually converge to the true curve. However, specification of a too smooth prior will slow down this convergence significantly. This is illustrated in Figure 2.2. Every one of its 10 panels is similarly constructed as before, but now with $n = 1000$ and $n = 10^8$ for the five panels on the left-hand and right-hand side, respectively, and with $\alpha = 1/2, 1, 2, 3, 5$ for the five panels from top to bottom. We note that the credible ball with $\alpha = 1$ has a difficulty capturing the bump in the true $\mu_0$, while the credible ball with $\alpha = 1/2$ captures the bump, but is also wider than the former one (see the right column in Figure 2.2). For $n = 10^8$ the posterior for this optimal prior has collapsed onto the true curve, whereas the smooth posterior for $\alpha = 5$ still has major difficulty in recovering the bump in the true curve (even though it “thinks” it has captured the correct curve, the bands having collapsed to a single curve in the figure).

2.5 Extremely ill-posed problems: heat equation

Suppose a differential equation describes the evolution of some feature of a system (e.g., heat conduction), depending on its initial value (at time $t = 0$). We observe the feature at time $T > 0$, in the presence of noise or measurement errors, and the aim is to recover the initial condition. In this section we consider the particular example of recovering the initial condition for the heat equation. Specifically, we assume we have noisy observations of the solution $u$ to the Dirichlet problem for the heat equation

$$\frac{\partial}{\partial t} u(x,t) = \frac{\partial^2}{\partial x^2} u(x,t), \quad u(x,0) = \mu(x), \quad u(0,t) = u(1,t) = 0,$$

(2.17)

where $u$ is defined on $[0,1] \times [0,T]$ and the function $\mu \in L^2[0,1]$ satisfies $\mu(0) = \mu(1) = 0$. The solution to (2.17) is given by (cf. [32])

$$u(x,t) = \sqrt{2} \sum_{i=1}^{\infty} \mu_i e^{-i^2 \pi^2 t} \sin(i\pi x),$$
where \((\mu_i)\) are the coordinates of \(\mu\) in the basis \(e_i = \sqrt{2} \sin(i\pi x)\), for \(i \geq 1\). In other words, it holds that \(u(\cdot, T) = K\mu\), for \(K\) the linear operator on \(L^2[0, 1]\) that is diagonalized by the basis \((e_i)\) and that has corresponding eigenvalues \(\kappa_i = \exp(-i^2\pi^2 T)\), for \(i \geq 1\). We assume we observe the solution \(K\mu\) in white noise of intensity \(1/n\). In other words, we observe a sequence of noisy, transformed Fourier coefficients \(Y = (Y_1, Y_2, \ldots)\) satisfying

\[
Y_i = \kappa_i \mu_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,
\]

for \((\mu_i)\) and \((\kappa_i)\) as above, and \(Z_1, Z_2, \ldots\) independent, standard normal random variables. The aim is to recover the coefficients \((\mu_i)\), or equivalently, the initial condition \(\mu = \sum_{i=1}^{\infty} \mu_i e_i\), under the assumption that the signal-to-noise ratio tends to infinity (so \(n \to \infty\)).

The decay of \(\kappa_i\) is faster than exponential, the inverse problem is considered extremely ill-posed. We note that the results of this section can be easily adapted for \(\kappa_i \approx \exp(-pi^2\gamma)\) for some \(p > 0\) and \(\gamma \geq 1\). However, we focus on the specific case of the heat equation problem to highlight practical motivation of this theoretical study of a class of extremely ill-posed inverse problems.

This heat conduction inverse problem has been studied in frequentist literature (see, e.g., [12, 20, 21, 46, 73, 74]) and has also been addressed in Bayesian framework (with additional assumptions on the noise), cf. [94]. For more background on how this backward heat conduction problem arises in practical problems, see, for instance, [3] or [32], and the references therein. Since the \(\kappa_i\) decay in a sub-Gaussian manner, the estimation of \(\mu\) is very hard in general. It is well known for instance that the minimax rate of estimation for \(\mu\) in a Sobolev ball of regularity \(\beta\) relative to the \(\ell_2\)-loss is only \((\log n)^{-\beta/2}\). This rate is attained by various methods, including generalized Tikhonov regularization and spectral cut-off [12, 46, 73, 74].

### 2.5.1 Main results

We first show that the posterior contracts as \(n \to \infty\) to the true parameter at a rate \(\varepsilon_n\) and quantify how this rate depends on the behavior of the sequence \((\lambda_i)\) of prior variances and the regularity \(\beta\) of the true parameter \(\mu_0\). The proof of the following theorem can be found in Section 2.6.

**Theorem 2.5.** (Contraction). If \(\lambda_i = \tau_n^2 i^{-1-2\alpha} \) for some \(\alpha > 0\) and \(\tau_n > 0\) such that \(n\tau_n^2 \to \infty\), then

\[
\sup_{\|\mu_0\|_\beta \leq R} \mathbb{E}_{\mu_0} \Pi_n(\mu : \|\mu - \mu_0\| \geq M_n \varepsilon_n | Y) \to 0, \quad (2.18)
\]

for every \(R > 0\) and \(M_n \to \infty\), where

\[
\varepsilon_n = \left(\log(n\tau_n^2)\right)^{-\frac{\beta}{2}} + \tau_n \left(\log(n\tau_n^2)\right)^{-\frac{\alpha}{2}}. \quad (2.19)
\]

In particular:

(i) If \(\tau_n \equiv 1\), then \(\varepsilon_n = \left(\log n\right)^{-\frac{\beta+\alpha}{2}}\).

(ii) If \(n^{-1/2+\delta} \leq \tau_n \leq \left(\log n\right)^{\frac{\alpha-\beta}{2}}\), for some \(\delta > 0\), then \(\varepsilon_n = \left(\log n\right)^{-\frac{\beta}{2}}\).
2.5. Extremely ill-posed problems: heat equation

\[ \lambda_i = e^{-\alpha i^2} \] for some \( \alpha > 0 \) then (2.18) holds with the rate

\[ \varepsilon_n = \left( \log n \right)^{-\frac{\beta}{2}}. \] (2.20)

We think of the parameters \( \beta \) and \( \alpha \) as the regularity of the true parameter \( \mu_0 \) and the prior, respectively. The first is validated by the fact that in the heat equation case \( (e_i) \) is the (sine) Fourier basis of \( L^2[0,1] \). Therefore, \( \beta \) quantifies the smoothness of \( \mu_0 \) in Sobolev sense. In the case of the polynomial decay of the variances of the prior (later referred to as the polynomial prior), the parameter \( \alpha \) is also closely related to Sobolev regularity, as already discussed after Assumption 2.2 (cf. (2.12)).

Recall that the frequentist minimax rate is of the order \( \left( \log n \right)^{-\beta/2} \). Now consider the case \( \lambda_i = \tau_n^{-2}\iota^{-1-2\alpha} \). By statement (i) of the theorem the posterior contracts at the optimal minimax rate if the regularity of the prior is at least the regularity of the truth \( (\alpha \geq \beta) \) and the scale \( \tau_n \) is fixed. Alternatively, the optimal rate is also attained by appropriately scaling a prior of any regularity. Note that if \( \alpha \geq \beta \) scaling is redundant. The theorem shows that “correct” specification of the prior regularity gives the optimal rate. In contrast to Theorem 2.3, however, the regularity of the prior does not have to match the regularity of the truth exactly. Moreover, even though rough priors still need to be scaled to give the optimal rate, there is no restriction on the “roughness”.

The second assertion of the theorem shows that for very smooth priors (where we take \( \lambda_i = e^{-\alpha i^2} \)) the contraction rate is always optimal. Since the prior does not depend on the unknown regularity \( \beta \), the procedure is rate-adaptive in this case.

Both choices of priors lead to the conclusion that oversmoothing yields the optimal rate, and this has been noted also in the frequentist literature (see [73]). A fully adaptive frequentist method is presented in [12], and in both situations the optimal performance is caused by the dominating bias.

However, in Bayesian inference one often takes the spread in the posterior distribution as a quantification of uncertainty. If \( \lambda_i = e^{-\alpha i^2} \) this spread is much smaller than the minimax rate. To understand the implications, we next consider the frequentist coverage of credible sets. As in the previous section, we again consider balls centered at the posterior mean.

Recall that a credible ball centered at the posterior mean \( \hat{\mu} \) takes the form

\[ \hat{\mu} + B(r_{n,\gamma}) := \{ \mu \in \ell_2 : \| \mu - \hat{\mu} \| < r_{n,\gamma} \}, \] (2.21)

where \( B(r) \) denotes an \( \ell_2 \)-ball of radius \( r \) around 0 and the radius \( r_{n,\gamma} \) is determined such that

\[ \Pi_n \left( \hat{\mu} + B(r_{n,\gamma}) \mid Y \right) = 1 - \gamma. \] (2.22)

As already noted, the spread of the posterior is not dependent on the data, neither is the radius \( r_{n,\gamma} \). The frequentist coverage or confidence of the set (2.21) is, by definition,

\[ P_{\mu_0} \left( \mu_0 \in \hat{\mu} + B(r_{n,\gamma}) \right), \] (2.23)

where under the probability measure \( P_{\mu_0} \) the variable \( Y \) follows (2.7) with \( \mu = \mu_0 \) and \( \kappa_i = \exp(-i^2\pi^2T) \) for \( i \geq 1 \). We shall consider the coverage as \( n \to \infty \) for fixed \( \mu_0 \), uniformly in Sobolev balls, and also along sequences \( \mu_0^n \) that change with \( n \).
The following theorem, proved in Section 2.7, shows that the relation of the coverage to the credibility level $1 - \gamma$ is mediated by the regularity of the true $\mu_0$ and the two parameters controlling the regularity of the prior — $\alpha$ and the scaling $\tau_n$ — for both types of priors. For further insight, the credible region is also compared to the “correct” frequentist confidence ball $\hat{\mu} + B(\hat{r}_{n,\gamma})$ chosen so that the probability in (2.23) is exactly equal to $1 - \gamma$.

**Theorem 2.6.** (Credibility). Suppose the true parameter $\mu_0$ belongs to $\mathbb{S}^\beta$ for $\beta > 0$. Let $\hat{\tau}_n = (\log n)^{(\alpha - \beta)/2}$.

If $\lambda_i = \tau_{n,i}^{-1 - 2\alpha}$ for some $\alpha > 0$ and $\tau_n > 0$ such that $n\tau_n^2 \to \infty$, then asymptotic coverage of the credible region (2.21) is

(i) $1$, uniformly in $\mu_0$ with $\|\mu_0\|_\beta \leq 1$, if $\tau_n \gg \hat{\tau}_n$; in this case $r_{n,\gamma}/\hat{r}_{n,\gamma} \to \infty$.

(ii) $1$, uniformly in $\mu_0$ with $\|\mu_0\|_\beta \leq r$ for $r$ small enough, if $\tau_n \approx \hat{\tau}_n$:

1. for every fixed $\mu_0 \in \mathbb{S}^\beta$, if $\tau_n \approx \hat{\tau}_n$.

(iii) $0$, along some $\mu_0^n$ with $\sup_n \|\mu_0^n\|_\beta < \infty$, if $\tau_n \lesssim \hat{\tau}_n$.

If $\lambda_i = e^{-\alpha i^2}$ for some $\alpha > 0$, then the asymptotic coverage of the credible region (2.21) is

(iv) $0$, for every $\mu_0$ such that $|\mu_{0,i}| \gtrsim e^{-\alpha i^2/2}$ for some $c < \alpha$.

If $\tau_n \equiv 1$, then the cases (i), (ii), and (iii) arise if $\alpha < \beta$, $\alpha = \beta$ and $\alpha \geq \beta$, respectively. If $\alpha > \beta$ in case (iii) the sequence $\mu_0^n$ can then be chosen a fixed element $\mu_0$.

The easiest interpretation of the theorem is in the situation without scaling ($\tau_n \equiv 1$). Then oversmoothing the prior (case (iii): polynomial prior with $\alpha > \beta$, and case (iv): exponential prior) has disastrous consequences for the coverage of the credible sets, whereas undersmoothing (case (i): polynomial prior with $\alpha < \beta$) leads to (very) conservative sets. Choosing a prior of correct regularity (case (ii) and (iii): polynomial prior with $\alpha = \beta$) gives mixed results, depending on the norm of the true $\mu_0$. These conclusions are analogous to the ones that can be drawn from Theorem 2.4 for the mildly ill-posed case.

There is one crucial difference, namely the radius of the conservative sets in case (i) are not of the correct order of magnitude. It means that the radius $\hat{r}_{n,\gamma}$ of the “correct” frequentist confidence ball is of strictly smaller order than the radius of the Bayesian credible ball.

By Theorem 2.5 the optimal contraction rate is obtained by smooth priors. Combining the two theorems leads to the conclusion that polynomial priors that slightly undersmooth the truth might be preferable. They attain a nearly optimal rate of contraction and the spread of their posterior gives a reasonable sense of uncertainty. Slightly undersmoothing is only possible, however, if an assumption about the regularity of the unknown true function is made. It is an important problem to devise methods that achieve this automatically, without knowledge about the true regularity. In Chapter 3, we consider a data-driven choice of the regularity of the prior and study the posterior contraction in the case of mildly-ill posed problems. In the case of the extremely ill-posed problem this is still an open question. Exponential priors, although adaptive and rate-optimal, often lead to very bad credible bands.
2.5. Extremely ill-posed problems: heat equation

2.5.2 Simulation example

To illustrate our results with simulated data we fix a time $T = 0.1$ and a true function $\mu_0$, which we expand as $\mu_0 = \sum_{i=1}^{\infty} \mu_{0,i} e_i$ in the basis $(e_i)$. The simulated data are the noisy and transformed coefficients

$$Y_i = \kappa_i \mu_{0,i} + \frac{1}{\sqrt{n}} Z_i.$$ 

The posterior distribution of $\mu$ is Gaussian, and can be described coordinate-wise (cf. (2.9)) by

$$\mu_i | Y \sim N \left( \frac{n \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Y_i, \frac{\lambda_i}{1 + n \lambda_i \kappa_i^2} \right).$$ 

We obtained posterior credible balls by drawing 100 realizations from the posterior distribution, and plotting 95 realizations with the smallest distance from the posterior mean. Overlapping draws from the posterior result in a darker color in the plot. We considered both types of priors.

Figure 2.3 illustrates these balls for $n = 10^4$ and the polynomial prior. In every of 10 panels in the figure the black curve represents the function $\mu_0$, defined by

$$\mu_0(x) = 4x(x - 1)(8x - 5), \quad \mu_{0,i} = \frac{8\sqrt{2}(13 + 11(-1)^i)}{\pi^{3/2}},$$

(2.24)

where $\mu_{0,i}$ are the coefficients relative to $e_i$, thus $\mu_0 \in S^\beta$ for every $\beta < 2.5$. The 10 panels represent 10 independent realizations of the data, yielding 10 different realizations of the posterior mean (the red curves) and the posterior credible balls. In the left five panels the prior is given by $\lambda_i = i^{-1 - 2\alpha}$ with $\alpha = 1$, whereas in the right panels the prior corresponds to $\alpha = 3$. This is also valid for Figure 2.4, with the exponential prior, so $\lambda_i = e^{-\alpha i^2}$. In the left panels $\alpha = 1$, and in the right panels $\alpha = 5$.

A comparison of the left and right panels in Figure 2.3 shows that the rough polynomial prior ($\alpha = 1$) is aware of the difficulty of inverse problem: it produces wide credible balls that in (almost) all cases contain nearly the whole true curve. Figure 2.3 together with Figure 2.4 show that smooth priors (polynomial with $\alpha = 3$ and both exponential priors) are overconfident: the spread of the posterior distribution poorly reflects the imprecision of estimation. Our theoretical results show that the inaccurate quantification of the estimation error (by the posterior spread) remains even as $n \to \infty$.

The reconstruction, by the posterior mean or any other posterior quantiles, will eventually converge to the true curve. The specification of the prior influences the speed of this convergence. This is illustrated in Figures 2.5 and 2.6. Every of 10 panels in each of the figures is similarly constructed as before, but now with $n = 10^4$ and $n = 10^8$ for the five panels on the left and right side, respectively, and with $\alpha = 1/2, 1, 2, 5, 10$ for the five panels from top to bottom ($\lambda_i = i^{-1 - 2\alpha}$ in Figure 2.5, and $\lambda_i = e^{-\alpha i^2}$ in Figure 2.6). As discussed above, all exponential priors give the optimal rate, but lead to bad credible balls. Also smooth polynomial priors give the optimal rate. This can be seen in Figure 2.5 for $n = 10^8$ and $\alpha = 2$ or 5, where credible balls are very close to the true curve. However, for $\alpha = 5$ it should be noted that the true curve is mostly outside the credible ball.
Figure 2.3: Polynomial prior. Realizations of the posterior mean (red) and 95 draws from the posterior (gray curves). In all ten panels $n = 10^4$. Left 5 panels: $\alpha = 1$; right 5 panels: $\alpha = 3$. True curve (black) given by (2.24).

Figure 2.4: Exponential prior. Realizations of the posterior mean (red) and 95 draws from the posterior (gray curves). In all ten panels $n = 10^4$. Left 5 panels: $\alpha = 1$; right 5 panels: $\alpha = 5$. True curve (black) given by (2.24).
2.5. Extremely ill-posed problems: heat equation

Figure 2.5: Polynomial prior. Realizations of the posterior mean (red) and 95 draws from the posterior (gray curves). Left 5 panels: $n = 10^4$, right 5 panels: $n = 10^5$; $\alpha = 1/2, 1, 2, 5, 10$ (top to bottom). True curve (black) given by (2.24).

Figure 2.6: Exponential prior. Realizations of the posterior mean (red) and 95 draws from the posterior (gray curves) Left 5 panels: $n = 10^4$, right 5 panels: $n = 10^5$; $\alpha = 1/2, 1, 2, 5, 10$ (top to bottom). True curve (black) given by (2.24).
2.6 Proofs of Theorems 2.3 and 2.5

In this section we prove contraction results in both inverse problem settings presented in this chapter. Technical details distinguishing both cases are presented in separate subsections.

Let the posterior distribution in (2.9) be denoted by $\bigotimes_{i=1}^\infty N\left(\sqrt{m_{i,n}Y_i}, s_{i,n}\right)$, where

$$s_{i,n} = \frac{\lambda_i}{1 + n\lambda_i \kappa_i^2}, \quad t_{i,n} = \frac{n\lambda_i^2 \kappa_i^2}{(1 + n\lambda_i \kappa_i^2)^2}.$$  

Because the posterior is Gaussian, it follows that

$$\int \|\mu - \mu_0\|^2 \, d\Pi_n(\mu \mid Y) = \|\hat{\mu} - \mu_0\|^2 + \sum_{i=1}^\infty s_{i,n},$$  

(2.25)

where $Y$ follows (2.7) with $\mu = \mu_0$, and

$$\mu = \left(\frac{n\lambda_i \kappa_i}{1 + n\lambda_i \kappa_i^2} Y_i\right)_i = \left(\frac{n\lambda_i \kappa_i^2 \mu_0, i + \sqrt{n\lambda_i \kappa_i Z_i}}{1 + n\lambda_i \kappa_i^2}\right)_i =: E_{\mu_0} \hat{\mu} + (\sqrt{t_{i,n} Z_i})_i.$$  

By Markov’s inequality the left side of (2.25) is an upper bound to $M_n^2 \varepsilon_n \Pi_n(\mu: \|\mu - \mu_0\| \geq M_n \varepsilon_n \mid Y)$. Therefore, it suffices to show that the expectation under $\mu_0$ of the right side of the display is bounded by a multiple of $\varepsilon_n^2$. The expectation of the first term is the mean square error of the posterior mean $\hat{\mu}$, and can be written as the sum $\|E_{\mu_0} \hat{\mu} - \mu_0\|^2 + \sum_{i=1}^\infty t_{i,n}$ of its square bias and “variance”. The second term $\sum_{i=1}^\infty s_{i,n}$ is deterministic.

2.6.1 Details of Theorem 2.3

Under Assumption 2.2 the three quantities are given by

$$\|E_{\mu_0} \hat{\mu} - \mu_0\|^2 = \sum_{i=1}^\infty \frac{\mu_{0,i}^2}{(1 + n\lambda_i \kappa_i^2)^2} \times \sum_{i=1}^\infty \frac{\mu_{0,i}^2}{(1 + n\tau_i^2 N^{-1-2\alpha-2p})^2},$$  

(2.26)

$$\sum_{i=1}^\infty t_{i,n} = \sum_{i=1}^\infty \frac{n\lambda_i^2 \kappa_i^2}{(1 + n\lambda_i \kappa_i^2)^2} \times \sum_{i=1}^\infty \frac{n\tau_i^4 N^{-1-2\alpha-2p}}{(1 + n\tau_i^2 N^{-1-2\alpha-2p})^2},$$  

(2.27)

$$\sum_{i=1}^\infty s_{i,n} = \sum_{i=1}^\infty \frac{\lambda_i}{1 + n\lambda_i \kappa_i^2} \times \sum_{i=1}^\infty \frac{\tau_i^2 N^{-1-2\alpha-2p}}{1 + n\tau_i^2 N^{-1-2\alpha-2p}}.$$  

(2.28)

By Lemma A.1 (applied with $q = \beta$, $t = 0$, $u = 1 + 2\alpha + 2p$, $v = 2$ and $N = n\tau_i^2$), the first can be bounded by $\|\mu_0\|^2 / (n\tau_n^2 (1 + 2\alpha + 2p)^2)$, which accounts for the first term in the definition of $\varepsilon_n$. By Lemma A.3 (applied with $S(i) = 1$, $q = -1/2$, $t = 2 + 4\alpha + 2p$, $u = 1 + 2\alpha + 2p$, $v = 2$, and $N = n\tau_n^2$), and again Lemma A.3 (applied with $S(i) = 1$, $q = -1/2$, $t = 1 + 2\alpha$, $u = 1 + 2\alpha + 2p$, $v = 1$ and $N = n\tau_n^2$), both the second and third expressions are of the order the square of the second term in the definition of $\varepsilon_n$.

The consequences (i) and (ii) follow by verification after substitution of $\tau_n$ as given. To prove consequence (iii), we note that the two terms in the definition of $\varepsilon_n$ are decreasing.
and increasing in \( \tau_n \), respectively. Therefore, the maximum of these two \textit{terms} is minimized with respect to \( \tau_n \) by equating the two \textit{terms}. This minimum (assumed at \( \tau_n = n^{-(1+\alpha+2p)/(3+4\alpha+6p)} \)) is much bigger than \( n^{-\beta/(1+2\beta+2p)} \) if \( \beta > 1 + 2\alpha + 2p \).

### 2.7. Proofs of Theorems 2.4 and 2.6

In this section we prove results on the frequentist coverage of credible balls in both inverse problem settings presented in this chapter. Technical details distinguishing both cases are presented in separate subsections.

Recall that the radius \( r_{n,\gamma} \) of a credible ball is determined so that

\[
\Pi_n\left( \hat{\mu} + B(r_{n,\gamma})|Y \right) = 1 - \gamma.
\]

Because the posterior distribution is \( \bigotimes_{i=1}^{\infty} N(\sqrt{m_i}Y_i, s_{i,n}) \), by (2.9), the radius \( r_{n,\gamma} \) satisfies \( P(U_n < r_{n,\gamma}^2) = 1 - \gamma \), for \( U_n \) a random variable distributed as the square norm of an
The variables for \( W \) work under Assumption 2.2. Note that 2.7.1 Details of Theorem 2.4 can be written as

\[
P(\mu_0 \in \hat{\mu} + B(r_{n,\gamma}))
\]

can be written as

\[
P\left( \|W_n + E_\mu \hat{\mu} - \mu_0\| \leq r_{n,\gamma} \right),
\]

for \( W_n \) possessing a \( \otimes_{i=1}^\infty N(0, t_{i,n}) \)-distribution. For ease of notation let \( V_n = \|W_n\|^2 \).

The variables \( U_n \) and \( V_n \) can be represented as

\[
U_n = \sum_{i=1}^\infty s_{i,n} Z_i^2 \quad \text{and} \quad V_n = \sum_{i=1}^\infty t_{i,n} Z_i^2,
\]

for \( Z_1, Z_2, \ldots \) independent standard normal variables.

2.7.1 Details of Theorem 2.4

We work under Assumption 2.2. Note that

\[
s_{i,n} - t_{i,n} = \frac{\lambda_i}{(1 + n \lambda_i k_i^2)^2} \sim \frac{r_{n,\gamma}^2 t_{i,n} - 2 \alpha - 1}{(1 + n \tau_n^2 t_{i,n} - 2 \alpha - 2 p - 1)^2}
\]

Therefore, by Lemma A.3 (applied with \( S \equiv 1 \) and \( q = -1/2 \); always the first case)

\[
EU_n = \sum_{i=1}^\infty s_{i,n} \tau_n^2 (n \tau_n^2)^{-\frac{2 \alpha}{1+2 \alpha + 2 p}},
\]

\[
EV_n = \sum_{i=1}^\infty t_{i,n} \tau_n^2 (n \tau_n^2)^{-\frac{2 \alpha}{1+2 \alpha + 2 p}},
\]

\[
E(U_n - V_n) = \sum_{i=1}^\infty (s_{i,n} - t_{i,n}) \tau_n^2 (n \tau_n^2)^{-\frac{2 \alpha}{1+2 \alpha + 2 p}},
\]

\[
\text{var } U_n = 2 \sum_{i=1}^\infty s_{i,n}^2 \tau_n^4 (n \tau_n^2)^{-\frac{1+4 \alpha}{1+2 \alpha + 2 p}},
\]

\[
\text{var } V_n = 2 \sum_{i=1}^\infty t_{i,n}^2 \tau_n^4 (n \tau_n^2)^{-\frac{1+4 \alpha}{1+2 \alpha + 2 p}}.
\]

We conclude that the standard deviations of \( U_n \) and \( V_n \) are negligible relative to their means, and also relative to the difference \( E(U_n - V_n) \) of their means. Because \( U_n \geq V_n \), we conclude that the distributions of \( U_n \) and \( V_n \) are asymptotically completely separated: \( P(V_n \leq v_n \leq U_n) \to 1 \) for some \( v_n \) (e.g., \( v_n = E(U_n + V_n)/2 \)). The numbers \( r_{n,\gamma}^2 \) are \( 1 - \gamma \)-quantiles of \( U_n \), and, hence, \( P(V_n \leq r_{n,\gamma}^2 (1 + o(1))) \to 1 \). Furthermore, it follows that

\[
r_{n,\gamma}^2 \sim r_n^2 (n \tau_n^2)^{-\frac{2 \alpha}{1+2 \alpha + 2 p}} \asymp EU_n \asymp EV_n.
\]

The square norm of the bias \( E_{\mu_0} \hat{\mu} - \mu_0 \) is given in (2.26), where it was noted that

\[
B_n := \sup_{\|\mu_0\|_2 \leq 1} \|E_{\mu_0} \hat{\mu} - \mu_0\| \asymp (n \tau_n^2)^{-\frac{\theta}{1+2 \alpha + 2 p}} \wedge 1.
\]

The bias \( B_n \) is decreasing in \( \tau_n \), whereas \( EU_n \) and \( \text{var } U_n \) are increasing. The scaling rate \( \tilde{\tau}_n \sim n^{(\alpha - \beta)/(1+2 \beta + 2 p)} \) balances the square bias \( B_n^2 \) with the variance \( EV_n \) of the posterior mean, and hence with \( r_{n,\gamma}^2 \).
Case (i). In this case, $B_n \ll r_{n,\gamma}$. Hence $P\left(\|W_n + E_{\mu_0}\hat{\mu} - \mu_0\| \leq r_{n,\gamma} - B_n\right) = P\left(V_n \leq r_{n,\gamma}^2(1 + o(1))\right) \to 1$, uniformly in the set of $\mu_0$ in the supremum defining $B_n$. Note that $\hat{r}_{n,\gamma}$ is such that the coverage in (2.32) is exactly $1 - \gamma$. Since $\|W_n\|^2 = V_n$, we have that $\hat{r}^2_{n,\gamma}$ is of the order $|B_n|^2 + r_{n,\gamma}^2\sigma^2(n) = |B_n|^2 - 2\alpha/\gamma(1 + 2\alpha + 2\beta)$, so of the order of $r^2_{n,\gamma}$.

Case (ii). In this case, $B_n \gg r_{n,\gamma}$. Hence, $P\left(\|W_n + E_{\mu_0}\hat{\mu} - \mu_0\| \leq r_{n,\gamma} - B_n\right) \to 0$ for any sequence $\mu_0^n$ (nearly) attaining the supremum in the definition of $B_n$. If $r_n \equiv 1$, then $B_n$ and $r_{n,\gamma}$ are both powers of $1/n$ and, hence, $B_n \gg r_{n,\gamma}$ implies that $B_n \gg r_{n,\gamma}n^{\delta}$, for some $\delta > 0$. The preceding argument then applies for a fixed $\mu_0$ of the form $\mu_0 = i^{-1/2-\beta-\varepsilon}$, for small $\varepsilon > 0$, that gives a bias that is much closer than $n^{\delta}$ to $B_n$.

Case (iii). In this case, $B_n \approx r_{n,\gamma}$. Hence, $P\left(\|W_n + E_{\mu_0}\hat{\mu} - \mu_0\| \leq r_{n,\gamma} - B_n\right) \to 0$ for any sequence $\mu_0^n$ (nearly) attaining the supremum in the definition of $B_n$. If $r_n \equiv 1$, then $B_n$ and $r_{n,\gamma}$ are both powers of $1/n$ and, hence, $B_n \gg r_{n,\gamma}$ implies that $B_n \gg r_{n,\gamma}n^{\delta}$, for some $\delta > 0$. The preceding argument then applies for a fixed $\mu_0$ of the form $\mu_0 = i^{-1/2-\beta-\varepsilon}$, for small $\varepsilon > 0$, that gives a bias that is much closer than $n^{\delta}$ to $B_n$.

Finally, we prove the existence of a sequence $\mu_0^n$ along which the coverage is a given $c \in [0,1)$. The coverage (2.32) with $\mu_0$ replaced by $\mu_0^n$ tends to $c$ if, for $b_n = E_{\mu_0}\hat{\mu} - \mu_0^n$ and $z_c$ a standard normal quantile,

$$
\frac{\|W_n + b_n\|^2 - E\|W_n + b_n\|^2}{sd\|W_n + b_n\|^2} \sim N(0,1),
$$

(2.33)

$$
\frac{\|W_n + b_n\|^2 - E\|W_n + b_n\|^2}{sd\|W_n + b_n\|^2} \rightarrow z_c.
$$

(2.34)

Because $W_n$ is mean-zero Gaussian, we have $E\|W_n + b_n\|^2 = E\|W_n\|^2 + E\|b_n\|^2$ and $\text{var}\|W_n + b_n\|^2 = \text{var}\|W_n\|^2 + 4\sum_{i=1}^{\infty} W_n,i b_{n,i}$. Here $\|W_n\|^2 = V_n$ and the distribution of $\sum_{i=1}^{\infty} W_n,i b_{n,i}$ is zero-mean Gaussian with variance $\sum_{i=1}^{\infty} t_{i,n} b_{n,i}^2$. Therefore, display (2.34) can be rewritten as

$$
\frac{\|W_n + b_n\|^2 - EV_n - \sum_{i=1}^{\infty} b_{n,i}^2}{\sqrt{\text{var} V_n + 4\sum_{i=1}^{\infty} t_{i,n} b_{n,i}^2}} \rightarrow z_c.
$$

(2.35)

We choose $(b_{n,i})$ differently in the cases that $\beta \leq 1 + 2\alpha + 2\beta$ and $\beta \geq 1 + 2\alpha + 2\beta$, respectively. In both cases the sequence has exactly one nonzero coordinate. We denote this coordinate by $b_{n,i,n}$, and set, for numbers $d_n$ to be determined,

$$
b_{n,i,n}^2 = \|W_n\|^2 - EV_n - d_n \text{sd } V_n.
$$

Because $r_{n,\gamma}^2$, $EV_n$ and $r_{n,\gamma}^2 - EV_n$ are of the same order of magnitude, and $\text{sd } V_n$ is of strictly smaller order, for bounded or slowly diverging $d_n$ the right-hand side of the preceding display is equivalent to $(r_{n,\gamma}^2 - EV_n)(1 + o(1))$. Consequently, the left-hand side of (2.35) is equivalent to

$$
d_n \text{sd } V_n
$$

$$
\sqrt{\text{var } V_n + 4t_{i,n}\sigma^2(n) r_{n,\gamma}^2 - EV_n}(1 + o(1))
$$

The remainder of the argument is different in the two cases.

Case $\beta \leq 1 + 2\alpha + 2\beta$. We choose $t_{i,n} \sim (n\sigma^2(n))^{1/(1+2\alpha+2\beta)}$. It can be verified that $t_{i,n} (r_{n,\gamma}^2 - EV_n)/ \text{var } V_n \approx 1$. Therefore, for $c \in [0,1]$, there exists a bounded or slowly diverging sequence $d_n$ such that the preceding display tends to $z_c$. [148]
The bias \( b_n \) results from a parameter \( \mu_0^n \) such that \( b_{n,i} = (1 + n\lambda_i\kappa_i^2)^{-1}(\mu_0^n)_i \), for every \( i \). Thus, \( \mu_0^n \) also has exactly one nonzero coordinate, and this is proportional to the corresponding coordinate of \( b_n \), by the definition of \( i_n \). It follows that

\[
i_n^{2\beta} (\mu_0^n)_{i_n}^2 \asymp i_n^{2\beta} b_{n,i_n}^2 \lesssim i_n^{2\beta} (r_{n,\gamma}^2 - \text{EV}_n) \asymp 1
\]

by the definition of \( \tau_n \). It follows that \( ||\mu_0^n||_{\beta} \lesssim 1 \).

Case \( \beta \geq 1 + 2\alpha + 2p \). We choose \( i_n = 1 \). In this case \( \tau_n \to 0 \) and it can be verified that \( t_{i_n,n}(r_{n,\gamma}^2 - \text{EV}_n)/\text{var} V_n \to 0 \). Also,

\[
(\mu_0^n)_{i_n}^2 \asymp (1 + n\tau_n^2 b_{n,1}^2) \lesssim (1 + n\tau_n^2)^2 \text{EV}_n.
\]

This is \( O(1) \), because \( \tau_n \) is chosen so that \( \text{EV}_n \) is of the same order as the square bias \( B_n^2 \), which is \( (n\tau_n^2)^{-2} \) in this case.

It remains to prove the asymptotic normality (2.33). We can write

\[
||W_n + b_n||^2 - \text{E}||W_n + b_n||^2 = \sum_{i=1}^{\infty} t_{i,n}(Z_i^2 - 2) + 2b_{n,i_n} \sqrt{t_{i,n}} Z_{i_n}
= \sum_{i \neq i_n} t_{i,n}(Z_i^2 - 1) + (2b_{n,i_n} \sqrt{t_{i,n}} Z_{i_n} + t_{i,n}(Z_i^2 - 1)).
\]  

(2.36)

The two terms above are independent and mean-zero. Consider the latter term, and note that since \( Z_i \) are standard normal we have

\[
\frac{2b_{n,i_n} \sqrt{t_{i,n}} Z_{i_n} + t_{i,n}(Z_i^2 - 1)}{\left(2t_{i,n}^2 + 4b_{n,i_n} b_{n,i_n}^2\right)^{1/2}}
\approx \frac{1 + t_{i,n}^2}{2b_{n,i_n}^2} \frac{1}{Z} + \frac{\sqrt{t_{i,n}}}{2b_{n,i_n}} \frac{1}{Z}^{-1/2} (Z^2 - 1).
\]

It can be seen that \( t_{i,n} = o(b_{n,i_n}^2) \). Therefore, the above sequence tends in distribution to the standard normal distribution.

The limit distribution of the remaining term in (2.36) can be established by a slight adaptation of the Lindeberg-Feller theorem (to infinite sums). We want to show that for every \( \varepsilon > 0 \)

\[
\lim_{n \to \infty} \frac{1}{\text{var}_n^2} \sum_{i \neq i_n} \text{E} \left( t_{i,n}^2 (Z_i^2 - 1)^2 1 \{|t_{i,n}(Z_i^2 - 1)| > \varepsilon \text{ var}_n\} \right) = 0,
\]

where

\[
\text{var}_n^2 = \text{var} \left( \sum_{i \neq i_n} t_{i,n}(Z_i^2 - 1) \right) = 2 \sum_{i \neq i_n} t_{i,n}^2.
\]

Let \( t_n^* = \sup_i t_{i,n} \). We have

\[
\frac{t_{i,n}^2}{\text{var}_n^2} \text{E} \left( (Z_i^2 - 1)^2 1 \{|t_{i,n}(Z_i^2 - 1)| > \varepsilon \text{ var}_n\} \right)
\leq \frac{(t_n^*)^2}{\text{var}_n^2} \text{E} \left( (Z_i^2 - 1)^2 1 \{|t_{i,n}(Z_i^2 - 1)| > \varepsilon \text{ var}_n\} \right).
\]  

(2.37)
Additionally
\[
E((Z_i^2 - 1)^2 \mathbb{1}\{|t_{i,n}(Z_i^2 - 1)| > \varepsilon \var_n\}) \leq E((Z_i^2 - 1)^2 \mathbb{1}\{|(Z_i^2 - 1)| > \varepsilon \var_n/t_n^*\}).
\]

(2.38)

With some effort it can be seen that \(t_n^*/\var_n \to 0\). Note that for \(Z \sim N(0,1)\) and every \(\delta_n \to \infty\)
\[
\lim_{n \to \infty} E((Z^2 - 1)^2 \mathbb{1}\{|(Z^2 - 1)| > \delta_n\}) = 0,
\]
where the expectation in the above limit depends only on the distribution of \(Z\). By the dominated convergence theorem, (2.37) and (2.38) together with the above observation verify the Lindeberg condition, since \(t_n^*/\var_n \to 0\). We conclude that the infinite sum in (2.36) without the \(i\)th term divided by its standard deviation tends in distribution to the standard normal distribution. Thus, the two terms converge jointly to asymptotically independent standard normal variables, if scaled separately by their standard deviations. Then their scaled sum is also asymptotically standard normally distributed.

2.7.2 Details of Theorem 2.6

Recall that \(U_n = \sum_{i=1}^{\infty} s_{i,n} Z_i^2\) and \(V_n = \sum_{i=1}^{\infty} t_{i,n} Z_i^2\), for \(Z_1, Z_2, \ldots\) independent standard normal variables, and \(\kappa_i = \exp(-i^2 \pi^2 T)\) for \(i \geq 1\). Let \(\lambda_i = \tau_{n,i}^2 i^{-1-2\alpha}\). By Lemma A.5 (cf. Subsection 2.6.2)
\[
EU_n = \sum_{i=1}^{\infty} s_{i,n} \asymp \tau_n^2 (\log n \tau_n^2)^{-\alpha},
\]
\[
EV_n = \sum_{i=1}^{\infty} t_{i,n} \asymp \tau_n^2 (\log n \tau_n^2)^{-\frac{1}{2}-\alpha},
\]
\[
\var U_n = 2 \sum_{i=1}^{\infty} s_{i,n}^2 \asymp \tau_n^4 (\log n \tau_n^2)^{-\frac{1}{2}-2\alpha},
\]
\[
\var V_n = 2 \sum_{i=1}^{\infty} t_{i,n}^2 \asymp \tau_n^4 (\log n \tau_n^2)^{-1-2\alpha}.
\]

It follows that
\[
r_{n,\gamma}^2 \asymp \tau_n^2 (\log n \tau_n^2)^{-\alpha} \asymp EU_n \gg EV_n \asymp \text{sd} V_n,
\]
and, therefore,
\[
P(V_n \leq \delta r_{n,\gamma}^2) = P\left(\frac{V_n - EV_n}{\text{sd} V_n} \leq \frac{\delta r_{n,\gamma}^2 - EV_n}{\text{sd} V_n}\right) \to 1,
\]
(2.39)

for every \(\delta > 0\). The square norm of the bias \(E_{\mu_0} \hat{\mu} - \mu_0\) is given in (2.29), where it was noted that
\[
B_n := \sup_{||\mu_0||_\beta \leq 1} ||E_{\mu_0} \hat{\mu} - \mu_0|| \asymp (\log n \tau_n^2)^{-\frac{\beta}{2}}.
\]
The bias $B_n$ is decreasing in $r_n$, whereas $EU_n$ is increasing. The optimal scaling rate $\tilde{r}_n \asymp (\log n)^{(\alpha - \beta)/2}$ balances the square bias $B_n^2$ with the posterior spread $EU_n$, and hence with $r_{n,\gamma}^2$.

Case (i). In this case, $B_n \ll r_{n,\gamma}$. Hence $P(\|W_n + E_{\mu_0} \hat{\mu} - \mu_0\| \leq r_{n,\gamma} - B_n) = P(V_n \leq r_{n,\gamma}^2(1 + o(1))) \to 1$, uniformly in the set of $\mu_0$ in the supremum defining $B_n$. Note that $\tilde{r}_{n,\gamma}$ is such that the coverage in (2.32) is exactly $1 - \gamma$. Since $\|W_n\|^2 = V_n$, we have that $\tilde{r}_{n,\gamma}^2$ is of the order $B_n^2 + \tau_n^2(\log(n\tau_n^2))^{-1/2 - \alpha}$, so of strictly smaller order than $r_{n,\gamma}^2$, and, therefore, $r_{n,\gamma}/\tilde{r}_{n,\gamma} \to \infty$.

Case (ii). In this case, $B_n \asymp r_{n,\gamma}$. By the second assertion of Lemma A.5 the bias $\|E_{\mu_0} \hat{\mu} - \mu_0\|$ at a fixed $\mu_0$ is of strictly smaller order than the supremum $B_n$. The argument of (i) shows that the asymptotic coverage then tends to 1. The maximal bias $B_n(r)$ over $\|\mu_0\|_\beta \leq r$ is of the order $r_{n,\gamma}$ and proportional to the radius $r$. Thus for small enough $r$ we have that $r_{n,\gamma} - B_n(r) \gtrsim r_{n,\gamma} \to \infty$. Then $P(\|W_n + E_{\mu_0} \hat{\mu} - \mu_0\| \leq r_{n,\gamma}) \geq P(\|W_n\| \leq r_{n,\gamma} - B_n(r)) \geq P(V_n \leq r_{n,\gamma}^2) \to 1$.

Case (iii). In this case, $B_n \gtrsim r_{n,\gamma}$. Hence any sequence $\mu_0^n$ that (nearly) attains the maximal bias over a sufficiently large ball $\|\mu_0\|_\beta \leq r$ such that $B_n(r) - r_{n,\gamma} \gtrsim r_{n,\gamma}$ satisfies $P(\|W_n + E_{\mu_0^n} \hat{\mu} - \mu_0^n\| \leq r_{n,\gamma}) \leq P(\|W_n\| \geq B_n(r) - r_{n,\gamma}) \leq P(V_n \gtrsim r_{n,\gamma}^2) \to 0$.

If $r_{n,\gamma} \equiv 1$, then $B_n$ and $r_{n,\gamma}$ are both powers of $1/\log n$ and hence $B_n \gg r_{n,\gamma}$ implies that $B_n \gtrsim r_{n,\gamma}(\log n)^\delta$, for some $\delta > 0$. The preceding argument then applies for a fixed $\mu_0$ of the form $\mu_{0,i} \asymp i^{-1/2 - \beta - \epsilon}$, for small $\epsilon > 0$, that gives a bias that is much closer than $(\log n)^\delta$ to $B_n$.

Case (iv). In the proof of Theorem 2.5, we obtained $EU_n \asymp EV_n \asymp n^{-\alpha/(\alpha + 2\pi^2 T)}$. It can be shown that $sd U_n \asymp n^{-\alpha/(\alpha + 2\pi^2 T)}$, so also $r_{n,\gamma}^2 \asymp n^{-\alpha/(\alpha + 2\pi^2 T)}$. If $|\mu_{0,i}| \gtrsim e^{-cT}/2$ for some $c < \alpha$, we have

$$\|E_{\mu_0} \hat{\mu} - \mu_0\|^2 = \sum_{i=1}^\infty \frac{\mu_{0,i}^2}{(1 + n\lambda_i\kappa_i^2)^2} \geq \sum_{i=1}^\infty \frac{e^{-cT}}{(1 + ne^{-2\pi^2 T}i^2)^2} \asymp n^{-\alpha/(\alpha + 2\pi^2 T)} \gg n^{-\alpha/(\alpha + 2\pi^2 T)},$$

by Lemma A.5 (applied with $t = 0$, $r = c$, $u = 0$, $p = \alpha + 2\pi^2 T$, $v = 2$, and $N = n$). Hence $P(\|W_n + E_{\mu_0} \hat{\mu} - \mu_0\| \leq r_{n,\gamma}) \leq P(V_n \geq \|E_{\mu_0} \hat{\mu} - \mu_0\|^2 - r_{n,\gamma}^2) \to 0$. 


Chapter 3

Adaptive recovery of the full parameter

3.1 Introduction

An important problem in nonparametric statistics is to devise methods that achieve good theoretical properties. Our focus in this chapter is on the ability of Bayesian methods to achieve adaptive, rate-optimal inference in mildly ill-posed nonparametric inverse problems introduced in Chapter 2. Nonparametric priors typically involve one or more tuning parameters, or hyper-parameters, that determine the degree of regularization. In practice there is widespread use of empirical Bayes and full, hierarchical Bayes methods to automatically select the appropriate values of such parameters. These methods are generally considered to be preferable to methods that use only a single, fixed value of the hyper-parameters.

The results of Chapter 2 indeed indicate that using a fixed prior can be undesirable, since it can lead to convergence rates that are sub-optimal, unless by chance the statistician has selected a prior that captures the fine properties of the unknown parameter (like its degree of smoothness, if it is a function). Theoretical work that supports the preference for empirical or hierarchical Bayes methods in nonparametric inverse problems does not exist at the present time however. It has until now been unknown whether these approaches can indeed robustify a procedure against prior mismatch. In this chapter we answer this question in the affirmative. We show that empirical and hierarchical Bayes methods can lead to adaptive, rate-optimal procedures in the context of nonparametric inverse problems, provided they are properly constructed.

We study this problem in the context introduced in Chapter 2: see Sections 2.3 and 2.4. Recall we assume that we observe a sequence of noisy coefficients $Y = (Y_1, Y_2, \ldots)$ satisfying

$$Y_i = \kappa_i \mu_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,$$

(3.1)

where $Z_1, Z_2, \ldots$ are independent, standard normal random variables, $\mu = (\mu_i) \in \ell_2$ is the infinite-dimensional parameter of interest, and $(\kappa_i)$ is a known sequence that may converge to 0 as $i \to \infty$, which complicates the inference. We suppose the problem is mildly ill-posed.
of order $p \geq 0$, in the sense that

$$C^{-1}i^{-p} \leq \kappa_i \leq Ci^{-p}, \quad i = 1, 2, \ldots, \tag{3.2}$$

for some $C \geq 1$. Minimax lower bounds for the rate of convergence of estimators for $\mu$ are well known in this setting. For instance, the lower bound over Sobolev balls of regularity $\beta > 0$ is given by $n^{-\beta/(1+2\beta+2p)}$ and over certain “analytic balls” the lower bound is of the order $n^{-1/2}(\log n)^{1/2+p}$ (see [20]). As already mentioned, there are several regularization methods which attain these rates, including classical Tikhonov regularization and Bayes procedures with Gaussian priors obtained in Chapter 2.

However, these methods are not adaptive, in the sense that they rely on knowledge of the regularity (e.g., in Sobolev sense) of the unknown parameter of interest to select the appropriate regularization (see Theorems 2.3 and 2.4). Another example of the Bayesian approach with fixed Gaussian priors that is not adaptive has been recently studied in [1].

In the last decade, however, several methods have been developed in frequentist literature that achieve the minimax convergence rate without such knowledge. This development parallels the earlier work on adaptive methods for the direct nonparametric problem (i.e., the case $p = 0$ in (3.1)) to some extent, although the inverse case is technically usually more demanding. The adaptive methods typically involve a data-driven choice of a tuning parameter in order to automatically achieve an optimal bias-variance trade-off, as in Lepski’s method for instance.

For nonparametric inverse problems, the construction of an adaptive estimator based on a properly penalized blockwise Stein’s rule has been studied in [24], cf. also [15]. This estimator is adaptive both over Sobolev and analytic scales. In [22], the data-driven choice of the regularizing parameters is based on unbiased risk estimation. The authors consider projection estimators and derive the corresponding oracle inequalities. For $\mu$ in the Sobolev scale they obtain asymptotically sharp adaptation in a minimax sense, whereas for $\mu$ in analytic scale, their rate is optimal up to a logarithmic term. Yet another approach to adaptation in inverse problems is the risk hull method studied in [23]. In this paper the authors consider spectral cut-off estimators and provide oracle inequalities. An extension of their approach is presented in [76]. The link between the penalized blockwise Stein’s rule and the risk hull method is presented in [77].

Adaptation properties of Bayes procedures for mildly ill-posed nonparametric inverse problems have until now not been studied in the literature. Results are only available for the direct problem, i.e., the case that $\kappa_i = 1$ for every $i$, or, equivalently, $p = 0$ in (3.2). In the paper [5], it is shown that in this case adaptive Bayesian inference is possible using a hierarchical, conditionally Gaussian prior. Other recent papers also exhibit priors that yield rate-adaptive procedures in the direct signal-in-white-noise problem (see, for instance, [28, 90] and [103]), but it is important to note that these papers use general theorems on contraction rates for posterior distributions (as given in [40], for instance) that are not suitable to deal with the truly ill-posed case in which $k_i \to 0$ as $i \to \infty$. As mentioned in Chapter 2, the reason for that lies in norms that are not equivalent. Obtaining rates relative to the $\ell_2$-norm is much more involved and requires a different approach, as seen in the previous chapter.

To obtain rate-adaptive Bayes procedures for the model (3.1) we consider a family of Gaussian priors $(\Pi_\alpha: \alpha > 0)$ for the parameter $\mu$. These priors are indexed by a parameter
3.1. Introduction

$\alpha > 0$ which quantifies the “regularity” of the prior $\Pi_\alpha$ ($\alpha$ plays exactly the same role as in the previous chapter, details in Section 3.2). Instead of choosing a fixed value for $\alpha$ as in Chapter 2, we view it as a tuning-, or hyper-parameter and consider two different methods for selecting it in a data-driven manner. The approach typically preferred by Bayesian statisticians is to endow the hyper-parameter with a prior distribution itself. This results in a full, hierarchical Bayes procedure. The paper [5] follows the same approach in the direct problem. We prove that under a mild assumption on the hyper-prior on $\alpha$, we obtain an adaptive procedure for the inverse problem using the hierarchical prior. Optimal convergence rates are obtained (up to lower order factors), uniformly over Sobolev and analytic scales.

A second approach we study consists in first “estimating” $\alpha$ from the data and then substituting the estimator $\hat{\alpha}_n$ for $\alpha$ in the posterior distribution for $\mu$ corresponding to the prior $\Pi_\alpha$. This empirical Bayes procedure is not really Bayesian in the strict sense of the word. However, for computational reasons empirical Bayes methods of this type are widely used in practice, making it relevant to study their theoretical performance. Rigorous results about the asymptotic behavior of empirical Bayes selectors of hyper-parameters in infinite-dimensional problems only exist for a limited number of special problems, see, e.g., [4, 52, 107]. In this chapter we prove that the likelihood-based empirical Bayes method that we propose has the same desirable adaptation and rate-optimality properties in nonparametric inverse problems as the hierarchical Bayes approach.

The estimator $\hat{\alpha}_n$ for $\alpha$ that we propose is the commonly used likelihood-based empirical Bayes estimator for the hyper-parameter. Concretely, it is the maximum likelihood estimator for $\alpha$ in the model in which the data is generated by first drawing $\mu$ from $\Pi_\alpha$ and then generating $Y = (Y_1, Y_2, \ldots)$ according to (3.1), i.e., $\mu|\alpha \sim \Pi_\alpha$, and $Y| (\mu, \alpha) \sim \bigotimes_{i=1}^{\infty} N(\kappa_i \mu_i, 1/n)$. A crucial element in the proof of the adaptation properties of both procedures we consider is understanding the asymptotic behavior of $\hat{\alpha}_n$. In contrast to the typical situation in parametric models (see [84]) this turns out to be rather delicate, since the likelihood for $\alpha$ can have complicated behavior. We are able, however, to derive deterministic asymptotic lower and upper bounds for $\hat{\alpha}_n$. In general these depend on the true parameter in a very complicated way. To get some insight into why our procedures work we show that if the true parameter has nice regular behavior of the form $\mu_{0,i} \asymp i^{-1/2-\beta}$ for some $\beta > 0$, then $\hat{\alpha}_n$ is essentially a consistent estimator for $\beta$ (see Lemma 3.1). This means that in some sense, the estimator $\hat{\alpha}_n$ correctly “estimates the regularity” of the true parameter (see [4] for work in a similar direction). Since the empirical Bayes procedure basically chooses the data-dependent prior $\Pi_{\hat{\alpha}_n}$ for $\mu$, this means that asymptotically, the procedure automatically succeeds in selecting among the priors $\Pi_\alpha, \alpha > 0$, the one for which the regularity of the prior and the truth are matched. This results in an optimal bias-variance trade-off and, hence, in optimal convergence rates.

The chapter is organized as follows. In Section 3.2, we first describe the empirical and hierarchical Bayes procedures in detail. Then in Section 3.3, we present a theorem on the asymptotic behavior of estimator $\hat{\alpha}_n$ for the hyper-parameter, followed by two results on the adaptation and rate of contraction of the empirical and hierarchical Bayes posteriors over Sobolev and analytic scales. The two approaches are illustrated numerically in Section 3.4. We apply them to simulated data from an inverse signal-in-white-noise problem, where the
Chapter 3. Adaptive recovery of the full parameter

problem is to recover a signal from a noisy observation of its primitive. This is exactly the same problem as in Section 2.4.2 of the previous chapter. Proofs of the main results are presented in Sections 3.5–3.8.

Notation

For $\beta \geq 0$ and $\gamma \geq 0$, the Sobolev norm $\|\mu\|_\beta$, the analytic norm $\|\mu\|_{A\gamma}$ and the $\ell_2$-norm $\|\mu\|$ of an element $\mu \in \ell_2$ are defined in a usual way by

$$
\|\mu\|_\beta^2 = \sum_{i=1}^{\infty} i^{2\beta} \mu_i^2,
$$

$$
\|\mu\|^2 = \sum_{i=1}^{\infty} \mu_i^2,
$$

$$
\|\mu\|_{A\gamma}^2 = \sum_{i=1}^{\infty} e^{2\gamma i} \mu_i^2,
$$

and the corresponding Sobolev space by $S^\beta = \{ \mu \in \ell_2 : \|\mu\|_\beta < \infty \}$, and the analytic space by $A^\gamma = \{ \mu \in \ell_2 : \|\mu\|_{A\gamma} < \infty \}$.

For two sequences $a_n$ and $b_n$ of numbers, $a_n \asymp b_n$ means that $|a_n/b_n|$ is bounded away from zero and infinity as $n \to \infty$, $a_n \lesssim b_n$ means that $a_n/b_n$ is bounded, $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$, and $a_n \ll b_n$ means that $a_n/b_n \to 0$ as $n \to \infty$. For two real numbers $a$ and $b$, we denote by $a \lor b$ their maximum, and by $a \land b$ their minimum.

3.2 Description of the empirical and hierarchical Bayes procedures

We assume that we observe the sequence of noisy coefficients $Y = (Y_1, Y_2, \ldots)$ satisfying (3.1), for $Z_1, Z_2, \ldots$ independent, standard normal random variables, $\mu = (\mu_1, \mu_2, \ldots) \in \ell_2$, and a known sequence $(\kappa_i)$ satisfying (3.2) for some $p \geq 0$ and $C \geq 1$. We denote the distribution of the sequence $Y$ corresponding to the “true” parameter $\mu_0$ by $P_{\mu_0}$, and the corresponding expectation by $E_{\mu_0}$.

For $\alpha > 0$, consider the product prior $\Pi_\alpha$ on $\ell_2$ given by

$$
\Pi_\alpha = \bigotimes_{i=1}^{\infty} N\left(0, i^{-1-2\alpha} \right).
$$

(3.3)

It is easy to see that this prior is “$\alpha$-regular”, in the sense that for every $\alpha' < \alpha$, it assigns mass 1 to the Sobolev space $S^{\alpha'}$ (cf. (2.12)). In Chapter 2, it was proved that if for the true parameter $\mu_0$ we have $\mu_0 \in S^\beta$ for $\beta > 0$, then the posterior distribution corresponding to the Gaussian prior $\Pi_\alpha$ contracts around $\mu_0$ at the optimal rate $n^{-\beta/(1+2\beta+2p)}$ if $\alpha = \beta$. If $\alpha \neq \beta$, only sub-optimal rates are attained in general, as shown in [16] for the direct problem. In other words, when using a Gaussian prior with a fixed regularity, optimal convergence rates are obtained if and only if the regularity of the prior and the truth are matched. Since the latter is unknown however, choosing the prior that is optimal from the point of view of convergence rates is typically not possible in practice. Therefore, we consider two data-driven methods for selecting the regularity of the prior.
3.2. Description of the empirical and hierarchical Bayes procedures

The first is a likelihood-based empirical Bayes method, which attempts to estimate the appropriate value of the hyper-parameter $\alpha$ from the data. In the Bayesian setting described by the conditional distributions

$$
\mu | \alpha \sim \Pi_{\alpha} \quad \text{and} \quad Y | (\mu, \alpha) \sim \bigotimes_{i=1}^{\infty} N \left( \kappa_i \mu_i, \frac{1}{n} \right),
$$

it holds that

$$
Y | \alpha \sim \bigotimes_{i=1}^{\infty} N \left( 0, i^{-1-2\alpha} \kappa_i^2 + \frac{1}{n} \right).
$$

The corresponding log-likelihood for $\alpha$ (relative to an infinite product of $N(0, 1/n)$-distributions) is easily seen to be given by

$$
\ell_n(\alpha) = -\frac{1}{2} \sum_{i=1}^{\infty} \left( \log \left( 1 + \frac{n}{i^{1+2\alpha} \kappa_i^2 - 2} \right) - \frac{n^2}{i^{1+2\alpha} \kappa_i^2 + n} Y_i^2 \right).
$$

(3.4)

The idea is to “estimate” $\alpha$ by the maximizer of $\ell_n$. The results ahead (Lemma 3.1 and Theorem 3.2) imply that with $P_{\mu_0}$-probability tending to one, $\ell_n$ has a global maximum on $[0, \log n)$ if $\mu_{0,i} \neq 0$ for some $i \geq 2$. (In fact, the cited results imply the maximum is attained on the slightly smaller interval $[0, (\log n)/(2 \log 2) - 1/2 - p]$, but for notational convenience we choose the larger interval). If the latter condition is not satisfied (if $\mu_0 = 0$, for instance), $\ell_n$ may attain its maximum only at $\infty$. Therefore, we truncate the maximizer at $\log n$ and define

$$
\hat{\alpha}_n = \arg\max_{\alpha \in [0, \log n]} \ell_n(\alpha).
$$

The continuity of $\ell_n$ ensures the argmax exists. If it is not unique, any value may be chosen.

We will always assume at least that $\mu_0$ has Sobolev regularity of some order $\beta > 0$. Lemma 3.1 and Theorem 3.2 imply that in this case $\hat{\alpha}_n > 0$ with probability tending to 1. An alternative to the truncation of the argmax of $\ell_n$ at $\log n$ could be to extend the definition of the priors $\Pi_{\alpha}$ to include the case $\alpha = \infty$. The prior $\Pi_{\infty}$ should then be defined as the product $N(0, 1) \otimes \delta_0 \otimes \delta_0 \otimes \cdots$, with $\delta_0$ the Dirac measure concentrated at 0. However, from a practical perspective it is more convenient to define $\hat{\alpha}_n$ as above.

The empirical Bayes procedure consists in computing the posterior distribution of $\mu$ corresponding to a fixed prior $\Pi_{\alpha}$ and then substituting $\hat{\alpha}_n$ for $\alpha$. Recall that under the model described above and the prior (3.3) the coordinates $(\mu_{0,i}, Y_i)$ of the vector $(\mu_0, Y)$ are independent, and hence the conditional distribution of $\mu_0$ given $Y$ factorizes over the coordinates as well. The computation of the posterior distribution reduces to countably many posterior computations in conjugate normal models (see also Section 2.3 of Chapter 2). Therefore, the posterior distribution corresponding to the prior $\Pi_{\alpha}$ is given by

$$
\Pi_{\alpha}(\cdot | Y) = \bigotimes_{i=1}^{\infty} N \left( \frac{nk_i^{-1}}{i^{1+2\alpha} \kappa_i^{-2} + n} Y_i, \frac{\kappa_i^{-2}}{i^{1+2\alpha} \kappa_i^{-2} + n} \right).
$$

(3.5)

Then the empirical Bayes posterior is the random measure $\Pi_{\hat{\alpha}_n}(\cdot | Y)$ defined by

$$
\Pi_{\hat{\alpha}_n}(B | Y) = \Pi_{\alpha}(B | Y) \bigg|_{\alpha = \hat{\alpha}_n}
$$

(3.6)
for measurable subsets $B \subset \ell_2$. Note that the construction of the empirical Bayes posterior does not use information about the regularity of the true parameter. In Theorem 3.3 below, we prove that it contracts around the truth at an optimal rate (up to lower order factors), uniformly over Sobolev and analytic scales.

The second method we consider is a full, hierarchical Bayes approach where we put a prior distribution on the hyper-parameter $\alpha$. We use a prior on $\alpha$ with a positive Lebesgue density $\lambda$ on $(0, \infty)$. The full, hierarchical prior for $\mu$ is then given by

$$
\Pi = \int_0^\infty \lambda(\alpha)\Pi_\alpha \, d\alpha. \tag{3.7}
$$

In Theorem 3.6 below, we prove that under mild assumptions on the prior density $\lambda$, the corresponding posterior distribution $\Pi(\cdot | Y)$ has the same desirable asymptotic properties as the empirical Bayes posterior (3.6).

### 3.3 Adaptation and contraction rates

Understanding of the asymptotic behavior of the maximum likelihood estimator $\hat{\alpha}_n$ is a crucial element in our proofs of the contraction rate results for the empirical and hierarchical Bayes procedures. The estimator somehow estimates the regularity of the true parameter $\mu_0$, but in a rather indirect and involved manner in general. Our first theorem gives deterministic upper and lower bounds for $\hat{\alpha}_n$, whose construction involves the function $h_n: (0, \infty) \to [0, \infty)$ defined by

$$
h_n(\alpha) = \frac{1 + 2\alpha + 2p}{n^{1/(1+2\alpha+2p)}} \log n \sum_{i=1}^\infty \frac{n^2 i^{1+2\alpha} \mu_{0,i}^2 \log i}{(i^{1+2\alpha} + n)^2}. \tag{3.8}
$$

For positive constants $0 < l < L$ we define the lower and upper bounds as

$$
\alpha_n = \inf\{\alpha > 0: h_n(\alpha) > l\} \wedge \sqrt{\log n}, \tag{3.9}
$$

$$
\bar{\alpha}_n = \inf\{\alpha > 0: h_n(\alpha) > L(\log n)^2\}. \tag{3.10}
$$

One can see that the function $h_n$ and hence the lower and upper bounds $\alpha_n$ and $\bar{\alpha}_n$ depend on the true $\mu_0$. We show in Theorem 3.2 that the maximum likelihood estimator $\hat{\alpha}_n$ is between these bounds with probability tending to one. In general, the true $\mu_0$ can have very complicated tail behavior, which makes it difficult to understand the behavior of the upper and lower bounds. If $\mu_0$ has regular tails, however, we can get some insight in the nature of the bounds. We have the following lemma, proved in Section 3.5.

**Lemma 3.1.** For any $l, L > 0$ in the definitions (3.9)–(3.10) the following statements hold.

(i) For all $\beta, R > 0$, there exists $c_0 > 0$ such that

$$
\inf_{\|\mu_0\|_\beta \leq R} \alpha_n \geq \beta - \frac{c_0}{\log n}
$$

for $n$ large enough.
(ii) For all $\gamma, R > 0$,

$$\inf_{\|\mu_0\|_{A^\gamma} \leq R} \alpha_n \geq \frac{\sqrt{\log n}}{\log \log n}$$

for $n$ large enough.

(iii) If $\mu_{0,i} \geq c i^{-\gamma - 1/2}$ for some $c, \gamma > 0$, then for a constant $C_0 > 0$ only depending on $c$ and $\gamma$, we have $\overline{\alpha}_n \leq \gamma + C_0 (\log \log n) / \log n$ for $n$ large enough.

(iv) If $\mu_{0,i} \neq 0$ for some $i \geq 2$, then $\overline{\alpha}_n \leq (\log n) / (2 \log 2) - 1/2 - p$ for $n$ large enough.

We note that items (i) and (iii) of the lemma imply that if $\mu_{0,i} \asymp i^{-1/2 - \beta}$, then the interval $[\underline{\alpha}_n, \overline{\alpha}_n]$ concentrates around the value $\beta$ asymptotically. In combination with Theorem 3.2 this shows that at least in this regular case, $\hat{\alpha}_n$ correctly estimates the regularity of the truth. The same is true in the analytic case, since item (ii) of the lemma shows that $\alpha_n \to \infty$ in that case, i.e., asymptotically, the procedure detects the fact that $\mu_0$ has infinite regularity.

Item (iv) implies that if $\mu_{0,i} \neq 0$ for some $i \geq 2$, then $\overline{\alpha}_n < \infty$ for large $n$. Conversely, the definitions of $h_n$ and $\overline{\alpha}_n$ show that if $\mu_{0,i} = 0$ for all $i \geq 2$, then $h_n \equiv 0$ and hence $\overline{\alpha}_n = \infty$.

The following theorem asserts that the point(s) where $\ell_n$ is maximal is (are) asymptotically between the bounds just defined, uniformly over Sobolev and analytic scales. The proof is given in Section 3.6.

**Theorem 3.2.** For every $R > 0$ the constants $l$ and $L$ in (3.9) and (3.10) can be chosen such that

$$\inf_{\mu_0 \in B(R)} P_{\mu_0} \left( \arg\max_{\alpha \geq 0} \ell_n(\alpha) \in [\underline{\alpha}_n, \overline{\alpha}_n] \right) \to 1,$$

where $B(R) = \{ \mu_0 \in \ell_2 : ||\mu_0||_{\beta} \leq R \}$ or $B(R) = \{ \mu_0 \in \ell_2 : ||\mu_0||_{A^\gamma} \leq R \}$.

With the help of Theorem 3.2, in Section 3.7 we can prove the following theorem, which states that the empirical Bayes posterior distribution (3.6) achieves optimal minimax contraction rates up to a slowly varying factor, uniformly over Sobolev and analytic scales.

**Theorem 3.3.** For every $\beta, \gamma, R > 0$ and $M_n \to \infty$ we have

$$\sup_{||\mu_0||_{\beta} \leq R} E_{\mu_0} \Pi_{\hat{\alpha}_n} \left( ||\mu - \mu_0|| \geq M_n L_n n^{-\beta/(1+2\beta+2p)} \bigg| Y \right) \to 0$$

and

$$\sup_{||\mu_0||_{A^\gamma} \leq R} E_{\mu_0} \Pi_{\hat{\alpha}_n} \left( ||\mu - \mu_0|| \geq M_n L_n (\log n)^{1/2 + p n^{-1/2}} \bigg| Y \right) \to 0,$$

where $(L_n)$ is a slowly varying sequence.

So indeed we see that both in the Sobolev and analytic cases, we obtain the optimal minimax rates up to a slowly varying factor. The proofs of the statements (given in Section 3.7) show that

$$L_n = \begin{cases} 
(\log n)^{3/2} (\log \log n)^{1/2} & \text{if } ||\mu_0||_{\beta} \leq R, \\
(\log n)^{(1/2+p)\sqrt{\log n}}/2 + 1 - p (\log \log n)^{1/2} & \text{if } ||\mu_0||_{A^\gamma} \leq R.
\end{cases}$$
These sequences converge to infinity, but they are slowly varying, hence they converge slower than any power of $n$. (The latter sequence is, however, faster than any power of $\log n$, the canonical example of a slowly varying sequence).

Suppose that we do not select the hyper-parameter of the prior with a data-driven procedure, but rather keep it fixed (as in Chapter 2, for example). One can then see that in the case of the analytic truth this prior leads to the suboptimal posterior contraction rate for any $\alpha$, whereas in the case when $\mu_0$ is in $S^\beta$, the optimal rate is achieved if $\alpha = \beta$ (see Theorem 2.3). The data-driven procedure overcomes this obstacle, although at the additional cost of a slowly varying term faster than any power of $\log n$. It is to be expected that the slowly varying term could be reduced if a more suitable type of prior was chosen (e.g., with variances $e^{-\tau_i}$ instead of $i^{-1-2\alpha}$).

**Remark 3.4.** The truncated maximizer $\hat{\alpha}_n$ in the empirical Bayes posterior in Theorem 3.3 can be replaced by any $\alpha_n$ in $[\underline{\alpha}_n, \overline{\alpha}_n]$ (see Section 3.7 for details). Inspection of the proof of Theorem 3.2 shows that the log-likelihood $\ell_n$ is strictly monotone on the interval $(0, \alpha_n]$. These two observations lead to the following practical conclusion: one should study the marginal likelihood for $\alpha$ starting from 0, and once a local maximum is found, it can be used in the empirical Bayes posterior.

The full Bayes procedure using the hierarchical prior (3.7) achieves the same results as the empirical Bayes method, under mild assumptions on the prior density $\lambda$ for $\alpha$.

**Assumption 3.5.** Assume that for every $c_1 > 0$ there exist $c_2, c_3 > 0$ and $c_4 \geq 1$ such that

$$c_4^{-1} \alpha^{-c_2} \leq \lambda(\alpha) \leq c_4 \alpha^{-c_2} \quad \text{or} \quad c_4^{-1} \exp(-c_3 \alpha) \leq \lambda(\alpha) \leq c_4 \exp(-c_3 \alpha)$$

for $\alpha \geq c_1$.

One can see that a many distributions satisfy this assumption. Careful inspection of the proof of the following theorem, given in Section 3.8, can lead to weaker assumptions, although these will be less attractive to formulate. Recall the notation $\Pi(\cdot \mid Y)$ for the posterior corresponding to the hierarchical prior (3.7).

**Theorem 3.6.** Suppose the prior density $\lambda$ satisfies Assumption 3.5. Then for every $\beta, \gamma, R > 0$ and $M_n \to \infty$ we have

$$\sup_{\|\mu_0\|_\beta \leq R} E_{\mu_0} \Pi(\|\mu - \mu_0\| \geq M_n L_n n^{-\beta/(1+2\beta+2p)} \mid Y) \to 0$$

and

$$\sup_{\|\mu_0\|_\beta \leq R} E_{\mu_0} \Pi(\|\mu - \mu_0\| \geq M_n L_n (\log n)^{1/2+p n^{-1/2}} \mid Y) \to 0,$$

where $(L_n)$ is a slowly varying sequence.

The hierarchical Bayes method thus yields exactly the same rates as the empirical method, and, therefore, the interpretation of this theorem is the same as before.
3.4 Simulation example: Volterra operator

In this chapter we work with the same example as in Chapter 2 in the mildly ill-posed setting. Recall that we consider the inverse signal-in-white-noise problem where we observe the process \( Y_t: t \in [0, 1] \) given by

\[
Y_t = \int_0^t \int_0^s \mu(u) \, du \, ds + \frac{1}{\sqrt{n}} W_t,
\]

with \( W \) a standard Brownian motion, and the aim is to recover the function \( \mu \). If, slightly abusing notation, we define \( Y_i = \int_0^1 e_i(t) \, dY_t \), for \( e_i \) the orthonormal basis functions given by \( e_i(t) = \sqrt{2} \cos((i-1/2)\pi t) \), then it is easily verified that the observations \( Y_i \) satisfy (3.1), with \( \kappa_i^2 = ((i-1/2)^2 \pi^2)^{-1} \), i.e., \( p = 1 \) in (3.2), and \( \mu_i \) the Fourier coefficients of \( \mu \) relative to the basis \( e_i \).

We consider simulated data from this model for \( \mu_0 \), the function with Fourier coefficients \( \mu_{0,i} = i^{-3/2} \sin(i) \), so we have a truth which essentially has regularity 1. In the following figure we plot the true function \( \mu_0 \) (black curve) and the empirical Bayes posterior mean (red curve) in the left panels, and the corresponding normalized likelihood \( \exp(\ell_n)/ \max(\exp(\ell_n)) \) in the right panels (we truncated the sum in (3.4) at a high level).

Figure 3.1 shows the results for the empirical Bayes procedure with simulated data for \( n = 10^3, 10^5, 10^7, 10^9 \), and \( 10^{11} \), from top to bottom. The figure shows that the estimator \( \hat{\alpha}_n \) does a good job in this case at estimating the regularity level 1, at least for large enough \( n \). We also see, however, that due to the ill-posedness of the problem, a large signal-to-noise ratio \( n \) is necessary for accurate recovery of the function \( \mu \).

We applied the hierarchical Bayes method to the simulated data as well. We chose a standard exponential prior distribution on \( \alpha \), which satisfies Assumption 3.5. Since the posterior can not be computed explicitly, we implemented an MCMC algorithm that generates (approximate) draws from the posterior distribution of the pair \( (\alpha, \mu) \). More precisely, we fixed a large index \( J \in \mathbb{N} \) and defined the vector \( \mu^J = (\mu_1, \ldots, \mu_J) \) consisting of the first \( J \) coefficients of \( \mu \). Then we devised a Metropolis-within-Gibbs algorithm for sampling from the posterior distribution of \( (\alpha, \mu^J) \) (e.g., [95]). The algorithm alternates between draws from the conditional distribution \( \mu^J|\alpha, Y \) and the conditional distribution \( \alpha|\mu^J, Y \). The former is explicitly given by (3.5). To sample from \( \alpha|\mu, Y \) we used a standard Metropolis-Hastings step. It is easily verified that the Metropolis-Hastings acceptance probability for a move from \( (\alpha, \mu) \) to \( (\alpha', \mu) \) is given by

\[
1 \wedge \frac{q(\alpha') \alpha) p(\mu^J|\alpha') \lambda(\alpha')}{q(\alpha|\alpha') p(\mu^J|\alpha) \lambda(\alpha)},
\]

where \( p(\cdot|\alpha) \) is the density of \( \mu^J \) if \( \mu \sim \Pi_\alpha \), i.e.,

\[
p(\mu^J|\alpha) \propto \prod_{j=1}^J j^{1/2 + \alpha} e^{-\frac{1}{2} j^{1+2\alpha} \mu_j^2},
\]

and \( q \) is the transition kernel of the proposal chain. We used a proposal chain that, if it is currently at location \( \alpha \), moves to a new \( N(\alpha, \sigma^2) \)-distributed location provided the latter is positive. We omit further details, the implementation is straightforward.
Figure 3.1: Left panels: the empirical Bayes posterior mean (red) and the true curve (black). Right panels: corresponding normalized likelihood for $\alpha$. We have $n = 10^3, 10^5, 10^7, 10^9$, and $10^{11}$, from top to bottom.

Figure 3.2: Left panels: the hierarchical Bayes posterior mean (red) and the true curve (black). Right panels: histograms of posterior for $\alpha$. We have $n = 10^3, 10^5, 10^7, 10^9$, and $10^{11}$ from top to bottom.
The results for the hierarchical Bayes procedure are given in Figure 3.2. The figure shows the results for simulated data with $n = 10^3, 10^5, 10^7, 10^9$ and $10^{11}$, from top to bottom. Every time we see the posterior mean (in red) and the true curve (black) on the left and a histogram for the posterior of $\alpha$ on the right. The results are comparable to what we found for the empirical Bayes procedure.

3.5 Proof of Lemma 3.1

Recall that $C^{-1}i^{-\gamma} \leq \kappa_i \leq C'i^{-\gamma}$. This implies for any $r > 0$ and $m$

$$C^{-m-2r} \frac{i^{-pm}}{(i^{1+2\alpha+2p} + n)^r} \leq \frac{\kappa_i^m}{(i^{1+2\alpha\kappa_i^{-2} + n})^r} \leq C^m+2r \frac{i^{-pm}}{(i^{1+2\alpha+2p} + n)^r}.$$ 

(i). We show that for all $\alpha \leq \beta - c_0 / \log n$, for some large enough constant $c_0 > 0$ that only depends on $\|\mu_0\|_\beta$, it holds that $h_n(\alpha) \leq l$, where $l$ is the given positive constant in the definition of $Q_n$.

The sum in the definition (3.8) of $h_n$ can be split into two sums, one over indices $i \leq n^{1/(1+2\alpha+2p)}$ and one over indices $i > n^{1/(1+2\alpha+2p)}$. The second sum is bounded by

$$C^4 n^2 \sum_{i \geq n^{1/(1+2\alpha+2p)}} i^{-1-2\alpha-4p-2\beta} (\log i)^{2\beta} \mu_0^2, i.$$ 

Since the function $x \mapsto x^{-\gamma} \log x$ is decreasing on $[e^{1/\gamma}, \infty)$, this is further bounded by

$$\frac{C^4 \|\mu_0\|_\beta^2}{1 + 2\alpha + 2p} n^{1+2\alpha-2\beta + 2\beta} \log n.$$ 

The sum over $i \leq n^{1/(1+2\alpha+2p)}$ is bounded by

$$C^4 \sum_{i \leq n^{1/(1+2\alpha+2p)}} i^{1+2\alpha-2\beta} i^{2\beta} \mu_0^2 \log i.$$ 

Since the logarithm is increasing we can take $(\log n)/(1 + 2\alpha + 2p)$ outside the sum and then bound $i^{1+2\alpha-2\beta}$ above by $n^{(1+2\alpha-2\beta)/(1+2\alpha+2p)}0$ to arrive at the subsequent bound

$$\frac{C^4 \|\mu_0\|_\beta^2}{1 + 2\alpha + 2p} n^{0 + \frac{1+2\alpha-2\beta}{1+2\alpha+2p}} \log n.$$ 

Combining the bounds for the two sums we obtain the upper bound

$$h_n(\alpha) \leq C^4 \|\mu_0\|_\beta^2 n^{-\frac{1+2(\beta-\alpha)}{1+2\alpha+2p}},$$

valid for all $\alpha > 0$. Now suppose that $\alpha \leq \beta - c_0 / \log n$. Then for $n$ large enough, the power of $n$ on the right-hand side is bounded by

$$n \frac{1+2(c_0 / \log n)}{1+2\alpha+2p} = e^{-\frac{2c_0}{1+2\alpha+2p}}.$$
Hence, given \( l > 0 \) we can choose \( c_0 \) so large, only depending on \( \| \mu_0 \|_0 \), that \( h_n(\alpha) \leq l \) for \( \alpha \leq \beta - c_0/\log n \).

(ii). We show that in this case we have \( h_n(\alpha) \leq l \) for \( \alpha_n \geq \sqrt{\log n/(\log \log n)} \) and \( n \geq n_0 \), where \( n_0 \) only depends on \( \| \mu_0 \|_A \). Again, we give an upper bound for \( h_n \) by splitting the sum in its definition into two smaller sums. The one over indices \( i > n^{1/(1+2\alpha+2p)} \) is bounded by

\[
C^4 n^2 \sum_{i > n^{1/(1+2\alpha+2p)}} i^{-1-2\alpha-4p} e^{-2\gamma i} (\log i)e^{2\gamma i} \mu_{0,i}.
\]

Using the fact that for \( \delta > 0 \) the function \( x \mapsto x^{-\delta} e^{-2\gamma x} \log x \) is decreasing on \([\epsilon^{1/\delta}, \infty)\) we can see that this is further bounded by

\[
\frac{C^4 \| \mu_0 \|^2_{A^\gamma}}{1+2\alpha+2p} e^{-2\gamma n^{1/(1+2\alpha+2p)} + \frac{1+2\alpha}{2\gamma}} \log n.
\]

The sum over indices \( i \leq n^{1/(1+2\alpha+2p)} \) is bounded by

\[
\frac{C^4 \log n}{1+2\alpha+2p} \sum_{i \leq n^{1/(1+2\alpha+2p)}} i^{1+2\alpha} e^{-2\gamma i} e^{2\gamma i} \mu_{0,i}^2.
\]

Since the maximum on \((0, \infty)\) of the function \( x \mapsto x^{1+2\alpha} \exp(-2\gamma x) \) equals \( \exp((1+2\alpha)(1+2\alpha)/2) - 1) \), we have the subsequent bound

\[
\frac{C^4 \| \mu_0 \|^2_{A^\gamma}}{1+2\alpha+2p} e^{(1+2\alpha) \log((1+2\alpha)/2\gamma)} \log n.
\]

Combining the two bounds we find that

\[
h_n(\alpha) \leq \| \mu_0 \|^2_{A^\gamma} \left( n^{1+2\alpha+2p} e^{-2\gamma n^{1+2\alpha+2p}} + n^{-\frac{1}{1+2\alpha+2p}} e^{(1+2\alpha) \log \frac{1+2\alpha}{2\gamma}} \right)
\]

for all \( \alpha > 0 \). If \( \alpha \leq \sqrt{\log n/(\log \log n)} \), then for \( n \) large enough

\[
n^{1+2\alpha+2p} e^{-2\gamma n^{1+2\alpha+2p}} \leq n \exp(-2\gamma n^{\log \log n/3}) = \exp(n^{\log \log n/3} - 2\gamma n^{\log \log n/3}),
\]

which tends to 0 as \( n \) goes to infinity, since \( \sqrt{\log n} \ll \log n \). Moreover, for \( n \) large enough

\[
- \frac{\log n}{1+2\alpha+2p} + (1+2\alpha) \log \frac{1+2\alpha}{2\gamma}
\]

\[
\leq - \frac{1}{3} \sqrt{\log n (\log \log n)} + 3 \sqrt{\frac{\log n}{\log \log n}} \log \left( \frac{3}{2\gamma \log \log n} \right),
\]

which tends to \(-\infty\) as \( n \) goes to infinity. It is then easily verified that for the given constant \( l > 0 \), we have \( h_n(\alpha) \leq l \) for \( n \geq n_0 \) if \( \alpha \leq \sqrt{\log n/\log \log n} \), where \( n_0 \) only depends on \( \| \mu_0 \|_{A^\gamma} \).

(iii). Let \( \gamma_n = \gamma + C_0(\log \log n)/(\log n) \). We will show that for \( n \) large enough, \( h_n(\gamma_n) \geq L(\log n)^2 \), provided \( C_0 \) is large enough. Note that

\[
\sum_{i=1}^{\infty} n^{2(\gamma_n - \gamma)} \mu_{0,i}^2 \log i \geq \frac{c^2}{4} \sum_{i \leq n^{1/(1+2\gamma_n+2p)}} i^{2(\gamma_n - \gamma) \log i}.
\]
By monotonicity and the fact that $|x| \geq x/2$ for $x$ large, the sum on the right is bounded from below by the integral

$$\int_0^{n^{1/(1+2\gamma_n+2p)/2}} x^{2\gamma_n-2\gamma} \log x \, dx.$$  

This integral can be computed explicitly (see also Theorem 1 on page 281 in [33]) and is for large $n$ bounded from below by a constant times

$$\frac{\log n}{1 + 2\gamma_n + 2p} \left( \frac{2\gamma_n-2\gamma+1}{1+4\gamma_n+4p} \right).$$  

It follows that, for large enough $n$, $h_n(\gamma_n)$ is bounded from below by a constant times $c^2 n^2(\gamma_n-\gamma)/(1+2\gamma_n+2p)$. Since $(\log \log n)/(\log n) \leq 1/4$ for $n$ large enough, we obtain

$$n^2(\gamma_n-\gamma)/(1+2\gamma_n+2p) \geq n \frac{1}{\log n} \left( \log \log n \right)^{1+2\gamma} C_0/(1+2\gamma+\gamma_0/2+2p).$$

Hence for $C_0$ large enough, only depending on $c$ and $\gamma$, we indeed have that and $h_n(\gamma_n) \geq L(\log n)^2$ for large $n$.

(iv). If $\mu_{0,i} \neq 0$ for $i \geq 2$, then

$$h_n(\alpha) \geq \frac{1 + 2\alpha + 2p}{n^{1/(1+2\alpha+2p)}} \log n \frac{n^2 i^{1+2\alpha}}{(i^{1+2\alpha+2p} + n)^2}.$$  

Now define $\alpha_n$ such that $i^2 + 2\alpha_n + 2p = n$. Then we have $h_n(\alpha_n) \geq n^{1-(1+2p)/(1+2\alpha_n+2p)}$ by construction. Since $\alpha_n \to \infty$, the right side is larger than $L(\log n)^2$ for $n$ large enough, irrespective of the value of $L$, hence $\overline{\alpha}_n \leq \alpha_n \leq (\log n)/(2 \log 2) - 1/2 - p$.

### 3.6 Proof of Theorem 3.2

With the help of the dominated convergence theorem, one can see that the random function $\ell_n$ is $(P_{\mu_0}-a.s.)$ differentiable and its derivative, which we denote by $M_n$, is given by

$$M_n(\alpha) = \sum_{i=1}^{\infty} \frac{n \log i}{i^{1+2\alpha} \kappa_i^{-2} + n} - \sum_{i=1}^{\infty} \frac{n^2 i^{1+2\alpha}}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} Y_i^2.$$  

We will show that on the interval $(0, \alpha_n + 1/\log n]$ the random function $M_n$ is positive and bounded away from 0 with probability tending to one, hence $\ell_n$ has no local maximum in this interval. Next we distinguish two cases according to the value of $\overline{\alpha}_n$. If $\overline{\alpha}_n = \infty$, then the inequality $\hat{\alpha}_n \leq \overline{\alpha}_n$ trivially holds. In the case $\overline{\alpha}_n < \infty$ we show that the integral of $M_n$ over the interval $[\overline{\alpha}_n, \infty)$ is a.s. upper bounded by a fixed positive constant times $n^{1/(1+2\alpha_n+2p)}(\log n)^2/(1 + 2\overline{\alpha}_n + 2p)$. Then we prove the constant $L$ can be set such that on the interval $[\overline{\alpha}_n - 1/\log n, \overline{\alpha}_n]$ it is bounded by an arbitrary negative constant times $n^{1/(1+2\alpha_n+2p)}(\log n)^3/(1 + 2\overline{\alpha}_n + 2p)$ with probability tending to one uniformly, hence the integral of $M_n$ over this interval is bounded above by an arbitrarily large negative constant times $n^{1/(1+2\alpha_n+2p)}(\log n)^2/(1 + 2\overline{\alpha}_n + 2p)$. This means that on the interval.
the function $\ell_n(\alpha)$ decreases more than it can possibly increase on the interval $[\alpha_n, \infty)$. Therefore, it holds with probability tending to one that $\ell_n$ has no global maximum on $(\alpha_n - 1/\log n, \infty)$.

We present the details of the proof for the case that $\mu_0 \in S^\beta$, but merely for the notational convenience. Note that in the case $\mu_0 \in A^\gamma$ by Lemma 3.1 $\alpha_n > \beta$ for any $\beta > 0$ and large enough $n$.

3.6.1 $M_n(\alpha)$ on $[\alpha_n, \infty)$

In this section we give a deterministic upper bound for the integral of $M_n(\alpha)$ on the interval $[\alpha_n, \infty)$. We can restrict to the case that $\mu_0, i \neq 0$ for some $i \geq 0$, since otherwise $\alpha_n = \infty$.

By Lemma 3.1, we have $\alpha_n \leq \alpha_n$ in this case, where $\alpha_n = (\log n)/(2 \log 2) - 1/2 - p$.

We have the trivial bound

$$M_n(\alpha) \leq C^2 \sum_{i=1}^{\infty} \frac{n \log i}{i^{1+2\alpha+2p} + n}.$$  

An application of Lemma A.9.(i) with $r = 1 + 2\alpha + 2p$ and $c = \beta + 2p$ shows that for $\beta/2 < \alpha \leq \alpha_n$,

$$M_n(\alpha) \lesssim \frac{1}{1 + 2\alpha + 2p} \frac{n^{1/(1+2\alpha+2p)}}{\log n}.$$  

For $\alpha \geq \alpha_n$ we apply Lemma A.9.(ii), and see that $M_n(\alpha) \lesssim n^{2-1-2\alpha-2p}$. Using the fact that $x \mapsto 2^{-x}x^3$ is decreasing for large $x$, it is easily seen that $n^{2-1-2\alpha-2p} \lesssim (\log n)^3/(1 + 2\alpha + 2p)^3$ for $\alpha \geq \alpha_n$, hence

$$M_n(\alpha) \lesssim \frac{(\log n)^3}{(1 + 2\alpha + 2p)^3}.$$  

By Lemma 3.1 we have $\beta/2 < \alpha_n$ for large enough $n$. It follows that the integral we want to bound is bounded by a constant times

$$n^{1/(1+2\alpha_n+2p)} \log n \int_{\pi_n}^{\alpha_n} \frac{1}{1 + 2\alpha + 2p} \frac{d\alpha}{(1 + 2\alpha + 2p)^3} + (\log n)^3 \int_{\alpha_n}^{\infty} \frac{1}{(1 + 2\alpha + 2p)^3} d\alpha.$$  

This quantity is bounded by a constant times

$$n^{1/(1+2\pi_n+2p)}(\log n)^2 \frac{1}{1 + 2\pi_n + 2p}.$$  

3.6.2 $M_n(\alpha)$ on $\alpha \in [\alpha_n - 1/\log n, \alpha_n]$

In this section we show that the process $M_n(\alpha)$ is with probability going to one smaller than a negative, arbitrary large constant times $n^{1/(1+2\pi_n+2p)}(\log n)^3/(1 + 2\pi_n + 2p)$ uniformly on the interval $[\alpha_n - 1/\log n, \alpha_n]$. More precisely, we show that for every $\beta, R, M > 0$, the
constant \( L > 0 \) in the definition of \( \tau_n \) can be chosen such that

\[
\limsup_{n \to \infty} \sup_{\|\mu_0\|_\beta \leq R} \sup_{\alpha \in [\tau_n - 1/\log n, \tau_n]} \mathbb{E}_{\mu_0} \left( \frac{(1 + 2\alpha + 2p)M_n(\alpha)}{n^{1/(1+2\alpha+2p)}(\log n)^3} \right) < -M \quad (3.11)
\]

\[
\sup_{\|\mu_0\|_\beta \leq R} \mathbb{E}_{\mu_0} \sup_{\alpha \in [\tau_n - 1/\log n, \tau_n]} \left( \frac{(1 + 2\alpha + 2p)[M_n(\alpha) - \mathbb{E}_{\mu_0} M_n(\alpha)]}{n^{1/(1+2\alpha+2p)}(\log n)^3} \right) \to 0. \quad (3.12)
\]

The expected value of the normalized version of the process \( M_n \) given on the left-hand side of (3.11) is equal to

\[
\frac{1 + 2\alpha + 2p}{n^{1/(1+2\alpha+2p)}(\log n)^3} \left( \sum_{i=1}^{\infty} \frac{n^2 \log i}{(i^{1+2\alpha}\kappa_i^{-2} + n)^2} - \sum_{i=1}^{\infty} \frac{n^2 i^{1+2\alpha} \mu_0 i \log i}{(i^{1+2\alpha}\kappa_i^{-2} + n)^2} \right). \quad (3.13)
\]

We write this as the sum of two terms and bound the first term by

\[
C^4 \frac{1 + 2\alpha + 2p}{n^{1/(1+2\alpha+2p)}(\log n)^3} \sum_{i=1}^{\infty} \frac{n \log i}{i^{1+2\alpha+2p} + n}.
\]

By Lemma A.9.(i), for all \( \alpha > c/2 - p \), where \( c \) is an arbitrary positive constant, this can be further bounded by a multiple of \( 1/(\log n)^2 \). By Lemma 3.1, \( \beta/4 < \tau_n - 1/\log n \) for large enough \( n \), hence by choosing \( c = \beta/2 + 2p \) we get that the right-hand side of the preceding display tends to zero uniformly over \([\tau_n - 1/\log n, \infty]\). We now consider the second term in (3.13), which is equal to \( h_n(\alpha)/(\log n)^2 \). By Lemma 3.7, for any \( \mu_0 \in \ell_2 \) and \( n \geq e^4 \), we have

\[
\frac{h_n(\alpha)}{(\log n)^2} \geq \frac{1}{(\log n)^2} h_n(\tau_n) = L,
\]

where the last equality holds by the definition of \( \tau_n \). This concludes the proof of (3.11).

To verify (3.12) it suffices, by Corollary 2.2.5 in [104] (applied with \( \psi(x) = x^2 \)), to show that

\[
\sup_{\|\mu_0\|_\beta \leq R} \sup_{\alpha \in [\tau_n - 1/\log n, \tau_n]} \text{var}_{\mu_0} \left( \frac{(1 + 2\alpha + 2p)M_n(\alpha)}{n^{1/(1+2\alpha+2p)}(\log n)^3} \right) \to 0, \quad (3.14)
\]

and

\[
\sup_{\|\mu_0\|_\beta \leq R} \int_0^{\text{diam}_n} \sqrt{N(\varepsilon, [\tau_n - 1/\log n, \tau_n], d_n)} \, d\varepsilon \to 0,
\]

where \( d_n \) is the semimetric defined by

\[
d^2(\alpha_1, \alpha_2) = \text{var}_{\mu_0} \left( \frac{(1 + 2\alpha_1 + 2p)M_n(\alpha_1)}{n^{1/(1+2\alpha_1+2p)}(\log n)^3} - \frac{(1 + 2\alpha_2 + 2p)M_n(\alpha_2)}{n^{1/(1+2\alpha_2+2p)}(\log n)^3} \right),
\]

\( \text{diam}_n \) is the diameter of \([\tau_n - 1/\log n, \tau_n]\) relative do \( d_n \), and \( N(\varepsilon, B, d) \) is the minimal number of \( d \)-balls of radius \( \varepsilon \) needed to cover the set \( B \).

By Lemma 3.8

\[
\text{var}_{\mu_0} \left( \frac{(1 + 2\alpha + 2p)M_n(\alpha)}{n^{1/(1+2\alpha+2p)}(\log n)^3} \right) \leq \frac{n^{-1/(1+2\alpha+2p)}}{(\log n)^4} \left( 1 + h_n(\alpha) \right), \quad (3.15)
\]
Next note that by definition of $h_n$, show that indeed ensure that \((3.16)\) is true.

The last bound also shows that the $d_n$-diameter of the set $[\overline{\alpha} - 1/\log n, \overline{\alpha}]$ is bounded above by a constant times $(\log n)^{-1}$, with a constant that does not depend on $\mu_0$ and $\alpha$. By Lemma 3.9 and the fact that $h_n(\alpha) \leq L(\log n)^2$ for $\alpha \in [\overline{\alpha} - 1/\log n, \overline{\alpha}]$, we get the following upper bound, for $\alpha_1, \alpha_2 \in [\overline{\alpha} - 1/\log n, \overline{\alpha}]$,

$$d_n(\alpha_1, \alpha_2) \lesssim |\alpha_1 - \alpha_2|,$$

with a constant that does not depend on $\mu_0$. Therefore, $N(\varepsilon, [\overline{\alpha} - 1/\log n, \overline{\alpha}], d_n) \lesssim 1/(\varepsilon \log n)$ and hence

$$\sup_{\|\mu_0\|_\beta \leq R} \int_0^{\text{diam}_n} \sqrt{N(\varepsilon, [\overline{\alpha} - 1/\log n, \overline{\alpha}], d_n)} \, d\varepsilon \lesssim \frac{1}{\log n} \to 0.$$  

### 3.6.3 $M_n(\alpha)$ on $(0, \overline{\alpha} + 1/\log n)$

In this subsection we prove that if the constant $l$ in the definition of $\alpha_n$ is small enough, then

$$\liminf_{n \to \infty} \inf_{\mu_0 \in \ell_2} \inf_{\alpha \in (0, \overline{\alpha} + 1/\log n]} \frac{E_{\mu_0} (1 + 2\alpha + 2p)M_n(\alpha)}{n^{1/(1+2\alpha+2p)} \log n} > 0 \tag{3.16}$$

$$\sup_{\mu_0 \in \ell_2} \sup_{\alpha \in (0, \overline{\alpha} + 1/\log n]} \frac{(1 + 2\alpha + 2p)M_n(\alpha) - E_{\mu_0} M_n(\alpha)}{n^{1/(1+2\alpha+2p)} \log n} \to 0. \tag{3.17}$$

This shows that $M_n$ is positive throughout $(0, \overline{\alpha} + 1/\log n]$ with probability tending to one uniformly over $\ell_2$.

Since $E_{\mu_0} Y_i^2 = \kappa_i^2 \mu_{0,i}^2 + 1/n$, the expected value on the left-hand side of (3.16) is equal to

$$\frac{1 + 2\alpha + 2p}{n^{1/(1+2\alpha+2p)} \log n} \sum_{i=1}^{\infty} \frac{n^2 \log i}{(i^{1+2\alpha} \kappa_i^2 + n^2)^2} - h_n(\alpha). \tag{3.18}$$

We first find a lower bound for the first term. Since $\alpha_n \leq \sqrt{\log n}$ by definition, we have $\alpha \ll \log n$ for all $\alpha \in (0, \overline{\alpha} + 1/\log n]$. Then it follows from Lemma A.11 that for $n$ large enough, the first term in (3.18) is bounded from below by $1/(12C^4)$ for all $\alpha \in (0, \overline{\alpha} + 1/\log n]$. Next note that by definition of $h_n$ and Lemma 3.7, we have

$$\sup_{\alpha \in (0, \overline{\alpha} + 1/\log n]} h_n(\alpha) \leq C' l,$$

where $C' > 0$ is a constant independent of $\mu_0$. So by choosing $l > 0$ small enough, we can indeed ensure that (3.16) is true.

To verify (3.17) it suffices again, by Corollary 2.2.5 in [104] applied with $\psi(x) = x^2$, to show that

$$\sup_{\mu_0 \in \ell_2} \sup_{\alpha \in (0, \overline{\alpha} + 1/\log n]} \frac{\var_{\mu_0} (1 + 2\alpha + 2p)M_n(\alpha)}{n^{1/(1+2\alpha+2p)} \log n} \to 0. \tag{3.19}$$
and
\[
\sup_{\mu_0 \in \ell_2} \int_0^{\text{diam}_n} \sqrt{N(\varepsilon, (0, \alpha_n + 1/\log n], d_n)} \, d\varepsilon \to 0,
\]
where \(d_n\) is the semimetric defined by
\[
d^2_n(\alpha_1, \alpha_2) = \text{var}_{\mu_0} \left( \frac{(1 + 2\alpha_1 + 2p)M_n(\alpha_1)}{n^{1/(1+2\alpha_1+2p)} \log n} - \frac{(1 + 2\alpha_2 + 2p)M_n(\alpha_2)}{n^{1/(1+2\alpha_2+2p)} \log n} \right),
\]
diam_n \(= \) the diameter of \((0, \alpha_n + 1/\log n]\) relative to \(d_n\), and \(N(\varepsilon, B, d)\) is the minimal number of \(d\)-balls of radius \(\varepsilon\) needed to cover the set \(B\).

By Lemma 3.8
\[
\text{var}_{\mu_0} \frac{(1 + 2\alpha + 2p)M_n(\alpha)}{n^{1/(1+2\alpha+2p)} \log n} \lesssim n^{-1/(1+2\alpha+2p)} \left( 1 + h_n(\alpha) \right),
\]
with a constant that does not depend on \(\mu_0\) and \(\alpha\). We have seen that on the interval \((0, \alpha_n + 1/\log n]\) the function \(h_n\) is bounded by a constant times \(l\), hence the variance in (3.19) is bounded by a multiple of \(n^{-1/(1+2\alpha_n+2\log n+2p)} \leq e^{-\left(1/3\right)\sqrt{\log n}} \to 0\), which proves (3.19).

The variance bound above also imply that the \(d_n\)-diameter of the set \((0, \alpha_n + 1/\log n]\) is bounded by a multiple of \(e^{-\left(1/6\right)\sqrt{\log n}}\). By Lemma 3.9, the definition of \(\alpha_n\) and Lemma 3.7,
\[
d_n(\alpha_1, \alpha_2) \lesssim |\alpha_1 - \alpha_2| (\log n) \sqrt{n^{-1/(1+2\alpha_n+2\log n+2p)} \lesssim |\alpha_1 - \alpha_2|},
\]
with constants that do not depend on \(\mu_0\). Hence for the covering number of \((0, \alpha_n + 1/\log n]\) \(\subset (0, 2\sqrt{\log n})\) we have
\[
N(\varepsilon, (0, \alpha_n + 1/\log n], d_n) \lesssim \frac{\sqrt{\log n}}{\varepsilon},
\]
and, therefore,
\[
\sup_{\mu_0 \in \ell_2} \int_0^{\text{diam}_n} \sqrt{N(\varepsilon, (0, \alpha_n + 1/\log n], d_n)} \, d\varepsilon \lesssim (\log n)^{1/4} e^{-\left(1/12\right)\sqrt{\log n}} \to 0.
\]

### 3.6.4 Bounds on \(h_n(\alpha)\), variances and distances

In this section we prove a number of auxiliary lemmas used in the preceding. The first one is about the behavior of the function \(h_n\) in a neighborhood of \(\alpha_n\) and \(\overline{\alpha}_n\).

**Lemma 3.7.** The function \(h_n\) satisfies the following bounds:

\[
h_n(\alpha) \gtrsim h_n(\overline{\alpha}_n), \quad \text{for } \alpha \in \left[ \overline{\alpha}_n - \frac{1}{\log n}, \alpha_n \right] \text{ and } n \geq e^4,
\]
\[
h_n(\alpha) \lesssim h_n(\alpha_n), \quad \text{for } \alpha \in \left[ \alpha_n, \alpha_n + \frac{1}{\log n} \right] \text{ and } n \geq e^2.
\]
Therefore, \( S_n(\alpha) \) yields the desired result.

**Proof.** We provide a detailed proof of the first inequality, the second one can be proved using similar arguments.

Let

\[
S_n(\alpha) = \sum_{i=1}^{\infty} \frac{n^2 i^{1+2\alpha} \mu_{0,i}^2 \log i}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2}
\]

be the sum in the definition of \( h_n \). Recall that \( C^{-1} i^{-p} \leq \kappa_i \leq C i^{-p} \). Splitting the sum into two parts we get, for \( \alpha \in [\bar{\alpha}_n - 1/\log n, \bar{\alpha}_n] \),

\[
4S_n(\alpha) \geq \sum_{i \leq n^{1/(1+2\alpha+2p)}} i^{1+2\bar{\alpha}_n-2/\log n} \mu_{0,i}^2 \log i + \frac{n^2}{C^4} \sum_{i > n^{1/(1+2\alpha+2p)}} i^{-1-2\bar{\alpha}_n-4p} \mu_{0,i}^2 \log i.
\]

In the first sum \( i^{-2/\log n} \) is bounded below by \( \exp(-2) \). Moreover, for \( i \in [n^{1/(1+2\bar{\alpha}_n+2p)}, n^{1/(1+2\alpha+2p)}] \), we have the inequality

\[
i^{1+2\bar{\alpha}_n} \mu_{0,i}^2 \log i \geq n^2 i^{-1-2\bar{\alpha}_n-4p} \mu_{0,i}^2 \log i.
\]

Therefore, \( S_n(\alpha) \) can be bounded from below by a constant times

\[
\sum_{i \leq n^{1/(1+2\bar{\alpha}_n+2p)}} i^{1+2\bar{\alpha}_n} \mu_{0,i}^2 \log i + \frac{n^2}{C^4} \sum_{i > n^{1/(1+2\alpha+2p)}} i^{-1-2\bar{\alpha}_n-4p} \mu_{0,i}^2 \log i \geq \sum_{i \leq n^{1/(1+2\bar{\alpha}_n+2p)}} \frac{n^2 i^{1+2\bar{\alpha}_n} \mu_{0,i}^2 \log i}{(i^{1+2\bar{\alpha}_n} \kappa_i^{-2} + n)^2} + \frac{1}{C^8} \sum_{i > n^{1/(1+2\alpha+2p)}} \frac{n^2 i^{1+2\bar{\alpha}_n} \mu_{0,i}^2 \log i}{(i^{1+2\bar{\alpha}_n} \kappa_i^{-2} + n)^2}.
\]

Hence, we have \( S_n(\alpha) \gtrsim S_n(\bar{\alpha}_n) \) for \( \alpha \in [\bar{\alpha}_n - 1/\log n, \bar{\alpha}_n] \).

Next note that for \( n \geq e^4 \) we have \( 2(1+2\bar{\alpha}_n - 2/\log n + 2p) \geq 1 + 2\bar{\alpha}_n + 2p \). Moreover, \( n^{-1/(1+2\bar{\alpha}_n-2/\log n+2p)} \gtrsim n^{-1/(1+2\bar{\alpha}_n+2p)} \). Therefore,

\[
\frac{1+2\alpha+2p}{n^{1/(1+2\alpha+2p)}} \log n \gtrsim \frac{1+2\bar{\alpha}_n+2p}{n^{1/(1+2\bar{\alpha}_n+2p)}} \log n
\]

for \( \alpha \in [\bar{\alpha}_n - 1/\log n, \bar{\alpha}_n] \) and for \( n \geq e^4 \). Combining this with the inequality for \( S_n(\alpha) \) yields the desired result. \( \square \)

Next we present two results on variances involving the random function \( \mathbb{M}_n \).

**Lemma 3.8.** For any \( \alpha > 0 \),

\[
\text{var}_{\mu_0} \left( \frac{1+2\alpha+2p}{n^{1/(1+2\alpha+2p)}} \mathbb{M}_n(\alpha) \right) \lesssim n^{-1/(1+2\alpha+2p)} (\log n)^2 (1 + h_n(\alpha)).
\]

**Proof.** The random variables \( Y_i^2 \) are independent and \( \text{var}_{\mu_0} Y_i^2 = 2/n^2 + 4k_i^2 \mu_{0,i}^2/n \), hence the variance in the statement of the lemma is equal to

\[
\frac{2n^2 (1+2\alpha+2p)^2}{n^{2/(1+2\alpha+2p)}} \sum_{i=1}^{\infty} \frac{i^{2+4\alpha} \kappa_i^{-4}(\log i)^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^4} + \frac{4n^2 (1+2\alpha+2p)^2}{n^{2/(1+2\alpha+2p)}} \sum_{i=1}^{\infty} \frac{i^{2+4\alpha} \kappa_i^{-2}(\log i)^2 \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^4}.
\]

(3.21)
3.6. Proof of Theorem 3.2

Recall that $C^{-1}i^{-p} \leq \kappa_i \leq Ci^{-p}$. By Lemma A.12, the first term is bounded by

$$\frac{2C^{12}n(1 + 2\alpha + 2p)\log n}{n^{2/(1+2\alpha+2p)}} \sum_{i=1}^{\infty} \frac{i^{1+2\alpha+2p}\log i}{(i^{1+2\alpha+2p} + n)^2} \leq \frac{2C^{12}(1 + 2\alpha + 2p)\log n}{n^{2/(1+2\alpha+2p)}} \sum_{i=1}^{\infty} \frac{n\log i}{i^{1+2\alpha+2p} + n}.$$  

Lemma A.9.(iii) further bounds the right-hand side of the above display by a multiple of $n^{-1/(1+2\alpha+2p)}(\log n)^2$, uniformly for $\alpha > c$, where $c > 0$ is an arbitrary constant. For $\alpha \leq c$ we get the same bound by applying Lemma A.10 (with $m = 2, l = 4, r = 1 + 2\alpha + 2p, r_0 = 1 + 2c + 2p$, and $s = 2r$) to

$$\frac{2C^{12}n^2(1 + 2\alpha + 2p)^2}{n^{2/(1+2\alpha+2p)}} \sum_{i=1}^{\infty} \frac{i^{2+4\alpha+4p}(\log i)^2}{(i^{1+2\alpha+2p} + n)^4}$$

which is an upper bound for the first term in (3.21). By Lemma A.12, the second term in (3.21) is bounded by

$$4C^6n^{-2/(1+2\alpha+2p)}(1 + 2\alpha + 2p)(\log n)\sum_{i=1}^{\infty} \frac{n^{2\mu_{0,i}}\log i}{(i^{1+2\alpha\kappa_i^{-2}} + n)^2} = 4C^6n^{-1/(1+2\alpha+2p)}(\log n)^2h_n(\alpha).$$

Combining the upper bounds for the two terms we arrive at the assertion of the lemma. \qed

**Lemma 3.9.** For any $0 < \alpha_1 < \alpha_2 < \infty$ we have that

$$\var_{\mu_0}\left(\frac{(1 + 2\alpha_1 + 2p)M_n(\alpha_1)}{n^{1/(1+2\alpha_1+2p)}} - \frac{(1 + 2\alpha_2 + 2p)M_n(\alpha_2)}{n^{1/(1+2\alpha_2+2p)}}\right) \lesssim (\alpha_1 - \alpha_2)^2(\log n)^4 \sup_{\alpha \in [\alpha_1, \alpha_2]} n^{-1/(1+2\alpha+2p)}(1 + h_n(\alpha)),$$

with a constant that does not depend on $\alpha$ and $\mu_0$.

**Proof.** The variance we have to bound can be written as

$$n^4 \sum_{i=1}^{\infty} (f_i(\alpha_1) - f_i(\alpha_2))^2 \var_{\mu_0} Y_i^2,$$

where $f_i(\alpha) = (1 + 2\alpha + 2p)i^{1+2\alpha\kappa_i^{-2}}n^{-1/(1+2\alpha+2p)}(i^{1+2\alpha\kappa_i^{-2}} + n)^{-2}$. Note that by the Cauchy–Schwarz inequality and by the properties of the integral

$$\sum_{i=1}^{\infty} (f_i(\alpha_1) - f_i(\alpha_2))^2 \leq \sum_{i=1}^{\infty} \left( \int_{\alpha_1}^{\alpha_2} |f_i'(\alpha)| \, d\alpha \right)^2 \leq \sum_{i=1}^{\infty} |\alpha_1 - \alpha_2| \int_{\alpha_1}^{\alpha_2} |f_i'(\alpha)|^2 \, d\alpha \leq (\alpha_1 - \alpha_2)^2 \sup_{\alpha \in [\alpha_1, \alpha_2]} \sum_{i=1}^{\infty} |f_i'(\alpha)|^2 \tag{3.22}.$$


For the derivative of \( f_i \) we have
\[
|f'_i(\alpha)| = \left| 2f_i(\alpha) \left( \frac{1}{1 + 2\alpha + 2p} + \log i + \frac{\log n}{(1 + 2\alpha + 2p)^2} - \frac{2i^{1+2\alpha} \kappa_i^{-2} \log i}{i^{1+2\alpha} \kappa_i^{-2} + n} \right) \right|
\leq 8f_i(\alpha) \left( \log i + \frac{\log n}{(1 + 2\alpha + 2p)^2} \right),
\]
hence the variance is bounded by a constant times
\[
(\alpha_1 - \alpha_2)^2 n^4 \sup_{\alpha \in [\alpha_1, \alpha_2]} \left( 1 + 2\alpha + 2p \right)^2 \sum_{i=1}^{\infty} i^{2+4\alpha} \kappa_i^{-4} \left( \log i \right)^2 \left( \log i + (\log n)/(1 + 2\alpha + 2p)^2 \right)^2 \left( 1 + 2\alpha + 2p \right)^2 \frac{n^2/(1+2\alpha+2p) (i^{1+2\alpha} \kappa_i^{-2} + n)^4}{\var_{\mu_0} Y_i^2}.
\]
Since \( \var_{\mu_0} Y_i^2 = 2/n^2 + 4\kappa_i^2 \mu_0^2/n \), it suffices to show that both
\[
n^2 \sup_{\alpha \in [\alpha_1, \alpha_2]} \left( 1 + 2\alpha + 2p \right)^2 \sum_{i=1}^{\infty} i^{2+4\alpha} \kappa_i^{-4} \left( \log i \right)^2 \left( \log i + (\log n)/(1 + 2\alpha + 2p)^2 \right)^2 \left( 1 + 2\alpha + 2p \right)^2 \frac{n^2/(1+2\alpha+2p) (i^{1+2\alpha} \kappa_i^{-2} + n)^4}{\var_{\mu_0} Y_i^2} = 2/n^2 + 4\kappa_i^2 \mu_0^2/n, \quad (3.23)
\]
and
\[
n^3 \sup_{\alpha \in [\alpha_1, \alpha_2]} \left( 1 + 2\alpha + 2p \right)^2 \sum_{i=1}^{\infty} i^{2+4\alpha} \kappa_i^{-2} \left( \log i \right)^2 \mu_0^2 \kappa_i^{-1} \left( \log i + (\log n)/(1 + 2\alpha + 2p)^2 \right)^2 \left( 1 + 2\alpha + 2p \right)^2 \frac{n^2/(1+2\alpha+2p) (i^{1+2\alpha} \kappa_i^{-2} + n)^4}{\var_{\mu_0} Y_i^2} = 2/n^2 + 4\kappa_i^2 \mu_0^2/n, \quad (3.24)
\]
are bounded by a constant times \((\log n)^4 \sup_{\alpha \in [\alpha_1, \alpha_2]} n^{-1/(1+2\alpha+2p)} (1 + h_n(\alpha))\).

By applying Lemma A.12 twice (once the first statement with \( r = 1 + 2\alpha + 2p \) and \( m = 1 \) and once the second one with the same \( r \) and \( m = 3 \) and \( \xi = 1 \)) the expression in (3.24) is seen to be bounded above by a constant times
\[
C^{16} (\log n)^3 \sup_{\alpha \in [\alpha_1, \alpha_2]} \left( n^{-2/(1+2\alpha+2p)} (1 + 2\alpha + 2p) \sum_{i=1}^{\infty} n^{2i^{1+2\alpha} \mu_0^2 \kappa_i^{-1} \log i} / (i^{1+2\alpha} \kappa_i^{-2} + n)^2 \right).
\]
The expression in the parentheses equals \( h_n(\alpha) n^{-1/(1+2\alpha+2p)} \log n \). Now fix \( c > 0 \). Again, applying Lemma A.12 twice implies that we get that (3.23) is bounded above by
\[
C^{12} (\log n)^3 \sup_{\alpha \in [\alpha_1, \alpha_2]} \left( 2n^{-2/(1+2\alpha+2p)} (1 + 2\alpha + 2p) \frac{\sum_{i=1}^{\infty} n^{i^{1+2\alpha+2p} \log i}} {i^{1+2\alpha+2p} + n} \right).
\]
Using the inequality \( x/(x + y) \leq 1 \) and Lemma A.9(iii), the expression in the parenthesis can be bounded by a constant times \( n^{-1/(1+2\alpha+2p)} \log n \) for \( \alpha > c \). For \( \alpha \leq c \), Lemma A.10 (with \( m = 2 \) or \( m = 4 \), \( l = 4 \), \( r = 1 + 2\alpha + 2p \), \( r_0 = 1 + 2c + 2p \), and \( s = 2r \)) gives the same bound (or even a better one) for (3.23). The proof is completed by combining the obtained bounds. \(\square\)
3.7 Proof of Theorem 3.3

As before, we only present the details of the proof for the Sobolev case \( \mu_0 \in S^\beta \). The analytic case can be dealt with similarly (see also Section 3.7.4). Again, recall that \( C^{-1}i^{-p} \leq \kappa_i \leq Ci^{-p} \) for \( i \geq 1 \).

By Markov’s inequality and Theorem 3.2,

\[
\sup_{\|\mu_0\|_\beta \leq R} \mathbb{E}_{\mu_0} \Pi_\alpha_n \left( \|\mu - \mu_0\| \geq M_n \varepsilon_n \bigm| Y \right) \leq \frac{1}{M_n^2 \varepsilon_n^2} \sup_{\|\mu_0\|_\beta \leq R} \mathbb{E}_{\mu_0} \sup_{\alpha_n \leq \alpha \leq \alpha_n \wedge \log n} R_n(\alpha) + o(1),
\]

where

\[
R_n(\alpha) = \int \|\mu - \mu_0\|^2 \Pi_\alpha(d\mu \mid Y)
\]

is the posterior risk. We will show that for \( \varepsilon_n = n^{-\beta/(1+2\beta+2p)}(\log n)^{3/2}(\log \log n)^{1/2} \) and arbitrary \( M_n \to \infty \), the first term on the right of (3.25) vanishes as \( n \to \infty \). Note that by the explicit posterior computation (3.5), we have

\[
R_n(\alpha) = \sum_{i=1}^{\infty} (\hat{\mu}_{\alpha,i} - \mu_{0,i})^2 + \sum_{i=1}^{\infty} \frac{\kappa_i^{-2}}{i^{1+2\alpha} \kappa_i^{-2} + n},
\]

where \( \hat{\mu}_{\alpha,i} = n\kappa_i^{-1}(i^{1+2\alpha} \kappa_i^{-2} + n)^{-1} Y_i \) is the \( i \)th coefficient of the posterior mean.

3.7.1 Bound for the expected posterior risk

In this section we prove that

\[
\sup_{\|\mu_0\|_\beta \leq R} \sup_{\alpha_n \leq \alpha \leq \alpha_n \wedge \log n} \mathbb{E}_{\mu_0} R_n(\alpha) = O(\varepsilon_n^2).
\]

To this end we define the sets

\[
P_n = \{\mu_0: \|\mu_0\|_\beta \leq R, \mu_{0,i} \neq 0 \text{ for some } i \geq 2\},
\]

\[
Q_n = \{\mu_0: \|\mu_0\|_\beta \leq R, \mu_{0,i} = 0 \text{ for all } i \geq 2\}.
\]

By Lemma 3.1.(iv), we have that \( \bar{\alpha}_n < \log n \) if \( \mu_0 \in P_n \). Hence, it suffices to show that

\[
\sup_{\mu_0 \in P_n} \sup_{\alpha_n \leq \alpha \leq \bar{\alpha}_n} \mathbb{E}_{\mu_0} R_n(\alpha) = O(\varepsilon_n^2), \quad \sup_{\mu_0 \in Q_n} \sup_{\alpha_n \leq \alpha \leq \log n} \mathbb{E}_{\mu_0} R_n(\alpha) = O(\varepsilon_n^2).
\]

The second term of (3.26) is deterministic. The expectation of the first term can be split into square bias and variance terms. We find that the expectation of (3.26) is given by

\[
\sum_{i=1}^{\infty} \frac{i^{2+4\alpha} \kappa_i^{-4} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} + \sum_{i=1}^{\infty} \frac{n\kappa_i^{-2}}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} + \sum_{i=1}^{\infty} \frac{\kappa_i^{-2}}{i^{1+2\alpha} \kappa_i^{-2} + n}.
\]

(3.27)
Note that the second and third terms in (3.27) are independent of $\mu_0$ and both bounded by

$$C^4 \sum_{i=1}^{\infty} \frac{i^{2p}}{i^{1+2\alpha+2p} + n}.$$ 

By Lemma A.10 (with $m = 0, l = 1, r = 1 + 2\alpha + 2p$ and $s = 2p$) this is for $\alpha \geq \frac{c}{\alpha}$ further bounded by

$$n - \frac{2\alpha}{1+2\alpha+2p} \leq n - \frac{2\alpha_n}{1+2\alpha_n+2p}.$$ 

In view of Lemma 3.1(i), the right-hand side is bounded by a constant times $n^{-2\beta/(1+2\beta+2p)}$ for large $n$. Indeed, $\beta - c_0/\log n < \alpha_n < \alpha$ and monotonicity of $x/(x+c)$ yield:

$$n - \frac{2\alpha_n}{1+2\alpha_n+2p} \leq n - \frac{2\beta - c_0/\log n}{1+2\beta - c_0/\log n + 2p} \leq n - \frac{2\beta - c_0/\log n}{1+2\beta - c_0/\log n + 2p} \leq e^{4c_0} n^{-2\beta/(1+2\beta+2p)}. \tag{3.28}$$

It remains to consider the first sum in (3.27). The supremum over $Q_n$ is easily dealt with. If $\mu_0 \in Q_n$, the first sum in (3.27) equals $\kappa_1^{-4} \mu_{0,1}^2 / (\kappa_1^{-2} + n^2)$, whence

$$\sup_{\mu_0 \in Q_n} \sup_{\alpha \geq \log n} \sum_{i=1}^{\infty} \frac{i^{2+4\alpha} \kappa_i^{-4} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq \frac{C^4 R^2}{n^2}.$$ 

For the supremum over $P_n$ we divide the first sum in (3.27) into three parts and show that each of the parts has the stated order. First we note that

$$\sum_{i > n^{1/(1+2\beta+2p)}} \frac{i^{2+4\alpha} \kappa_i^{-4} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq \sum_{i > n^{1/(1+2\beta+2p)}} \mu_{0,i}^2 \leq \|\mu_0\|^2 n^{-2\beta/(1+2\beta+2p)}. \tag{3.29}$$

Next, observe that elementary calculus shows that for $\alpha > 0$ and $n \geq e$, the maximum of the function $i \mapsto i^{1+2\alpha+4p} / \log i$ over the interval $[2, n^{1/(1+2\alpha+2p)}]$ is taken at $i = n^{1/(1+2\alpha+2p)}$. It follows that for $\alpha > 0$,

$$\sum_{i \leq n^{1/(1+2\alpha+2p)}} \frac{i^{2+4\alpha} \kappa_i^{-4} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq \frac{C^4 \mu_{0,1}^2}{n^2} + \frac{C^4}{n^2} \sum_{2 \leq i \leq n^{1/(1+2\alpha+2p)}} \frac{(i^{1+2\alpha+4p}/\log i) n^{2+4\alpha} \mu_{0,i}^2 \log i}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2}$$

$$\leq \frac{C^4 \mu_{0,1}^2}{n^2} + C^4 n^{-2+2\alpha \over 1+\alpha} h_n(\alpha).$$

Hence, since $x \mapsto x/(x+c)$ is increasing for every $c > 0$, we have

$$\sup_{\mu_0 \in P_n} \sup_{\alpha \geq \log n} \sum_{i \leq n^{1/(1+2\alpha+2p)}} \frac{i^{2+4\alpha} \kappa_i^{-4} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq \frac{R}{n^2} + n^{-2+2\alpha \over 1+\alpha} h_n(\alpha) = \frac{R}{n^2} + L n^{-2+2\alpha \over 1+2\alpha+2p} (\log n)^2.$$
By (3.28) this is further bounded by a constant times \( n^{-2\beta/(1+2\beta+2p)}(\log n)^2 \).

To complete the proof, we now deal with the terms between \( 2 \sqrt[2]{1/(1+2\alpha_n+2p)} \) and \( n^{1/(1+2\beta+2p)} \). (For \( \alpha_n \geq (\log n)/(2 \log 2) - 1/2 - p \) the expression \( n^{1/(1+2\alpha_n+2p)} \) is not greater than 2.) Let \( J = J(n) \) be the smallest integer such that \( (\alpha_n \land ( (\log n)/(2 \log 2 - 1/2 - p) ))/(1 + 1/(\log n) \leq \beta \). One can see that \( J \) is bounded above by a multiple of \( (\log n)(\log \log n) \) for any positive \( \beta \). We partition the summation range under consideration into \( J \) pieces using the auxiliary numbers

\[
b_j = 1 + 2 \alpha_n \land ( (\log n)/(2 \log 2 - 1/2 - p)) + 2p, \quad j = 0, \ldots, J.
\]

Note that the sequence \( b_j \) is decreasing. Now we have

\[
\sum_{i = 2 \sqrt[2]{1/(1+2\alpha_n+2p)}}^{n^{1/(1+2\beta+2p)}} i^{2+4\alpha_n \land -4\mu_{0,i}} (i^{1+2\alpha_n \land -2 + n})^2 \leq \sum_{j = 0}^{J-1} \sum_{i = n^{1/b_j}}^{n^{1/b_j+1}} \mu_{0,i}^2 \leq 4 \sum_{j = 0}^{J-1} \sum_{i = n^{1/b_j}}^{n^{1/b_j+1}} (n^{b_j} \mu_{0,i}^2) (i^{b_j + 1} + n)^2,
\]

and the upper bound is uniform in \( \alpha \). Since \( (b_j - b_{j+1}) \log n = b_{j+1} - 1 - 2p \), it holds for \( n^{1/b_j} \leq i \leq n^{1/b_{j+1}} \) that \( i^{b_j - b_{j+1}} \leq n^{1/\log n} = e \). On the same interval \( i^{2p} \) is bounded by \( n^{2p/b_{j+1}} \). Therefore, the right-hand side of the preceding display is further bounded by a constant times

\[
\sum_{j = 0}^{J-1} \sum_{i = n^{1/b_j}}^{n^{1/b_j+1}} n^{b_{j+1}} \mu_{0,i}^2 \log i (i^{b_{j+1} + 1} + n)^2 \leq C^4 \sum_{j = 0}^{J-1} n^{2p/b_{j+1} - 1} \sum_{i = n^{1/b_j}}^{n^{1/b_{j+1}}} n^{2p/b_{j+1} - 2p} \mu_{0,i}^2 \log i (i^{b_{j+1} + 1} + n)^2 \leq C^4 (\log n) \sum_{j = 0}^{J-1} n^{(1+2p)/b_{j+1} - 1} \log \frac{(\alpha_n \land ( (\log n)/(2 \log 2 - 1/2 - p))}{(1 + 1/(\log n))^{j+1}} n^{1/b_{j+1}} \log n \frac{(b_{j+1} - 1 - 2/p)}{b_{j+1}} \leq C^4 (\log n) \sum_{j = 0}^{J-1} n^{(1+2p)/b_{j+1} - 1} \log n \frac{(b_{j+1} - 1 - 2/p)}{b_{j+1}} \leq (\log n) n^{-2\beta/(1+2\beta+2p)} \sum_{j = 0}^{J-1} h_n(b_{j+1}/2 - 1/2 - p).
\]

In the last step we used the fact that, by construction, \( b_j/2 - 1/2 - p \geq \beta/(1 + 1/(\log n)) \). It follows from the definition of \( \alpha_n \) and \( b_j \) that \( h_n(b_{j+1}/2 - 1/2 - p) \) is bounded above by \( L(\log n)^2 \) for every \( j \leq J - 1 \), and we recall that \( J = J(n) \) is bounded above by a multiple of \( (\log n)(\log \log n) \). Finally, we note that

\[
n^{-2\beta/(1+2\beta+2p)} \leq c n^{-2\beta/(1+2\beta+2p)}.n^{-2\beta/(1+2\beta+2p)}(\log n)^3(\log \log n), \text{ in the appropriate uniform sense over } P_n.
\]
3.7.2 Bound for the centered posterior risk

To complete the proof of Theorem 3.3, we show in this section that we also have

\[
\sup_{\|\mu\|_2 \leq R} \mathbb{E}_{\mu_0} \sup_{\alpha \in [\alpha_n, \alpha_n^\prime] \land \log n} \left| \sum_{i=1}^{\infty} (\hat{\mu}_{\alpha,i} - \mu_{0,i})^2 - \mathbb{E}_{\mu_0} \sum_{i=1}^{\infty} (\hat{\mu}_{\alpha,i} - \mu_{0,i})^2 \right| = O(\varepsilon_n^2),
\]

for \( \varepsilon_n = n^{-\beta/(1+2\beta+2p)}(\log n)^{3/2}(\log \log n)^{1/2} \). Using the explicit expression for the posterior mean \( \hat{\mu}_{\alpha,i} \), we see that the random variable in the supremum is the absolute value of \( \mathcal{V}(\alpha)/n - 2\mathcal{W}(\alpha)/\sqrt{n} \), where

\[
\mathcal{V}(\alpha) = \sum_{i=1}^{\infty} \frac{n^2\kappa_i^{-2}}{(1+2\alpha\kappa_i^{-2} + n)^2} (Z_i^2 - 1), \quad \mathcal{W}(\alpha) = \sum_{i=1}^{\infty} \frac{n \kappa_i^{-3}}{(1+2\alpha\kappa_i^{-2} + n)^2} \mu_{0,i} Z_i.
\]

We deal with the two processes separately.

For the process \( \mathcal{V} \), Corollary 2.2.5 in [104] implies that

\[
\mathbb{E}_{\mu_0} \sup_{\alpha \in [\alpha_n, \infty)} \sqrt{\sum_{\alpha \in [\alpha_n, \infty)} \sqrt{\var_\mu \mathcal{V}(\alpha) + \int_{0}^{\text{diam}_n} \sqrt{N(\varepsilon, [\alpha_n, \infty)}, d_\varepsilon} d\alpha},
\]

where \( \text{diam}_n^2(\alpha_1, \alpha_2) = \text{var}(\mathcal{V}(\alpha_1) - \mathcal{V}(\alpha_2)) \) and diam\(_n\) is the \( d_n \)-diameter of \([\alpha_n, \infty]\). Now the variance of \( \mathcal{V}(\alpha) \) is equal to

\[
\var_\mu \mathcal{V}(\alpha) = 2n^4 \sum_{i=1}^{\infty} \frac{\kappa_i^{-4}}{(1+2\alpha\kappa_i^{-2} + n)^4} \leq 2C^4 n^4 \sum_{i=1}^{\infty} \frac{\kappa_i^{-4}}{(1+2\alpha+2p + n)^4}.
\]

since \( \var Z_i^2 = 2 \). Using Lemma A.10 (with \( m = 0, l = 4, r = 1 + 2\alpha + 2p \) and \( s = 4p \)), we can conclude that the variance of \( \mathcal{V}(\alpha) \) is bounded above by a multiple of \( n^{(1+4p)/(1+2\alpha+2p)} \).

It follows that the diameter of the interval \( \text{diam}_n \lesssim n^{(1+4p)/(1+2\alpha+2p)} \). To compute the covering number of the interval \([\alpha_n, \alpha_n^\prime]\), we first note that for \( 0 \leq \alpha_1 < \alpha_2 \),

\[
\var_\mu (\mathcal{V}(\alpha_1) - \mathcal{V}(\alpha_2)) \leq 2 \sum_{i=2}^{\infty} \left( \frac{n^2\kappa_i^{-2}}{(1+2\alpha_1\kappa_i^{-2} + n)^2} - \frac{n^2\kappa_i^{-2}}{(1+2\alpha_2\kappa_i^{-2} + n)^2} \right)^2 \var Z_i^2.
\]

Hence, for \( \varepsilon > 0 \), a single \( \varepsilon \)-ball covers the whole interval \([C’ \log(n/\varepsilon), \infty)\) for some constant \( C’ > 0 \). By Lemma 3.10, the distance \( d_n(\alpha_1, \alpha_2) \) is bounded above by a multiple of \( |\alpha_1 - \alpha_2|n^{(1+4p)/(2+4\alpha_n+4p)}(\log n) \). Therefore, we find that the covering number of the interval \([\alpha_n, C’ \log(n/\varepsilon)]\) relative to the metric \( d_n \) is bounded above by a multiple of \((\log n)n^{(1+4p)/(2+4\alpha_n+4p)}(\log(n/\varepsilon))/\varepsilon \). Combining everything, we see that

\[
\mathbb{E}_{\mu_0} \sup_{\alpha \in [\alpha_n, \infty)} |\mathcal{V}(\alpha)| \lesssim n^{\frac{1+4p-n}{2+4\alpha_n+4p}}(\log n).
\]
By the fact that \( x \mapsto x/(x + c) \) is increasing and Lemma 3.1.(i) (cf. (3.28)), the right-hand side divided by \( n \) is bounded by

\[
n^{- \frac{2c}{1 + 2\alpha + 2p}} (\log n) \lesssim n^{- \frac{2\beta}{1 + 2\beta + 2p}} (\log n).
\]

It remains to deal with the process \( \mathbb{W} \). The basic line of reasoning is the same as followed above for \( \mathbb{V} \). An essential difference, however, is the derivation of a bound for the variance of \( \mathbb{W} \), of which we provide the details. The rest of the proof is left to the reader. The variance \( \text{var}_{\mu_0} (\mathbb{W}(\alpha)/\sqrt{n}) \) is given by

\[
\text{var}_{\mu_0} \left( \frac{\mathbb{W}(\alpha)}{\sqrt{n}} \right) = \sum_{i=1}^{\infty} \frac{n i^{2+4\alpha} \kappa_i^{-6} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^4}.
\]

We show that uniformly for \( \alpha \in [\alpha_n, \overline{\alpha}_n] \), this variance is bounded above by a constant (which depends only on \( \|\mu_0\|_\beta \)) times \( n^{-1/(1+2\alpha+2p)} (\log n)^2 \).

For the sum over \( i \leq n^{1/(1+2\alpha+2p)} \) we have

\[
\sum_{i \leq n^{1/(1+2\alpha+2p)}} \frac{n i^{2+4\alpha} \kappa_i^{-6} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^4} \leq C^6 \mu_{0,1}^2 + C^6 \frac{n^3}{n^3} \sum_{2 \leq i \leq n^{1/(1+2\alpha+2p)}} \frac{n^2 i^{1+2\alpha+6p} (\log i)^{-1} i^{1+2\alpha} \mu_{0,i}^2 \log i}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2}
\]

\[
\leq C^6 \|\mu_0\|_\beta^2 + C^6 (1 + 2\alpha + 2p) \frac{n^3}{n^3} \sum_{i \leq n^{1/(1+2\alpha+2p)}} \frac{n^2 i^{1+2\alpha} \mu_{0,i}^2 \log i}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq C^6 \|\mu_0\|_\beta^2 + C^6 n^{-1+4\alpha+2p} h_n(\alpha).
\]

We have used again the fact that on the range \( i \leq n^{1/(1+2\alpha+2p)} \), the quantity \( i^{1+2\alpha+6p} (\log i)^{-1} \) is maximal for the largest \( i \). Now the function \( x \mapsto -(1 + 2\alpha)/(x + c) \) is decreasing on \((0, \infty)\) for any \( c > 1/2 \). Moreover \( h_n(\alpha) \leq L(\log n)^2 \) for any \( \alpha \leq \overline{\alpha}_n \), thus the preceding display is bounded above by a multiple of \( n^{-1+4\alpha+2p} (\log n)^2 \). Using Lemma 3.1.(i) this is further bounded by a constant times \( n^{-1+4\beta+2p} (\log n)^2 \) (cf. (3.28)).

Next we consider sum over the range \( i > n^{1/(1+2\alpha+2p)} \). We distinguish two cases according to the value of \( \alpha \). First suppose that \( 1 + 2\alpha > 2p \). Then \( i^{-1+2\alpha+2p} (\log i)^{-1} \) is decreasing in \( i \), hence

\[
\sum_{i > n^{1/(1+2\alpha+2p)}} \frac{n i^{2+4\alpha} \kappa_i^{-6} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^4} \leq \sum_{i > n^{1/(1+2\alpha+2p)}} \frac{n \kappa_i^{-2} \mu_{0,i}^2}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq C^2 \frac{n^3}{n^3} \sum_{i > n^{1/(1+2\alpha+2p)}} \frac{n^2 i^{-1+2\alpha+2p} (\log i)^{-1} i^{1+2\alpha} \mu_{0,i}^2 \log i}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq C^2 (1 + 2\alpha + 2p) \frac{n^{4p}}{n^{4p+2\alpha+2p} (\log n)} \sum_{i > n^{1/(1+2\alpha+2p)}} \frac{n^2 i^{1+2\alpha} \mu_{0,i}^2 \log i}{(i^{1+2\alpha} \kappa_i^{-2} + n)^2} \leq C^2 n^{-1+4\alpha+2p} h_n(\alpha).
\]
As above, this is further bounded by a multiple of the desired rate. If $1 + 2\alpha < 2p$, then
\[
\sum_{i > n^{1/(1 + 2\alpha + 2p)}} n^{2 + 4\alpha} \kappa_i^{-6} \mu_0^2 \frac{1}{(i + 2\alpha \kappa_i^2 + n)^4} \leq n C^2 \sum_{i > n^{1/(1 + 2\alpha + 2p)}} i^{-2 - 4\alpha - 2\beta} i^{2\beta} \mu_0^2
\]
\[
\leq C^2 \|\mu_0\|_2^2 n^{2p - 2\beta} n^{1 + 4\beta - 1}. 
\]
Since $\alpha_n \geq \beta - c_0 / \log n$, we have $1 + 2\alpha > 2\beta$ for large enough $n$, for any $\alpha \in [\alpha_n, \overline{\alpha}_n]$. Since we have assumed $1 + 2\alpha < 2p$, this implies that $2p > 2\beta$. Therefore, the right-hand side of the preceding display attains its maximum at $\alpha = \overline{\alpha}_n$. Using again that $\alpha_n \geq \beta - c_0 / \log n$, it is straightforward to show that for $\alpha \in [\alpha_n, \overline{\alpha}_n]$, \[
\frac{1}{1 + 2\alpha + 2p} - 1 \leq \frac{1}{n^{2p - 2\beta}} - 1 \leq e^{4c_0} n^{-1 + 4\beta - 2p}.
\]
This completes the proof.

### 3.7.3 Bounds for the semimetrics associated to $\mathcal{V}$ and $\mathcal{W}$

The following lemma is used in Section 3.7.2.

**Lemma 3.10.** For any $\alpha_n \leq \alpha_1 < \alpha_2$ the following inequalities hold:
\[
\text{var}_{\mu_0}(\mathcal{V}(\alpha_1) - \mathcal{V}(\alpha_2)) \lesssim (\alpha_1 - \alpha_2)^2 n \frac{1 + 4p}{1 + 2\alpha_2 + 2p} (\log n)^2,
\]
\[
\text{var}_{\mu_0} \left( \frac{\mathcal{W}(\alpha_1)}{\sqrt{n}} - \frac{\mathcal{W}(\alpha_2)}{\sqrt{n}} \right) \lesssim (\alpha_1 - \alpha_2)^2 n^{-1 + 4\beta} (\log n)^4,
\]
with a constant that does not depend on $\alpha$ and $\mu_0$.

**PROOF.** Recall that $C^{-1} i^{-p} \leq \kappa_i \leq C i^{-p}$. The left-hand side of the first inequality is equal to
\[
n^4 \sum_{i=1}^{\infty} (f_i(\alpha_1) - f_i(\alpha_2))^2 \kappa_i^{-4} \text{var} Z_i^2,
\]
where $f_i(\alpha) = (i^{1 + 2\alpha} \kappa_i^{-2} + n)^{-2}$. The derivative of $f_i$ is given by
\[
f_i'(\alpha) = -4i^{1 + 2\alpha} \kappa_i^{-2} (\log i) / (i^{1 + 2\alpha} \kappa_i^{-2} + n)^3,
\]
hence the preceding display is bounded above by a multiple of
\[
C^{20} (\alpha_1 - \alpha_2)^2 n^4 \sup_{\alpha \in [\alpha_1, \alpha_2]} \sum_{i=1}^{\infty} \frac{i^{2 + 4\alpha + 8p} (\log i)^2}{(i^{1 + 2\alpha + 2p} + n)^6}
\]
\[
\leq C^{20} (\alpha_1 - \alpha_2)^2 n^3 (\log n)^2 \sup_{\alpha \in [\alpha_1, \alpha_2]} \frac{1}{(1 + 2\alpha + 2p)^2} \sum_{i=1}^{\infty} \frac{i^{1 + 2\alpha + 6p}}{(i^{1 + 2\alpha + 2p} + n)^4}
\]
\[
\lesssim (\alpha_1 - \alpha_2)^2 (\log n)^2 \sup_{\alpha \in [\alpha_1, \alpha_2]} n^{1 + 4p},
\]
3.8. Proof of Theorem 3.6

with the help of Lemma A.12 (with \( r = 1 + 2\alpha + 2p \), and \( m = 2 \)), and Lemma A.10 (with \( m = 0, l = 4, r = 1 + 2\alpha + 2p \), and \( s = r + 4p \), (cf. (3.22)). Since \( \alpha \geq \alpha_n \), we get the first assertion of the lemma.

We next consider \( \mathbb{W}/\sqrt{n} \). The left-hand side of the second inequality in the statement of the lemma is equal to

\[
\sum_{i=1}^{\infty} (f_i(\alpha_1) - f_i(\alpha_2))^2 n \mu_{\alpha,i}^2 \var Z_i,
\]

where now \( f_i(\alpha) = i^{1+2\alpha-3} / (i^{1+2\alpha^2} + n)^2 \). The derivative of this \( f_i \) satisfies \( |f_i'(\alpha)| \leq 2(\log i) f_i(\alpha) \), hence we get the upper bound

\[
4C^{14} (\alpha_2 - \alpha_1)^2 \sup_{\alpha \in [\alpha_1,\alpha_2]} \sum_{i=1}^{\infty} n i^{2n+6p} \mu_{\alpha,i}^2 \log^2 i.
\]

The proof is completed by arguing as in the case of the upper bound for the variance of \( \mathbb{W}(\alpha)/\sqrt{n} \) (see page 71).

3.7.4 Proof of Theorem 3.3 for the analytic case

The assertion of Theorem 3.3 in the case of the analytic truth \( \mu_0 \in A^\gamma \) can be proven along the lines of the proof presented above. In view of Lemma 3.1.(ii),

\[
\sqrt{\log n} / (\log \log n) < \alpha_n < \alpha_n,
\]

and whence

\[
n^{-1/2(\alpha_n+2p)} < n^{-1/2(\alpha_n+2p)} \leq n^{1/2(\log n)/(\log \log n)+2p}
\]

\[
= n^{-1}(1+2\log n)/(\log \log n)+2p \leq n^{-1}(\log n)^{(1/2+p)\sqrt{\log n}}.
\]

3.8 Proof of Theorem 3.6

Let \( B(R) \) denote a Sobolev or analytic ball of radius \( R \), and \( \varepsilon_{n,B} \) the corresponding contraction rate. Let \( A_n \) be the event that \( \hat{\alpha}_n \in [\alpha_n, \alpha_n] \). Then with \( \alpha \mapsto \lambda_n(\alpha|Y) \) denoting the posterior Lebesgue density of \( \alpha \), we have

\[
\sup_{\mu_0 \in B(R)} E_{\mu_0} \Pi(||\mu - \mu_0|| \geq M_n \varepsilon_{n,B}|Y)
\]

\[
\leq \sup_{\mu_0 \in B(R)} \mathbb{P}_0(A_n^\varepsilon) + \sup_{\mu_0 \in B(R)} E_{\mu_0} \int_{\Omega_n} \lambda_n(\alpha|Y) d\alpha 1_{A_n}
\]

\[+ \sup_{\mu_0 \in B(R)} E_{\mu_0} \int_{\Omega_n} \lambda_n(\alpha|Y) \Pi_\alpha(\|\mu - \mu_0|| \geq M_n \varepsilon_{n,B}|Y) d\alpha 1_{A_n}. \tag{3.30}
\]

By Theorem 3.2, the first term on the right vanishes as \( n \to \infty \), provided \( l \) and \( L \) in the definitions of \( \alpha_n \) and \( \alpha_n \) are chosen small and large enough, respectively. We will show that the other terms tend to 0 as well.
Observe that \( \lambda_n(\alpha|Y) \propto L_n(\alpha)\lambda(\alpha) \), where \( L_n(\alpha) = \exp(\ell_n(\alpha)) \), for \( \ell_n \) the random function defined by (3.4). In Section 3.6.3, we have shown that on the interval \((0, \alpha_n + 1/\log n)\]

\[
\ell'_n(\alpha) = M_n(\alpha) \gtrsim \frac{n^{1+2\alpha_n+2p} \log n}{1+2\alpha_n + 2p},
\]

on the event \( A_n \). Therefore, on the interval \((0, \alpha_n]\) we have

\[
\ell_n(\alpha) < \ell_n(\alpha_n) \leq \ell_n(\alpha_n + \frac{1}{2\log n}) - \frac{C' n^{1+2\alpha_n+2p}}{1+2\alpha_n + 2p}
\]

for some \( C' > 0 \) and on the interval \([\alpha_n + 1/(2\log n), \alpha_n + 1/\log n]\),

\[
\ell_n(\alpha) \geq \ell_n(\alpha_n + \frac{1}{2\log n})
\]

For the likelihood \( L_n \) we have the corresponding bounds

\[
L_n(\alpha) < \exp\left(-\frac{C' n^{1+2\alpha_n+2p}}{1+2\alpha_n + 2p}\right)L_n(\alpha_n + \frac{1}{2\log n})
\]

for \( \alpha \in (0, \alpha_n] \) and

\[
L_n(\alpha) \geq L_n(\alpha_n + \frac{1}{2\log n})
\]

for \( \alpha \in [\alpha_n + 1/(2\log n), \alpha_n + 1/\log n] \) on the event \( A_n \). Using these estimates for \( L_n \), we obtain the following upper bound for the second term on the right-hand side of (3.30):

\[
E_{\mu_0} \frac{\int_{\alpha_n}^{\alpha_n + 1/\log n} \lambda(\alpha)L_n(\alpha) d\alpha}{\int_{0}^{\infty} \lambda(\alpha)L_n(\alpha) d\alpha} \leq E_{\mu_0} \exp\left(-\frac{C' n^{1+2\alpha_n+2p}}{1+2\alpha_n + 2p}\right) \frac{L_n(\alpha_n + \frac{1}{2\log n}) \int_{0}^{\alpha_n + 1/\log n} \lambda(\alpha) d\alpha}{L_n(\alpha_n + \frac{1}{2\log n}) \int_{\alpha_n + 1/(2\log n)}^{\alpha_n + 1/\log n} \lambda(\alpha) d\alpha}
\]

\[
\leq \exp\left(-\frac{C' n^{1+2\alpha_n+2p}}{1+2\alpha_n + 2p}\right) \left( \int_{\alpha_n + 1/(2\log n)}^{\alpha_n + 1/\log n} \lambda(\alpha) d\alpha \right)^{-1}.
\]

From Lemma 3.1, we know that \( \alpha_n \geq \beta/2 \) for large enough \( n \), hence by Assumption 3.5, Lemma 3.11, and the definition of \( \alpha_n \),

\[
\int_{\alpha_n + 1/(2\log n)}^{\alpha_n + 1/\log n} \lambda(\alpha) d\alpha \geq C_1 (2\log n)^{-C_2} \exp\left(-C_3 \exp\left(\sqrt{\log n}/3\right)\right)
\]

for some \( C_1, C_2, C_3 > 0 \). Therefore, the right-hand side of (3.31) is bounded above by a constant times

\[
\exp\left(-\frac{C' n^{1+2\sqrt{\log n}+2p}}{1+2\sqrt{\log n} + 2p}\right)(\log n)^{C_2} \exp\left(C_3 \exp\left(\sqrt{\log n}/3\right)\right).
\]

It is easy to see that this quantity tends to 0 as \( n \to \infty \).
In bounding the third term on the right-hand side of (3.30) we may replace the supremum over \( B(R) \) by the supremum over the set \( P_n \) defined in Section 3.7.1, since otherwise \( \alpha_n = \infty \). For \( \mu_0 \in P_n \) we have \( \alpha_n \leq \log n/(2 \log 2) - 1/2 - p \) (Lemma 3.1). We then write the third term as

\[
\sup_{\mu_0 \in P_n} \mathbb{E}_{\mu_0} \left( \int_{\Omega_n} \lambda_n(\alpha|Y) \Pi_\alpha(\|\mu - \mu_0\| \geq M_n \varepsilon_{n,B}|Y) \, d\alpha \right) + \int_{\Omega_n} \lambda_n(\alpha|Y) \Pi_\alpha(\|\mu - \mu_0\| \geq M_n \varepsilon_{n,B}|Y) \, d\alpha \right) \mathbb{1}_{A_n}.
\]

(3.32)

The first term in (3.32) is bounded above by

\[
\sup_{\mu_0 \in P_n} \mathbb{E}_{\mu_0} \sup_{\alpha \in [\alpha_n, \infty]} \Pi_\alpha(\|\mu - \mu_0\| \geq M_n \varepsilon_{n,B}|Y).
\]

This goes to zero, as we have shown in the proof of Theorem 3.3. In Section 3.6.1, we have shown that the differentiated log-likelihood function \( M_n(\alpha) \) on the interval \( [\alpha_n, \infty) \) can increase maximally by a multiple of

\[
\frac{n^{1/2 + \alpha_n + 2p} (\log n)^2}{1 + 2\alpha_n + 2p}.
\]

Moreover, in Section 3.6.2, we have shown that for \( \alpha \in [\alpha_n - 1/\log n, \alpha_n] \),

\[
\ell_n(\alpha) = M_n(\alpha) < -M \frac{n^{1/2 + \alpha_n + 2p} (\log n)^3}{1 + 2\alpha_n + 2p}
\]

on the event \( A_n \), and \( M \) can be made arbitrarily large by increasing the constant \( L \) in the definition of \( \alpha_n \). Therefore, the integral of \( M_n(\alpha) \) on \([\alpha_n - 1/\log n, \alpha_n - 1/(2 \log n)] \) is bounded above by

\[
-\frac{M n^{1/2 + \alpha_n + 2p} (\log n)^2}{2(1 + 2\alpha_n + 2p)},
\]

and by choosing a large enough constant \( L \) in the definition of \( \alpha_n \) it holds that for some \( N > 0 \),

\[
\ell_n(\alpha) \leq \ell_n(\alpha_n - \frac{1}{2 \log n}) - N \frac{n^{1/2 + \alpha_n + 2p} (\log n)^2}{1 + 2\alpha_n + 2p}
\]

for \( \alpha \in [\alpha_n, \infty) \), and

\[
\ell_n(\alpha) \geq \ell_n(\alpha_n - \frac{1}{2 \log n})
\]

for \( \alpha \in [\alpha_n - 1/\log n, \alpha_n - 1/(2 \log n)] \). These bounds lead to the following bounds for the likelihood:

\[
L_n(\alpha) \leq L_n(\alpha_n - \frac{1}{2 \log n}) \exp\left(-N \frac{n^{1/2 + \alpha_n + 2p} (\log n)^2}{1 + 2\alpha_n + 2p}\right)
\]

for \( \alpha \in [\alpha_n, \infty) \), and

\[
L_n(\alpha) \geq L_n(\alpha_n - \frac{1}{2 \log n})
\]
for $\alpha \in [\overline{\alpha}_n - 1/(2 \log n), \overline{\alpha}_n - 1/(2 \log n)]$. Similarly to the upper bound for the second term of (3.30), we now write

$$\sup_{\mu_0 \in P_n} E_{\mu_0} \int_{\overline{\alpha}_n}^{\infty} \lambda_n(\alpha | Y) \ d\alpha \leq \sup_{\mu_0 \in P_n} E_{\mu_0} \int_{\overline{\alpha}_n}^{\infty} \lambda(\alpha) L_n(\alpha) \ d\alpha \leq \sup_{\mu_0 \in P_n} \exp \left(-N \frac{n+2\overline{\alpha}_n}{1+2\overline{\alpha}_n+2p} \right) \frac{\int_{\overline{\alpha}_n}^{\infty} \lambda(\alpha) \ d\alpha}{\int_{\overline{\alpha}_n-1/(2 \log n)}^{\infty} \lambda(\alpha) \ d\alpha}.$$  

Since $\overline{\alpha}_n \geq \underline{\alpha}_n \geq \beta/2$ for $n$ large enough, Assumption 3.5 and Lemma 3.11 imply that

$$\int_{\overline{\alpha}_n}^{\infty} \lambda(\alpha) \ d\alpha \leq C_4 \log n \ C_5 \exp \left(C_6 \frac{C_7}{2 \log 2} \right).$$

Since $\overline{\alpha}_n \leq \log n/(2 \log 2) - 1/2 - p$, the right-hand side of the preceding display is bounded above by

$$C_4 \exp \left( -2 C_9 \log 2 \right) \log n \ C_5 \exp \left( C_6 \left( \frac{\log n}{2 \log 2} - \frac{1}{2} - p \right)^C_7 \right),$$

which tends to zero for any fixed constant $C_7$ smaller than 1.

**Lemma 3.11.** Suppose that for $c_1, c_2, c_3 > 0$ and $c_4 \geq 1$, the prior density $\lambda$ satisfies

$$c_4^{-\alpha} \leq \lambda(\alpha) \leq c_4 \alpha^{-c_2} \quad \text{or} \quad c_4^{-\alpha} \leq \lambda(\alpha) \leq c_4 \exp(-c_3 \alpha),$$

for $\alpha \geq c_1$. Then there exist positive constants $C_1, \ldots, C_6$ and $C_7 < 1$ depending on $c_1$ only such that for all $x \geq c_1$, every $\delta_n \to 0$, and $n$ large enough

$$\int_{x+\delta_n}^{x+2\delta_n} \lambda(\alpha) \ d\alpha \geq C_1 \delta_n^{C_2} \exp \left(-C_3 \exp \left( \frac{x}{\delta_n} \right) \right)$$

and

$$\int_{x-\delta_n}^{x} \lambda(\alpha) \ d\alpha \leq C_4 \delta_n^{-C_5} \exp(C_6 x^{C_7}).$$

**Proof.** Consider a polynomially decaying prior. It is easy to see that

$$\int_{x+\delta_n}^{x+2\delta_n} \lambda(\alpha) \ d\alpha \gtrsim \delta_n (x + 2\delta_n)^{-c_2},$$

which is bounded below like in the assertion of the lemma for $x \geq c_1$, large enough $n$ and appropriately chosen constants, where $c_1$ is some fixed positive constant. The left-hand side of the latter inequality is bounded above by a constant times

$$\frac{1}{c_2} x^{1-c_2} \frac{1}{\delta_n} (x - \delta_n)^{c_2} \lesssim \frac{1}{\delta_n} x.$$

Furthermore, one can see that the right-hand side of the preceding display is bounded above like in the assertion of the lemma for appropriately chosen constants.
Next consider exponentially decaying priors. We see that for \( n \) large enough
\[
\int_{x+\delta_n}^{x+2\delta_n} \lambda(\alpha) \, d\alpha \gtrsim \int_{x+\delta_n}^{x+2\delta_n} \exp(-c_3(x + 2\delta_n)) \, d\alpha \geq \delta_n \exp(-2c_3x),
\]
which is bounded below like in the assertion of the lemma for large enough \( n \) and \( C_33 \) and small enough \( C_1 \). The left-hand side of the latter inequality is bounded above by a multiple of
\[
\frac{\int_x^\infty \exp(-c_3\alpha) \, d\alpha}{\int_x^{x-\delta_n} \exp(-c_3(x - \delta_n)) \, d\alpha} \lesssim \exp(-c_3x) \frac{1}{\delta_n} \exp(c_3x) = (\delta_n)^{-1}.
\]
\( \square \)
Chapter 4

Recovery of linear functionals of the parameter

4.1 Introduction

In the nonparametric inverse problem setting, besides the fully nonparametric problem studied in the preceding chapters, the statistician might be also interested in some finite-dimensional aspects of the infinite-dimensional parameter of interest. To this end, we consider estimating a linear functional $L\mu$ of the parameter for $L: H_1 \rightarrow \mathbb{R}$ in the inverse problem model introduced in Chapter 2:

$$Y = K\mu + \frac{1}{\sqrt{n}}Z,$$

(4.1)

where the unknown parameter $\mu$ is an element of a separable Hilbert space $H_1$, and is mapped into another Hilbert space $H_2$ by a known, injective, continuous linear operator $K: H_1 \rightarrow H_2$, and further perturbed by unobserved, scaled Gaussian white noise $Z$.

Estimating a linear functional of the parameter of interest in the inverse problem setting is often studied in the frequentist literature, see, e.g., [14, 43–45, 55, 78]. We study a Bayesian approach to this problem. We again assume a mean-zero Gaussian prior for the parameter $\mu$, and marginalize the posterior distribution for $\mu$ to obtain the posterior distribution of the linear functional $L\mu$.

Since the parameter of interest is infinite-dimensional, and we estimate a finite-dimensional linear functional of it, the problem in this chapter is semiparametric. A continuous, linear functional of the parameter of interest $L\mu$ can be denoted by $\langle \mu, l \rangle_1$ for some $l \in H_1$. Other linear functionals, including a value at a point in case the parameter $\mu$ is a function, can be also represented by some $l$ in a possibly bigger space. This is precisely discussed in Section 4.2, but in here we mention it to indicate that the linear functional $L\mu$ can be associated with an infinite-dimensional parameter $l$. The behavior of this representer $l$ describes the regularity of the linear functional $L\mu$, which plays an important role in the description of the accuracy in estimating the linear functional. Even though the problem is semiparametric, the rates we obtain can be slower than $n^{-1/2}$. This is caused by both the ill-posedness of the inverse problem, and the regularity of the linear functional.
In Chapter 4, we formally introduce the setting of linear functionals, introducing so-called measurable linear functionals relative to the prior, and obtain the posterior distribution. The linear functional problem in the mildly ill-posed setting is presented in Section 4.3, where we present rates of contraction and study the frequentist coverage of credible intervals. In this setting we also obtain a Bernstein–von Mises theorem, and discuss asymptotic efficiency in the linear functional setting. Linear functionals of the initial condition for the heat equation are studied in Section 4.4. We again obtain rates of contraction and study coverage of credible intervals. The extreme ill-posedness of the problem results in worse performance of credible intervals than in the mildly ill-posed setting. In both sections we illustrate our results with the corresponding simulation examples of Chapter 2. Technical proofs are placed in Sections 4.5–4.7.

**Notation**

Throughout the chapter $\langle \cdot , \cdot \rangle_1$ and $\| \cdot \|_1$, and $\langle \cdot , \cdot \rangle_2$ and $\| \cdot \|_2$ denote the inner products and norms of the Hilbert spaces $H_1$ and $H_2$. The adjoint of an operator $A$ between two Hilbert spaces is denoted by $A^T$.

For $\beta \in \mathbb{R}$, the Sobolev norm $\| \mu \|_\beta$ of an element $\mu \in \mathbb{R}^\infty$ and the $\ell_2$-norm $\| \mu \|$ of an element $\mu \in \ell_2$ are defined in a usual way by

$$\| \mu \|_\beta^2 = \sum_{i=1}^{\infty} i^{2\beta} \mu_i^2, \quad \| \mu \|^2 = \sum_{i=1}^{\infty} \mu_i^2,$$

and the corresponding Sobolev space by $S^\beta = \{ \mu \in \mathbb{R}^\infty : \| \mu \|_\beta < \infty \}$.

We say that $S: [0, \infty) \to \mathbb{R}$ is slowly varying if $S(tx)/S(t) \to 1$ as $t \to \infty$, for every $x > 0$.

For two sequences $(a_n)$ and $(b_n)$ of numbers, $a_n \asymp b_n$ means that $|a_n/b_n|$ is bounded away from zero and infinity as $n \to \infty$, $a_n \lessgtr b_n$ means that $a_n/b_n$ is bounded, $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$, and $a_n \ll b_n$ means that $a_n/b_n \to 0$ as $n \to \infty$. For two real numbers $a$ and $b$, we denote by $a \lor b$ their maximum, and by $a \land b$ their minimum. For $m \in \mathbb{R}$, $\sigma > 0$ and $B$ a measurable subset of $\mathbb{R}$ let $N(m, \sigma^2)(B)$ denote $P(X \in B)$ for a $N(m, \sigma^2)$-distributed random variable $X$. 

4.2 Marginal posterior distributions. \(\Pi\)-measurable linear functionals.

We assume a mean-zero Gaussian prior \(N(0, \Lambda)\) for the parameter \(\mu\). Recall (Proposition 2.1 of Chapter 2) the conditional distribution of \(\mu\) given \(Y\) is Gaussian \(N(A_nY, S_n)\) on \(H_1\) where

\[
S_n = \Lambda - A_n(n^{-1}I + K\Lambda K^T)A_n^T,
\]

and \(A_n: H_2 \to H_2\) is the continuous linear operator

\[
A_n = \Lambda^{1/2} \left( \frac{1}{n} I + \Lambda^{1/2} K^T K \Lambda^{1/2} \right)^{-1} \Lambda^{1/2} K^T \left( \frac{1}{n} I + K\Lambda K^T \right)^{-1}.
\]

Next, consider the posterior distribution of a linear functional \(L\mu\) of the parameter. We are not only interested in continuous, linear functionals \(L\mu = \langle \mu, \ell \rangle_1\), for some given \(\ell \in H_1\), but also in certain discontinuous functionals, such as point evaluation in a Hilbert space of functions. The latter entail some technicalities. We consider \emph{measurable linear functionals relative to the prior} \(N(0, \Lambda)\), defined in [92], pages 27–29, as Borel measurable maps \(L: H_1 \to \mathbb{R}\) that are linear on a measurable linear subspace \(H_1 \subset H_1\) such that \(N(0, \Lambda)(H_1) = 1\). This definition is exactly right to make the marginal posterior Gaussian.

**Proposition 4.1.** (Marginal posterior).

If \(\mu\) is \(N(0, \Lambda)\)-distributed and \(Y\) given \(\mu\) is \(N(K\mu, n^{-1}I)\)-distributed, then the conditional distribution of \(L\mu\) given \(Y\) for a \(N(0, \Lambda)\)-measurable linear functional \(L: H_1 \to \mathbb{R}\) is a Gaussian distribution \(N(LA_nY, s_n^2)\) on \(\mathbb{R}\), where

\[
s_n^2 = (LA^{1/2})(LA^{1/2})^T - LA_n(n^{-1}I + K\Lambda K^T)(LA_n)^T,
\]

and \(A_n: H_2 \to H_2\) is the continuous linear operator defined in (4.3).

**Proof.** As in the proof of Proposition 2.1, the first term in the decomposition \(L\mu = L(\mu - A_nY) + LA_nY\) is independent of \(Y\). Therefore, the posterior distribution is the marginal distribution of \(L(\mu - A_nY)\) shifted by \(LA_nY\). It suffices to show that this marginal distribution is \(N(0, s_n^2)\).

By Theorem 1 on page 28 in [92], there exists a sequence of continuous linear maps \(L_m H_1 \to \mathbb{R}\) such that \(L_mh \to Lh\) for all \(h\) in a set with probability one under the prior \(\Pi = N(0, \Lambda)\). This implies that \(L_m\Lambda^{1/2}h \to \Lambda^{1/2}h\) for every \(h \in H_1\). Indeed, if \(V = \{h \in H_1: L_mh \to Lh\}\) and \(g \notin V\), then \(V_1 = V + g\) and \(V\) are disjoint measurable, affine subspaces of \(H_1\), where \(\Pi(V) = 1\). The range of \(\Lambda^{1/2}\) is the RKHS of \(\Pi\) and, hence, if \(g\) is in this range, then \(\Pi(V_1) > 0\), as \(\Pi\) shifted over an element from its RKHS is equivalent to \(\Pi\). But then \(V\) and \(V_1\) are not disjoint.

Therefore, from the first definition of \(A_n\) in (4.3) we see that \(L_mA_n \to LA_n\), and, hence, \(L_m(\mu - A_nY) \to L(\mu - A_nY)\), almost surely. As \(L_m\) is continuous, the variable \(L_m(\mu - A_nY)\) is normally distributed with mean zero and variance \(L_mS_mL_m^T = (L_m\Lambda^{1/2})(L_m\Lambda^{1/2})^T - L_mA_n(n^{-1}I + K\Lambda K^T)(L_mA_n)^T\), for \(S_n\) given by (2.4). The desired result follows upon taking the limit as \(m \to \infty\). \(\square\)
As shown in the preceding proof, \( N(0, \Lambda) \)-measurable linear functionals \( L \) automatically have the further property that \( L \Lambda^{1/2}; H_1 \to \mathbb{R} \) is a continuous linear map. This shows that \( LA_n \) and the adjoint operators \( (LA_1)^T \) and \( (LA_n)^T \) are well defined, so that the formula for \( s_n^2 \) makes sense. If \( L \) is a continuous linear functional, one can also write these adjoints in terms of the adjoint \( L^T \) of \( L \), and express \( s_n^2 \) in the covariance operator \( S_n \) (4.2) as \( s_n^2 = Ls_nL^T \). This is exactly as expected.

We again employ the singular value decomposition of the operator \( K \). We therefore have the eigenbasis \( (e_i) \) of \( H_1 \) and the sequence of eigenvalues \( \kappa_i^2 \), and a sequence

\[
Y_i = \kappa_i \mu_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots ,
\]

for \( Z_1, Z_2, \ldots \) independent, standard normal random variables, and aim to recover \( L\mu \). Note that we again identify the Hilbert space \( H_1 \) with \( \ell_2 \), thus \( L\mu \) can be both viewed as a linear functional on \( H_1 \) and \( \ell_2 \). We consider product priors \( \Pi \) on \( \ell_2 \) given by

\[
\Pi = \bigotimes_{i=1}^{\infty} N(0, \lambda_i).
\]

A continuous, linear functional \( L; H_1 \to \mathbb{R} \) can be identified with an inner product \( L\mu = \langle \mu, l \rangle_1 \), for some \( l \in H_1 \), and hence with a sequence \( (l_i) \) in \( \ell_2 \) giving its coordinates in the eigenbasis \( (e_i) \).

As shown in the proof of Proposition 4.1, for \( L \) in the larger class of \( N(0, \Lambda) \)-measurable linear functionals, the functional \( L \Lambda^{1/2} \) is a continuous linear map on \( H_1 \) and hence can be identified with an element of \( H_1 \). For such a functional \( L \) we define a sequence \( (l_i) \) by \( l_i = (L \Lambda^{1/2})_i/\sqrt{\lambda_i} \), for \( ((L \Lambda^{1/2})_i) \) the coordinates of \( L \Lambda^{1/2} \) in the eigenbasis. The assumption that \( L \) is a \( N(0, \Lambda) \)-measurable linear functional implies that \( \sum_{i=1}^{\infty} l_i^2 \lambda_i < \infty \), but \( (l_i) \) need not be contained in \( \ell_2 \); if \( (l_i) \in \ell_2 \), then \( L \) is continuous and the definition of \( (l_i) \) agrees with the definition in the preceding paragraph.

Since we also consider the sequence formulation of inverse problems, we can also start with a sequence \( (l_i) \) and show that under certain conditions it gives rise to a measurable linear functional relative to the prior. Let \( (l_i) \in \mathbb{R}^\infty \) satisfy \( \sum_{i=1}^{\infty} l_i^2 \lambda_i < \infty \). Then similarly to the proof of Proposition 4.1, we can show that \( L\mu = \lim_{n \to \infty} \sum_{i=1}^{n} l_i \mu_i \) exists for all \( \mu = (\mu_i) \) in a (measurable) subspace of \( \ell^2 \) with \( \bigotimes_{i=1}^{\infty} N(0, \lambda_i) \)-probability one. We define \( L\mu = 0 \) if the limit does not exist.

Given above consideration, the posterior distribution of the linear functional \( L \), denoted by \( \Pi_n(\mu: L\mu \in \cdot | Y) \), is given by

\[
\Pi_n(\mu: L\mu \in | Y) = N \left( \sum_{i=1}^{\infty} \frac{nl_i \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Y_i, \sum_{i=1}^{\infty} \frac{l_i^2 \lambda_i}{1 + n \lambda_i \kappa_i^2} \right). \tag{4.7}
\]

We denote the posterior mean by \( \hat{L}\mu \) (cf. (2.10)).

In the remainder of this chapter we study the marginal distribution of the linear functional \( L \) for various choices of a type of regularity of the representer \( (l_i) \) of the linear functional \( L \). We consider both types of inverse problems. Similar to the problem of full recovery, correct
combinations of the regularity of the true parameter $\mu_0$ and the smoothness and scale of the prior, together with the behavior of the representer $(l_i)$ and the degree of ill-posedness of the inverse problem lead to the minimax rates. Decay of the coefficients $(l_i)$ of $L$ may alleviate the level of ill-posedness, with rapid decay even bringing the functional in the domain of “regular” $n^{-1/2}$-rate estimation.

4.3 Mildly ill-posed problems

In this section we again consider the mildly ill-posed problem. Recall the setting studied in Section 2.4:

**Assumption 4.2.** (Mildly ill-posed problem). The sequences $(\kappa_i)$ and $(\lambda_i)$ in (4.5) and (4.6) satisfy

$$\lambda_i = \tau_n^{2i} - 1 - 2\alpha, \quad C^{-1}i^{-p} \leq \kappa_i \leq Ci^{-p}$$

for some $\alpha > 0$, $p \geq 0$, $C \geq 1$ and $\tau_n > 0$ such that $n\tau_n^2 \to \infty$. Furthermore, the true parameter $\mu_0$ belongs to $S^{\beta}$ for some $\beta > 0$: that is, it satisfies $\sum_{i=1}^{\infty} \mu_{0,i}^2i^{2\beta} < \infty$.

4.3.1 Main results

Our first theorems show contraction of the marginal posterior distribution of the linear functional $L\mu$. We then study the frequentist coverage of posterior credible intervals.

We measure the smoothness of the functional $L$ by the size of the coefficients $l_i$, as $i \to \infty$. First we assume that the sequence is in $S^q$, for some $q$. In this chapter we extend the definition of the Sobolev space $S^q$ to negative $q$, by considering elements of $\mathbb{R}^\infty$ rather than $\ell_2$ in the definition of $S^{\beta}$ (cf. Chapters 2 and 3). We note that the dual space of $S^{\beta}$ is isomorphic with $S^{-\beta}$. The proof of the following theorem is given in Section 4.5.

**Theorem 4.3.** (Contraction). If $\mu_0$, $(\lambda_i)$, $(\kappa_i)$ and $(\tau_n)$ are as in Assumption 4.2 and the representer $(l_i)$ of the linear functional $L$ is contained in $S^q$ for $q \geq -\beta$, then

$$\sup_{\|\mu_0\|_p \leq R} E_{\mu_0}\Pi_n(\mu; |L\mu - L\mu_0| \geq M_n\varepsilon_n|Y) \to 0,$$

for every $R > 0$ and $M_n \to \infty$, where

$$\varepsilon_n = (n\tau_n^2)^{-\frac{\beta + q}{2 + 2\alpha + 2p} + 1} + \tau_n(n\tau_n^2)^{-\frac{1/2 + \alpha + q}{2 + 2\alpha + 2p} + \frac{1}{2}}.$$

In particular:

(i) If $\tau_n \equiv 1$, then $\varepsilon_n = n^{-\frac{\beta(1/2 + \alpha + q)}{2 + 2\alpha + 2p} + \frac{1}{2}}$.

(ii) If $q \leq p$ and $\beta + q \leq 1 + 2\alpha + 2p$ and $\tau_n \asymp n^{-\frac{1/2 + \alpha - \beta}{2\beta + 2p}}$, then $\varepsilon_n \asymp n^{-\frac{\beta + q}{2\beta + 2p}}$.

(iii) If $q \leq p$ and $\beta + q > 1 + 2\alpha + 2p$, then $\varepsilon_n \asymp n^{-\frac{\beta + q}{2\beta + 2p}}$ for every scaling $\tau_n$.

(iv) If $q \geq p$ and $\tau_n \gtrsim n^{-\frac{1/2 + \alpha - \beta + p - q}{2\beta + 2p}}$, where $\beta = \beta \wedge (1 + 2\alpha + 2p - q)$, then $\varepsilon_n = n^{-1/2}$.
If \( q \geq p \), then the smoothness of the functional \( \mathcal{L} \) cancels the ill-posedness of the operator \( K \), and estimating \( L\mu \) becomes a “regular” problem with an \( n^{-1/2} \) rate of convergence. Without scaling the prior (\( \tau_n \equiv 1 \)), the posterior contracts at this rate (see (i) or (iv)) if the prior is not too smooth (\( \beta = 1/2 + q - p \)). With scaling, the rate is also attained, with any prior, provided the scaling parameter \( \tau_n \) does not tend to zero too fast (see (iv)). Inspection of the proof shows that too smooth priors or too small scale creates a bias that slows down the rate.

If \( q < p \), where we take \( q \) the “biggest” value such that \( (l_i) \in S^q \), estimating \( L\mu \) is still an inverse problem. The minimax rate over a ball in the Sobolev space \( S^q \) is known to be of the order \( n^{-(\beta+q)/(2\beta+2p)} \) in this case (see [30, 31, 45]).

This rate is attained without scaling (see (i): \( \tau_n \equiv 1 \)) if and only if the prior smoothness \( \alpha \) is equal to the true smoothness \( \beta \) minus \( 1/2 \) (\( \alpha + 1/2 = \beta \)). An intuitive explanation for this apparent mismatch of prior and truth is that regularity of the parameter in the Sobolev scale (\( \mu_0 \in S^q \)) is not the appropriate type of regularity for estimating a linear functional \( L\mu \). For instance, the difficulty of estimating a function at a point is determined by the regularity in a neighborhood of the point, whereas the Sobolev scale measures global regularity over the domain. The fact that a Sobolev space of order \( \beta \) embeds continuously in a Hölder space of regularity \( \beta - 1/2 \) might give a quantitative explanation of the “loss” in smoothness by \( 1/2 \) in the special case that the eigenbasis is the Fourier basis. In our Bayesian context we draw the conclusion that the prior must be adapted to the inference problem if we want to obtain the optimal frequentist rate: for estimating the global parameter, a good prior must match the truth (\( \alpha = \beta \)), but for estimating a linear functional a good prior must consider a Sobolev truth of order \( \beta \) as having regularity \( \alpha = \beta - 1/2 \).

If the prior smoothness \( \alpha \) is not \( \beta - 1/2 \), then the minimax rate may still be attained by scaling the prior. As in the global problem, this is possible only if the prior is not too rough (\( \beta + q \leq 1 + 2\alpha + 2p \), cases (ii) and (iii)). The optimal scaling when this is possible (case (ii)) is the same as the optimal scaling for the global problem (Theorem 2.3.(ii)) after decreasing \( \beta \) by \( 1/2 \). So the “loss in regularity” persists in the scaling rate. Heuristically this seems to imply that a simple data-based procedure to set the scaling, such as empirical or hierarchical Bayes, cannot attain simultaneous optimality in both the global and local senses.

In the application of the preceding theorem, the functional \( \mathcal{L} \), and hence the sequence \( (l_i) \), will be given. Naturally, we apply the theorem with \( q \) equal to the largest value such that \( (l_i) \in S^q \). Unfortunately, this lacks precision for the sequences \( (l_i) \) that decrease exactly at some polynomial order: a sequence \( l_i \sim i^{-q-1/2} \) is in \( S^{q'} \) for every \( q' < q \), but not in \( S^q \). In the following theorem, proved in Section 4.5, we consider these sequences, and the slightly more general ones such that \( |l_i| \leq i^{-q-1/2} S(i) \), for some slowly varying sequence \( S(i) \). For these sequences \( (l_i) \in S^{q'} \) for every \( q' < q \), \( (l_i) \notin S^{q'} \) for \( q' > q \), and \( (l_i) \in S^q \) if and only if \( \sum_i S^2(i)/i < \infty \).

**Theorem 4.4.** (Contraction). If \( \mu_0 \), \( (\lambda_i) \), \( (\kappa_i) \) and \( (\tau_n) \) are as in Assumption 4.2 and the representer \( (l_i) \) of the linear functional \( \mathcal{L} \) satisfies \( |l_i| \leq i^{-q-1/2} S(i) \) for a slowly varying function \( S \) and \( q > -\beta \), then the result of Theorem 4.3 is valid with \( \varepsilon_n = (n\tau_n^2)^{-\frac{\beta+q}{1+2\alpha+2p}} \frac{\gamma_n}{n} + \tau_n(n\tau_n^2)^{-\frac{1/2+\alpha+q}{1+2\alpha+2p}} \frac{1}{2} \delta_n \), (4.8)
where, for $\rho_n = \left(n\tau_n^2\right)^{1+2\alpha+2p}$ and $r = 1+2\alpha+2p$,
\[
\gamma_n^2 = \begin{cases} 
S^2(\rho_n), & \text{if } \beta + q < r, \\
\sum_{i \leq \rho_n} S^2(i)/i, & \text{if } \beta + q = r, \\
1, & \text{if } \beta + q > r,
\end{cases}
\delta_n^2 = \begin{cases} 
S^2(\rho_n), & \text{if } q < p, \\
\sum_{i \leq \rho_n} S^2(i)/i, & \text{if } q = p, \\
1, & \text{if } q > p.
\end{cases}
\]

This has the same consequences as in Theorem 4.3, up to the addition of slowly varying terms.

Because the posterior distribution for the linear functional $L\mu$ is the one-dimensional normal distribution $N(L\hat{\mu}, s_n^2)$ (cf. (4.7)), the natural credible interval for $L\mu$ has endpoints $L\hat{\mu} \pm z_{\gamma/2}s_n$, for $z_\gamma$ the (lower) standard normal $\gamma$-quantile. The coverage of this interval is
\[
P_{\mu_0}(L\hat{\mu} + z_{\gamma/2}s_n \leq L\mu_0 \leq L\hat{\mu} - z_{\gamma/2}s_n),
\]
where $Y$ follows (4.5) with $\mu = \mu_0$. To obtain precise results concerning coverage, we assume that $(l_i)$ behaves polynomially up to a slowly varying term, first in the situation $q < p$ that estimating $L\mu$ is an (ill-posed) inverse problem. Let $\tilde{\tau}_n$ be the (optimal) scaling $\tau_n$ that equates the two terms in the right-hand side of (4.8). This satisfies $\tilde{\tau}_n \equiv n^{(1/2+\alpha-\beta)/(2\beta+2p)}\eta_n$, for a slowly varying factor $\eta_n$, where $\tilde{\beta} = \beta \wedge (1+2\alpha+2p-q)$. The proof of the following theorem is given in Section 4.6.

**Theorem 4.5.** (Credibility). Let $\mu_0$, $(\lambda_i)$, $(\kappa_i)$ and $(\tau_n)$ be as in Assumption 4.2, let $\tilde{\tau}_n$ be as above, and let $|l_i| = i^{-q-1/2}S(i)$ for $q < p$ and a slowly varying function $S$. Then the asymptotic coverage of the interval $L\hat{\mu} \pm z_{\gamma/2}s_n$ is:

(i) in $(1-\gamma, 1)$, uniformly in $\mu_0$ such that $\|\mu_0\|_\beta \leq 1$ if $\tau_n \gg \tilde{\tau}_n$.

(ii) in $(1-\gamma, 1)$, for every $\mu_0 \in S^\delta$, if $\tau_n \asymp \tilde{\tau}_n$ and $\beta + q < 1 + 2\alpha + 2p$; in $(0, c)$, along some $\mu_0^n$ with $\sup_n \|\mu_0^n\|_\beta < \infty$ if $\tau_n \asymp \tilde{\tau}_n$ (any $c \in (0, 1)$).

(iii) $0$ along some $\mu_0^n$ with $\limsup_n \|\mu_0^n\|_\beta < \infty$ if $\tau_n \ll \tilde{\tau}_n$.

In case (iii) the sequence $\mu_0^n$ can be taken a fixed element $\mu_0$ in $S^\delta$ if $\tau_n \lesssim n^{-\delta}\tilde{\tau}_n$ for some $\delta > 0$.

Furthermore, if $\tau_n \equiv 1$, then the coverage takes the form as in (i), (ii) and (iii) if $\alpha < \beta - 1/2$, $\alpha = \beta - 1/2$, and $\alpha > \beta - 1/2$, respectively, where in case (iii) the sequence $\mu_0^n$ can be taken a fixed element $\mu_0$.

Similarly as in the corresponding nonparametric problem (see Section 2.4 of Chapter 2), oversmoothing leads to coverage 0, while undersmoothing gives conservative intervals. Without scaling the cut-off for under- or oversmoothing is at $\alpha = \beta - 1/2$; with scaling the cut-off for the scaling rate is at the optimal rate $\tilde{\tau}_n$.

The conservativeness in the case of undersmoothing is less extreme for functionals than for the full parameter, as the coverage is strictly between the credibility level $1 - \gamma$ and 1. The general message is the same: oversmoothing is disastrous for the interpretation of
credible interval, whereas undersmoothing gives intervals that at least have the correct order of magnitude, in the sense that their width is of the same order as the variance of the posterior mean (see the proof). Too much undersmoothing is also undesirable, as it leads to very wide confidence intervals, and may cause that \( \sum_{i=1}^{\infty} l_i^2 \lambda_i \) is no longer finite, thus the linear functional \( L\mu \) is no longer measurable with respect to the prior.

The results (i) and (ii) are the same for every \( q < p \), even if \( \tau_n \equiv 1 \). Closer inspection would reveal that for a given \( \mu_0 \) the exact coverage depends on \( q \) (and \( \mathcal{S}(i) \)) in a complicated way.

If \( q \geq p \), then the smoothness of the functional \( L \) compensates the lack of smoothness of \( K^{-1} \), and estimating \( L\hat{\mu} \) is not a true (ill-posed) inverse problem. This drastically changes the performance of credible intervals. Although oversmoothing again destroys their coverage, credible intervals are exact confidence sets if the prior is not too smooth. We formulate this in terms of a Bernstein–von Mises theorem.

The Bernstein–von Mises theorem for parametric models asserts that the posterior distribution approaches a normal distribution centered at an efficient estimator of the parameter and with variance equal to its asymptotic variance. It is the ultimate link between Bayesian and frequentist procedures. There is no version of this theorem for infinite-dimensional parameters [37], but the theorem may hold for “smooth” finite-dimensional projections, such as the linear functional \( L\mu \) (see [7, 18, 85]).

In the present situation the posterior distribution of \( L\hat{\mu} \) is already normal by the normality of the model and the prior: it is a \( N(L\hat{\mu}, s_n^2) \)-distribution by Proposition 4.1. To speak of a Bernstein–von Mises theorem, we also require the following:

(i) That the (root of the) spread \( s_n \) of the posterior distribution is asymptotically equivalent to the standard deviation \( t_n \) of the centering variable \( L\hat{\mu} \).

(ii) That the sequence \( (L\hat{\mu} - L\mu_0)/t_n \) tends in distribution to a standard normal distribution.

(iii) That the centering \( L\hat{\mu} \) is an asymptotically efficient estimator of \( L\mu \).

We shall show that (i) happens if and only if the functional \( L \) cancels the ill-posedness of the operator \( K \), that is, if \( q \geq p \) in Theorem 4.4. Interestingly, the rate of convergence \( t_n \) must be \( n^{-1/2} \) up to a slowly varying factor in this case, but it could be strictly slower than \( n^{-1/2} \) by a slowly varying factor increasing to infinity.

Because \( L\hat{\mu} \) is normally distributed by the normality of the model, assertion (ii) is equivalent to saying that its bias tends to zero faster than \( t_n \). This happens provided the prior does not oversmooth the truth too much. For very smooth functionals \( q > p \) there is some extra “space” in the cut-off for the smoothness, which (if the prior is not scaled: \( \tau_n \equiv 1 \)) is at \( \alpha = \beta - 1/2 + q - p \), rather than at \( \alpha = \beta - 1/2 \) as for the (global) inverse estimating problem. Thus, the prior may be considerably smoother than the truth if the functional is very smooth.

Say that \( l \in \mathcal{R}^q \) if \( |l_i| = i^{-q-1/2} \mathcal{S}(i) \) for a slowly varying function \( \mathcal{S} \). Write

\[
B_n = \sup_{\|\mu_0\|_ \beta \leq 1} |E_{\mu_0}L\hat{\mu} - L\mu_0|
\]
for the maximal bias of $L\bar{\mu}$ over a ball in the Sobolev space $S^\beta$. Finally, let $\tau_n$ be the (optimal) scaling $\tau_n$ in that it equates the two terms in the right-hand side of (4.8).

**Theorem 4.6.** (Bernstein–von Mises). Let $\mu_0$, $(\lambda_i)$, and $(\kappa_i)$ be as in Assumption 4.2, and let $l$ be the representor of the linear functional $L$:

(i) If $l \in S^p$, then $s_n/t_n \to 1$. If $l \in \mathcal{R}^q$, then $s_n/t_n \to 1$ if and only if $q \geq p$.

(ii) If $l \in S^p$, then $nt_n^2 \to \sum_{i=1}^{\infty} l_i^2/\kappa_i^2$. If $l \in \mathcal{R}^q$ and $q \geq p$, then $n \to nt_n^2$ is slowly varying.

(iii) If $l \in S^q$ for $q \geq p$, then $B_n = o(t_n)$ if either $\tau_n \gg n^{(\alpha+1/2-\beta)/(2\beta+2\alpha)}$ or $(\tau_n \equiv 1$ and $\alpha < \beta - 1/2 + q - p)$. If $l \in \mathcal{R}^q$ for $q \geq p$, then $B_n = o(t_n)$ if $(\tau_n \gg \tau_n)$ or $(\tau_n \equiv 1$ and $\alpha < \beta - 1/2 + q - p)$ or $(q = p, \tau_n \equiv 1$ and $\alpha = \beta - 1/2 + q - p)$ or $(q > p, \tau_n \equiv 1$ and $\alpha = \beta - 1/2 + q - p$ and $S(i) \to 0$ as $i \to \infty$).

(iv) If $l \in S^p$ or $l \in \mathcal{R}^p$ and $B_n = o(t_n)$, then

$$E_{\mu_0} \sup_B |\Pi_n(L\mu \in B \mid Y) - N(L\bar{\mu}, t_n^2)(B)| \to 0,$$

and $(L\bar{\mu} - L\mu_0)/t_n$ converges under $\mu_0$ in distribution to a standard normal distribution, uniformly in $\|\mu_0\|_\beta \lesssim 1$. If $l \in S^p$, then also

$$E_{\mu_0} \sup_B |\Pi_n(L\mu \in B \mid Y) - N\left(\sum_{i=1}^{\infty} Y_i l_i/\kappa_i, 1/n \sum_{i=1}^{\infty} l_i^2/\kappa_i^2\right)(B)| \to 0.$$

Here $B$ denotes any measurable subset of $\mathbb{R}$. In both cases (iv), the asymptotic coverage of the credible interval $L\bar{\mu} \pm z_{\gamma/2}s_n$ is $1 - \gamma$, uniformly in $\|\mu_0\|_\beta \lesssim 1$. Finally, if the conditions under (iii) fail, then there exists $\mu_0^c$ with $\sup_n \|\mu_n^c\|_\beta < \infty$ along which the coverage tends to an arbitrarily low value.

The observation $Y$ in (4.1) can be viewed as a reduction by sufficiency of a random sample of size $n$ from the distribution $N(K\mu, I)$. Therefore, the model fits in the framework of i.i.d. observations, and “asymptotic efficiency” can be defined in the sense of semiparametric models discussed in, for example, [8, 98] and [99]. Because the model is shift-equivariant, it suffices to consider local efficiency at $\mu_0 = 0$. The one-dimensional submodels $N(K(th), I)$ on the sample space $\mathbb{R}^{H^1}$, for $t \in \mathbb{R}$ and a fixed “direction” $h \in H_1$, have likelihood ratios

$$\log \frac{dN(tKh, I)}{dN(0, I)}(Y) = tY_{Kh} - \frac{1}{2} t^2 \|Kh\|^2_2.$$

Thus, their score function at $t = 0$ is the $(Kh)$th coordinate of a single observation $Y = (Y_h; h \in H_2)$, the score operator is the map $\tilde{K}: H_1 \to L_2(N(0, I))$ given by $\tilde{K}h(Y) = Y_{Kh}$, and the tangent space is the range of $\tilde{K}$. We denote the score operator by the same symbol $K$ as in (2.1); if the observation $Y$ were realizable in $H_2$, and not just in the bigger sample.
Chapter 4. Recovery of linear functionals of the parameter

space $\mathbb{R}^{H_2}$, then $Y_{Kh}$ would correspond to $(Y, Kh)_2$ and, hence, the score would be exactly $Kh$ for the operator in (4.1) after identifying $H_2$ and its dual space. The adjoint of the score operator restricted to the closure of the tangent space is the operator $\tilde{K}_T: KH_1 \subset L_2(N(0, I)) \rightarrow H_1$ that satisfies $\tilde{K}_T(Y_g) = K^T g$, where $K^T$ on the right is the adjoint of $K: H_1 \rightarrow H_2$. The functional $L\mu = (l, \mu)_1$ has derivative $l$. Therefore, by [97] asymptotically regular sequences of estimators exist, and the local asymptotic minimax bound for estimating $L\mu$ is finite, if and only if $l$ is contained in the range of $K^T$. Furthermore, the variance bound is $\|m\|^2_2$ for $m \in H_2$ such that $K^T m = l$.

In our situation the range of $K^T$ is $S^p$, and if $l \in S^p$, then by Theorem 4.6.(iv) the variance of the posterior is asymptotically equivalent to the variance bound and its centering can be taken equal to the estimator $\sum_{i=1}^{\infty} Y_i l_i / \kappa_i$, which attains this variance bound. Thus, the theorem gives a semiparametric Bernstein–von Mises theorem, satisfying every of (i), (ii), (iii) in this case. If only $l \in \mathbb{R}^p$ and not $l \in S^p$, the theorem still gives a Bernstein–von Mises type theorem, but the rate of convergence is slower than $n^{-1/2}$, and the standard efficiency theory does not apply.

4.3.2 Volterra operator: pointwise credible bands

In Section 2.4.2, we presented the Bayesian recovery of a function from a noisy version of its primitive. More specifically, we considered the simulated data example, where for a chosen sequence of true coefficients $(\mu_{0,i})$ we put

$$Y_i = \kappa_i \mu_{0,i} + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,$$

with $Z_1, Z_2, \ldots$ independent standard normal random variables. Recall $\kappa_i = ((i - 1/2)\pi)^{-1}$. We then obtained posterior credible balls and the posterior mean, and compared them with the true function $\mu_0$.

Another way to quantify uncertainty of the recovery is to apply the framework of linear functionals and plot pointwise credible bands. The posterior distribution of $\mu$ is Gaussian, and can be described coordinate-wise (cf. (2.9)) by

$$\mu_i | Y \sim N \left( \frac{n \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Y_i, \frac{\lambda_i}{1 + n \lambda_i \kappa_i^2} \right).$$

The (marginal) posterior distribution for the function $\mu$ at a point $x$ is obtained by expanding $\mu(x) = \sum_{i=1}^{\infty} \mu_i e_i(x)$, and applying the linear functional $L\mu = \sum_{i=1}^{\infty} l_i \mu_i$ with $l_i = e_i(x)$. This shows that

$$\mu(x) | Y \sim N \left( \sum_{i=1}^{\infty} \frac{n \lambda_i \kappa_i e_i(x)}{1 + n \lambda_i \kappa_i^2} Y_i, \sum_{i=1}^{\infty} \frac{\lambda_i e_i^2(x)}{1 + n \lambda_i \kappa_i^2} \right).$$

We obtained pointwise credible bands by computing for every $x$ a central 95% interval in the normal distribution on the right-hand side.

Figure 4.1 illustrates these bands for $n = 1000$. In every one of the 10 panels in the figure the black curve represents the function $\mu_0$, defined by the coefficients $i^{-3/2} \sin(i)$ relative
Figure 4.1: Realizations of the posterior mean (red) and pointwise credible bands (green), and 20 draws from the posterior (dashed curves). In all ten panels $n = 10^3$ and $\beta = 1$. Left 5 panels: $\alpha = 1$; right 5 panels: $\alpha = 5$.

Figure 4.2: Realizations of the posterior mean (red) and pointwise credible bands (green), and 20 draws from the posterior (dashed curves). In all ten panels $\beta = 1$ and $\alpha = 1/2, 1, 2, 3, 5$ (top to bottom). Left 5 panels: $n = 10^3$; right 5 panels: $n = 10^8$. 
to $e_i$ (thus $\mu_0 \in S^\beta$ for every $\beta < 1$). The 10 panels represent 10 independent realizations of the data, yielding 10 different realizations of the posterior mean (the red curves) and the pointwise credible bands (the green curves). In the left five panels the prior is given by $\lambda_i = i - 2\alpha - 1$ with $\alpha = 1$, whereas in the right panels the prior corresponds to $\alpha = 5$. Each of the 10 panels also show 20 realizations from the posterior distribution.

These plots are similar to those with credible balls, and indicate the intrinsic difficulty of the inverse problem: better estimation requires bigger sample size. A comparison of the left and right panels shows that the rough prior ($\alpha = 1$) is aware of the difficulty: it produces credible bands that in (almost) all cases contain the true curve. On the other hand, the smooth prior ($\alpha = 5$) is overconfident; the spread of the posterior distribution poorly reflects the imprecision of estimation.

Specifying a prior that is too smooth relative to the true curve yields a posterior distribution which gives both a bad reconstruction and a misguided sense of uncertainty, also when illustrated by pointwise credible bands rather than credible balls. Our theoretical results show that the inaccurate quantification of estimation error remains even as $n \to \infty$. This is illustrated in Figure 4.2. Every one of its 10 panels is similarly constructed as before, but now with $n = 1000$ and $n = 10^8$ for the five panels on the left-hand and right-hand side, respectively, and with $\alpha = 1/2, 1, 2, 3, 5$ for the five panels from top to bottom.

## 4.4 Extremely ill-posed problems: heat equation

Recall the Dirichlet problem for the heat equation,

$$\frac{\partial}{\partial t} u(x,t) = \frac{\partial^2}{\partial x^2} u(x,t), \quad u(x,0) = \mu(x), \quad u(0,t) = u(1,t) = 0,$$

where $u$ is defined on $[0, 1] \times [0, T]$ and the function $\mu \in L^2[0, 1]$ satisfies $\mu(0) = \mu(1) = 0$. The solution is to this problem is then given by

$$u(x,t) = \sqrt{2} \sum_{i=1}^{\infty} \mu_i e^{-i^2 \pi^2 T} \sin(i\pi x),$$

where $(\mu_i)$ are the coordinates of $\mu$ in the basis $e_i = \sqrt{2} \sin(i\pi x)$, for $i \geq 1$. Therefore, $\kappa_i = \exp(-i^2 \pi^2 T)$, for $i \geq 1$. We assume we observe (a sequence version of) the solution in white noise of intensity $1/n$:

$$Y_i = \kappa_i \mu_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \ldots,$$

for $(\mu_i)$ and $(\kappa_i)$ as above, and $Z_1, Z_2, \ldots$ independent, standard normal random variables.

### 4.4.1 Main results

As in the previous section, we measure the smoothness of the functional $L$ by the size of the coefficients $l_i$, as $i \to \infty$. It is natural to assume that the sequence $(l_i)$ is in the Sobolev space $S^q$ for some $q$, but also more controlled behavior will be assumed in following theorems. The proof of the following theorem can be found in Section 4.5.
4.4. Extremely ill-posed problems: heat equation

Theorem 4.7. (Contraction). If \( \lambda_i = \tau_n^{2i-1-2\alpha} \) for some \( \alpha > 0 \) and \( \tau_n > 0 \) such that \( n\tau_n^2 \to \infty \), and the representor \( (l_i) \) of the linear functional \( L \) is contained in \( S^q \), or \( |l_i| \lesssim i^{-q-1/2} \) for some \( q \geq -\beta \), then

\[
\sup_{\|\mu_0\|_\beta \leq R} E_{\mu_0} \Pi_n(\mu : |L\mu - L\mu_0| \geq M_n \varepsilon_n \mid Y) \to 0 \tag{4.9}
\]
for every \( R > 0 \) and \( M_n \to \infty \), where

\[
\varepsilon_n = \left( \log(n\tau_n^2) \right)^{-\frac{\beta+q}{2}} + \tau_n \left( \log(n\tau_n^2) \right)^{-\frac{1/2+\alpha+q}{2}}. \tag{4.10}
\]

In particular:

(i) If \( \tau_n \equiv 1 \), then \( \varepsilon_n = \left( \log n \right)^{-\frac{\beta+1}{2}} \).

(ii) If \( n^{-1/2+\delta} \lesssim \tau_n \lesssim (\log n)^{-\frac{1/2+\alpha-\beta}{2}} \), for some \( \delta > 0 \), then \( \varepsilon_n = \left( \log n \right)^{-\frac{\beta+q}{2}} \).

If \( \lambda_i = e^{-\alpha i^2} \) for some \( \alpha > 0 \) then (4.9) holds with the rate

\[
\varepsilon_n = \left( \log n \right)^{-\frac{\beta+\delta}{2}}. \tag{4.11}
\]

The minimax rate over a ball in the Sobolev space \( S^\beta \) is known to be of the order \( (\log n)^{-\frac{1}{2}+\alpha} \) (for the case of \( q = -1/2 \) see [43], and for general \( q \) in a closely related model see [14]). In view of Theorem 2.5, (see Chapter 2) it is not surprising that exponential priors yield this optimal rate. In case of a polynomial prior this rate is attained without scaling if and only if the prior smoothness \( \alpha \) is greater than or equal to \( \beta - 1/2 \). This phenomenon was already observed in the previous section, and the intuitive explanation was given: the regularity of the parameter in the Sobolev scale is not the appropriate type of regularity to consider for estimating a linear functional \( L\mu \).

If the polynomial prior is too rough, then the minimax rate may still be attained by scaling the prior. The upper bound on the scaling is the same as in the global case (see Theorem 2.5.(ii)) after decreasing \( \beta \) by 1/2. So the “loss in regularity” persists in the scaling.

The posterior distribution for the linear functional \( L\mu \) is a one-dimensional normal distribution (cf. (4.7)). Recall that the natural credible interval for \( L\mu \) has endpoints \( L\hat{\mu} \pm z_{\gamma/2}s_n \), for \( z_{\gamma} \) the (lower) standard normal \( \gamma \)-quantile. We study the coverage of this interval

\[
P_{\mu_0}(L\hat{\mu} + z_{\gamma/2}s_n \leq L\mu_0 \leq L\hat{\mu} - z_{\gamma/2}s_n),
\]
where \( Y \) follows (4.5) with \( \mu = \mu_0 \).

In the following theorem we restrict \( (l_i) \) to sequences that behave polynomially.

Theorem 4.8. (Credibility). Suppose the true parameter \( \mu_0 \) belongs to \( S^\beta \) for \( \beta > 0 \). Let \( \tilde{\tau}_n = (\log n)^{1/2+\alpha-\beta}/2 \).

If \( \lambda_i = \tau_n^{2i-1-2\alpha} \) for some \( \alpha > 0 \) and \( \tau_n > 0 \) such that \( n\tau_n^2 \to \infty \), and \( |l_i| \lesssim i^{-q-1/2} \), then the asymptotic coverage of the interval \( L\hat{\mu} \pm z_{\gamma/2}s_n \) is:

(i) 1, uniformly in \( \mu_0 \) such that \( \|\mu_0\|_\beta \leq 1 \) if \( \tau_n \gg \tilde{\tau}_n \).
(ii) \( \| \mu_0 \|_\beta \leq r \) for \( r \) small enough, if \( \tau_n \asymp \tilde{\tau}_n \);

1, for every fixed \( \mu_0 \in S^\beta \), if \( \tau_n \asymp \tilde{\tau}_n \).

(iii) 0, along some \( \mu_0^n \) with \( \sup_n \| \mu_0^n \|_\beta < \infty \), if \( \tau_n \lesssim \tilde{\tau}_n \).

If \( \lambda_i = e^{-\alpha i^2} \) for some \( \alpha > 0 \), then the asymptotic coverage of the interval \( L \hat{\mu} \pm z_{\gamma/2} s_n \) is:

(iv) 0, for every \( \mu_0 \) such that \( \mu_0, i \gtrsim e^{-c i^2 / 2} i^{-q-1/2} \) for some \( c < \alpha \).

In case (iii) the sequence \( \mu_0^n \) can be taken a fixed element \( \mu_0 \) in \( S^\beta \) if the scaling \( \tau_n \lesssim (\log n)^{(1/2+\alpha-\beta)/2-\delta} \) for some \( \delta > 0 \). Furthermore, if \( \tau_n \equiv 1 \), then the cases (i), (ii) and (iii) arise if \( \alpha < \beta - 1/2 \), \( \alpha = \beta - 1/2 \) and \( \alpha \geq \beta - 1/2 \), respectively. If \( \alpha > \beta - 1/2 \) in case (iii) the sequence \( \mu_0^n \) can then be chosen a fixed \( \mu_0 \).

Similarly as in the problem of full recovery of the parameter \( \mu \), oversmoothing leads to coverage 0, while undersmoothing gives (extremely) conservative intervals. In the case of a polynomial prior without scaling the cut-off for under- or oversmoothing is at \( \alpha = \beta - 1/2 \), while the cut-off for scaling is at the optimal rate \( \tilde{\tau}_n \). Exponential priors are bad even for very smooth \( \mu_0 \), and the asymptotic coverage in this case is always 0. It should be noted that too much undersmoothing is also undesirable, as it leads to very wide credible intervals, and may cause that \( \sum_{i=1}^\infty l_i^2 \lambda_i \) is no longer finite.

In contrast with Theorem 4.5, the conservativeness in case of undersmoothing is extreme, as the coverage is 1. Since it holds for every linear functional that can be considered in this setting, we do not have a Bernstein–von Mises theorem. The linear functionals considered in this section are not smooth enough to cancel the ill-posedness of the problem (cf. discussion after Theorem 4.6). Inspection of the proof of Theorem 4.6 suggests that if we consider “supersmooth” functionals (i.e., \( |l_i| \lesssim e^{-qi^2} \) for \( q \) large enough), we can obtain a Bernstein–von Mises theorem.

### 4.4.2 Simulation example: pointwise credible bands

We present pointwise credible bands in the simulated example studied in Section 2.5.2. Recall that the simulated data are the noisy and transformed coefficients

\[
Y_i = \kappa_i \mu_{0,i} + \frac{1}{\sqrt{n}} Z_i.
\]

The (marginal) posterior distribution for the function \( \mu \) at a point \( x \) is obtained by expanding \( \mu(x) = \sum_{i=1}^\infty \mu_i e_i(x) \), and applying the linear functional \( L\mu = \sum_{i=1}^\infty l_i \mu_i \) with \( l_i = e_i(x) \) (so \( |l_i| \lesssim 1 \) and \( q = -1/2 \)). Recall

\[
\mu(x) \mid Y \sim N \left( \sum_{i=1}^\infty \frac{n \lambda_i \kappa_i e_i(x)}{1 + n \lambda_i \kappa_i^2} Y_i, \sum_{i=1}^\infty \frac{\lambda_i e_i^2(x)}{1 + n \lambda_i \kappa_i^2} \right).
\]

We obtained pointwise credible bands by computing for every \( x \) a central 95\% interval for the normal distribution on the right side of the above display. We considered both types
Figure 4.3: Polynomial prior. Realizations of the posterior mean (red) and pointwise credible bands (green), and 20 draws from the posterior (dashed curves). In all ten panels \( n = 10^4 \). Left 5 panels: \( \alpha = 1 \); right 5 panels: \( \alpha = 3 \). True curve (black) given by (4.12).

Figure 4.4: Exponential prior. Realizations of the posterior mean (red) and pointwise credible bands (green), and 20 draws from the posterior (dashed curves). In all ten panels \( n = 10^4 \). Left 5 panels: \( \alpha = 1 \); right 5 panels: \( \alpha = 5 \). True curve (black) given by (4.12).
Figure 4.5: Polynomial prior. Realizations of the posterior mean (red) and pointwise credible bands (green), and 20 draws from the posterior (dashed curves). Left 5 panels: $n = 10^4$, right 5 panels: $n = 10^8$; $\alpha = 1/2, 1, 2, 5, 10$ (top to bottom). True curve (black) given by (4.12).

Figure 4.6: Exponential prior. Realizations of the posterior mean (red) and pointwise credible bands (green), and 20 draws from the posterior (dashed curves). Left 5 panels: $n = 10^4$, right 5 panels: $n = 10^8$; $\alpha = 1/2, 1, 2, 5, 10$ (top to bottom). True curve (black) given by (4.12).
of priors. Figure 4.3 illustrates these bands for \( n = 10^4 \) and the polynomial prior. In every of 10 panels in the figure the black curve represents the function \( \mu_0 \), defined by

\[
\mu_0(x) = 4x(x - 1)(8x - 5), \quad \mu_{0,i} = \frac{8\sqrt{2}(13 + 11(-1)^i)}{\pi^3 i^3},
\]

where \( \mu_{0,i} \) are the coefficients relative to \( e_i \), thus \( \mu_0 \in S^\beta \) for every \( \beta < 2.5 \). The 10 panels represent 10 independent realizations of the data, yielding 10 different realizations of the posterior mean (the red curves) and the posterior credible bands (the green curves). In the left five panels the prior is given by \( \lambda_i = i^{-1 - 2\alpha} \) with \( \alpha = 1 \), whereas in the right panels the prior corresponds to \( \alpha = 3 \). This is also valid for Figure 4.4, with the exponential prior, so \( \lambda_i = e^{-\alpha i^2} \). In the left panels \( \alpha = 1 \), and in the right panels \( \alpha = 5 \).

Credible bands are again similar to credible balls and yield analogous conclusions. A comparison of the left and right panels in Figure 4.3 shows that the rough polynomial prior (\( \alpha = 1 \)) is aware of the difficulty of inverse problem: it produces wide pointwise credible bands that in (almost) all cases contain nearly the whole true curve. Figure 4.3 together with Figure 4.4 show that smooth priors (polynomial with \( \alpha = 3 \) and both exponential priors) are overconfident: the spread of the (marginal) posterior distribution poorly reflects the imprecision of estimation. Our theoretical results show that the inaccurate quantification of the estimation error (by the posterior spread) remains even as \( n \to \infty \). This is illustrated in Figures 4.5 and 4.6. Every of 10 panels in each of the figures is similarly constructed as before, but now with \( n = 10^4 \) and \( n = 10^8 \) for the five panels on the left and right side, respectively, and with \( \alpha = 1/2, 1, 2, 5, 10 \) for the five panels from top to bottom (\( \lambda_i = i^{-1 - 2\alpha} \) in Figure 4.5, and \( \lambda_i = e^{-\alpha i^2} \) in Figure 4.6). As discussed above, all exponential priors give the optimal rate, but lead to bad credible bands. Also smooth polynomial priors give the optimal rate. This can be seen in Figure 4.5 for \( n = 10^8 \) and \( \alpha = 2 \) or 5, where pointwise credible bands are very close to the true curve. Since for \( \alpha = 10 \) pointwise credible bands are relatively far from the true curve, we see that our result are indeed asymptotic, i.e., the reconstruction, by the posterior mean or other posterior quantiles will occur for \( n \) large enough, depending on the regularity of the prior.

4.5 Proofs of Theorems 4.3, 4.4, and 4.7

In this section we prove contraction results for the linear functionals of the parameter in both inverse problem settings presented in this chapter. Technical details distinguishing both cases are presented in separate subsections.

By (4.7) the posterior distribution is \( N(L\hat{\mu}, s_n^2) \), and hence, similarly as in the proofs of Theorems 2.3 and 2.5 in Chapter 2, it suffices to show that

\[
E_{\mu_0}|L\hat{\mu} - L\mu_0|^2 + s_n^2 = |E_{\mu_0}L\hat{\mu} - L\mu_0|^2 + \sum_{i=1}^{\infty} \frac{l_i^2 n\lambda_i^2 \kappa_i^2}{(1 + n\lambda_i \kappa_i^2)^2} + s_n^2
\]

is bounded above by a multiple of \( \varepsilon_n^2 \).
### 4.5.1 Details of Theorem 4.3

Under Assumption 4.2 the expression on the right can be written

\[
|E_{\mu_0} L \hat{\mu} - L \mu_0| = \left| \sum_{i=1}^{\infty} \frac{l_i \mu_{0,i}}{1 + n \lambda_i \kappa_i^2} \right| \leq \sum_{i=1}^{\infty} \frac{|l_i \mu_{0,i}|}{1 + n \tau_n^2 i^{-1-2\alpha-2p}},
\]

(4.14)

\[
t_n^2 = \sum_{i=1}^{\infty} \frac{l_i^2 n \lambda_i^2 \kappa_i^2}{(1 + n \lambda_i \kappa_i^2)^2} \leq n \tau_n^4 \sum_{i=1}^{\infty} \frac{l_i^2 i^{-2-4\alpha-2p}}{(1 + n \tau_n^2 i^{-1-2\alpha-2p})^2},
\]

(4.15)

\[
s_n^2 = \sum_{i=1}^{\infty} \frac{l_i^2 \lambda_i}{1 + n \lambda_i \kappa_i^2} \leq \tau_n^2 \sum_{i=1}^{\infty} \frac{l_i^2 i^{-1-2\alpha}}{1 + n \tau_n^2 i^{-1-2\alpha-2p}}.
\]

(4.16)

By the Cauchy–Schwarz inequality the square of the bias (4.14) satisfies

\[
|E_{\mu_0} L \hat{\mu} - L \mu_0|^2 \lesssim \|\mu_0\|_2^2 \sum_{i=1}^{\infty} \frac{l_i^2 i^{-2\beta}}{(1 + n \tau_n^2 i^{-1-2\alpha-2p})^2}.
\]

(4.17)

By Lemma A.1 (applied with \( q = q, t = 2, u = 1 + 2\alpha + 2p, v = 2 \) and \( N = n \tau_n^2 \)) the right-hand side of this display can be further bounded by \( \|\mu_0\|_2^2 \|l\|_q^2 \) times the square of the first term in the sum of two terms that defines \( \varepsilon_n \). By Lemma A.1 (applied with \( q = q, t = 2, u = 1 + 2\alpha + 2p, v = 2 \) and \( N = n \tau_n^2 \)) and again Lemma A.1 (applied with \( q = q, t = 1 + 2\alpha, u = 1 + 2\alpha + 2p, v = 1 \) and \( N = n \tau_n^2 \)), the right-hand sides of (4.15) and (4.16) are bounded above by \( \|l\|_q^2 \) times the square of the second term in the definition of \( \varepsilon_n \).

Consequences (i)–(iv) follow by substitution, and, in the case of (iii), optimization over the scale parameter \( \tau_n \).

### 4.5.2 Details of Theorem 4.4

This follows the same lines as the proof of Theorem 4.3, except that we use Lemma A.3 (with \( q = q, t = 2, u = 1 + 2\alpha + 2p, v = 2 \) and \( N = n \tau_n^2 \)) and Lemma A.3 (with \( q = q, t = 2 + 4\alpha + 2p, u = 1 + 2\alpha + 2p, v = 2 \) and \( N = n \tau_n^2 \)) and again Lemma A.3 (with \( q = q, t = 1 + 2\alpha, u = 1 + 2\alpha + 2p, v = 1 \) and \( N = n \tau_n^2 \)) to bound the three terms (4.15)–(4.17).

### 4.5.3 Details of Theorem 4.7

If \( \lambda_i = \tau_n^2 i^{-1-2\alpha} \) the three quantities in (4.13) are given by

\[
|E_{\mu_0} L \hat{\mu} - L \mu_0| = \left| \sum_{i=1}^{\infty} \frac{l_i \mu_{0,i}}{1 + n \lambda_i \kappa_i^2} \right| \leq \sum_{i=1}^{\infty} \frac{|l_i \mu_{0,i}|}{1 + n \tau_n^2 i^{-1-2\alpha} e^{-2\pi^2 T i^2}},
\]

(4.18)

\[
t_n^2 = \sum_{i=1}^{\infty} \frac{l_i^2 n \lambda_i^2 \kappa_i^2}{(1 + n \lambda_i \kappa_i^2)^2} = n \tau_n^4 \sum_{i=1}^{\infty} \frac{l_i^2 i^{-2-4\alpha} e^{-2\pi^2 T i^2}}{(1 + n \tau_n^2 i^{-1-2\alpha} e^{-2\pi^2 T i^2})^2},
\]

(4.19)

\[
s_n^2 = \sum_{i=1}^{\infty} \frac{l_i^2 \lambda_i}{1 + n \lambda_i \kappa_i^2} = \tau_n^2 \sum_{i=1}^{\infty} \frac{l_i^2 i^{-1-2\alpha}}{1 + n \tau_n^2 i^{-1-2\alpha} e^{-2\pi^2 T i^2}}.
\]

(4.20)
By the Cauchy–Schwarz inequality the square of the bias (4.18) satisfies
\[
|E_{\mu_0} L\hat{\mu} - L\mu_0|^2 \leq \|\mu_0\|_{2}^{2} \sum_{i=1}^{\infty} \frac{t_{i}^2 i^{-2\beta}}{\left(1 + n\tau_n^2 i^{-1 - 2\alpha} e^{-2\pi^2 T_i^2}\right)^2}.
\] (4.21)

Consider \((l_i) \in S^q\). By Lemma A.2 (applied with \(q = q, t = 2\beta, r = 0, u = 1 + 2\alpha, p = 2\pi^2 T, v = 2,\) and \(N = n\tau_n^2\)) the right side of this display can be further bounded by \(\|\mu_0\|_{2}^{2} \left\|l_i\right\|_q^2\) times the square of the first term in the sum of two terms that defines \(\varepsilon_n\). By Lemma (applied with \(q = q, t = 2 + 4\alpha, r = 2\pi^2 T, u = 1 + 2\alpha, p = 2\pi^2 T, v = 2,\) and \(N = n\tau_n^2\)), and again by Lemma A.2 (applied with \(q = q, t = 1 + 2\alpha, r = 0, u = 1 + 2\alpha, p = 2\pi^2 T, v = 1,\) and \(N = n\tau_n^2\)) the right sides of (4.19) and (4.20) are bounded above by \(\|l_i\|_q^2\) times the square of the second term in the definition of \(\varepsilon_n\).

Consider \(|l_i| \lesssim i^{-q - 1/2}\). This follows the same lines as in the case of \((l_i) \in S^q\), except that we use Lemma A.5 instead of Lemma A.2. In this case the upper bound for the standard deviation of the posterior mean \(t_n\) is of the order \(\tau_n (\log (n\tau_n^2))^{-(1 + \alpha + q)/2}\).

Consequences (i)–(ii) follow by substitution.

If \(\lambda_i = e^{-\alpha i^2}\), then in case \((l_i) \in S^q\) we use Lemma A.5 (with \(q = q, t = 2\beta, r = 0, u = 0, p = \alpha + 2\pi^2 T, v = 2,\) and \(N = n\)), and Lemma A.5 (with \(q = q, t = 0, r = 2\alpha + 2\pi^2 T, u = 0, p = \alpha + 2\pi^2 T, v = 2,\) and \(N = n\)), and again Lemma A.5 (with \(q = q, t = 0, r = \alpha, u = 0, p = \alpha + 2\pi^2 T, v = 2,\) and \(N = n\)) to bound (4.21) by a multiple of \((\log n)^{-(\beta + q)}\), and (4.19)–(4.20) by a multiple of \(n^{-\alpha/(\alpha + 2\pi^2 T)} (\log n)^{-q}\).

If \(|l_i| \lesssim i^{-q - 1/2}\), we use Lemma A.2 (with \(t = 1 + 2q + 2\beta, r = 0, u = 0, p = \alpha + 2\pi^2 T, v = 2,\) and \(N = n\)), and Lemma A.2 (with \(t = 1 + 2q, r = 2\alpha + 2\pi^2 T, u = 0, p = \alpha + 2\pi^2 T, v = 2,\) and \(N = n\)), and again Lemma A.2 (with \(t = 1 + 2q, r = \alpha, u = 0, p = \alpha + 2\pi^2 T, v = 1,\) and \(N = n\)) to bound (4.21) by a multiple of \((\log n)^{-(\beta + q)}\), and (4.19)–(4.20) by a multiple of \(n^{-\alpha/(\alpha + 2\pi^2 T)} (\log n)^{-1/2 - q}\).

\[\text{4.6 Proofs of Theorems 4.5 and 4.8}\]

In this section we prove results on frequentist coverage of credible bands in both inverse problem settings presented in this chapter. Technical details distinguishing both cases are presented in separate subsections.

Under (4.5) the variable \(L\hat{\mu}\) is \(N(E_{\mu_0} L\hat{\mu}, t_n^2)\)-distributed, for \(t_n^2\) given in (4.15) and (4.19). It follows that the coverage can be written, with \(W\) a standard normal variable,
\[
P\left(|W_{t_n} + E_{\mu_0} L\hat{\mu} - L\mu_0| \leq -s_n z_{\alpha/2}\right).
\] (4.22)

The bias \(|E_{\mu_0} L\hat{\mu} - L\mu_0|\) and posterior spread \(s_n^2\) are expressed as series in (4.14) and (4.16), or (4.18) and (4.20), for the mildly and the extremely ill-posed problems, respectively. Note that \(t_n \leq s_n\) for every \(n\), as every term in the infinite series (4.15) and (4.19) is \(\propto n\lambda_i \kappa_i^2 / (1 + n\lambda_i \kappa_i^2) \leq 1\) times the corresponding term in (4.16) and (4.20), respectively.

Because \(W\) is centered, the coverage (4.22) is largest if the bias \(|E_{\mu_0} L\hat{\mu} - L\mu_0|\) is zero. It is then at least \(1 - \gamma\), because \(t_n \leq s_n\). It tends to exactly 1, if \(t_n \ll s_n\); or remains strictly smaller than 1, if \(t_n \approx s_n\), and tends to exactly \(1 - \gamma\) iff \(s_n/t_n \rightarrow 1\).
4.6.1 Details of Theorem 4.5

In the proof of Theorem 4.4 $s_n$ and $t_n$ were seen to have the same order of magnitude, given by the second term in $\varepsilon_n$ given in (4.8), with a slowly varying term $\delta_n$ as given in the theorem,

$$s_n \asymp t_n \asymp t_n (n \tau_n^2) \frac{1/2 + \alpha + q}{1 + 2\alpha + 2p} \delta_n.$$  \hspace{1cm} (4.23)

Because of that the coverage (4.22) in case $\mu_0 = 0$ remains strictly smaller than 1, and tends to exactly $1 - \gamma$ iff $s_n/t_n \to 1$. By Theorem 4.6.(ii) the latter is impossible if $q < p$. The analysis for nonzero $\mu_0$ depends strongly on the size of the bias relative to $t_n$.

The supremum of the bias satisfies, for $\gamma_n$, the slowly varying term given in Theorem 4.4,

$$B_n = \sup_{\|\mu\|_{\beta} \leq 1} |E_{\mu_0} L \hat{\mu} - L \mu_0| \asymp (n \tau_n^2)^{-\frac{\beta + q}{1 + 2\alpha + 2p} \land 1} \gamma_n.$$  \hspace{1cm} (4.24)

That the left-hand side of (4.24) is smaller than the right-hand side was already shown in the proof of Theorem 4.4, with the help of Lemma A.3. That this upper bound is sharp follows by considering the sequence $\mu_0^n$ defined by, with $B_n$ the right-hand side of the preceding display,

$$\mu_0^n = \frac{1}{B_n} \frac{i^{-2\beta} l_i}{1 + n \tau_n^{2} i^{-1 - 2\alpha - 2p}}.$$  

This is the sequence that gives equality in the application of the Cauchy–Schwarz inequality to derive (4.17). Using Lemma A.3, it can be seen that $\|\mu_0^n\|_{\beta} \lesssim 1$ and that the bias at $\mu_0^n$ is of the order $B_n$.

By Lemma A.4, the bias at a fixed $\mu_0 \in S^\beta$ is of strictly smaller order than the supremum $B_n$ if $\beta + q < 1 + 2\alpha + 2p$.

The maximal bias $B_n$ is a decreasing function of the scaling parameter $\tau_n$, while the standard deviation $t_n$ and root-spread $s_n$ increase with $\tau_n$. The scaling rate $\tilde{\tau}_n$ in the statement of the theorem balances $B_n$ with $s_n \asymp t_n$.

Case (i). If $\tau_n \gg \tilde{\tau}_n$, then $B_n \ll t_n$. Hence, the bias $E_{\mu_0} L \hat{\mu} - L \mu_0$ in (4.22) is negligible relative to $t_n \asymp s_n$, uniformly in $\|\mu_0\|_{\beta} \lesssim 1$, and the coverage is asymptotic to $P(|W t_n| \leq -s_n z_{\gamma/2})$, which is asymptotically strictly between $1 - \gamma$ and 1.

Case (iii). If $\tau_n \ll \tilde{\tau}_n$, then $B_n \gg t_n$. If $b_n = E_{\mu_0} L \hat{\mu} - L \mu_0$ is the bias at a sequence $\mu_0^n$ that (nearly) attains the supremum in the definition of $B_n$, then the coverage at $\mu_0^n$ satisfies $P(|W t_n + b_n| \leq -s_n z_{\gamma/2}) \leq P(|W t_n| \geq b_n - s_n|z_{\gamma/2}|) \to 0$, as $b_n \asymp B_n \gg s_n$. By the same argument, the coverage also tends to zero for a fixed $\mu_0$ in $S^\beta$ with bias $b_n = E_{\mu_0} L \hat{\mu} - L \mu_0 \gg t_n$. For this we choose $\mu_{0,i} = i^{-\beta} l_i S(i)$ for a slowly varying function $S$ such that $\sum_{i=1}^{\infty} S^2(i) \hat{S}^2(i)/i < \infty$. The latter condition ensures that $\|\mu_0\|_{\beta} < \infty$. By another application of Lemma A.3, the bias at $\mu_0$ is of the order (cf. (4.14))

$$\sum_{i=1}^{\infty} \frac{l_i \mu_{0,i}}{1 + n \tau_n^{2} i^{-1 - 2\alpha - 2p}} = \sum_{i=1}^{\infty} \frac{(l_i \hat{S}^{1/2}(i))^2 i^{-\beta + q}}{1 + n \tau_n^{2} i^{-1 - 2\alpha - 2p}} \asymp (n \tau_n^2)^{-\frac{\beta + q}{1 + 2\alpha + 2p} \land 1} \gamma_n,$$  

where $S$ is a slowly varying function.
where, for $\rho_n = (n\tau_n^2)^{1/(1+2\alpha+2p)}$,

$$\hat{\gamma}_n^2 = \begin{cases} S^2(\rho_n)\hat{S}(\rho_n), & \text{if } \beta + q < 1 + 2\alpha + 2p, \\
\sum_{i \leq \rho_n} \frac{S^2(i)\hat{S}(i)}{i}, & \text{if } \beta + q = 1 + 2\alpha + 2p, \\
1, & \text{if } \beta + q > 1 + 2\alpha + 2p. \end{cases}$$

Therefore, the bias at $\mu_0$ has the same form as the maximal bias $B_n$; the difference is in the slowly varying factor $\hat{\gamma}_n$. If $\tau_n \leq \tilde{\tau}_n n^{-\delta}$, then $B_n \gtrsim t_n n^{\delta'}$ for some $\delta' > 0$ and, hence, $b_n \gtrsim B_n \hat{\gamma}_n / \gamma_n \gg t_n$.

Case (ii). If $\tau_n \gg \tilde{\tau}_n$, then $B_n \gg t_n$. If $b_n = E_{\mu_0} L\hat{\mu} - L\mu_0^n$ is again the bias at a sequence $\mu_0^n$ that nearly assumes the supremum in the definition of $B_n$, we have that $P(|Wt_n + db_n| \leq -s_n z_{\gamma/2}) \leq P(|Wt_n| \geq db_n - s_n |z_{\gamma/2}|)$ attains an arbitrarily small value if $d$ is chosen sufficiently large. This is the coverage at the sequence $d\mu_0^n$, which is bounded in $S^\beta$. On the other hand, the bias at a fixed $\mu_0 \in S^\beta$ is of strictly smaller order than the supremum $B_n$, and, hence, the coverage at a fixed $\mu_0$ is as in case (i).

If the scaling rate is fixed to $\tau_n \equiv 1$, then it can be checked from (4.23) and (4.24) that $B_n \ll t_n$, $B_n \ll t_n$ and $B_n \gg t_n$ in the three cases $\alpha < \beta - 1/2$, $\alpha = \beta - 1/2$ and $\alpha > \beta - 1/2$, respectively. In the first and third cases the maximal bias and the spread differ by more than a polynomial term $n^{\delta}$; in the second case it must be noted that the slowly varying terms $\gamma_n$ and $\delta_n$ are equal (to $S(\rho_n))$. It follows that the preceding analysis (i), (ii), (iii) extends to this situation.

### 4.6.2 Details of Theorem 4.8

Let $\lambda_i = \tau_i^{2i-1-2\alpha}$. Recall that in this case $t_n \ll s_n$ (cf. the proof of Theorem 4.7). The supremum of the bias satisfies

$$B_n = \sup_{\|\mu_0\|_\beta \lesssim 1} |E_{\mu_0} L\hat{\mu} - L\mu_0| \asymp (\log(n\tau_n^2))^{-\frac{\alpha+1}{2}}. \quad (4.25)$$

The maximal bias $B_n$ is a decreasing function of the scaling parameter $\tau_n$, while the root spread $s_n$ increases with $\tau_n$. The scaling rate $\tilde{\tau}_n = (\log n)^{1/(2+\alpha-\beta)/2}$ in the statement of the theorem balances $B_n$ with $s_n$.

Case (i). If $\tau_n \gg \tilde{\tau}_n$, then $B_n \ll s_n$. Hence, the bias $|E_{\mu_0} L\hat{\mu} - L\mu_0|$ in (4.22) is negligible relative to $s_n$, uniformly in $|\mu_0|_\beta \lesssim 1$, and $P(|Wt_n + E_{\mu_0} L\hat{\mu} - L\mu_0| \leq -s_n z_{\gamma/2}) \geq P(|Wt_n| \leq -s_n z_{\gamma/2} - |E_{\mu_0} L\hat{\mu} - L\mu_0|) \to 1$.

Case (ii). If $\tau_n \asymp \tilde{\tau}_n$, then $B_n \asymp s_n$. If $b_n = |E_{\mu_0} L\hat{\mu} - L\mu_0^n|$ is the bias at a sequence $\mu_0^n$ that nearly assumes the supremum in the definition of $B_n$, we have that $P(|Wt_n + db_n| \leq -s_n z_{\gamma/2}) \geq P(|Wt_n| \leq s_n |z_{\gamma/2}| - db_n) \to 1$ if $d$ is chosen sufficiently small. This is the coverage at the sequence $d\mu_0^n$, which is bounded in $S^\beta$. On the other hand, using Lemma A.6 it can be seen that the bias at a fixed $\mu_0 \in S^\beta$ is of strictly smaller order than the supremum $B_n$, and hence the coverage at a fixed $\mu_0$ is as in case (i).
Case (iii). If \( \tau_n \lesssim \tilde{\tau}_n \), then \( B_n \gtrsim s_n \). If \( b_n = |E_{\mu_0}L\tilde{\mu} - L\mu_0^\alpha| \) is again the bias at a sequence \( \mu_0^n \) that (nearly) attains the supremum in the definition of \( B_n \), we have that

\[
P\left( |Wt_n + db_n| \leq -s_n z_{\gamma/2} \right) \leq P\left( |Wt_n| \geq db_n - s_n |z_{\gamma/2}| \right) \to 0 \text{ if } d \text{ is chosen sufficiently large.}
\]

This is the coverage at the sequence \( d\mu_0^n \), which is bounded in \( S^3 \). By the same argument the coverage also tends to zero for a fixed \( \mu_0 \) in \( S^3 \) with bias \( b_n = |E_{\mu_0}L\tilde{\mu} - L\mu_0| \gtrsim s_n \gg t_n \). For this we choose \( \mu_0, i = i^{-\beta - 1/2 - \delta^i} \) for some \( \delta^i > 0 \). By another application of Lemma A.5, the bias at \( \mu_0 \) is of the order

\[
\sum_{i=1}^{\infty} \frac{l_i \mu_{0,i}}{1 + n\tau_i^2 - 1 - 2\alpha e^{-2\pi T i^2}} \approx \sum_{i=1}^{\infty} \frac{i^{-\beta - q - \delta - 1}}{1 + n\tau_i^2 - 1 - 2\alpha e^{-2\pi T i^2}} \approx (\log(n\tau_i^2))^{-\beta + q + \delta}.
\]

Therefore, if \( \tau_n \leq \tilde{\tau}_n (\log n)^{-\delta} \), then \( B_n \gtrsim s_n (\log(n\tau_i^2))^{-\delta} \) for some \( \delta > 0 \), and hence taking \( \delta' = \delta'' \) we have \( b_n \asymp B_n (\log(n\tau_i^2))^{-\delta''/2} \gg s_n \gg t_n \).

Case (iv). In the proof of Theorem 4.7, we obtained \( s_n \asymp t_n \approx n^{-\alpha/(\alpha + 2\pi^2 T)} (\log n)^{-q} \). If \( \mu_{0,i} \gtrsim e^{-\alpha^2/2} i^{-q - 1/2} \) for some \( c < \alpha \), we have

\[
|E_{\mu_0}L\tilde{\mu} - L\mu_0| \gtrsim \sum_{i=1}^{\infty} \frac{l_i \mu_{0,i}}{1 + n\lambda_i \kappa_i^2} \gtrsim \sum_{i=1}^{\infty} \frac{e^{-\alpha^2 i^{-2q - 1}}}{(1 + n e^{-(\alpha + 2\pi^2 T)i^2})^2} \gtrsim n^{-\frac{\delta}{s + \alpha + 2\pi^2 T}} (\log n)^{-1/2 - q} \gtrsim n^{-\frac{\delta}{s + \alpha + 2\pi^2 T}} (\log n)^{-1/2 - q},
\]

by Lemma A.5 (applied with \( t = 1 + 2q, r = c, u = 0, p = \alpha + 2\pi^2 T, v = 2 \), and \( N = n \)). Hence \( P\left( |Wt_n + E_{\mu_0}L\tilde{\mu} - L\mu_0| \leq -s_n z_{\gamma/2} \right) \leq P\left( |Wt_n| \geq |E_{\mu_0}L\tilde{\mu} - L\mu_0| - s_n z_{\gamma/2} \right) \to 0 \).

If the scaling rate is fixed to \( \tau_n \equiv 1 \), then it can be checked from 4.25 and the proof of Theorem 4.7 that \( B_n \ll s_n, B_n \asymp s_n \) and \( B_n \gg s_n \) in the three cases \( \alpha < \beta - 1/2, \alpha = \beta - 1/2 \) and \( \alpha \geq \beta - 1/2 \), respectively. In the first and third cases the maximal bias and the root spread differ by more than a logarithmic term \( (\log n)^{\delta} \). It follows that the preceding analysis (i), (ii), (iii) extends to this situation.

### 4.7 Proof of Theorem 4.6

(i). The two quantities \( s_n \) and \( t_n \) are given by

\[
s_n^2 = \sum_{i=1}^{\infty} \frac{t_i^2 \lambda_i}{1 + n\lambda_i \kappa_i^2}, \quad t_n^2 = \sum_{i=1}^{\infty} \frac{t_i^2 n\lambda_i \kappa_i^2}{(1 + n\lambda_i \kappa_i^2)^2}
\]

Every term in the series \( t_n^2 \) is \( n\lambda_i \kappa_i^2 / (1 + n\lambda_i \kappa_i^2) \leq 1 \) times the corresponding term in the series \( s_n^2 \). Therefore, \( s_n/t_n \to 1 \) if and only if the series are determined by the terms for which these numbers are “close to” 1, that is, \( n\lambda_i \kappa_i^2 \) is large. More precisely, we show below that \( s_n/t_n \to 1 \) if and only if, for every \( c > 0 \),

\[
\sum_{n\lambda_i \kappa_i^2 \leq c} \frac{t_i^2 \lambda_i}{1 + n\lambda_i \kappa_i^2} = o \left( \sum_i \frac{t_i^2 \lambda_i}{1 + n\lambda_i \kappa_i^2} \right)
\]

(4.26)
If \( l \in S^p \), then the series on the left is as in Lemma A.1 with \( q = p, u = 1 + 2\alpha + 2p, v = 1, N = n\tau_n^2 \) and \( t = 1 + 2\alpha \). Hence, \((t + 2q)/u \geq v\), and the display follows from the final assertion of the lemma. If \( l_i = i^{-q-1/2}S(i)\) for a slowly varying function \( S\), then the series is as in Lemma A.3, with the same parameters, and by the last statement of the lemma the display is true if and only if \((t + 2q)/u \geq v\), that is, \( q \geq p\).

To prove that (4.26) holds iff \( s_n/t_n \to 1\), write \( s_n^2 = A_n + B_n\), for \( A_n \) and \( B_n \) the sums over the terms in \( s_n^2 \) with \( n\lambda_i\kappa_i^2 > c \) and \( n\lambda_i\kappa_i^2 \leq c\), respectively, and, similarly, \( t_n^2 = C_n + D_n\). Then
\[
\frac{D_n}{B_n} \leq \frac{c}{1 + c} \leq \frac{C_n}{A_n} \leq 1.
\]
It follows that
\[
\frac{t_n^2}{s_n^2} = \frac{C_n + D_n}{A_n + B_n} = \frac{C_n/A_n + (D_n/B_n)(B_n/A_n)}{1 + B_n/A_n} \leq \frac{1 + c/(1 + c)(B_n/A_n)}{1 + B_n/A_n}.
\]
Because \( x \mapsto (1 + rx)/(1 + x) \) is strictly decreasing from 1 at \( x = 0 \) to \( r < 1 \) at \( x = \infty \) (if \( 0 < r < 1 \)), the right-hand side of the equation is asymptotically 1 if and only if \( B_n/A_n \to 0\), and otherwise its liminf is strictly smaller. Thus, \( t_n/s_n \to 1 \) implies that \( B_n/A_n \to 0\).

Second,
\[
\frac{t_n^2}{s_n^2} \geq \frac{C_n}{A_n + B_n} = \frac{C_n/A_n}{1 + B_n/A_n} \geq \frac{c/(1 + c)}{1 + B_n/A_n}.
\]
It follows that \( \lim \inf t_n^2/s_n^2 \geq c/(1 + c) \) if \( B_n/A_n \to 0\). This being true for every \( c > 0 \) implies that \( t_n/s_n \to 1\).

(ii). If \( l \in S^p \), then \( \sum_{i=1}^{\infty} l_i^2\kappa_i^{-2} < \infty \). Clearly
\[
nl^2_n = \sum_{i=1}^{\infty} \frac{l_i^2n^2\lambda_i^2\kappa_i^2}{(1 + n\lambda_i\kappa_i^2)^2} \leq \sum_{i=1}^{\infty} \frac{l_i^2}{\kappa_i^2}.
\]
For \( i \leq \ell_n := (C\sqrt{n\tau_n})^{1/(1+2\alpha+2p)} \) we have \( \sqrt{n\lambda_i\kappa_i^2} > 1 \), and thus
\[
\frac{n}{(1 + \sqrt{n})^2} \left( \sum_{i=1}^{\infty} \frac{l_i^2}{\kappa_i^2} - o(1) \right) = \sum_{i \leq \ell_n} \frac{l_i^2n^2\lambda_i^2\kappa_i^2}{(\sqrt{n\lambda_i\kappa_i^2} + n\lambda_i\kappa_i^2)^2} \leq nl^2_n.
\]
If \( l \in R^q \), then we apply Lemma A.3 with \( q = p \) (\( q > p \) is included in \( l \in S^p \)), \( t = 1 + 2\alpha \), \( u = 1 + 2\alpha + 2p \), \( v = 1 \) and \( N = n\tau_n^2 \) to see that \( s_n^2 \leq n^{-1} \sum_{i \leq N^{1/2}v^1} S^2(i)/i \).

(iii). If \( l \in S^q \), then the bias is bounded above in (4.17), and in the proof of Theorem 4.3 its supremum \( B_n \) over \( \| \mu_0 \|_{\mathcal{B}} \lesssim 1 \) is seen to be bounded by \( (n\tau_n^2)^{-((\beta + q)/(1+2\alpha+2p))\wedge 1} \), the first term in the definition of \( \varepsilon_n \) in the statement of this theorem. This upper bound is \( o(n^{-1/2}) \) iff the stated conditions hold. Here we use that \( S^2(N) \ll \sum_{i \leq N} S^2(i)/i \) as \( N \to \infty \), as noted in the proof of Lemma A.3.

The supremum of the bias \( B_n \) in the case that \( l \in R^q \) is given in (4.24). It was already seen to be \( o(t_n) \) if \( \tau \gg \tau_n \) in the proof of case (i) of Theorem 4.5. If \( \tau_n = 1 \), we have that \( B_n \simeq n^{-(\beta + q)/(1+2\alpha+2p)}\wedge 1 \), for \( \gamma_n \) the slowly varying factor given in the statement of Theorem 4.4. Furthermore, we have \( s_n \asymp t_n \asymp n^{-1/2} \), for \( \delta_n \) the slowly varying factor in
the same statement. Under the present conditions, \( \delta_n \approx 1 \) if \( q > p \) and \( \delta_n^2 \approx \sum_{i \leq \rho_n} S^2(i)/i \) if \( q = p \). We can now verify that \( B_n = o(t_n) \) if and only if the conditions as stated hold.

(iv). The total variation distance between two Gaussian distributions with the same expectation and standard deviations \( s_n \) and \( t_n \) tends to zero if and only if \( s_n/t_n \to 1 \). Therefore,

\[
E_{\mu_0} \sup_B |\Pi_n(L_\mu \in B \mid Y) - N(L_\mu, t_n^2(B))| \to 0,
\]

and it suffices to show that \((L_\mu - L_{\mu_0})/s_n\) converges under \( \mu_0 \) in distribution to a standard normal distribution, uniformly in \( \|\mu_0\|_\beta \lesssim 1 \). The total variance distance between two Gaussians with the same standard deviation \( \sigma_n \) and means \( \mu_n \) and \( \nu_n \) tends to zero if and only if \( \mu_n - \nu_n = o(\sigma_n) \). Note that \( \mu_n = B_n, \nu_n = 0 \) and \( \sigma_n = t_n \). The uniform convergence in the total variance distance implies the desired convergence in distribution.

Now let \( l \in S^p \). By (ii) the rescaled variance of the posterior mean \( nt_n^2 \) converges to \( \sum_{i=1}^{\infty} l_i^2 \kappa_i^{-2} \). It suffices to show that \((L_\mu - \sum_{i=1}^{\infty} Y_i l_i/\kappa_i)/s_n \to 0 \). Under Assumption 4.2 this difference is equal to

\[
\frac{E_{\mu_0} L_\mu - L_{\mu_0}}{s_n} + \frac{1}{s_n \sqrt{n}} \left( \sum_{i=1}^{\infty} \frac{nl_i \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Z_i - \sum_{i=1}^{\infty} \frac{l_i}{\kappa_i} Z_i \right).
\]

Because the bias is \( o(t_n) \) and \( s_n \approx n^{-1/2} \), it suffices to show that

\[
\sum_{i=1}^{\infty} \frac{nl_i \lambda_i \kappa_i}{1 + n \lambda_i \kappa_i^2} Z_i - \sum_{i=1}^{\infty} \frac{l_i}{\kappa_i} Z_i = \sum_{i=1}^{\infty} \frac{Z_i l_i}{\kappa_i} \left( \frac{1}{1 + n \lambda_i \kappa_i^2} \right)
\]

converges to zero. If \( \sum_{i=1}^{\infty} l_i^2 / \kappa_i^2 < \infty \), then the variance of this expression is seen to tend to zero by dominated convergence.

The final assertion of the theorem follows along the lines of the proof of Theorem 4.5.
Part II

Irregular models
Chapter 5

Semiparametric posterior limits under LAE

5.1 Introduction

Posterior limits are often studied in finite-dimensional settings under some kind of regularity condition, usually associated with the local asymptotic normality of the model. Connection between asymptotic behaviors of posterior distribution and sampling distribution of an efficient estimator has attracted the biggest attention, since as such is considered to be an attractive link between frequentist and Bayesian statistics. Indeed, if the sample size is large enough, posterior credible sets coincide with frequentist confidence sets, and vice versa. Due to the efficiency of the underlying frequentist estimator, the corresponding confidence sets, and hence the credible sets, are of the optimal size.

In recent years, asymptotic efficiency of Bayesian semiparametric methods has enjoyed much attention. The general question concerns a nonparametric model $\mathcal{P}$ in which exclusive interest goes to the estimation of a sufficiently smooth, finite-dimensional functional of interest. Asymptotically, regularity of the estimator combined with the Cramér–Rao bound in the Gaussian location model that forms the limit experiment (see [67]) fixes the rate of convergence to $n^{-1/2}$ and poses a bound to the accuracy of regular estimators expressed through Hajék’s convolution (see, e.g., [47]) and asymptotic minimax theorems (see, e.g., [48]). As seen already in Chapter 1, in Bayesian context, efficiency of estimation in regular (semi-)parametric models is best captured by a so-called Bernstein–von Mises limit.

In the previous chapter we studied a semiparametric model derived from the nonparametric inverse problem setting. The rates associated with recovery of linear functionals of the parameter of interest in the ill-posed inverse problem model can be slower than the regular parametric $n^{-1/2}$ rate. Due to conjugacy, the limiting shape of the posterior distribution was known to be Gaussian. We obtained a Bernstein–von Mises-type theorem stating that the posterior distribution and the distribution of the posterior mean (roughly speaking an efficient estimator in the setting of previous chapter, see discussion after Theorem 4.6) are close in the total variation distance. The associated rate was shown to be exactly of the order $n^{-1/2}$ (in case of efficient estimation), or of the order $n^{-1/2}$ decelerated by a slowly varying factor.
Chapter 5. Semiparametric posterior limits under LAE

In the present chapter we also study a semiparametric model, however now related to rates faster than \( n^{-1/2} \). We focus on an irregular problem that stems from the following classical example of estimation of a point of discontinuity of a density: to be a bit more specific, consider an almost-everywhere differentiable Lebesgue density on \( \mathbb{R} \) that displays a jump at some point \( \theta \in \mathbb{R} \); estimators for \( \theta \) exist that converge at rate \( n^{-1} \) with exponential limit distributions (cf. [50]). To illustrate the form that this conclusion takes in Bayesian context, consider the following theorem.

**Theorem 5.1.** For \( \theta \in \mathbb{R} \), let \( F_\theta(x) = (1 - e^{-(x-\theta)}) \vee 0 \). Assume that \( X_1, X_2, \ldots \) form an i.i.d. sample from \( F_{\theta_0} \), for some \( \theta_0 \). Let a prior density \( \pi: \mathbb{R} \to (0, \infty) \) be a continuous Lebesgue probability density. Then the associated posterior distribution satisfies,

\[
\sup_A \left| \Pi_n(\theta \in A \mid X_1, \ldots, X_n) - \text{Exp}_{\hat{\theta}_n,n}(A) \right| \to 0,
\]

in \( P_{\theta_0} \)-probability, where \( \hat{\theta}_n = X_{(1)} \) is the maximum likelihood estimate for \( \theta_0 \), and the supremum is taken over all measurable subsets of \( \Theta \).

The proof of this Bernstein–von Mises limit is elementary and does not depend in any crucial way on the particular parametric family of distributions that we chose. Any distribution with a density \( \eta \) that has a discontinuity at \( \theta \) and is of bounded variation with finite Fisher information for location can be used instead (see also the proof of Theorem 5.10).

If we consider the underlying distribution to be unknown, but also irrelevant to us, the problem can be viewed as semiparametric. As a frequentist semiparametric problem, estimation of a support boundary point is a well-understood problem (see [50]): assuming that the distribution \( P_\theta \) of \( X \) is supported on the half-line \( [\theta, \infty) \) and an i.i.d. sample \( X_1, X_2, \ldots, X_n \) is given, we follow [50] and estimate \( \theta \) with the first order statistic \( X_{(1)} = \min_i \{ X_i \} \). If \( P_\theta \) has an absolutely continuous Lebesgue density of the form \( p_\theta(x) = \eta(x - \theta) 1_{\{x \geq \theta\}} \), its rate of convergence is determined by the behavior of the quantity \( \epsilon \mapsto \int_0^\epsilon \eta(x) \, dx \) for small values of \( \epsilon \). If \( \eta(x) = x^\alpha (1 + o(1)) \) as \( x \downarrow 0 \), for some \( \alpha \in (-1, 1) \), then,

\[
n^{1/(1+\alpha)} \left( X_{(1)} - \theta \right) = O_{P_\theta}(1).
\]

For densities of this form, for any sequence \( \theta_n \) that converges to \( \theta \) at rate \( n^{-1/(1+\alpha)} \), Hellinger distances obey (see Theorem VI.1.1 in [50]):

\[
n^{1/2} H(P_{\theta_n}, P_\theta) = O(1).
\]

If we substitute the estimators \( \theta_n = \hat{\theta}_n(X_1, \ldots, X_n) = X_{(1)} \), uniform tightness of the sequence in the above display signifies rate optimality of the estimator (cf. [68, 69]). Regarding asymptotic efficiency beyond rate-optimality, e.g., in the sense of minimal asymptotic variance (or other measures of dispersion of the limit distribution), one notices that the (one-sided) limit distributions one obtains for \( X_{(1)} \) can always be improved upon by de-biasing (see Section VI.6, examples 1–3 in [50], and [70]).

As a semiparametric Bayesian question, the matter of estimating support boundaries is not settled by the above: for the posterior, it is the local limiting behavior of the likelihood
around the point of convergence (rather than just its point of maximum; see, e.g., Theorems VI.2.1–VI.2.3 in [50]) that determines convergence rather than the behavior of any particular statistic. In this chapter we shed some light on the behavior of marginal posteriors for the parameter of interest in semiparametric, irregular estimation problems, through a study of the Bernstein–von Mises phenomenon. Only the prototypical case of a density of bounded variation, supported on the half-line \([\theta, \infty)\) or on the interval \([0, \theta]\), with a finite jump at \(\theta\), is analyzed in detail. We offer a slight abstraction from the prototypical case, by considering the class of models that exhibit a weakly converging expansion of the likelihood called local asymptotic exponentiality (LAE) (cf. [50]), to be compared with local asymptotic normality (cf. [66]) in regular problems. Like in the parametric case of Theorem 5.1, this type of asymptotic behavior of the likelihood is expected to give rise to a (negative-)exponential marginal posterior limit satisfying the irregular Bernstein–von Mises limit:

\[
\sup_A \left| \Pi_n \left( h \in A \mid X_1, \ldots, X_n \right) - \text{Exp}_{\Delta_n, \gamma_{\theta_0, \eta_0}} \right| \to 0, \tag{5.3}
\]

in \(P_0^n\)-probability, where \(h = n(\theta - \theta_0)\), the supremum is taken over all measurable subsets of the properly rescaled \(\Theta\), and the random sequence \(\Delta_n\) converges weakly to exponentiality (see Definition 5.2). Like in the regular case, the limit (5.3) allows for the asymptotic identification of credible sets with confidence intervals associated with the maximum likelihood estimator. The constant \(\gamma_{\theta_0, \eta_0}\) determines the scale in the limiting exponential distribution and, as such, the width of credible sets. In this chapter, we explore general sufficient conditions on model and prior to conclude that the limit (5.3) obtains. The main theorem is applied in two semiparametric LAE example models, one for a shift parameter and one for a scale parameter (compare with the two regular semiparametric questions in [93]).

The chapter is structured as follows: in Section 5.2 we give the main theorem and a corollary that simplifies the formulation. In Section 5.3, the proof of the main theorem is built up in several steps, from consistency under perturbation introduced in [7], to an LAE expansion for integrated likelihoods and on to posterior exponentiality of the type described by (5.3). Section 5.4 discusses two semiparametric LAE models to demonstrate that they satisfy the exponential Bernstein–von Mises property (5.3) asymptotically.

**Notation and conventions**

The (frequentist) true distribution of each of the data points in the i.i.d. sample \(X_n = (X_1, \ldots, X_n)\) is denoted \(P_0\) and assumed to lie in the model \(\mathcal{P}\). Associated order statistics are denoted \(X_{(1)}, X_{(2)}, \ldots\). The location-scale family associated with the exponential distribution is denoted \(\text{Exp}_{\Delta, \lambda}^+\) and its negative version by \(\text{Exp}_{\Delta, \lambda}^-\). We localize \(\theta\) by introducing \(h = n(\theta - \theta_0)\) with inverse \(\theta_n(h) = \theta_0 + n^{-1}h\). The expectation of a random variable \(f\) with respect to a probability measure \(P\) is denoted \(Pf\); the sample average of \(g(X)\) is denoted

\[
\mathbb{P}_ng(X) = \frac{1}{n} \sum_{i=1}^n g(X_i), \quad \text{and} \quad G_ng(X) = \sqrt{n}(\mathbb{P}_ng(X) - Pg(X)).
\]
If $h_n$ is stochastic, $P_{\theta_n(h_n),\eta}^n f$ denotes the integral
\[ \int f(\omega) \frac{dP_{\theta_n(h_n)(\omega),\eta}(\omega)}{dP_0^n(\omega)} dP_0^n(\omega). \]

The Hellinger distance between $P$ and $P'$ is denoted $H(P, P')$ and induces a metric $d_H$ on the space of nuisance parameters $H$ by $d_H(\eta, \eta') = H(P_{\theta_0,\eta}, P_{\theta_0,\eta'})$, for all $\eta, \eta' \in H$. A prior on (a subset $\Theta$ of) $\mathbb{R}^k$ is said to be thick (at $\theta \in \Theta$) if it is Lebesgue absolutely continuous with a density that is continuous and strictly positive (at $\theta$).

## 5.2 Main results

Throughout this chapter we consider estimation of a functional $\theta: \mathcal{P} \rightarrow \mathbb{R}$ on a nonparametric model $\mathcal{P}$ based on a sample $X_1, X_2, \ldots$, distributed i.i.d. according to some unknown $P_0 \in \mathcal{P}$. We assume that $\mathcal{P}$ is parametrized in terms of a one-dimensional parameter of interest $\theta \in \Theta$ and a nuisance parameter $\eta \in H$ so that we can write $\mathcal{P} = \{P_{\theta,\eta}; \theta \in \Theta, \eta \in H\}$, and that $\mathcal{P}$ is dominated by a $\sigma$-finite measure on the sample space with densities $p_{\theta,\eta}$, $(\theta \in \Theta, \eta \in H)$. The set $\Theta$ is open in $\mathbb{R}$, and $(H, d_H)$ is an infinite dimensional metric space (to be specified further at later stages). Assuming identifiability, there exist unique $(\theta_0, \eta_0) \in \Theta \times H$ such that $P_0 = P_{\theta_0,\eta_0}$. Assuming measurability of the map $(\theta, \eta) \rightarrow P_{\theta,\eta}$ and priors $\Pi_\Theta$ on $\Theta$ and $\Pi_H$ on $H$, the prior $\Pi$ on $\mathcal{P}$ is defined as the product prior $\Pi_\Theta \times \Pi_H$ on $\Theta \times H$, lifted to $\mathcal{P}$. The subsequent sequence of posteriors takes the form,
\[ \Pi_n(A|X_1, \ldots, X_n) = \int_A \prod_{i=1}^{n} p(X_i) d\Pi(P) \bigg/ \int_{\mathcal{P}} \prod_{i=1}^{n} p(X_i) d\Pi(P), \quad (5.4) \]
where $A$ is any measurable model subset.

Throughout most of this chapter, the parameter of interest $\theta$ is represented in localized form, by centering on $\theta_0$ and rescaling: $h = n(\theta - \theta_0) \in \mathbb{R}$. (We also make use of the inverse $\theta_n(h) = \theta_0 + n^{-1} h$.) The following (irregular) local expansion of the likelihood is due to Ibragimov and Has’minskii [50].

**Definition 5.2.** (Local asymptotic exponentiality) A one-dimensional parametric model $\theta \mapsto P_{\theta,\eta}$ is said to be locally asymptotically exponential (LAE) at $\theta_0 \in \Theta$ if there exists a sequence of random variables $(\Delta_n)$ and a positive constant $\gamma_{\theta_0,\eta}$ such that for all $(h_n)$, $h_n \rightarrow h$,
\[ \prod_{i=1}^{n} \frac{p_{\theta_0 + nh_n,\eta}(X_i)}{p_{\theta_0,\eta}} = \exp(h \gamma_{\theta_0,\eta} + o_{P_{\theta_0}}(1)) 1\{h \leq \Delta_n\}, \]
with $\Delta_n$ converging weakly to $\text{Exp}^+_{0,\gamma_{\theta_0,\eta}}$.

We use the above definition in a semiparametric context, therefore we introduce the $\eta$-dependence in the notation. In many examples, e.g., that of Subsection 5.4.1, $\Delta_n$ and its weak limit are independent of $\theta_0$. This definition should be viewed as an irregular variation on the one-dimensional version of Le Cam’s local asymptotic normality (LAN, see [66]),
which forms the smoothness requirement in the context of the semiparametric Bernstein–von Mises theorem (see [7]). Like the LAN expansion gives rise to asymptotic normality of the marginal posterior for the parameter of interest, an LAE expansion is expected to give rise to a one-sided, exponential marginal posterior limit, cf. (5.3). We are interested in general sufficient conditions on model and prior to conclude that the limit (5.3) obtains.

In order to establish the limit (5.3), we study posterior convergence of a particular type, termed consistency under perturbation in [7]. One can compare this type of consistency with ordinary posterior consistency in nonparametric models, except here the nonparametric component is the nuisance parameter \( \eta \) and we allow for (stochastic) perturbation by (local) deformations of the parameter of interest \( \theta_n(h_n) = \theta_0 + n^{-1}h_n \). In regular situations, this gives rise to accumulation of posterior mass around so-called least-favorable submodels (see [7]), but here the parameter of interest is irregular and the situation is less involved: accumulation of posterior mass occurs around \((\theta_n(h_n), \eta_0)\). Therefore, posterior consistency under perturbation describes concentration in \(d_H\)-neighborhoods of the form,

\[
D(\rho) = \{ \eta \in H : d_H(\eta, \eta_0) < \rho \}.
\] (5.5)

To guarantee sufficiency of prior mass around the point of convergence, we use Kullback–Leibler-type neighborhoods of the form,

\[
K_n(\rho, M) = \left\{ \eta \in H : P_0 \left( \sup_{|h| \leq M} 1_{A_{\theta_n(h)}} \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} \right) \leq \rho^2, P_0 \left( \sup_{|h| \leq M} 1_{A_{\theta_n(h)}} \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} \right)^2 \leq \rho^2 \right\},
\] (5.6)

where, in the present LAE setting,

\[
A_{\theta_n(h)} = \left\{ x : \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}}(x) > 0 \right\}.
\]

Note that \( \prod_{i=1}^n 1_{A_{\theta_n(h)}}(X_i) = 1_{\{ h_\leq \Delta_n \}} \), as in the LAE expansion.

Suppose that \( A \) in (5.4) is of the form \( A = B \times H \) for some measurable \( B \subset \Theta \). Since we use a product prior \( \Pi_\Theta \times \Pi_H \), the marginal posterior of the parameter \( \theta \in \Theta \) depends on the nuisance factor only through the integrated likelihood,

\[
S_n : \Theta \to \mathbb{R} : \theta \mapsto \int_H \prod_{i=1}^n \frac{p_{\theta, \eta}}{p_{\theta_0, \eta_0}}(X_i) d\Pi_H(\eta),
\] (5.7)

and its localized version, \( h \mapsto s_n(h) = S_n(\theta_0 + n^{-1}h) \). One of the conditions of the subsequent theorem is a domination condition based on the quantities,

\[
U_n(\rho, h_n) = \sup_{\eta \in D(\rho)} P_{\theta_0, \eta}^n \left( \prod_{i=1}^n \frac{p_{\theta_n(h_n), \eta}}{p_{\theta_0, \eta}}(X_i) \right).
\]

Another condition required in the irregular version of the semiparametric Bernstein–von Mises theorem is one-sided contiguity (cf. condition (iv) of Theorem 5.3 below). Lemma 5.11
shows that such one-sided contiguity and domination as in (5.8) are closely related and provides two different sufficient conditions for both to hold in general. The log-Lipschitz construction is used in the examples of Section 5.4; in other applications of the theorem it may be more convenient to by-pass Lemma 5.11 and prove (5.8) and contiguity directly from the model definition.

The main result of this chapter is the following irregular version of the semiparametric Bernstein–von Mises theorem:

**Theorem 5.3.** (LAE: Semiparametric Bernstein–von Mises). Let $X_1, X_2, \ldots$ be distributed i.i.d.-$P_0$, with $P_0 \in \mathcal{P}$. Let $\Pi_H$ and $\Pi_\Theta$ be priors on $H$ and $\Theta$ and assume that $\Pi_\Theta$ is thick at $\theta_0$. Suppose that $\theta \mapsto P_\theta, \eta$ is stochastically LAE in the $\theta$-direction, for all $\eta$ in a $d_H$-neighborhood of $\eta_0$ and that $\gamma_{\theta_0, \eta_0} > 0$. Assume also that for large enough $n$, the map $h \mapsto s_n(h)$ is continuous on $(-\infty, \Delta_n]$, $P_0^n$-almost-surely. Furthermore, assume that there exists a sequence $(\rho_n)$ with $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$ such that,

(i) for all $M > 0$, there exists a $K > 0$ such that for large enough $n$,

$$
\Pi_H(K_n(\rho_n, M)) \geq e^{-Kn\rho_n^2},
$$

(ii) for all $n$ large enough, the Hellinger metric entropy satisfies,

$$
N(\rho_n, H, d_H) \leq e^{n\rho_n^2},
$$

and, for every bounded, stochastic $(h_n)$,

(iii) the model satisfies the domination condition,

$$
U_n(\rho_n, h_n) = O(1),
$$

(iv) for every $\eta \in D(\rho)$ for $\rho > 0$ small enough, the sequence $P_{\theta_n(\rho_n), \eta}^n$ is contiguous with respect to the sequence $P_{\theta_0, \eta}^n$,

(v) and for all $L > 0$, Hellinger distances satisfy the uniform bound,

$$
\sup_{\eta \in D^c(L, \rho_n)} \frac{H(P_{\theta_n(\rho_n), \eta}, P_{\theta_0, \eta})}{H(P_{\theta_0, \eta}, P_0)} = o(1).
$$

Finally, suppose that,

(vi) for every $(M_n)$, $M_n \to \infty$, the posterior satisfies

$$
\Pi_n(|h| \leq M_n | X_1, \ldots, X_n) \xrightarrow{P_0} 1.
$$

Then the sequence of marginal posteriors for $\theta$ converges in total variation to a negative exponential distribution,

$$
\sup_A \left| \Pi_n(h \in A | X_1, \ldots, X_n) - \text{Exp}_{\Delta_n, \gamma_{\theta_0, \eta_0}, \eta_0}(A) \right| \xrightarrow{P_0} 0.
$$

(5.9)
5.2. Main results

Regarding the nuisance rate of convergence $\rho_n$, conditions (i) and (ii) are expected in some form or other in order to achieve consistency under perturbation. As stated, they almost co-incide with requirements for nonparametric convergence at rate $(\rho_n)$ without a parameter of interest (see [39]). A simplified version of Theorem 5.3 that does not refer to any specific nuisance $\rho_n$ is stated as Corollary 5.4. In the rate-free case of Corollary 5.4, conditions on prior mass and entropy numbers ((i) and (ii)) essentially require nuisance consistency (at some rate rather than a specific one), thus weakening requirements on model and prior. Concerning conditions (iii)–(v), note that, typically, the numerator in condition (v) converges to zero at rate $O(n^{-1/2})$ (cf. (5.2)) while the denominator goes to zero at slower, nonparametric rate. As such, condition (v) is to be viewed as a weak condition that rarely poses a true restriction on the applicability of the theorem. Furthermore, Lemma 5.11 formulates two slightly stronger conditions to validate both (iii) and (iv) above for any rate $\rho_n$.

Condition (vi) of Theorem 5.3 appears to be the hardest to verify in applications. On the other hand it cannot be weakened (since (vi) also follows from (5.9)). Besides condition (i), only condition (vi) implies a requirement on the nuisance prior $\Pi_H$. Experience with the LAN version (see [7]) suggests that conditions (i)–(v) are relatively weak in applications, while (vi) harbors the potential for negative surprises, mainly due to semiparametric bias leading to sub-optimal asymptotic variance, sub-optimal marginal rate or even marginal inconsistency. On the other hand, there are conditions under which condition (vi) is easily seen to be valid: in Section 5.3.3 we present a model condition that guarantees marginal posterior convergence according to (vi) for any choice of the nuisance prior $\Pi_H$.

As discussed already after Theorem 5.3, in many situations the domination condition holds for any rate $(\rho_n)$. This circumstance simplifies the result substantially, leading to the conditions that are comparable to those of Schwartz’ consistency theorem (see [89]).

**Corollary 5.4.** (Rate-free Semiparametric Bernstein–von Mises). Let $X_1, X_2, \ldots$ be distributed i.i.d.-$P_0$, with $P_0 \in \mathcal{P}$ and let $\Pi_\Theta$ be thick at $\theta_0$. Suppose that $\theta \mapsto P_{\theta,\eta}$ is stochastically LAE in the $\theta$-direction, for all $\eta$ in a $d_H$-neighborhood of $\eta_0$ and that $\gamma_{\theta_0,\eta_0}$ is strictly positive. Also assume that for large enough $n$, the map $h \mapsto s_n(h)$ is continuous on $(\theta_0, \Delta_n] P^n_\eta$-almost-surely. Furthermore, assume that,

(i) for all $\rho > 0$, the Hellinger metric entropy satisfies $N(\rho, H, d_H) < \infty$, and the nuisance prior satisfies $\Pi_H(K(\rho)) > 0$,

(ii) for every $M > 0$, there exists an $L > 0$ such that for all $\rho > 0$ and large enough $n$ $K(\rho) \subset K_n(L\rho, M)$,

and that for every bounded, stochastic $(h_n)$,

(iii) there exists an $r > 0$ such that $U_n(r, h_n) = O(1)$,

(iv) for every $\eta \in D(r)$ the sequence $P^n_{\theta_n(h_n),\eta}$ is contiguous to the sequence $P^n_{\theta_0,\eta}$,

(v) and that Hellinger distances satisfy, $\sup_{\eta \in H} H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}) = O(n^{-1/2})$.

Finally, assume that,
(vi) for every \((M_n)\), \(M_n \to \infty\), the posterior satisfies,

\[
\Pi_n( |h| \leq M_n | X_1, \ldots, X_n ) \xrightarrow{P_0} 1.
\]

Then marginal posteriors for \(\theta\) converge in total variation to a negative exponential distribution,

\[
\sup_A \left| \Pi_n( h \in A | X_1, \ldots, X_n ) - \text{Exp}_{\Delta_n, \gamma_{\theta_0}, \eta_0}(A) \right| \xrightarrow{P_0} 0.
\]

**Proof.** Under conditions (i), (ii), (v), and the stochastic LAE assumption, the assertion of Corollary 5.8 holds. Due to conditions (iii) and (iv), conditions (iii) and (iv) in Theorem 5.3 are satisfied for large enough \(n\). Condition (vi) then suffices for the assertion of Theorem 5.10. \(\square\)

### 5.3 Asymptotic posterior exponentiality

In this section we give the proof of Theorem 5.3 in several steps: the first step (Subsection 5.3.1) is a proof of consistency under perturbation under a condition on the nuisance prior \(\Pi_H\) and a testing condition. In Subsection 5.3.2 we show that the integral of the likelihood with respect to the nuisance prior displays an LAE expansion, if consistency under perturbation obtains and contiguity/domination conditions are satisfied. In the third step, also discussed in Subsection 5.3.2, we show that an LAE expansion of the integrated likelihood gives rise to a semiparametric exponential limit for the posterior in total variation, if the marginal posterior for the parameter of interest converges at \(n^{-1}\) rate. The rate of marginal convergence depends on the existence of a suitable test sequence, which is discussed in Subsection 5.3.3. Put together, the results constitute a proof of Theorem 5.3. Stated conditions are verified in two examples in Section 5.4.

#### 5.3.1 Posterior convergence under perturbation

Given a rate sequence \((\rho_n)\), \(\rho_n \downarrow 0\), we say that the conditioned nuisance posterior is *consistent under \(n^{-1}\)-perturbation at rate \(\rho_n\)*, if, for all bounded, stochastic sequences \((h_n)\),

\[
\Pi_n \left( D^c(\rho_n) \mid \theta = \theta_0 + n^{-1} h_n, X_1, \ldots, X_n \right) \xrightarrow{P_0} 0,
\]

(For a more elaborate discussion of this property, the reader is referred to [7]).

**Theorem 5.5.** (Posterior convergence under perturbation). Assume there is a sequence \((\rho_n)\), \(\rho_n \downarrow 0\), \(n\rho_n^2 \to \infty\) with the property that for all \(M > 0\) there exists a \(K > 0\) such that,

\[
\Pi_H( K_n(\rho_n, M) ) \geq e^{-Kn\rho_n^2}, \quad N(\rho_n, H, d_H) \leq e^{n\rho_n^2},
\]

for large enough \(n\). Assume also that for all \(L > 0\) and all bounded, stochastic \((h_n)\),

\[
\sup_{\eta \in D^c(L\rho_n)} \frac{H(P_{\theta_n(h_n), \eta}, P_{\theta_0, \eta})}{H(P_{\theta_0, \eta}, P_0)} = o(1).
\]
Then, for every bounded, stochastic \((h_n)\), there exists an \(L > 0\) such that,

\[
\Pi_n \left( D^c(L\rho_n) \mid \theta = \theta_0 + n^{-1}h_n, X_1, \ldots, X_n \right) = o_{P_0}(1).
\]

The proof of this theorem, given in Subsection 5.5.1, can be broken down into two separate steps, with the following testing condition in between: for every bounded, stochastic \((h_n)\) and all \(L > 0\) large enough, a test sequence \((\phi_n)\) and constant \(C > 0\) must exist, such that,

\[
P_0^n(\phi_n) \to 0, \quad \sup_{\eta \in D^c(L\rho_n)} P_{\theta_n(h_n),\eta}^n(1 - \phi_n) \leq e^{-CL^2n^2}, \quad (5.11)
\]

for large enough \(n\). According to Lemma 5.7, the metric entropy condition and “cone condition” (5.10) suffice for the existence of such a test sequence. Here, we concatenate and refer to [7] for a full discussion. While the above testing argument is instrumental in the control of the numerator of (5.4), the denominator of the posterior is lower bounded with the help of the following lemma, which adapts Lemma 8.1 in [39] to \(n^{-1}\)-perturbed, irregular setting. For the proofs of the following lemmas we refer the reader to Subsection 5.5.1.

**Lemma 5.6.** Let \((h_n)\) be stochastic and bounded by some \(M > 0\). Then \(P_0^n(B_n) \leq (C^2n^2)^{-1}\), for all \(C > 0\), \(\rho > 0\) and \(n \geq 1\), where \(\theta_n(h_n) = \theta_0 + n^{-1}h_n\), and,

\[
B_n = \left\{ \left( \int_H \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}(X_i)}{p_0}(X_i) d\Pi_H(\eta) < e^{-(1+C)n\rho^2} \Pi_H(K_n(\rho,M)) \right) \cap \left\{ h_n \leq \Delta_n \right\} \right\}.
\]

**Lemma 5.7.** (Testing under perturbation: Lemma 3.2 in [7]). If \((\rho_n)\) satisfies \(\rho_n \downarrow 0\), \(n^2\rho_n^2 \to \infty\) and the following requirements are met:

(i) For all \(n\) large enough, \(N(\rho_n, H, d_H) \leq e^{n\rho^2}\).

(ii) For all \(L > 0\) and all bounded, stochastic \((h_n)\),

\[
\sup_{\eta \in D^c(L\rho_n)} \frac{H(p_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} = o(1).
\]

Then for all \(L \geq 4\), there exists a test sequence \((\phi_n)\) such that for all bounded, stochastic \((h_n)\),

\[
P_0^n(\phi_n) \to 0, \quad \sup_{\eta \in D^c(L\rho_n)} P_{\theta_n(h_n),\eta}^n(1 - \phi_n) \leq e^{-L^2n^2/4}
\]

for large enough \(n\).

In many applications, \((\rho_n)\) does not play an explicit role because consistency at some rate is sufficient. The following provides a possible formulation of weakened conditions guaranteeing consistency under perturbation. Corollary 5.8 is based on the family of Kullback–Leibler neighborhoods that would also play a role for marginal posterior consistency of the nuisance with known \(\theta_0\) (as in [39]):

\[
K(\rho) = \left\{ \eta \in H : -P_0 \log \frac{p_{\theta_0,\eta}}{p_0} \leq \rho^2, P_0 \left( \log \frac{p_{\theta_0,\eta}}{p_0} \right)^2 \leq \rho^2 \right\},
\]

for \(\rho > 0\).
Corollary 5.8. Assume that for all $\rho > 0$, $N(\rho, H, d_H) < \infty$ and $\Pi_H(K(\rho)) > 0$. Furthermore, assume that for every stochastic, bounded $(h_n)$,

(i) for every $M > 0$, there exists an $L > 0$ such that for all $\rho > 0$ and large enough $n$, $K(\rho) \subset K_n(L\rho, M)$.

(ii) Hellinger distances satisfy $\sup_{\eta \in H} H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}) = O(n^{-1/2})$.

Then there exists a sequence $(\rho_n)$, $\rho_n \downarrow 0$, $n\rho_n^2 \to \infty$, such that the conditional nuisance posterior converges under $n^{-1}$-perturbation at rate $\rho_n$.

The following proof is based on the proof of Corollary 3.3 in [7].

**Proof.** Define functions $g_1$, $g_2$ and $g_n$ as follows:

$$g_1(\rho) = \Pi_H(K(\rho)), \quad g_2(\rho) = N(\rho, H, d_H), \quad g_n(\rho) = e^{-n\rho^2} \left( \frac{1}{g_1(\rho)} + g_2(\rho) \right).$$

For large enough $n$, the functions $g_n$ are well defined and finite by the assumptions and $g_n(\rho) \to 0$ as $n \to \infty$, for every fixed $\rho > 0$. Therefore, there exists a sequence $(\rho_n)$ such that $\rho_n \downarrow 0$ and $n\rho_n^2 \to \infty$, with $g_n(\rho_n) \to 0$ (e.g., fix $n_1 < n_2 < \cdots$ large enough, such that $g_n(1/k) \leq 1/k$ for all $n \geq n_k$; next define $\rho_n = 1/k$ for $n_k \leq n < n_{k+1}$). In particular, there exists an $N$ such that $g_n(\rho_n) \leq 1$ for all $n \geq N$. This implies that for all $n$ large enough, $g_1(\rho_n) \geq e^{-n\rho_n^2}$, and $g_2(\rho_n) \leq e^{n\rho_n^2}$, Under condition (ii), (5.10) of Theorem 5.5 is satisfied. Then, the assertion of Theorem 5.5 holds. \qed

### 5.3.2 Marginal posterior asymptotic exponentiality

To see how the irregular Bernstein–von Mises assertion (5.3) arises, we note the following: the marginal posterior density $\pi_n: \Theta \to \mathbb{R}$ for the parameter of interest with respect to the prior $\Pi_\Theta$ is given by,

$$\pi_n(\theta) = \left[ \int_H \prod_{i=1}^n p_{\theta,\eta_0}(X_i) d\Pi_H(\eta) \right] \left[ \int_\Theta \int_H \prod_{i=1}^n p_{\theta,\eta_0}(X_i) d\Pi_H(\eta) d\Pi_\Theta(\theta) \right],$$

$P_0$-almost-surely. This form resembles that of a parametric posterior density on $\Theta$ if one replaces the ordinary, parametric likelihood by the integral of the semiparametric likelihood with respect to the nuisance prior, c.f. $S_n(\theta)$ in (5.7). If $S_n(\theta)$ displays properties similar to those that lead to posterior asymptotic normality in the smooth parametric case, we may hope that in the irregular, semiparametric setting the classical proof can be largely maintained. More specifically, we shall replace the LAN expansion of the parametric likelihood by a stochastic LAE expansion of the likelihood integrated over the nuisance as in (5.7). Theorem 5.10 uses this observation to reduce the proof of the main theorem of this chapter to a strictly parametric discussion, much in the way the proof of asymptotic posterior normality in [7] mimics the parametric proof of Le Cam and Yang [71].

In this subsection, we prove marginal posterior asymptotic exponentiality in two parts: first we show that $S_n(\theta)$ satisfies an LAE expansion of its own, and second, we use this to
obtain Bernstein–von Mises assertion (5.3), proceeding along the lines of proofs presented in [56, 58] and [71]. The proofs are given in Subsection 5.5.2. We restrict attention to the case in which the model itself is stochastically LAE and the posterior is consistent under $n^{-1}$-perturbation (although other, less stringent formulations are conceivable).

**Theorem 5.9.** (Integrated local asymptotic exponentiality). Suppose that the model is stochastically locally asymptotically exponential in the $\theta$-direction at all points $(\theta_0, \eta), (\eta \in H)$ and that conditions (iii) and (iv) of Theorem 5.3 are satisfied. Furthermore, assume that model and prior $\Pi$ are such that for some rate $(\rho_n)$ and every bounded, stochastic $(h_n)$,

$$\Pi_n\left( D^c(\rho_n) \mid \theta = \theta_0 + n^{-1}h_n; X_1, \ldots, X_n \right) \to 0.$$  

Then the integral LAE expansion holds, i.e.,

$$\int \prod_{i=1}^n \frac{p_{\theta_0}(h_n, \eta)}{p_0}(X_i) d\Pi_H(\eta) = \int \prod_{i=1}^n \frac{p_{\theta_0}(h_n, \eta)}{p_0}(X_i) d\Pi_H(\eta) \exp(h_n\gamma_{\theta_0, \eta_0} + o_P(1)) \mathbb{1}_{\{h_n \leq \Delta_n\}},$$

for any stochastic sequence $(h_n) \subset \mathbb{R}$ that is bounded in $P_0$-probability.

The following theorem uses the above integrated LAE expansion in conjunction with a marginal posterior convergence condition to derive the exponential Bernstein–von Mises assertion. Marginal posterior convergence forms the subject of the next subsection.

**Theorem 5.10.** (Posterior asymptotic exponentiality). Let $\Theta$ be open in $\mathbb{R}$ with thick prior $\Pi_\Theta$. Suppose that for every $n \geq 1$, $h \mapsto s_n(h)$ is continuous on $(-\infty, \Delta_n]$, $P_0$-almost-surely. Assume that for every stochastic sequence $(h_n) \subset \mathbb{R}$ that is bounded in probability,

$$\frac{s_n(h_n)}{s_n(0)} = \exp(h_n\gamma_{\theta_0, \eta_0} + o_P(1)) \mathbb{1}_{\{h_n \leq \Delta_n\}}, \tag{5.12}$$

for some positive constant $\gamma_{\theta_0, \eta_0}$. Suppose that for every $M_n \to \infty$, we have,

$$\Pi_n\left( |h| \leq M_n \mid X_1, \ldots, X_n \right) \to 1. \tag{5.13}$$

Then the sequence of marginal posteriors for $\theta$ is asymptotically exponential in $P_0$-probability, converging in total variation to a negative exponential distribution,

$$\sup_A \left| \Pi_n\left( h \in A \mid X_1, \ldots, X_n \right) - \text{Exp}_{\Delta_n, \gamma_{\theta_0, \eta_0}}(A) \right| \to 0. \tag{5.14}$$

Conditions (iii) and (iv) of Theorem 5.3 are crucial in the derivation of the two theorems presented above. In the following lemma, proved in Subsection 5.5.2, we present two sufficient conditions for both the domination and the one-sided contiguity condition to hold. The first method poses the domination condition in slightly stronger form (see “$q$-domination” below); the second relies on a log-Lipschitz condition for model densities and uniform finiteness of exponential moments of the Lipschitz constant.
Lemma 5.11. Suppose that the model satisfies at least one of the following two conditions:

(i) ("q-domination" condition)
for every bounded, stochastic \((h_n)\), small enough \( \rho > 0 \), and some \( q > 1 \),
\[
\sup_{\eta \in D(\rho)} P^n_{\theta_0, \eta} \left( \prod_{i=1}^{n} \frac{p_{\theta_n}(h_n), \eta}{p_{\theta_0}, \eta} (X_i) \right)^q = O(1),
\] (5.15)

(ii) (log-Lipschitz condition)
or, for all \( \eta \in H \) there exists a measurable \( m_{\theta_0, \eta} > 0 \) such that for every \( x \in A_{\theta_0} \) and for every \( \theta \) in a neighborhood of \( \theta_0 \),
\[
\frac{p_{\theta, \eta}}{p_{\theta_0, \eta}} (x) \leq e^{m_{\theta_0, \eta}(|\theta_0 - \theta|)},
\] (5.16)
and for small enough \( \rho > 0 \) and all \( K > 0 \),
\[
\sup_{\eta \in D(\rho)} P^n_{\theta_0, \eta} e^{K m_{\theta_0, \eta}} < \infty.
\]
Then, for fixed \( \rho > 0 \) small enough,

(i) the model satisfies the domination condition
\[
\sup_{\eta \in D(\rho)} P^n_{\theta_0, \eta} \left( \prod_{i=1}^{n} \frac{p_{\theta_n}(h_n), \eta}{p_{\theta_0}, \eta} (X_i) \right) = O(1),
\]

(ii) and, for every \( \eta \in D(\rho) \), the \( (P^n_{\theta_n(h_n), \eta}) \) is contiguous with respect to the \( (P^n_{\theta_0, \eta}) \).

The log-Lipschitz version of this lemma is used in both examples of Section 5.4 to satisfy conditions (iii) and (iv) of Theorem 5.3.

5.3.3 Marginal posterior convergence at \( n^{-1} \) rate

One of the conditions in the main theorem is marginal consistency at rate \( n^{-1} \), so that the posterior measure of a sequence of model subsets of the form
\[
\Theta_n \times H = \{ (\theta, \eta) \in \Theta \times H : n|\theta - \theta_0| \leq M_n \},
\]
converge to one in \( P_0 \)-probability, for every sequence \((M_n)\) such that \( M_n \to \infty \). As mentioned in [7], (semiparametric) marginal posteriors have not been studied extensively or systematically in the literature. As a result fundamental questions (e.g., semiparametric bias) concerning marginal posterior consistency have not yet received the attention they deserve. Here, we present a straightforward formulation of sufficient conditions, based solely on bounded likelihood ratios. This has the advantage of leaving the nuisance prior completely unrestricted but may prove to be too stringent a condition on the model in some applications. Conceivably [17], the nuisance prior has a much more significant role to play in questions on marginal consistency. The inadequacy of Lemma 5.12 manifests itself primarily through the occurrence of a supremum over the nuisance space \( H \) in condition (5.17), a uniformity that is too coarse. It can be refined somewhat by requiring uniform bound on the likelihood ratios on a sequence of model subsets, capturing the most of the full nonparametric posterior mass. Reservations aside, it appears from the examples of Section 5.4 that the lemma is also useful in the form stated.
5.3. Asymptotic posterior exponentiality

Lemma 5.12. Let the sequence of maps \( \theta \mapsto S_n(\theta) \) be \( P_0 \)-almost surely continuous on \( (-\infty, \Delta_n] \) and exhibit the stochastic integral LAE property. Furthermore, assume that there exists a constant \( C > 0 \) such that for any \( (M_n) \), \( M_n \to \infty \), \( M_n \leq n \) for \( n \geq 1 \), and \( M_n = o(n) \).

\[
P_0^n \left( \sup_{\eta \in H} \sup_{\theta \in \Theta_n} \mathbb{P}_n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} \leq - \frac{CM_n}{n} \right) \to 1. \tag{5.17}
\]

Then, for any nuisance prior \( \Pi_H \) and \( \Pi_\Theta \) that is thick at \( \theta_0 \),

\[
\Pi_n \left( n|\theta - \theta_0| > M_n \mid X_1, \ldots, X_n \right) \overset{P_0}{\to} 0,
\]

for any \( (M_n) \), \( M_n \to \infty \).

**Proof.** Let us first note, that if marginal consistency holds for a sequence \( M_n \), then it also holds for any sequence \( M'_n \) that diverges faster (i.e., if \( M_n = O(M'_n) \)). Without loss of generality, we therefore assume that \( M_n \) diverges more slowly than \( n \), i.e., \( M_n = o(n) \). We can also assume \( M_n \leq n \) for \( n \geq 1 \). Define \( F_n \) to be the events in (5.17) so that \( P_0^n(F_n^c) = o(1) \) by assumption. In addition, let

\[
G_n = \left\{ (X_1, \ldots, X_n) \mid \int_{\Theta} S_n(\theta) d\Pi_\Theta(\theta) \geq e^{-CM_n/2} S_n(\theta_0) \right\}.
\]

By Lemma 5.13, \( P_0^n(G_n^c) = o(1) \) as well. Hence,

\[
P_0^n \Pi_n \left( n|\theta - \theta_0| > M_n \mid X_1, \ldots, X_n \right)
\leq P_0^n \Pi_n \left( n|\theta - \theta_0| > M_n \mid X_n \right) 1_{F_n \cap G_n} (X_n) + o(1)
\leq e^{CM_n/2} P_0^n \left( \frac{1}{S_n(\theta_0)} \int_{H} \prod_{i=1}^{n} \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} (X_i) \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}} (X_i) d\Pi_\Theta d\Pi_H 1_{F_n} (X_n) \right) + o(1).
\]

On the events \( F_n \) we have

\[
\int_{H} \int_{\Theta_n} \prod_{i=1}^{n} \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} (X_i) \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}} (X_i) d\Pi_\Theta d\Pi_H
= \int_{H} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}} (X_i) \int_{\Theta_n} \exp \left( n\mathbb{P}_n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} \right) d\Pi_\Theta d\Pi_H
\leq \int_{H} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_{\theta_0,\eta_0}} (X_i) d\Pi_H \sup_{\eta \in H} \sup_{\theta \in \Theta_n^c} \exp \left( n\mathbb{P}_n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} \right)
\leq S_n(\theta_0) \exp \left( \sup_{\eta \in H} \sup_{\theta \in \Theta_n^c} n\mathbb{P}_n \log \frac{p_{\theta,\eta}}{p_{\theta_0,\eta}} \right),
\]

which ultimately proves marginal consistency at rate \( n^{-1} \). \( \square \)

In the proof of Lemma 5.12 the lower bound for the denominator of the marginal posterior comes from the following lemma, proved in Subsection 5.5.2. (Let \( \Pi_n \) denote the prior \( \Pi_\Theta \) in the local parametrization in terms of \( h = n(\theta - \theta_0) \).)
Lemma 5.13. Let the sequence of maps \( \theta \mapsto s_n(\theta) \) exhibit the LAE property of (5.12). Assume that the prior \( \Pi_\Theta \) is thick at \( \theta_0 \) (and denoted by \( \Pi_n \) in the local parametrization in terms of \( h \)). Then
\[
P_n(\int s_n(h) d\Pi_n(h) < a_n s_n(0)) \to 0,
\]
for every sequence \((a_n), a_n \downarrow 0\).

5.4 Estimation of support boundary points

In this section we discuss two examples of support boundary estimation for which the likelihood displays an LAE expansion. In Subsection 5.4.1 the parameter of interest is a shift parameter, while in Subsection 5.4.2 we consider a semiparametric scaling family.

5.4.1 Semiparametric shifts

The so-called location problem is one of the classical problems in statistical inference: let \( X_1, X_2, \ldots \) be i.i.d. real-valued random variables, each with marginal \( F_\mu: \mathbb{R} \to [0, 1] \), where \( \mu \in \mathbb{R} \) is the location, i.e., the distribution function \( F_\mu \) is some fixed distribution \( F \) shifted over \( \mu \):
\[
F_\mu(x) = F(x - \mu).
\]
Depending on the nature of \( F \), the corresponding location estimation problem can take various forms: for instance, in case \( F \) possesses a density \( f: \mathbb{R} \to [0, \infty) \) that is symmetric around 0 (and satisfies the regularity condition \( \int (f'/f)^2(x) dF(x) < \infty \)), the location \( \mu \) is estimated at rate \( n^{-1/2} \) (equally well whether we know \( f \) or not [93]). If \( F \) has a support that is contained in a half-line in \( \mathbb{R} \) (i.e., if there is a domain boundary), the problem of estimating the location might become easier. Examples have been given in the introduction where we considered support boundaries of varying degrees of steepness and concluded that the steeper the boundary, the faster the minimax rate of convergence for it.

In this subsection we consider a model of densities with a steep type of boundary, a true discontinuity at \( \mu \): we assume that \( p(x) = 0 \) for \( x < \mu \) and \( p(\mu) > 0 \) while \( p: \mathbb{R} \to [0, \infty) \) is continuous at all \( x \geq \mu \). Observed is an i.i.d. sample \( X_1, X_2, \ldots \) with marginal \( P_0 \). The distribution \( P_0 \) is assumed to have a density of above form, i.e., with unknown location \( \theta \) for a nuisance density \( \eta \) in some space \( H \). Model distributions \( P_{\theta, \eta} \) are then described by densities,
\[
p_{\theta, \eta}: [\theta, \infty) \to [0, \infty): x \mapsto \eta(x - \theta),
\]
for \( \eta \in H \) and \( \theta \in \Theta \subset \mathbb{R} \). As for the family \( H \) of nuisance densities, our interest does not lie in modeling of the tail, we concentrate on specifying the behavior at the discontinuity. For that reason (and in order to connect with Theorem 5.3), we impose some conditions on the nuisance space \( H \): fix \( \alpha > 0 \) and assume that \( \eta: [0, \infty) \to [0, \infty) \) is differentiable and that \( \ell(t) = \eta'(t)/\eta(t) + \alpha \) is a bounded continuous function with a limit at infinity. For given \( S > 0 \), let \( \mathcal{L} \) denote the ball of radius \( S \) in the space \( (C[0, \infty], \| \cdot \|_\infty) \) of continuous functions from the extended half-line (thus with a limit at infinity) to \( \mathbb{R} \) with uniform norm. The following lemma maps \( \mathcal{L} \) to the space \( H \) which we choose to model the nuisance.
Lemma 5.14. Let $\alpha > S$ be fixed. Define $H$ as the image of $\mathcal{L}$ under the map that takes $\ell \in \mathcal{L}$ into densities $\eta_\ell$ by an Esscher transform of the form,

$$
\eta_\ell(x) = \frac{e^{-\alpha x + \int_0^x \ell(t) \, dt}}{\int_0^\infty e^{-\alpha y + \int_0^y \ell(t) \, dt} \, dy},
$$

for $x \geq 0$. This map is uniform-to-Hellinger continuous and the space $H$ is a collection of probability densities that are (i) monotone decreasing with subexponential tails, (ii) continuously differentiable on $[0, \infty)$ and (iii) log-Lipschitz with constant $\alpha + S$.

PROOF. One easily shows that $\ell \mapsto \exp(-\alpha x + \int_0^x \ell) \ell$ is uniform-to-uniform continuous and that $\exp(-\alpha x + \int_0^x \ell) > 0$, which implies uniform-to-Hellinger continuity of the Esscher transform. For the properties of $\eta_\ell$, note that $\int_0^x \ell(y) \, dy \leq S \cdot x < \alpha x$, so that $x \mapsto \exp(-\alpha x + \int_0^x \ell(t) \, dt)$ is subexponential, which implies that $\ell \mapsto \eta_\ell$ gives rise to a probability density. The density $\eta$ is differentiable and monotone decreasing. Furthermore, for all $\theta, \theta_0 \in \Theta$ and all $x \geq \theta_0$,

$$
\frac{\eta_\ell(x - \theta)}{\eta_\ell(x - \theta_0)} \leq \exp\left(\alpha(\theta - \theta_0) + \int_{x - \theta_0}^{x - \theta} \ell(t) \, dt\right) \leq e^{(\alpha + S)|\theta - \theta_0|},
$$

proving the log-Lipschitz property. \qed

Since $H$ consists of functions of bounded variation, Theorem V.2.2 in [50] confirms that the model exhibits local asymptotic exponentiality in the $\theta$-direction for every fixed $\eta$. In the notation of Definition 5.2, $\gamma_{\theta_0, \eta} = \eta(0)$, i.e., the size of the discontinuity at zero. Since it is not difficult to find a prior on a space of bounded continuous functions (see, e.g., Lemma 5.20 below), (Borel) measurability of the Esscher transform as a map between $\mathcal{L}$ and $H$ enables a push-forward prior on $H$.

Theorem 5.15. Let $X_1, X_2, \ldots$ be an i.i.d. sample from the location model introduced above with $P_0 = P_{\theta_0, \eta_0}$ for some $\theta_0 \in \Theta$, $\eta_0 \in H$. Endow $\Theta$ with a prior that is thick at $\theta_0$ and $\mathcal{L}$ with a prior $\Pi_\mathcal{L}$ such that $\mathcal{L} \subset \text{supp}(\Pi_\mathcal{L})$. Then the marginal posterior for $\theta$ satisfies,

$$
\sup_A \left| \Pi(n(\theta - \theta_0) \in A \mid X_1, \ldots, X_n) - \exp_{\Delta_n, \gamma_{\theta_0, \eta_0}}(A) \right| \xrightarrow{P_0} 0, \quad (5.19)
$$

where $\Delta_n$ is exponentially distributed with scale $\gamma_{\theta_0, \eta_0} = \eta_0(0)$.

The proof of Theorem 5.15 consists of a verification of the conditions of Corollary 5.4. The following lemmas make the most elaborate steps explicit. Their proofs can be found in Subsection 5.5.3.

Lemma 5.16. Hellinger covering numbers for $H$ are finite, i.e., $N(\rho, H, d_H) < \infty$ for all $\rho > 0$.

Assuming that the nuisance prior is such that $\mathcal{L} \subset \text{supp}(\Pi_\mathcal{L})$, the following lemma establishes that $\Pi_H(K(\rho)) > 0$, and that condition (ii) of Corollary 5.4 is satisfied.
Lemma 5.17. For every $M > 0$ there exist constants $L_1, L_2 > 0$ such that for small enough $\rho > 0$, \{\(\eta_{i} \in H : \|\tilde{\ell} - \ell_0\|_\infty \leq \rho^2\}\} \subset K(L_1 \rho) \subset K_n(L_2 \rho, M).

By Lemma 5.14, the log-Lipschitz constant $m_{\theta_0, \eta}$ of Lemma 5.11 equals $\alpha + S$ for every $\eta \in H$, so that the domination condition (iii) and contiguity requirement (iv) of Corollary 5.4 are satisfied. The following lemma shows that condition (v) of Corollary 5.4 is also satisfied.

Lemma 5.18. For all bounded, stochastic $(h_n)$, Hellinger distances between $P_{\theta, (h_n), \eta}$ and $P_{\theta_0, \eta}$ are of order $n^{1/2}$ uniformly in $\eta$, i.e.,

\[
\sup_{\eta \in H} n^{1/2} H(P_{\theta, (h_n), \eta}, P_{\theta_0, \eta}) = O(1).
\]

To verify condition (vi) of Corollary 5.4 we now check condition (5.17) of Lemma 5.12.

Lemma 5.19. Let $(M_n)$, $M_n \to \infty$, $M_n \leq n$ for $n \geq 1$, $M_n = o(n)$ be given. Then there exists a constant $C > 0$ such that the condition of Lemma 5.12 is satisfied.

Proof. Note first that for fixed $x$ and $\eta$, the map $\theta \mapsto p_{\theta, \eta}(x)$ is monotone increasing. Therefore,

\[
\sup_{\theta \in \Theta_n} \frac{1}{n} \log \prod_{i=1}^{n} P_{\theta, \eta}(X_i) \leq \frac{1}{n} \log \prod_{i=1}^{n} \frac{\eta(X_i - \theta^*)}{\eta(X_i - \theta_0)} 1_{\{X_i \geq \theta^*\}}(X_i),
\]

where $\theta^* = X(1)$ if $X(1) \geq \theta_0 + M_n/n$, or $\theta_0 - M_n/n$ otherwise. We first note that $X(1) < \theta_0 + M_n/n$ with probability tending to one. Indeed, shifting the distribution to $\theta = 0$, we calculate,

\[
P_{0, \eta_0}(X(1) \geq \frac{M_n}{n}) = \left(1 - \int_{0}^{\frac{M_n}{n}} \eta_0(x) \, dx\right)^n \leq \exp\left(-n \int_{0}^{\frac{M_n}{n}} \eta_0(x) \, dx\right).
\]

By Lemma A.13, the right-hand side of the above display is bounded further as follows,

\[
\exp\left(-\gamma_{\theta_0, \eta_0} M_n + M_n \int_{0}^{\frac{M_n}{n}} |\eta_0'(x)| \, dx\right) \leq \exp\left(-\frac{\gamma_{\theta_0, \eta_0}}{2} M_n\right),
\]

for large enough $n$. We continue with $\theta^* = \theta_0 - M_n/n$. By absolute continuity of $\eta$ we have

\[
\eta(X_i - \theta^*) = \eta(X_i - \theta_0) + \int_{X_i - \theta_0}^{X_i - \theta^*} \eta'(y) \, dy,
\]

and the conditions on the nuisance $\eta$ yield the following bound,

\[
\int_{X_i - \theta_0}^{X_i - \theta^*} \eta'(y) \, dy \leq (\theta_0 - \theta^*)(S - \alpha)\eta(X_i - \theta_0).
\]

Therefore,

\[
\frac{1}{n} \log \prod_{i=1}^{n} \frac{\eta(X_i - \theta^*)}{\eta(X_i - \theta_0)} 1_{\{X_i \geq \theta^*\}}(X_i) \leq \frac{1}{n} \log \left(1 - \frac{(\alpha-S)M_n}{n}\right)^n \leq -\frac{(\alpha-S)M_n}{n}.
\]

If $C < \alpha - S$, the condition of Lemma 5.12 is clearly satisfied.
To demonstrate that priors exist such that \( \mathcal{L} \subset \text{supp}(\Pi_{\mathcal{L}}) \), an explicit construction based on the distribution of Brownian sample paths is provided in the following lemma.

**Lemma 5.20.** Let \( S > 0 \) be given. Let \( \{W_t: t \in [0, 1]\} \) be Brownian motion on \([0, 1]\) and let \( Z \) be independent and distributed \( N(0, 1) \). We define the prior \( \Pi_{\mathcal{L}} \) on \( \mathcal{L} \) as the distribution of the process,

\[
\hat{\ell}(t) = S \Psi(Z + W_{\Psi(t)}),
\]

where \( \Psi: [-\infty, \infty] \to [-1, 1]; x \mapsto 2 \arctan(x)/\pi. \) Then \( \mathcal{L} \subset \text{supp}(\Pi_{\mathcal{L}}). \)

**Proof.** Consider \( C[0, 1] \) with the uniform norm and its Borel \( \sigma \)-algebra, equipped with the law \( \Pi_{t \mapsto Z + W_t} \), as a probability space. Since \( \Psi \) is Lipschitz, the map \( f \) that takes \( C[0, 1] \) into \( C[0, \infty] \), \( Z + W \mapsto Z + W_{\Psi(\cdot)} \) is continuous, norm-preserving, and Borel-to-Borel measurable. This enables the view of \( C[0, \infty] \) with its Borel \( \sigma \)-algebra as a probability space, with probability measure \( \Pi'(B) = \Pi(f^{-1}(B)) \). Similarly, the map \( g \) that takes \( C[0, \infty] \) into \( \mathcal{L} \), \( Z + W_{\Psi(\cdot)} \mapsto S \Psi(Z + W_{\Psi(\cdot)}) \) is continuous and Borel-to-Borel measurable. We view \( \mathcal{L} \) with its Borel \( \sigma \)-algebra as a probability space, with probability measure \( \Pi_{\mathcal{L}}(C) = \Pi'(g^{-1}(C)) \). Let \( T \) denote a closed set in \( \mathcal{L} \) such that \( \Pi_{\mathcal{L}}(T) = 1 \). Note that \( f^{-1}(g^{-1}(T)) \) is closed and \( \Pi(f^{-1}(g^{-1}(T))) = 1 \), so that \( \text{supp}(\Pi) \subset f^{-1}(g^{-1}(T)). \) Since the support of \( \Pi_{\mathcal{L}} \) equals the intersection of all such \( T \), \( \text{supp}(\Pi) \subset \bigcap_T f^{-1}(g^{-1}(\text{supp}(\Pi_{\mathcal{L}}))) \). Since \( \text{supp}(\Pi) = C[0, 1] \), for every \( y \in C[0, 1], f(g(y)) \in \text{supp}(\Pi_{\mathcal{L}}) \). The continuity does not change under \( g \circ f \), so \( \text{supp}(\Pi_{\mathcal{L}}) \) includes \( \mathcal{L} \). \( \square \)

### 5.4.2 Semiparametric scaling

Another important statistical problem is related to the *scale* or *dispersion* of the probability distribution: let \( X_1, X_2, \ldots \) be i.i.d. real-valued random variables, each with marginal \( F_\lambda: \mathbb{R} \to [0, 1] \), where \( \lambda \in (0, \infty) \) is the scale, i.e., the distribution function \( F_\lambda \) is some fixed distribution \( F \) scaled by \( \lambda: F_\lambda(x) = F(x/\lambda) \).

Again, depending on the nature of \( F \), the corresponding scale estimation problem can take various forms: for instance, in case \( F \) possesses a density \( f: \mathbb{R} \to [0, \infty) \) with support \( \mathbb{R} \) that is absolutely continuous and satisfies the regularity condition \( \int (1 + x^2)(f'/f)^2(x) \, dF(x) < \infty \), the scale \( \lambda \) is estimated at rate \( n^{-1/2} \) (equally well whether we know \( f \) or not, as conjectured in [93], and studied later in [106] and [83]). If \( F \) is supported on \([0, \infty)\) (or \((-\infty, 0]\)) the problem can be reparametrized and viewed as a regular location problem. When \( F \) has a support that is a closed interval with one non-zero endpoint (i.e., only one point of the support varies with scale), the problem of estimating the scale might become easier. Probably the best known example of this type is estimation of the scale parameter in the family of the uniform distributions on the interval \([0, \lambda], (\lambda > 0)\).

In this subsection we consider an extension of this uniform example: we assume that \( p(x) > 0 \) for \( x \in [0, \lambda] \) and 0 otherwise while \( p: [0, \lambda] \to [0, \infty) \) is continuous at all \( x \in (0, \lambda) \). Observed is an i.i.d. sample \( X_1, X_2, \ldots \) with marginal \( P_0 \). The distribution \( P_0 \) is assumed to have a density of above form, i.e., with unknown scale \( \theta \) for a nuisance density \( \eta \).
in some space $H$. Model distributions $p_{\theta, \eta}$ are then described by densities,

$$p_{\theta, \eta}: [0, \theta] \rightarrow [0, \infty): x \mapsto \frac{1}{\theta} \eta\left(\frac{x}{\theta}\right),$$  

(5.20)

for $\eta \in H$ and $\theta \in \Theta \subset (0, \infty)$. Fix $S > 0$ and assume that $\eta: [0, 1] \rightarrow [0, \infty)$ is monotone increasing, differentiable and bounded, and that $\dot{\ell}(t) = \eta'(t)/\eta(t) - S$ is a bounded continuous function. For given $S > 0$, let $L$ denote the ball of radius $S$ in the normed space $(C[0, 1], \| \cdot \|_{\infty})$ of continuous functions from the unit interval to $\mathbb{R}$ with uniform norm. The following lemma maps $L$ to the space $H$ with which we choose to model the nuisance.

**Lemma 5.21.** Define $H$ as the image of $L$ under the map that takes $\dot{\ell} \in L$ into densities $\eta_\dot{\ell}$ by an Esscher transform of the form,

$$\eta_\dot{\ell}(x) = \frac{e^{Sx + \int_0^x \dot{\ell}(t) \, dt}}{\int_0^1 e^{Sy + \int_0^y \dot{\ell}(t) \, dt} \, dy},$$  

(5.21)

for $x \in [0, 1]$. This map is uniform-to-Hellinger continuous and the space $H$ is a collection of probability densities that are (i) monotone increasing and bounded away from zero and infinity and (ii) continuously differentiable on $[0, 1]$. Moreover, the resulting densities $p_{\theta, \eta}$ satisfy the log-Lipschitz condition (5.16).

**Proof.** The uniform-to-Hellinger continuity of the Esscher transform is proven in the previous section and $\dot{\ell} \mapsto \eta_\dot{\ell}$ gives rise to a probability density trivially. For the properties of $\eta_\dot{\ell}$, note that

$$-Sx \leq \int_0^x \dot{\ell}(y) \, dy \leq Sx,$$

so that for $x \in [0, 1]$

$$\frac{2S}{e^{2S} - 1} \leq \eta_\dot{\ell}(x) \leq e^{2S} \leq e^{2S}.$$

The density $\eta$ is obviously monotone increasing and continuously differentiable. To check the log-Lipschitz condition fix a neighborhood of $\theta_0$, say $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ for some $0 < \varepsilon < \theta_0/2$, and let $x \in [0, \theta_0]$. For any differentiable function $f$, some $y_0$, and $y \in (y_0 - \varepsilon, y_0 + \varepsilon)$ we have that

$$|f(y_0) - f(y)| \leq \sup_{\varepsilon \in [y_0 - \varepsilon, y_0 + \varepsilon]} |f'(z)| \cdot |y_0 - y|,$$

which implies that

$$\log \frac{\theta_0}{\theta} \leq \frac{1}{\theta_0 - \varepsilon} |\theta_0 - \theta|,$$

and

$$\log \left(\frac{\eta(x/\theta)}{\eta(x/\theta_0)}\right) \leq \frac{x}{(\theta_0 - \varepsilon)^2} \sup_{\theta' \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]} \left|\frac{\eta'(x/\theta')}{\eta(x/\theta')}\right| \cdot |\theta_0 - \theta|.$$

(Note that if $p_{\theta, \eta}(x)$ is 0, the log-Lipschitz condition is trivially satisfied). By the definition of $H$ and since $\varepsilon < \theta_0/2$ we can put $m_{\theta_0, \eta}(x) = 2(1 + 4S)/\theta_0$.)

Theorem V.2.2 in [50] verifies local asymptotic exponentiality in the $\theta$-direction for every fixed $\eta$, although in its positive version. This does not pose problems in applying results of
5.4. Estimation of support boundary points

previous sections: we maintain the sign for $h$ and write $\Delta_n = -\nabla_n$, where $\nabla_n = n(\theta_0 - X(n))$. In the notation of Definition 5.2, $\gamma_{\theta_0,\eta} = \eta(1)/\theta_0$, i.e., the scale of the limiting exponential distribution is the size of the discontinuity at the varying endpoint of the support. Again, we use a push-forward prior on $H$ based on a prior for $L$.

As already noted, our scaling and location problems are both LAE and the parametrizations and solutions we formulate are closely related. However, the nuisance parametrizations are quite different and the relation between the models is a subtle one. Therefore, the location theorem of the previous subsection and the scaling theorem that follows are very similar in appearance, but form the answers to quite distinct questions.

Theorem 5.22. Let $X_1, X_2, \ldots$ be an i.i.d. sample from the scale model introduced above with $P_0 = P_{\theta_0,\eta_0}$ for some $\theta_0 \in \Theta, \eta_0 \in H$. Endow $\Theta$ with a prior that is thick at $\theta_0$, and $L$ with a prior $\Pi_L$ such that $L \subset \text{supp}(\Pi_L)$. Then the marginal posterior for $\theta$ satisfies,

$$\sup_A \left| \Pi(n(\theta - \theta_0) \in A \mid X_1, \ldots, X_n) - \text{Exp}_{-\nabla_n, \gamma_{\theta_0,\eta_0}}(A) \right| \xrightarrow{P_0} 0,$$

(5.22)

where $\nabla_n$ is exponentially distributed with scale $\gamma_{\theta_0,\eta_0} = \eta_0(1)/\theta_0$.

The proof of Theorem 5.22 consists of a verification of the conditions of Corollary 5.4 (after the aforementioned modification to comply with the positive version of the LAE expansion). The following lemmas make the most elaborate steps explicit, as in the proof of Theorem 5.15. Their proofs are collected in Subsection 5.5.4.

Lemma 5.23. Hellinger covering numbers for $H$ are finite, i.e., $N(\rho, H, d_H) < \infty$ for all $\rho > 0$.

Assuming that the nuisance prior is such that $L \subset \text{supp}(\Pi_L)$, the following lemma establishes that $\Pi_H(K(\rho)) > 0$, and that condition (ii) of Corollary 5.4 is satisfied.

Lemma 5.24. For every $M > 0$ there exist constants $L_1, L_2 > 0$ such that for small enough $\rho > 0$, $\{\eta \in H : \|\eta - \eta_0\|\infty \leq \rho^2\} \subset K(L_1 \rho) \subset K_n(L_2 \rho, M)$.

By Lemma 5.21 the model satisfies the log-Lipschitz condition of Lemma 5.11 with the same log-Lipschitz constant for every $\eta \in H$, so that the domination condition (iii) and contiguity requirement (iv) of Corollary 5.4 are satisfied. The following lemma shows that condition (v) of Corollary 5.4 is also satisfied.

Lemma 5.25. For all bounded, stochastic $(h_n)$, Hellinger distances between $P_{\theta_n(h_n),\eta}$ and $P_{\theta_0,\eta}$ are of order $n^{1/2}$ uniformly in $\eta$, i.e.,

$$\sup_{\eta \in H} n^{1/2} H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}) = O(1).$$

To verify condition (vi) of Corollary 5.4 we now check condition (5.17) of Lemma 5.12.

Lemma 5.26. Let $(M_n)$, $M_n \to \infty$, $M_n \leq n$ for $n \geq 1, M_n = o(n)$ be given. Then there exists a constant $C > 0$ such that the condition of Lemma 5.12 is satisfied.
Proof. Note first that for fixed $x$ and $\eta$, the map $\theta \mapsto p_{\theta, \eta}(x)$ is monotone decreasing. Therefore,

$$
\sup_{\theta \in \Theta_n} \frac{1}{n} \log \prod_{i=1}^{n} \frac{p_{\theta, \eta}}{p_{\theta_0, \eta}}(X_i) \leq \frac{1}{n} \log \prod_{i=1}^{n} \frac{\eta(X_i/\theta^*)/\theta^*}{\eta(X_i/\theta_0)/\theta_0} 1_{\{X_{(n)} \leq \theta^*\}}(X_n),
$$

where $\theta^* = X_{(n)}$ if $X_{(n)} \leq \theta_0 - M_n/n$, or $\theta_0 + M_n/n$ otherwise. We first note that $X_{(n)} > \theta_0 - M_n/n$ with probability tending to one. We calculate,

$$
P_0^n \left( X_{(n)} \leq \theta_0 - \frac{M_n}{n} \right) = \left( 1 - \int_{1 - \frac{M_n}{\theta_0 n}}^{1} \eta_0(x) \, dx \right)^n \leq \exp \left( -n \int_{1 - \frac{M_n}{\theta_0 n}}^{1} \eta_0(x) \, dx \right).
$$

By the monotonicity of $\eta_0$, the right-hand side of the above display is bounded further by,

$$
\exp \left( -\gamma_{\theta_0, \eta_0} M_n \right),
$$

for $n \geq 1$. We continue with $\theta^* = \theta_0 + M_n/n$. By absolute continuity of $\eta$ we have

$$
\frac{\eta(X_i/\theta^*)}{\theta^*} = \frac{\eta(X_i/\theta_0)}{\theta_0} + \int_{\theta_0}^{\theta^*} g'(y) \, dy,
$$

where $g(y) = \eta(X_i/y)/y$. We note that

$$
g'(y) = \frac{1}{y} \eta'(X_i/y) \left( -\frac{X_i}{y^2} \right) + \eta(X_i/y) \left( -\frac{1}{y^2} \right) \leq \eta(X_i/y) \left( -\frac{1}{y^2} \right).
$$

Monotonicity of $\eta$ yields the following bound,

$$
\int_{\theta_0}^{\theta^*} g'(y) \, dy \leq (\theta^* - \theta_0) \frac{\eta(X_i/\theta_0)}{\theta_0} \left( -\frac{1}{\theta^*} \right).
$$

Therefore,

$$
\frac{1}{n} \log \prod_{i=1}^{n} \frac{\eta(X_i/\theta^*)/\theta^*}{\eta(X_i/\theta_0)/\theta_0} 1_{\{X_{(n)} \leq \theta^*\}}(X_n) \leq \frac{1}{n} \log \left( 1 - \frac{M_n}{n} \frac{1}{\theta_0 + M_n/n} \right)^n.
$$

If $C < 1/(\theta_0 + 1)$, the condition of Lemma 5.12 is clearly satisfied. \qed

To demonstrate that priors exist such that $\mathcal{L} \subset \text{supp}(\Pi_{\mathcal{L}})$, an explicit construction based on the distribution of Brownian sample paths is provided in the following simplified version of Lemma 5.20.

**Lemma 5.27.** Let $S > 0$ be given. Let $\{W_t : t \in [0, 1]\}$ be Brownian motion on $[0, 1]$ and let $Z$ be independent and distributed $N(0, 1)$. We define the prior $\Pi_{\mathcal{L}}$ on $\mathcal{L}$ as the distribution of the process,

$$
\dot{\ell}(t) = S \Psi(Z + W_t),
$$

where $\Psi : [-\infty, \infty] \to [-1, 1] : x \mapsto 2 \arctan(x)/\pi$. Then $\mathcal{L} \subset \text{supp}(\Pi_{\mathcal{L}})$. 
5.5. Proofs

In this section, several longer proofs of theorems and lemmas in the main text have been collected.

5.5.1 Proofs of Theorem 5.5 and Lemmas 5.6 and 5.7

Proof of Theorem 5.5. Let \((h_n)\) be bounded by \(M\). Let \(\tilde{C} > 0\) be given. Given the Hellinger metric entropy and cone-conditions, Lemma 5.7 guarantees that for any \(L \geq 4\), there exists a test sequence \((\phi_n)\) such that for all bounded, stochastic \((h_n)\), (5.11) is satisfied. Based on \(C, K\) and \(\tilde{C}\) choose \(L > 0\) such that \(L^2 > (1 + K + \tilde{C})/C \lor 16\). By Lemma 5.6, the events,

\[
F_n = \left\{ X_n: \prod_{i=1}^{n} \frac{p_{\theta_n(h_n),\eta}(X_i)}{p_0} d\Pi_H(\eta) \geq e^{-(1+\tilde{C})n\rho_n^2} \Pi_H(K_n(\rho_n, M)) \right\},
\]

satisfy \(P_0^n(F_n \cap \{h_n \leq \Delta_n\}) \leq (\tilde{C}^2n\rho_n^2)^{-1} \to 0\). Using also the first property of the test sequence in (5.11), we see that,

\[
P_0^n \Pi_n(D^c(L\rho_n) | \theta = \theta_n(h_n); X_n) = P_0^n \Pi_n(D^c(L\rho_n) | \theta = \theta_n(h_n); X_n) \mathbf{1}_{\{h_n \leq \Delta_n\}} \\
\leq P_0^n \Pi_n(D^c(L\rho_n) | \theta = \theta_n(h_n); X_n) \times \mathbf{1}_{F_n}(X_n) \mathbf{1}_{\{h_n \leq \Delta_n\}} (1 - \phi_n(X_n)) + o(1).
\]

(5.23)

Based on the definition of the events \(F_n\), the first term on the right is bounded further,

\[
P_0^n \Pi_n(D^c(L\rho_n) | \theta = \theta_n(h_n); X_n) \mathbf{1}_{F_n}(X_n) \mathbf{1}_{\{h_n \leq \Delta_n\}} (1 - \phi_n)(X_n) \\
\leq e^{(1+\tilde{C})n\rho_n^2} \frac{1}{\Pi_H(K_n(\rho_n, M))} \\
\times P_0^n \int_{D^c(L\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_n(h_n),\eta}(X_i)}{p_0} (1 - \phi_n)(X_n) d\Pi_H(\eta).
\]

(5.24)
By Fubini’s theorem and the second property of the test sequence in (5.11), we obtain,

\[
P^n_0 \int_{D^c(L\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_0}(X_i) (1 - \phi_n)(X_n) \, d\Pi_H(\eta)
\]

\[
\leq P^n_0 \int_{D^c(L\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_0}(X_i) (1 - \phi_n)(X_n) \, d\Pi_H(\eta)
\]

\[
= \int_{D^c(L\rho_n)} P^n_{\theta_n(h_n),\eta}(1 - \phi_n)(X_n) \, d\Pi_H(\eta) \leq e^{-CL^2n\rho^2}.
\] (5.25)

Combining (5.23) with (5.24) and (5.25), we find that,

\[
P^n_0 \Pi_n(D^c(L\rho_n) | \theta = \theta_n(h_n); X_n) \leq \frac{e^{(1+C)n\rho^2}}{\Pi_H(K_n(\rho_n,M))}e^{-CL^2n\rho^2} = o(1).
\]

by the choice we made for \(L\) above.

PROOF OF LEMMA 5.6. Let \(C > 0, \rho > 0, \) and \(n \geq 1\) be given. If \(\Pi_H(K_n(\rho, M)) = 0\), the assertion holds trivially, so we assume \(\Pi_H(K_n(\rho, M)) > 0\) without loss of generality and consider the conditional prior \(\Pi_n(A) = \Pi_H(A|K_n(\rho, M))\) (for measurable \(A \subset H\)). Since,

\[
\int_{H} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_0}(X_i) \, d\Pi_H(\eta) \geq \Pi_H(K_n(\rho, M)) \int_{H} \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_0}(X_i) \, d\Pi_n(\eta),
\]

we may choose to consider only the neighborhoods \(K_n\). Restricting attention to the event \(\{h_n \leq \Delta_n\}\), we obtain,

\[
\log \int \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_0}(X_i) \, d\Pi_n(\eta) \geq \int \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_0}(X_i) \, d\Pi_n(\eta)
\]

\[
\geq \int \inf_{|h| \leq M} n_{\mathbb{P}_n} 1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h),\eta}}{p_0} \, d\Pi_n(\eta)
\]

\[
\geq \int n_{\mathbb{P}_n} \inf_{|h| \leq M} 1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h),\eta}}{p_0} \, d\Pi_n(\eta)
\]

\[
\geq \sqrt{n} \int -\mathbb{G}_n \left( \sup_{|h| \leq M} -1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h),\eta}}{p_0} \right) \, d\Pi_n(\eta) - n\rho^2,
\]

using the definition of \(K_n\) in the last step (see (5.6)). Then,

\[
P^n_0 \left( \left\{ \int \prod_{i=1}^n \frac{p_{\theta_n(h_n),\eta}}{p_0}(X_i) \, d\Pi_n(\eta) < e^{-(1+C)n\rho^2} \right\} \cap \{h_n \leq \Delta_n\} \right)
\]

\[
\leq P^n_0 \left( \int -\mathbb{G}_n \left( \sup_{|h| \leq M} -1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h),\eta}}{p_0} \right) \, d\Pi_n(\eta) < -\sqrt{nC\rho^2} \right).
\]

By Chebyshev’s inequality, Jensen’s inequality, Fubini’s theorem and the fact that for any
$P_0$-square-integrable random variables $Z_n$, $P^n_0 \left( \mathbb{G}_n Z_n \right)^2 \leq P^n_0 Z_n^2$, 

$$P^n_0 \left( \int - \mathbb{G}_n \left( \sup_{|h| \leq M} -1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h),\eta}}{p_0} \right) \, d\Pi_n(\eta) < -\sqrt{n C^2 \rho^2} \right) \leq \frac{1}{nC^2 \rho^2} \int P^n_0 \left( \mathbb{G}_n \sup_{|h| \leq M} -1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h),\eta}}{p_0} \right)^2 \, d\Pi_n(\eta) \leq \frac{1}{nC^2 \rho^2},$$

where the last step follows again from definition (5.6). \hfill \Box

Here we reproduce the proof of Lemma 3.2 in [7].

**PROOF OF LEMMA 5.7.** Let $(\rho_n)$ be such that the conditions of Lemma 5.7 are satisfied. Let $(h_n)$ and $L \geq 4$ be given. For all $j \geq 1$, define $H_{i,j,n} = \{ \eta \in H: jL\rho_n \leq d_H(\eta_0, \eta) \leq (j + 1)L\rho_n \}$ and $\mathcal{P}_{i,j,n} = \{ P_{\theta_0,\eta}; \eta \in H_{i,j,n} \}$. Cover $\mathcal{P}_{i,j,n}$ with Hellinger balls $B_{i,j,n}(jL\rho_n/4)$, where

$$B_{i,j,n}(r) = \{ P: H(P_{i,j,n}, P) \leq r \}$$

and $P_{i,j,n} \in \mathcal{P}_{i,j,n}$, that is, there exists an $\eta_{i,j,n} \in H_{i,j,n}$ such that $P_{i,j,n} = P_{\theta_0,\eta_{i,j,n}}$. Denote $H_{i,j,n} = \{ \eta \in H_{i,j,n}; P_{\theta_0,\eta} \in B_{i,j,n}(jL\rho_n/4) \}$. By assumption, the minimal number of such balls needed to cover $\mathcal{P}_{i,j}$ is finite; we denote the corresponding covering number by $N_{i,j,n}$, that is, $1 \leq l \leq N_{i,j,n}$.

Let $\eta \in H_{i,j,n}$ be given. There exists an $i$ ($1 \leq i \leq N_{i,j,n}$) such that $d_H(\eta_0, \eta_{i,j,n}) \leq jL\rho_n/4$. Then, by the triangle inequality, the definition of $H_{i,j,n}$ and condition (ii) of the lemma,

$$H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}) \leq H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta}) + H(P_{\theta_0,\eta}, P_{\theta_0,\eta_{i,j,n}}) \leq \frac{H(P_{\theta_n(h_n),\eta}, P_{\theta_0,\eta})}{H(P_{\theta_0,\eta}, P_0)} H(P_{\theta_0,\eta}, P_0) + \frac{1}{4} jL\rho_n \leq \frac{1}{2} jL\rho_n$$

for large enough $n$. We conclude that there exists an $N \geq 1$ such that for all $n \geq N$, $j \geq 1$, $1 \leq i \leq N_{i,j,n}$, $\eta \in H_{i,j,n}$, $P_{\theta_n(h_n),\eta} \in B_{i,j,n}(jL\rho_n/2)$. Moreover, Hellinger balls are convex and for all $P \in B_{i,j,n}(jL\rho_n/2)$, $H(P, P_0) \geq jL\rho_n/2$. As a consequence of the minimax theorem (see [69], or [9, 10]), there exists a test sequence $(\phi_{i,j,n})_{n \geq 1}$ such that

$$P^n_0 \phi_{i,j,n} \vee \sup_P P^n(1 - \phi_{i,j,n}) \leq e^{-nH^2(B_{i,j,n}(jL\rho_n/2), P_0)} \leq e^{-n^2L^2j^2\rho_n^2/4},$$

where the supremum runs over all $P \in B_{i,j,n}(jL\rho_n/2)$. Defining, for all $n \geq 1$, $\phi_n = \sup_{j \geq 1} \max_{1 \leq i \leq N_{j,n}} \phi_{i,j,n}$, we find

$$P^n_0 \phi_n \leq \sum_{j=1}^{\infty} \sum_{i=1}^{N_{j,n}} P^n_0 \phi_{i,j,n} \leq \sum_{j=1}^{\infty} N_{j,n} e^{-L^2j^2n\rho_n^2/4},$$

$$P^n(1 - \phi_n) \leq P^n(1 - \phi_{i,j,n}) \leq e^{-L^2n\rho_n^2/4},$$

$$P^n(1 - \phi_n) \leq e^{-L^2n\rho_n^2/4},$$

(5.27)
for all $P = P_{\theta_n(h_n), \eta}$ and $\eta \in D^c(L\rho_n)$. Since $L \geq 4$, we have for all $j \geq 1$,
\[
N_{j,n} = N\left(Lj\rho_n/4, \mathcal{P}_{j,n}, H\right) \leq N\left(Lj\rho_n/4, \mathcal{P}, H\right)
\leq N(\rho_n, \mathcal{P}, H) \leq e^{n\rho_n^2}
\tag{5.28}
\]
by condition (i) of the lemma. Upon substitution of (5.28) into (5.27), we obtain the following bounds:
\[
P_0^n\phi_n \leq \frac{e^{(1-L^2/4)n\rho_n^2}}{1 - e^{-L^2n\rho_n^2/4}}, \quad \sup_{\eta \in D^c(L\rho_n)} P_{\theta_n(h_n), \eta}^n (1 - \phi_n) \leq e^{-L^2n\rho_n^2/4}
\]
for large enough $n$, which implies the assertion of the lemma.

\section*{5.5.2 Proof of Theorems 5.9 and 5.10, and Lemmas 5.11 and 5.13}

\textbf{Proof of Theorem 5.9.} Let $(h_n)$ be bounded in $P_0$-probability. Throughout this proof we write $\theta_n(h_n) = \theta_0 + n^{-1}h_n$. Let $\delta, \varepsilon > 0$ be given. There exists a constant $M > 0$ such that $P_0^n(|h_n| > M) < \delta/2$ for all $n \geq 1$. By the consistency assumption, for large enough $n$,
\[
P_0^n\left( \log \Pi_n\left(D(\rho_n) \mid \theta = \theta_n; X_1, \ldots, X_n\right) \geq -\varepsilon \right) > 1 - \frac{\delta}{2}.
\]
This implies that the posterior’s numerator and denominator are related through,
\[
P_0^n\left(\int_H \prod_{i=1}^n \frac{p_{\theta_n(h_n), \eta}(X_i)}{p_0}(X_i) d\Pi_H(\eta)\right)
\leq e^\varepsilon 1_{|h_n| \leq M} \int_{D(\rho_n)} \prod_{i=1}^n \frac{p_{\theta_n(h_n), \eta}(X_i)}{p_0}(X_i) d\Pi_H(\eta) > 1 - \delta,
\]
for this $M$ and all $n$ large enough. We continue with the integral over $D(\rho_n)$ under the restriction $|h_n| \leq M$. By stochastic local asymptotic exponentiality for every fixed $\eta$, we have,
\[
\prod_{i=1}^n \frac{p_{\theta_n(h_n), \eta}(X_i)}{p_0}(X_i) = \prod_{i=1}^n \frac{p_{\theta_0, \eta}(X_i)}{p_0}(X_i) \exp(h_n\gamma_{\theta_0, \eta} + R_n(h_n, \eta; X_n)),
\]
where the rest-term $R_n(h_n, \eta; X_n)$ converges to zero in $P_{\theta_0, \eta}$-probability. Define for all $\varepsilon > 0$ the events,
\[
F_n(\eta, \varepsilon) = \left\{ X_n : \sup_{|h| \leq M} |h\gamma_{\theta_0, \eta} - h\gamma_{\theta_0, \eta_0}| \leq \varepsilon \right\},
\]
and note that $F_n^c(0, \varepsilon) = \emptyset$. With the domination condition (iii) of Theorem 5.3, Fatou’s lemma yields:
\[
\limsup_{n \to \infty} \int_{D(\rho_n)} P_{\theta_n(h_n), \eta}^n\left(F_n^c(\eta, \varepsilon)\right) d\Pi_H(\eta)
\leq \int \limsup_{n \to \infty} 1_{D(\rho_n) \setminus \{0\}} P_{\theta_n(h_n), \eta}^n\left(F_n^c(\eta, \varepsilon)\right) d\Pi_H(\eta) = 0.
\]
Combined with Fubini’s theorem, this suffices to conclude that

\[
\int_{D(\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_n(n),\eta}}{p_0}(X_i) \, d\Pi_H(\eta) \]

\[= \int_{D(\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_n(n),\eta}}{p_0}(X_i) 1_{F_n(\eta,\varepsilon)}(X_n) \, d\Pi_H(\eta) + o_{P_0}(1), \tag{5.29}
\]

and we continue with the first term on the right-hand side. For every \( \eta \in H \), define the events,

\[G_n(\eta, \varepsilon) = \left\{ X_n : \sup_{|h| \leq M} |R_n(h, \eta; X_n)| \leq \varepsilon/2 \right\},\]

and note that \( P^n_{\theta_0,\eta}(G'_n(\eta, \varepsilon)) \to 0 \). By the contiguity condition (iv) of Theorem 5.3, the probabilities \( P^n_{\theta_0,\eta}(G'_n(\eta, \varepsilon)) \) converge to zero as well. Reasoning as with the events \( F_n(\eta, \varepsilon) \), we conclude that,

\[
\int_{D(\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_n(n),\eta}}{p_0}(X_i) 1_{F_n(\eta,\varepsilon)}(X_n) \, d\Pi_H(\eta) \]

\[= \int_{D(\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_n(n),\eta}}{p_0}(X_i) 1_{G_n(\eta,\varepsilon) \cap F_n(\eta,\varepsilon)}(X_n) \, d\Pi_H(\eta) + o_{P_0}(1). \]

For fixed \( n \) and \( \eta \) and for all \( X_n \in G_n(\eta, \varepsilon) \cap F_n(\eta, \varepsilon) \), and by stochastic local asymptotic exponentiality,

\[
\left| \log \prod_{i=1}^{n} \frac{p_{\theta(n),\eta}}{p_0}(X_i) - \log \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_0}(X_i) - h_n \gamma_{\theta_0,\eta_0} \right| \leq |R_n(h_n, \eta; X_n)| + |h_n(\gamma_{\theta_0,\eta_0} - \gamma_{\theta_0,\eta})| \leq 2\varepsilon,
\]

from which it follows that,

\[
\exp(h_n \gamma_{\theta_0,\eta_0} - 2\varepsilon) \int_{D(\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_0}(X_i) 1_{G_n(\eta,\varepsilon) \cap F_n(\eta,\varepsilon)}(X_n) \, d\Pi_H(\eta) \]

\[\leq \int_{D(\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_n(n),\eta}}{p_0}(X_i) 1_{G_n(\eta,\varepsilon) \cap F_n(\eta,\varepsilon)}(X_n) \, d\Pi_H(\eta) \]

\[\leq \exp(h_n \gamma_{\theta_0,\eta_0} + 2\varepsilon) \int_{D(\rho_n)} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_0}(X_i) 1_{G_n(\eta,\varepsilon) \cap F_n(\eta,\varepsilon)}(X_n) \, d\Pi_H(\eta).\]

The integrals can be relieved of indicators for \( G_n \cap F_n \) by reversing preceding arguments (with \( \theta_0 \) replacing \( \theta_n \), at the expense of an \( \exp(o_{P_0}(1)) \)-factor, leading to,

\[
\exp(h_n \gamma_{\theta_0,\eta_0} - 3\varepsilon + o_{P_0}(1)) \int_{H} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_0}(X_i) \, d\Pi_H(\eta) \]

\[\leq \int_{H} \prod_{i=1}^{n} \frac{p_{\theta_n(n),\eta}}{p_0}(X_i) \, d\Pi_H(\eta) \]

\[\leq \exp(h_n \gamma_{\theta_0,\eta_0} + 3\varepsilon + o_{P_0}(1)) \int_{H} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_0}(X_i) \, d\Pi_H(\eta).\]
for all $h_n \leq \Delta_n$. Since this holds for arbitrarily small $\varepsilon > 0$, it proves desired result. \hfill \Box \\

**Proof of Theorem 5.10.** Let $C$ be an arbitrary compact subset of $\mathbb{R}$ containing an open neighborhood of the origin. Denote the (randomly located) distribution $\text{Exp}_n\gamma_{\theta_0,n_0}$ by $\Xi_n$. The prior and marginal posterior for the local parameter $h$ are denoted $\Pi_n$ and $\Pi_n(\cdot|\underline{X}_n)$. Conditioned on $C \subset \mathbb{R}$, these measures are denoted $\Xi_n^C$, $\Pi_n^C$ and $\Pi_n^C(\cdot|\underline{X}_n)$ respectively. Define the functions $\xi_n^*, \xi_n: \mathbb{R} \to \mathbb{R}$ as,

$$
\xi_n^*(x) = \gamma_{\theta_0,n_0} \exp^{\gamma_{\theta_0,n_0}}(x - \Delta_n), \quad \xi_n(x) = \xi_n^*(x) \mathbf{1}_{\{x \leq \Delta_n\}},
$$

noting that $\xi_n$ is the Lebesgue density for $\Xi_n$. Also define $s_n^*(h) = s_n(h)$ on $(-\infty, \Delta_n]$ and $s_n^*(h) = s_n(0) \exp(h\gamma_{\theta_0,n_0} + d_n)$ elsewhere. Finally, define, for every $g, h \in C$ and large enough $n$,

$$
f_n(g, h) = \left(1 - \frac{\xi_n(h) s_n(g) \pi_n(g)}{\xi_n(g) s_n(h) \pi_n(h)}\right)_+ \mathbf{1}_{\{g \leq \Delta_n\}} \mathbf{1}_{\{h \leq \Delta_n\}},
$$

and

$$
f_n^*(g, h) = \left(1 - \frac{\xi_n^*(h) s_n^*(g) \pi_n(g)}{\xi_n^*(g) s_n^*(h) \pi_n(h)}\right)_+.
$$

By (5.12) we know that $d_n = \log s_n(\Delta_n) - \log s_n(0) - \Delta_n \gamma_{\theta_0,n_0} = o_P(1)$. Furthermore, for every stochastic sequence $(h_n)$ in $C$,

$$
\log s_n^*(h_n) = \log s_n^*(0) + h_n \gamma_{\theta_0,n_0} + o_P(1), \\
\log \xi_n^*(h_n) = (h_n - \Delta_n) \gamma_{\theta_0,n_0} + \log \gamma_{\theta_0,n_0}.
$$

Since $\xi_n^*(h)$ and $\xi_n(h)$ ($s_n^*(h)$ and $s_n(h)$, respectively) coincide on $\{h \leq \Delta_n\}$, $f_n(g, h) \leq f_n^*(g, h)$. For any two stochastic sequences $(h_n), (g_n)$ in $C$, we observe $\pi_n(g_n)/\pi_n(h_n) \to 1$ as $n \to \infty$ since $\pi$ is continuous and non-zero at $\theta_0$. Combination with the above display leads to,

$$
\log \frac{\xi_n^*(h) s_n^*(g) \pi_n(g)}{\xi_n^*(g) s_n^*(h) \pi_n(h)} = (h_n - \Delta_n) \gamma_{\theta_0,n_0} - (g_n - \Delta_n) \gamma_{\theta_0,n_0} + g_n \gamma_{\theta_0,n_0} - h_n \gamma_{\theta_0,n_0} + o_P(1) = o_P(1).
$$

Since $x \mapsto (1 - e^x)_+$ is continuous on $(-\infty, \infty)$, we conclude that for any stochastic sequence $(g_n, h_n)$ in $C \times C$, $f_n(g_n, h_n) \xrightarrow{P_0} 0$. To render this limit uniform over $C \times C$, continuity is enough: $(g, h) \mapsto \pi_n(g)/\pi_n(h)$ is continuous since the prior is thick. Note that $\xi_n^*(h)/s_n^*(h)$ is of the form $\gamma_{\theta_0,n_0} \exp(\gamma_{\theta_0,n_0}(\Delta_n + R_n(h)))$ for all $h, n \geq 1$, and $R_n(h_n) = o_P(1)$. Tightness of $\Delta_n$ and $R_n$ implies that $\xi_n^*(h)/s_n^*(h) \in (0, \infty), (P_0^0 - a.s.)$. Continuity of $h \mapsto s_n(h)$ and $h \mapsto \xi_n^*(h)$ then implies continuity of $(g, h) \mapsto (\xi_n^*(h) s_n^*(g))/\xi_n^*(g) s_n^*(h))$, $(P_0^0 - a.s.)$. Hence we conclude that,

$$
\sup_{(g,h) \in C \times C} f_n(g, h) \leq \sup_{(g,h) \in C \times C} f_n^*(g, h) \xrightarrow{P_0} 0.
$$

Since $s_n(h)$ is supported on $(-\infty, \Delta_n]$, since $C$ contains a neighborhood of the origin and since $\Delta_n$ is tight and positive, $\Xi_n(C) > 0$ and $\Pi_n(C|\underline{X}_n) > 0, (P_0^0 - a.s.)$. So conditioning
on $C$ is well-defined (for the relevant cases where $h \leq \Delta_n$). Let $\delta > 0$ be given and define events,
\[ \Omega_n = \left\{ X_n : \sup_{(g,h) \in C \times C} f_n(g, h) \leq \delta \right\}. \]
Based on $\Omega_n$ and (5.30), write,
\[ P_0^n \sup_A \left| \Pi_n^C(h \in A | X_n) - \Xi_n^C(A) \right| \leq P_0^n \sup_A \left| \Pi_n^C(h \in A | X_n) - \Xi_n^C(A) \right| 1_{\Omega_n} + o(1). \]
Note that both $\Xi_n^C$ and $\Pi_n^C(\cdot | X_n)$ have strictly positive densities on $C$. Therefore, $\Xi_n^C$ is dominated by $\Pi_n^C(\cdot | X_n)$ for all $n$ large enough. With that observation, the first term on the right-hand side of the above display is calculated to be,
\[ \frac{1}{2} P_0^n \sup_A \left| \Pi_n^C(h \in A | X_n) - \Xi_n^C(A) \right| 1_{\Omega_n}(X_n) \]
\[ = P_0^n \int_C \left(1 - \frac{d\Xi_n^C}{d\Pi_n^C(\cdot | X_n)}\right) + 1_{\{h \leq \Delta_n\}} d\Pi_n^C(h | X_n) 1_{\Omega_n}(X_n) \]
\[ = P_0^n \left( \int_C \left(1 - \xi_n(h) \right) \frac{d\Xi_n^C}{d\Pi_n^C(\cdot | X_n)} \right) + \]
\[ \times 1_{\{h \leq \Delta_n\}} d\Pi_n^C(h | X_n) 1_{\Omega_n}(X_n) \]
\[ = P_0^n \left( \int_C \left(1 - \int_C \frac{s_n(g) \pi_n(g)}{s_n(h) \pi_n(h)} \xi_n(g) \right) 1_{\{g \leq \Delta_n\}} d\Xi_n^C(g) \right) + \]
\[ \times 1_{\{h \leq \Delta_n\}} d\Pi_n^C(h | X_n) 1_{\Omega_n}(X_n) \]
for large enough $n$. Jensen’s inequality leads to
\[ \frac{1}{2} P_0^n \sup_A \left| \Pi_n^C(h \in A | X_n) - \Xi_n^C(A) \right| 1_{\Omega_n}(X_n) \]
\[ \leq P_0^n \left( \int C \left(1 - \frac{s_n(g) \pi_n(g)}{s_n(h) \pi_n(h)} \xi_n(g) \right) + \]
\[ \times 1_{\{h \leq \Delta_n\}} 1_{\{g \leq \Delta_n\}} d\Xi_n^C(g) d\Pi_n^C(h | X_n) 1_{\Omega_n}(X_n) \right) \]
\[ \leq P_0^n \int_{(g,h) \in C \times C} f_n(g, h) d\Xi_n^C(g) d\Pi_n^C(h | X_n) 1_{\Omega_n}(X_n) \leq \delta. \]
We conclude that for all compact $C \subset \mathbb{R}$ containing a neighborhood of the origin, $P_0^n \|\Pi_n^C - \Xi_n^C\| \to 0$. To finish the argument, let $(C_m)$ be a sequence of closed balls centered at the origin with radii $M_m \to \infty$. For each fixed $m \geq 1$ the above display holds with $C = C_m$, so if we traverse the sequence $(C_m)$ slowly enough, convergence to zero can still be guaranteed, i.e., there exist $(M_m)$, $M_m \to \infty$ such that, $P_0^n \|\Pi_n^{B_m} - \Xi_n^{B_m}\| \to 0$, where $B_n$ has radius $M_n$. Using Lemma 2.11 in [56] and Lemma 5.28 below we conclude that (5.14) holds. \[ \square \]
Lemma 5.28. Let $K_n$ be a sequence of balls centered on the origin with radii $M_n \to \infty$. Let $\text{Exp}_{\lambda_m}$ be a sequence of negative exponential distributions (with fixed scale $a$) located respectively at the (random) points $(\Delta_m)$. If the sequence $\Delta_n$ is uniformly tight, then

$$
\Xi_{\Delta_n,a}(\mathbb{R} \setminus K_n) \xrightarrow{P} 0.
$$

**Proof.** Fix $\delta > 0$ and $a > 0$. Uniform tightness of $\Delta_n$ implies the existence of $L > 0$ such that

$$
\sup_{n \geq 1} P_0^n(|\Delta_n| \geq L) \leq \delta.
$$

Define $A_n = \{\omega: \Delta_n(\omega) \geq L\}$. Let $\lambda \in \mathbb{R}$ be given. Denote by $\Xi_{\lambda,a}$ a negative exponential distribution with location parameter $\lambda$ and scale parameter $a$. Since $\Xi_{\lambda,a}$ is tight, there exists for every $\varepsilon > 0$ a constant $L' > 0$ such that $\Xi_{\lambda,a}(B(\lambda, L')) \geq 1 - \varepsilon$. If $|\lambda| \leq L$, then $B(\lambda, L') \subset B(0, L' + L)$. Therefore, with $M = L' + L$, $\Xi_{\lambda,a}(B(0, M)) \geq 1 - \varepsilon$ for all $\lambda$ such that $|\lambda| \leq L$. Choose $N \geq 1$ such that $M_n \geq M$ for all $n \geq N$. Let $n \geq N$ be given. Then

$$
P_0^n(\Xi_{\Delta_n,a}(\mathbb{R} \setminus B(0, M_n)) > \varepsilon)
\leq P_0^n(A_n) + P_0^n(\Xi_{\Delta_n,a}(\mathbb{R} \setminus B(0, M_n)) > \varepsilon) \cap A_n^c)
\leq \delta + P_0^n(\Xi_{\Delta_n,a}(B(0, M))^c) > \varepsilon) \cap A_n^c).
$$

(5.31)

Note that on the complement of $A_n$, we observe $|\Delta_n| < L$, so

$$
\Xi_{\Delta_n,a}(B(0, M))^c) \leq 1 - \Xi_{\lambda,a}(B(0, M)) \leq 1 - \inf_{|\lambda| < L} \Xi_{\lambda,a}(B(0, M)) \leq \varepsilon,
$$

and we conclude that the last term on the right-hand side of (5.31) equals zero. \hfill \Box

**Proof of Lemma 5.11.** Assume first that the “$q$-domination” condition is satisfied. Assertion (i) follows from Jensen’s inequality. For the second assertion, fix $\eta \in D(\rho)$ and take a sequence of events $(F_n)$ such that $P^n_{\theta_0,\eta}(F_n) \to 0$. Contiguity now follows from Hölder’s inequality (with $1/p + 1/q = 1$),

$$
P^n_{\theta_0,\eta}(F_n) \leq \left(\int \left(\prod_{i=1}^n \frac{p^n_{\theta_0}(h_n; \eta)}{p^n_{\theta_0,\eta}}(X_i)\right)^q dP^n_{\theta_0,\eta}\right)^{1/q} \left(\int 1_{F_n} dP^n_{\theta_0,\eta}\right)^{1/p}
\lesssim P^n_{\theta_0,\eta}(F_n)^{1/p} \to 0.
$$

Next, assume that the log-Lipschitz condition is satisfied. Let $(h_n)$ be a stochastic sequence bounded by $M \geq 0$. By (5.16),

$$
\prod_{i=1}^n \frac{p^n_{\theta_0}(h_n; \eta)}{p^n_{\theta_0,\eta}}(X_i) \leq \exp\left(\sum_{i=1}^n m_{\theta_0,\eta}(X_i) \frac{|h_n|}{n}\right) \leq \exp\left(\frac{M}{n} \sum_{i=1}^n m_{\theta_0,\eta}(X_i)\right),
$$

for $X_i$ in $A_{\theta_0}$, which holds with $P_{\theta_0,\eta}$-probability one. Therefore,

$$
P^n_{\theta_0,\eta}\left(\prod_{i=1}^n \frac{p^n_{\theta_0}(h_n; \eta)}{p^n_{\theta_0,\eta}}(X_i)\right) \leq P^n_{\theta_0,\eta}\left(\exp\left(\frac{M}{n} \sum_{i=1}^n m_{\theta_0,\eta}(X_i)\right)\right)
\leq P_{\theta_0,\eta}\exp(Mm_{\theta_0,\eta}).$$
Due to the uniformity of the assumed bound on \( P_{\theta_0, \eta} \exp(Km_{\theta_0, \eta}) \), this proves (i). For the second assertion, fix \( \eta \in D(\rho) \) for some \( \rho > 0 \) small enough, and take a sequence of events \( F_n \) such that \( P_{\theta_0, \eta}^n(F_n) \to 0 \). Then,

\[
P_{\theta_0(h_n), \eta}^n(F_n) \leq \int \exp \left( \frac{M}{n} \sum_{i=1}^{n} m_{\theta_0, \eta}(X_i) \right) 1_{F_n}(X_n) dP_{\theta_0, \eta}^n
\]

\[
\leq \left( \int \exp \left( \frac{qM}{n} \sum_{i=1}^{n} m_{\theta_0, \eta}(X_i) \right) dP_{\theta_0, \eta}^n \right)^{1/q} \left( \int 1_{F_n} dP_{\theta_0, \eta}^n \right)^{1/p}
\]

\[
\leq \left( P_{\theta_0, \eta} \exp(qMm_{\theta_0, \eta}) \right)^{1/q} P_{\theta_0, \eta}^n(F_n)^{1/p} \to 0,
\]

where we have used Hölder’s inequality (with \( 1/p + 1/q = 1 \)) and Jensen’s inequality. The uniform bound on \( P_{\theta_0, \eta} \exp(Km_{\theta_0, \eta}) \) implies that the first term on the r.h.s. of the above display \( \left( P_{\theta_0, \eta} \exp(qMm_{\theta_0, \eta}) \right)^{1/q} \) is finite for any \( \eta \in D(\rho) \) and \( q > 1 \).

**Proof of Lemma 5.13.** Let \( M > 0 \) be given and define the set \( C = \{ h: -M \leq h \leq 0 \} \). Denote the \( o_{P_0}(1) \) rest-term in the integral LAE expansion (5.12) by \( h \mapsto R_n(h) \). By continuity of \( \theta \mapsto S_n(\theta) \), the expansion holds uniformly over compacta for large enough \( n \) and in particular, \( \sup_{h \in C} |R_n(h)| \) converges to zero in \( P_0 \)-probability. Let \( (K_n, K_n) \to \infty \) be given. The events \( B_n = \{ \sup_{C} |R_n(h)| \leq K_n/2 \} \) satisfy \( P_0^n(B_n) \to 1 \). Since \( \Pi_{\theta} \) is thick at \( \theta_0 \), there exists a \( \pi > 0 \) such that \( \inf_{h \in C} d\Pi_{n}/dh \geq \pi \), for large enough \( n \).

Therefore,

\[
P_0^n \left( \int_{C} \frac{s_n(h)}{s_n(0)} d\Pi_{n}(h) \leq e^{-K_n} \right) \leq P_0^n \left( \left\{ \int_{C} \frac{s_n(h)}{s_n(0)} dh \leq \pi^{-1} e^{-K_n} \right\} \cap B_n \right) + o(1).
\]

On \( B_n \), the integral LAE expansion is lower bounded so that, for large enough \( n \),

\[
P_0^n \left( \left\{ \int_{C} \frac{s_n(h)}{s_n(0)} d\Pi_{n}(h) \leq \pi^{-1} e^{-K_n} \right\} \cap B_n \right) \leq P_0^n \left( \int_{\Omega} e^{h \gamma_{0, 0}} dh \leq \pi^{-1} e^{-K_n} \right).
\]

Since \( \int_{\Omega} e^{h \gamma_{0, 0}} dh \geq M e^{-M \gamma_{0, 0}} \) and \( K_n \to \infty \), \( e^{-K_n/\pi} \leq \pi M e^{-M \gamma_{0, 0}} \) for large enough \( n \). Combination of the above with \( K_n = -\log a_n \) proves the desired result.

**5.5.3 Proofs of Subsection 5.4.1**

**Proof of Lemma 5.16.** Given \( 0 < S < \alpha \), we define \( \rho_0^2 = \alpha - S > 0 \). Consider the distribution \( Q \) with Lebesgue density \( q > 0 \) given by \( q(x) = \rho_0^2 e^{-\rho_0^2 x} \) for \( x \geq 0 \). Then the family \( \mathcal{F} = \{ x \mapsto \sqrt{\eta_i/q(x)} : \ell \in \mathcal{L} \} \) forms a subset of the collection of all monotone functions \( \mathbb{R} \mapsto [0, C] \), where \( C \) is fixed and depends on \( \alpha \), and \( S \). Referring to Theorem 2.7.5 in [104], we conclude that the \( L_2(Q) \)-bracketing entropy \( N_{[]}(\varepsilon, \mathcal{F}, L_2(Q)) \) of \( \mathcal{F} \) is finite for all \( \varepsilon > 0 \). Noting that,

\[
d_H(\eta, \eta_0)^2 = d_H(\eta_\ell, \eta_\ell_0)^2 = \int_{\mathbb{R}} \left( \sqrt{\frac{\eta_\ell}{q}(x)} - \sqrt{\frac{\eta_\ell_0}{q}(x)} \right)^2 dQ(x),
\]

it follows that \( N(\rho, H, d_H) = N(\rho, \mathcal{F}, L_2(Q)) \leq N_{[]}(2\rho, \mathcal{F}, L_2(Q)) < \infty \).
**Proof of Lemma 5.17.** Let \( \rho, 0 < \rho < \rho_0 \) and \( \hat{\ell} \in \mathcal{L} \) such that \( \|\hat{\ell} - \hat{\ell}_0\|_\infty \leq \rho^2 \) be given. Then,

\[
\log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}}(x) - \int_0^{x-\theta_0} (\hat{\ell} - \hat{\ell}_0)(t) \, dt \leq \rho^2 P_0(X - \theta_0) + O(\rho^4),
\]

(5.32)

for all \( x \geq \theta_0 \). Define, for all \( \alpha > S \) and \( \hat{\ell} \in \mathcal{L} \), the logarithm \( z \) of the normalizing factor in (5.18). Then the relevant log-density-ratio can be written as,

\[
\log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}}(x) = \int_0^{x-\theta_0} (\hat{\ell} - \hat{\ell}_0)(t) \, dt - z(\alpha, \ell) + z(\alpha, \ell_0),
\]

where only the first term is \( x \)-dependent. Assume that \( \hat{\ell} \in \mathcal{L} \) is such that \( \|\hat{\ell} - \hat{\ell}_0\|_\infty < \rho^2 \). Then, \( |\int_0^{y-\theta_0} (\hat{\ell} - \hat{\ell}_0)(t) \, dt| \leq \rho^2 (y - \theta_0) \), so that \( z(\alpha - \rho^2, \hat{\ell}_0) \leq z(\alpha, \ell) \leq z(\alpha + \rho^2, \hat{\ell}_0) \).

Noting that \( d^k z / d\alpha^k(\alpha, \hat{\ell}_0) = (-1)^k P_0(X - \theta_0)^k \) \( < \infty \) and using the first-order Taylor expansion of \( z \) in \( \alpha \), we find, \( z(\alpha \pm \rho^2, \hat{\ell}_0) = z(\alpha, \hat{\ell}_0) + \rho^2 P_0(X - \theta_0) + O(\rho^4) \), and (5.32) follows.

Next note that, for every \( k \geq 1 \),

\[
\left| P_0 \left( \int_0^{x-\theta_0} (\hat{\ell} - \hat{\ell}_0)(t) \, dt \right)^k \right| \leq \rho^{2k} \int_\theta^\infty \left( \int_0^{x-\theta_0} dy \right)^k dP_0 = \rho^{2k} P_0(X - \theta_0)^k,
\]

(5.33)

Using (5.32) we bound the differences between KL divergences and integrals of scores as follows:

\[
\left| \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}}(x) \right) - \left( \int_0^{x-\theta_0} (\hat{\ell} - \hat{\ell}_0)(t) \, dt \right) \right| \leq \rho^2 \left( P_0(X - \theta_0) + O(\rho^2) \right),
\]

\[
\left| \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}}(x) \right)^2 - \left( \int_0^{x-\theta_0} (\hat{\ell} - \hat{\ell}_0)(t) \, dt \right)^2 \right| \leq \rho^2 \left( P_0(X - \theta_0) + O(\rho^2) \right)
\]

\[
\times \left| 2 \int_0^{x-\theta_0} (\hat{\ell} - \hat{\ell}_0)(t) \, dt + \rho^2 \left( P_0(X - \theta_0) + O(\rho^2) \right) \right|,
\]

and, combining with the bounds (5.33), we see that,

\[
-P_0 \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \leq 2\rho^2 \left( P_0(X - \theta_0) + O(\rho^2) \right),
\]

and

\[
P_0 \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \right)^2 \leq \rho^4 \left( P_0(X - \theta_0)^2 + 3P_0(X - \theta_0) + O(\rho^2) \right),
\]
which proves the first inclusion. Let \( M > 0 \). Note that \( A_\theta = [\theta, \infty) \), and that

\[
\sup_{|h| \leq M} -1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} = \sup_{|h| \leq M} 1_{A_{\theta_n}(h)} \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_n(h), \eta}}
\]

\[
= \sup_{|h| \leq M} 1_{A_{\theta_n}(h)} \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_0, \eta}} + \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_0, \eta}}
\]

\[
\leq (\alpha + S) \frac{M}{n} + \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_0, \eta}},
\]

so that,

\[
P_0 \left( \sup_{|h| \leq M} -1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} \right) \leq -P_0 \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_0, \eta}} + \frac{(\alpha + S) M}{n},
\]

\[
P_0 \left( \sup_{|h| \leq M} -1_{A_{\theta_n}(h)} \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} \right)^2 \leq P_0 \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \right)^2 + \frac{2(\alpha + S) M}{n} \left[ P_0 \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \right)^2 \right]^{1/2} + \frac{(\alpha + S)^2 M^2}{n^2},
\]

implying the existence of a constant \( L_2 \).

\[\Box\]

**Proof of Lemma 5.18.** Fix \( n \) and \( \omega \); write \( h_n \) for \( h_n(\omega) \). First we consider the case that \( h_n \geq 0 \), for \( x \geq \theta_0 \),

\[
(\eta^{1/2}(x - \theta_n(h_n)) - \eta^{1/2}(x - \theta_0))^2 = \eta(x - \theta_0)1_{[\theta_0, \theta_n(h_n)]}(x) + \eta^{1/2}(x - \theta_n(h_n)) - \eta^{1/2}(x - \theta_0))^2 1_{[\theta_n(h_n), \infty)}(x)
\]

To upper bound the second term, we use the absolute continuity of \( \eta^{1/2} \),

\[
|\eta^{1/2}(x - \theta_0) - \eta^{1/2}(x - \theta_n(h_n))| = \frac{1}{2} \left| \int_{x - \theta_0 - \frac{M}{\eta}}^{x - \theta_0} \eta' \eta^{1/2}(y) \, dy \right|
\]

\[
\leq \frac{1}{2} \int_{0}^{\frac{M}{\eta}} \left| \frac{\eta'}{\eta^{1/2}} \right| (z + x - \theta_n(h_n)) \, dz,
\]

and then by Jensen’s inequality,

\[
(\eta^{1/2}(x - \theta_0) - \eta^{1/2}(x - \theta_n(h_n)))^2 \leq \frac{M}{4n} \int_{0}^{\frac{M}{\eta}} (\eta')^2 \eta (z + x - \theta_n(h_n)) \, dz.
\]

Similarly for \( h_n < 0 \) and \( x \geq \theta_n(h_n) \),

\[
(\eta^{1/2}(x - \theta_0) - \eta^{1/2}(x - \theta_n(h_n)))^2 \leq \eta(x - \theta_n(h_n))1_{[\theta_n(h_n), \theta_0]}(x) - \eta(x - \theta_n(-M))1_{[\theta_n(-M), \theta_0]}(x)
\]

\[
+ \eta(x - \theta_n(-M))1_{[\theta_n(-M), \theta_0]} + \frac{M}{4n} \int_{0}^{\frac{M}{\eta}} (\eta')^2 \eta (z + x - \theta_0) \, dz 1_{[\theta_0, \infty)}(x).
\]
Combining these results, we obtain a bound for the squared Hellinger distance:

\[
H^2(P_{\theta_0(h_n), \eta}, P_{\theta_0, \eta}) \leq \int_{\theta_0}^{\theta_0(M)} \eta(x - \theta_0) \, dx + \int_{\theta_0(-M)}^{\theta_0} \eta(x - \theta_0(-M)) \, dx 
\]

\[
\quad + 1_{\{h_n < 0\}} \int_{\theta_0(h_n)}^{\theta_0} \eta(x - \theta_0(h_n)) \, dx 
\]

\[
- 1_{\{h_n < 0\}} \int_{\theta_0(-M)}^{\theta_0} \eta(x - \theta_0(-M)) \, dx 
\]

\[
1_{\{h_n \geq 0\}} \int_{\theta_0(h_n)}^{\infty} \frac{M}{4n} \int_0^{\frac{\eta'}{\eta}} (\eta')^2 (z + x - \theta_0(h_n)) \, dz \, dx 
\]

\[
+ 1_{\{h_n < 0\}} \int_{\theta_0}^{\infty} \frac{M}{4n} \int_0^{\frac{\eta'}{\eta}} (\eta')^2 (z + x - \theta_0) \, dz \, dx. 
\]

As for the first two terms on the right-hand side of (5.34), we note the following inequality:

\[
\int_{\theta_0}^{\theta_0(M)} \eta(x - \theta_0) \, dx + \int_{\theta_0(-M)}^{\theta_0} \eta(x - \theta_0(-M)) \, dx \leq 2\gamma_{\theta_0, \eta} \frac{M}{n} + \frac{M^2}{n^2} \int_0^\infty |\eta'(y)| \, dy, 
\]

by Lemma A.13. Furthermore, by shifting appropriately, we find that the third and fourth term of (5.34) satisfy the bound,

\[
1_{\{h_n < 0\}} \left( \int_{\theta_0(h_n)}^{\theta_0} \eta(x - \theta_0(h_n)) \, dx - \int_{\theta_0(-M)}^{\theta_0} \eta(x - \theta_0(-M)) \, dx \right) 
\]

\[
= 1_{\{h_n < 0\}} \left( \int_0^{\frac{h_n}{\eta}} \eta(y) \, dy - \int_0^{\frac{M}{\eta}} \eta(y) \, dy \right) 
\]

\[
= -1_{\{h_n < 0\}} \int_0^{\frac{M}{\eta}} \eta(y) \, dy \leq 0, 
\]

(\textit{where it is noted that the } h_n \text{ dependent integral in the above display is well defined for any } h_n). \textit{Finally, the fifth and sixth term of (5.34) are bounded by the Fisher information for location associated with } \eta: \]

\[
\int_0^\infty \frac{M}{n} \int_0^{\frac{\eta'}{\eta}} (z + x) \, dz \, dx = \int_0^{\frac{M}{\eta}} \int_0^\infty \frac{M}{n} \frac{(\eta')^2}{\eta} (x) \, dx \, dz \leq \frac{M}{n} \int_0^\infty \frac{(\eta')^2}{\eta} (x) \, dx. 
\]

Combining, we obtain the following upper bound for the relevant Hellinger distance,

\[
H^2(P_{\theta_0(h_n), \eta}, P_{\theta_0, \eta}) \leq 2\gamma_{\theta_0, \eta} \frac{M}{n} + 2\frac{M^2}{n^2} \left( \int_0^\infty |\eta'(x)| \frac{\eta(x)}{\eta(x)} \, dx + \int_0^\infty \left( \frac{\eta'(x)}{\eta(x)} \right)^2 \eta(x) \, dx \right), 
\]

which proves the lemma upon noting that \(|\eta'(x)| = \eta(x)|\dot{\ell}(x) - \alpha| \leq \eta(x)(\alpha - S)\). \hfill \Box
5.5.4 Proofs of Subsection 5.4.2

PROOF OF LEMMA 5.23. Denote by $Q$ the distribution with density $\eta_0 = \eta_{\ell_0}$. Then the family $\mathcal{F} = \{x \mapsto \sqrt{\eta_\ell/\eta_0} : \ell \in \mathcal{L}\}$ forms a subset of the collection $C^1_M([0,1])$, where $M$ is fixed and depends on $S$. Referring to Corollary 2.7.2 in [104], we conclude that the $L_2(Q)$-bracketing entropy $N_{[\epsilon]}(\mathcal{F}, L_2(Q))$ of $\mathcal{F}$ is finite for all $\epsilon > 0$. Noting that,

$$d_H(\eta, \eta_0)^2 = d_H(\eta_{\ell}, \eta_0)^2 = \int_0^1 \left(\sqrt{\frac{\eta_{\ell}(x)}{\eta_0}(x)} - \sqrt{\frac{\eta_0(x)}{\eta_0}(x)}\right)^2 dQ(x),$$

it follows that $N(\rho, \mathcal{H}, d_H) = N(\rho, \mathcal{F}, L_2(Q)) \leq N_{[\epsilon]}(2\rho, \mathcal{F}, L_2(Q)) < \infty$. □

PROOF OF LEMMA 5.24. Let $\rho > 0$ and $\hat{\ell} \in \mathcal{L}$ such that $\|\hat{\ell} - \ell_0\|_\infty \leq \rho^2$ be given. Then,

$$\left| \log \frac{p_{\theta_0,\eta}(x)}{p_{\theta_0,\eta_0}(x)} - \int_0^{x/\theta_0} (\hat{\ell} - \ell_0(t)) \, dt \right| \leq \rho^2 P_0(X/\theta_0) + O(\rho^4), \quad (5.35)$$

for all $x \in [0, \theta_0]$. Define, for all $\alpha \in \mathbb{R}$ and $\ell \in \mathcal{L}$,

$$z(\alpha, \hat{\ell}) = \log \int_0^1 e^{\alpha y + \int_0^y \hat{\ell}(t) \, dt} \, dy.$$

Then the relevant log-density-ratio can be written as,

$$\log \frac{p_{\theta_0,\eta}(x)}{p_{\theta_0,\eta_0}(x)} = \int_0^{x/\theta_0} (\hat{\ell} - \ell_0(t)) \, dt - z(S, \hat{\ell}) + z(S, \ell_0),$$

where only the first term is $x$-dependent. Assume that $\hat{\ell} \in \mathcal{L}$ is such that $\|\hat{\ell} - \ell_0\|_\infty < \rho^2$. Then, $|\int_0^y (\hat{\ell} - \ell_0(t)) \, dt| \leq \rho^2 y$, so that $z(S - \rho^2, \ell_0) \leq z(S, \hat{\ell}) \leq z(S + \rho^2, \ell_0)$. Noting that $d^k z/d\alpha^k(S, \ell_0) = P_0(X/\theta_0)^k < \infty$ and using the first-order Taylor expansion of $z$ in $\alpha$, we find, $z(S \pm \rho^2, \ell_0) = z(S, \ell_0) \pm \rho^2 P_0(X/\theta_0) + O(\rho^4)$, and (5.35) follows.

Next note that, for every $k \geq 1,

$$P_0\left(\int_0^{x/\theta_0} (\hat{\ell} - \ell_0(t)) \, dt\right)^k \leq \rho^{2k} \int_0^{\theta_0} \left(\int_0^{x/\theta_0} dy\right)^k dP_0 = \rho^{2k} P_0(X/\theta_0)^k, \quad (5.36)$$

Using (5.35) we bound the differences between KL divergences and integrals of scores as follows:

$$\left| \left( \log \frac{p_{\theta_0,\eta}(x)}{p_{\theta_0,\eta_0}(x)} \right) - \left( \int_0^{x/\theta_0} (\hat{\ell} - \ell_0(t)) \, dt \right) \right| \leq \rho^2 (P_0(X/\theta_0) + O(\rho^2)), \quad \left| \frac{\log p_{\theta_0,\eta}(x)}{\log p_{\theta_0,\eta_0}(x)} \right|^2 - \left( \int_0^{x/\theta_0} (\hat{\ell} - \ell_0(t)) \, dt \right)^2 \leq \rho^2 (P_0(X/\theta_0) + O(\rho^2))$$

$$\times 2 \int_0^{x/\theta_0} (\hat{\ell} - \ell_0(t)) \, dt + \rho^2 (P_0(X/\theta_0) + O(\rho^2)).$$
and, combining with the bounds (5.36), we see that,

\[ -P_0 \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \leq 2 \rho^2 \left( P_0(X/\theta_0) + O(\rho^2) \right), \]

\[ P_0 \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \right)^2 \leq \rho^4 \left( P_0(X/\theta_0)^2 + 3P_0(X/\theta_0) + O(\rho^2) \right), \]

which proves the first inclusion. Let \( M > 0 \). Note that \( A_\theta = [0, \theta] \), and that for large enough \( n \),

\[
\sup_{|h| \leq M} \left( 1 - A_{\theta_n(h)} \right) \left( \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} \right) = \sup_{|h| \leq M} \left( 1 - A_{\theta_n(h)} \right) \left( \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_n(h), \eta}} \right) \\
= \sup_{|h| \leq M} \left( 1 - A_{\theta_n(h)} \right) \left( \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_0, \eta}} \cdot \frac{p_{\theta_0, \eta}}{p_{\theta_n(h), \eta}} \right) \\
= \sup_{|h| \leq M} \left( 1 - A_{\theta_n(h)} \right) \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} + \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_n(h), \eta}} \\
\leq \frac{2 + 8S}{\theta_0} \frac{M}{n} + \log \frac{p_{\theta_0, \eta_0}}{p_{\theta_0, \eta}},
\]

so that,

\[
P_0 \left( \sup_{|h| \leq M} \left( 1 - A_{\theta_n(h)} \right) \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} \right) \leq -P_0 \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} + \frac{2 + 8S}{\theta_0} \frac{M}{n},
\]

\[
P_0 \left( \sup_{|h| \leq M} \left( 1 - A_{\theta_n(h)} \right) \log \frac{p_{\theta_n(h), \eta}}{p_{\theta_0, \eta_0}} \right)^2 \leq P_0 \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \right)^2 + \frac{4 + 16S}{\theta_0} \frac{M}{n} \left[ P_0 \left( \log \frac{p_{\theta_0, \eta}}{p_{\theta_0, \eta_0}} \right)^2 \right]^{1/2} + \frac{(2 + 8S)^2 M^2}{\theta_0^2 n^2},
\]

implying the existence of a constant \( L_2 \).

**Proof of Lemma 5.25.** Note that the elements of the nuisance space \( H \) are uniformly bounded by \( e^{2S} \). Fix \( n \) and \( \omega \); write \( h_n \) for \( h_n(\omega) \). First we consider the case that \( h_n \geq 0 \),

\[
\left( \frac{\eta^{1/2}(x/\theta_n(h_n))}{\theta_n^{1/2}(h_n)} - \frac{\eta^{1/2}(x/\theta_0)}{\theta_0^{1/2}} \right)^2 = \frac{\eta(x/\theta_n(h_n))}{\theta_n(h_n)} 1_{[\theta_0, \theta_n(h_n)]}(x) \\
+ \left( \frac{\eta^{1/2}(x/\theta_n(h_n))}{\theta_n^{1/2}(h_n)} - \frac{\eta^{1/2}(x/\theta_0)}{\theta_0^{1/2}} \right)^2 1_{[0, \theta_0]}(x).
\]

Note that the first term is bounded from above by \( (e^{2S}/\theta_0) 1_{[\theta_0, \theta_n(M)]}(x) \). To upper bound the second term, we use the absolute continuity of \( \eta^{1/2} \). Let \( g(y) = (\eta(x/y)/y)^{1/2} \),

\[
\left| \frac{\eta^{1/2}(x/\theta_n(h_n))}{\theta_n^{1/2}(h_n)} - \frac{\eta^{1/2}(x/\theta_0)}{\theta_0^{1/2}} \right| = \left| \int_{\theta_0}^{\theta_n(h_n)} g(y) \ dy \right| \leq \int_{\theta_0}^{\theta_n(M)} \left| g'(y) \right| \ dy.
\]
Note that
\[ g'(y) = \frac{1}{2g(y)} \left( \frac{1}{y} \eta'(x) \left( - \frac{x}{y^2} \right) + \left( - \frac{1}{y^2} \right) \eta \left( \frac{x}{y} \right) \right). \]

By the definition of the nuisance space, for \( y \in [\theta_0, \theta_n(M)] \), and \( x \leq \theta_0 \),
\[ |g'(y)| \leq \frac{e^S}{\theta_0^{3/2}} (S + 1), \]
and then,
\[ \left( \frac{\eta^{1/2}(x/\theta_n(h_n))}{\theta_n^{1/2}(h_n)} - \frac{\eta^{1/2}(x/\theta_0)}{\theta_0^{1/2}} \right)^2 \leq \frac{M^2 e^{2S}}{n^2} \frac{e^{2S}}{\theta_0^{3/2}} (S + 1)^2. \]

Similarly for \( h_n < 0 \),
\[ \left( \frac{\eta^{1/2}(x/\theta_n(h_n))}{\theta_n^{1/2}(h_n)} - \frac{\eta^{1/2}(x/\theta_0)}{\theta_0^{1/2}} \right)^2 \leq \frac{e^{2S}}{\theta_0 - M/n} 1_{[\theta_n(-M), \theta_0]}(x) \]
\[ + \frac{M^2 e^{2S}}{n^2} \frac{e^{2S}}{(\theta_0 - M/n)^3} \left( \frac{S\theta_0}{\theta_0 - M/n} + 1 \right)^2 1_{[0, \theta_0]}(x). \]

Combining these results, we obtain a bound for the squared Hellinger distance:
\[ H^2(P_{\theta_n(h_n)}, \eta, P_{\theta_0, \eta}) \leq \frac{Me^{2S}}{n\theta_0} + \frac{Me^{2S}}{n\theta_0 - M} + \frac{M^2 e^{2S}}{n^2} \frac{e^{2S}}{\theta_0^{3/2}} (S + 1)^2 + \frac{M^2 e^{2S} \theta_0}{n^2} \frac{e^{2S} \theta_0}{(\theta_0 - M/n)^3} \left( \frac{S\theta_0}{\theta_0 - M/n} + 1 \right)^2, \]
uniformly in \( \eta \). \( \square \)
Appendix A

Technical lemmas

For $\beta \in \mathbb{R}$, the Sobolev norm $\|\mu\|_\beta$ of an element $\mu \in \mathbb{R}^\infty$ and the $\ell_2$-norm $\|\mu\|$ of an element $\mu \in \ell_2$ are defined in a usual way by $\|\mu\|_\beta^2 = \sum_{i=1}^{\infty} i^{2\beta} \mu_i^2$, $\|\mu\|^2 = \sum_{i=1}^{\infty} \mu_i^2$, and the corresponding Sobolev space by $S^\beta = \{ \mu \in \mathbb{R}^\infty : \|\mu\|_\beta < \infty \}$.

Recall that $S : [0, \infty) \to \mathbb{R}$ is slowly varying if $S(tx)/S(t) \to 1$ as $t \to \infty$, for every $x > 0$.

For two sequences $(a_n)$ and $(b_n)$ of numbers, $a_n \asymp b_n$ means that $|a_n/b_n|$ is bounded away from zero and infinity as $n \to \infty$, $a_n \lesssim b_n$ means that $a_n/b_n$ is bounded, $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$, and $a_n \ll b_n$ means that $a_n/b_n \to 0$ as $n \to \infty$. For two real numbers $a$ and $b$, we denote by $a \vee b$ their maximum, and by $a \wedge b$ their minimum.

Lemma A.1. For any $q \in \mathbb{R}$, $t \geq -2q$, $u > 0$ and $v \geq 0$, as $N \to \infty$,

$$\sup_{\|\xi\| \leq 1} \sum_{i=1}^{\infty} \xi_i^2 i^{-t} \left(1 + Ni^{-u}\right)^v \asymp N^{-\frac{t+2q}{u} \wedge v}.$$

Moreover, for every fixed $\xi \in S^q$, as $N \to \infty$,

$$N^{-\frac{t+2q}{u} \wedge v} \sum_{i=1}^{\infty} \xi_i^2 i^{-t} \left(1 + Ni^{-u}\right)^v \to \begin{cases} 0, & \text{if } (t+2q)/u < v, \\ \|\xi\|_2^2 (uv-t)/2, & \text{if } (t+2q)/u \geq v. \end{cases}$$

The last assertion remains true if the sum is limited to the terms $i \leq cN^{1/u}$, for any $c > 0$.

Proof. In the range $i \leq N^{1/u}$ we have $Ni^{-u} \leq 1 + Ni^{-u} \leq 2Ni^{-u}$, and $1 \leq 1 + Ni^{-u} \leq 2$ in the range $i > N^{1/u}$. Thus, deleting either the first or second term, we obtain

$$\sum_{i \leq N^{1/u}} \frac{\xi_i^2 t^{-t}}{(1 + Ni^{-u})^v} \asymp \sum_{i \leq N^{1/u}} \xi_i^2 t^{-2q} i^{iuv-t-2q} N^v \leq \|\xi\|_q^2 N^{-\frac{t+2q}{u} \wedge v},$$

$$\sum_{i > N^{1/u}} \frac{\xi_i^2 t^{-t}}{(1 + Ni^{-u})^v} \asymp \sum_{i > N^{1/u}} \xi_i^2 t^{-2q} i^{-t-2q} \leq N^{-\frac{t+2q}{u}} \sum_{i > N^{1/u}} \xi_i^2 t^{2q}. $$
The inequality in the first line follows by bounding \( i \) in \( i^{uv-t-2q} \) by \( N^{1/u} \) if \( uv-t-2q > 0 \), and by 1 otherwise. This proves the upper bound for the supremum.

The lower bound follows by considering the two sequences \((\xi_i)\) given by \( \xi_i = i^{-q} \) for \( i \sim N^{1/u} \) and \( \xi_i = 0 \) otherwise (showing that the supremum is bigger than \( N^{-(t+2q)/u} \)), and given by \( \xi_1 = 1 \) and \( \xi_i = 0 \) otherwise (showing that the supremum is bigger than \( N^{-v} \)).

The second line of the preceding display shows that the sum over the terms \( i > N^{1/u} \) is \( o(N^{-(t+2q)/u}) \). Furthermore, the first line can be multiplied by \( N^{(t+2q)/u} \) to obtain

\[
N^{\frac{t+2q}{u}} \sum_{i \leq N^{1/u}} \frac{\xi_i^{2-t}}{(1 + Ni^{-u})^v} \times \sum_{i \leq N^{1/u}} \xi_i^{2q} \left( \frac{i}{N^{1/u}} \right)^{uv-t-2q}.
\]

If \( (t+2q)/u < v \), then \( uv-t-2q > 0 \) and this tends to zero by dominated convergence. Also,

\[
N^v \sum_{i = 1}^{\infty} \frac{\xi_i^{2-t}}{(1 + Ni^{-u})^v} = \sum_{i = 1}^{\infty} \xi_i^{2q} \left( \frac{Ni^{-u}}{1 + Ni^{-u}} \right)^v.
\]

If \( (t+2q)/u \geq v \), then \( q \geq (uv-t)/2 \) and, hence, \( \xi \in S^{(uv-t)/2} \), and the right-hand side tends to \( \sum_i \xi_i^{2q} \) by dominated convergence.

The final assertion needs to be proved only in the case that \((t+2q)/u \geq v\), as in the other case the whole sum tends to 0. The sum over the terms \( i \leq N^{1/u} \) was seen to be always \( o(N^{-(t+2q)/u}) \), which is \( o(N^{-v}) \) if \( (t+2q)/u \geq v \). The final assertion for \( c = 1 \) follows, because the sum over the terms \( i \leq N^{1/u} \) was seen to have the exact order \( N^{-v} \) (if \( \xi \neq 0 \)). For general \( c \) the proof is analogous, or follows by scaling \( N \).

**Lemma A.2.** For any \( q \in \mathbb{R}, u, v, 0 \geq v, t \geq -2q, p > 0, \) and \( 0 \leq r < vp \), as \( N \to \infty \),

\[
\sup_{\|\xi\|_1 \leq 1} \sum_{i = 1}^{\infty} \frac{\xi_i^{2-t} e^{-ri^2}}{(1 + Ni^{-u} e^{-pi^2})^v} \asymp N^{-\frac{v}{p}} (\log N)^{-\frac{1}{2} - q + \frac{u-r}{p}}.
\]

Moreover, for every fixed \( \xi \in S^q \), as \( N \to \infty \),

\[
N^{-\frac{v}{p}} (\log N)^{\frac{1}{2} + q - \frac{u-r}{p}} \sum_{i = 1}^{\infty} \frac{\xi_i^{2-t} e^{-ri^2}}{(1 + Ni^{-u} e^{-pi^2})^v} \to 0.
\]

**Proof.** Let \( I_N \) be the solution to \( Ni^{-u} e^{-pi^2} = 1 \). In the range \( i \leq I_N \) we have \( Ni^{-u} e^{-pi^2} \leq 1 + Ni^{-u} e^{-pi^2} \leq 2Ni^{-u} e^{-pi^2} \), while \( 1 \leq 1 + Ni^{-u} e^{-pi^2} \leq 2 \) in the range \( i \geq I_N \). Thus

\[
\sum_{i \leq I_N} \frac{\xi_i^{2-t} e^{-ri^2}}{(1 + Ni^{-u} e^{-pi^2})^v} \asymp \sum_{i \leq I_N} \xi_i^{2q} i^{uv-t-2q} e^{(vp-r)i^2} \frac{1}{N^v} \leq \|\xi\|^2 N^{-\frac{v}{p}} I_N^{-t-2q + \frac{u-r}{p}},
\]

since for \( N \) large enough all terms \( i^{uv-t-2q} e^{(vp-r)i^2} \) in this range will be dominated by \( I_N^{uv-t-2q} e^{(vp-r)i^2} \) and \( I_N \) solves the equation \( Ni^{-u} e^{-pi^2} = 1 \). Similarly for the second range, we have

\[
\sum_{i \geq I_N} \frac{\xi_i^{2-t} e^{-ri^2}}{(1 + Ni^{-u} e^{-pi^2})^v} \asymp \sum_{i \geq I_N} \xi_i^{2q} i^{t-2q} e^{-ri^2} \leq N^{-\frac{v}{p}} I_N^{-t-2q + \frac{u-r}{p}} \sum_{i \geq I_N} \xi_i^{2q}.
\]
Lemma A.7 yields the upper bound for the supremum.

The lower bound follows by considering the sequence \( (\xi_i) \) given by \( \xi_i = i^{-q} \) for \( i \sim I_N \) and \( \xi_i = 0 \) otherwise, showing that the supremum in the assertion of the lemma is bigger than \( N^{-r/p} (\log N)^{-t/2-q+ur/(2p)} \).

The preceding display shows that the sum over the terms \( i \geq I_N \) is of the smaller order than \( N^{-r/p} (\log N)^{-t/2-q+ur/(2p)} \). Furthermore,

\[
N^{r/p} (\log N)^{-\frac{1}{2}+q-\frac{v}{2p}} \sum_{i \leq I_N} \frac{\xi_i^{2i-t} e^{-ri^2}}{(1 + Ni^{-u} e^{-pi^2}) v} \leq \sum_{i \leq I_N} \frac{\xi_i^{2i-t} e^{-ri^2}}{N^v I_N^{1-2q+ur/p} e^{-rI_N^2}},
\]

and this tends to zero by dominated convergence. Indeed, as noted before, for \( N \) large enough all terms \( i^{uv-t-2q} e^{(vp-r)i^2} \) in the range \( i \leq I_N \) are upper bounded by \( I_N^{2v-t-2q} e^{(vp-r)I_N^2} = N^{v-r/p} I_N^{1-2q+ur/p} \), and by Lemma A.7

\[
N^{v-\frac{t}{p}} I_N^{1-2q+ur/p} \leq N^{v-\frac{t}{p}} (\log N)^{-\frac{1}{2}-q+\frac{v}{2p}} \to \infty,
\]

since \( v-r/p > 0 \).

\[\square\]

**Lemma A.3.** For any \( t, v \geq 0, u > 0 \), and \( (\xi_i) \) such that \( |\xi_i| = i^{-q-1/2}S(i) \) for \( q > -t/2 \) and a slowly varying function \( S: (0, \infty) \to (0, \infty) \), as \( N \to \infty \),

\[
\sum_{i=1}^{\infty} \frac{\xi_i^{2i-t}}{(1 + Ni^{-u})^v} \leq \begin{cases} N^{-\frac{t+2q}{u}} S^2(N^{1/u}), & \text{if } (t+2q)/u < v, \\ N^{-v} \sum_{i \leq N^{1/u}} S^2(i) / i, & \text{if } (t+2q)/u = v, \\ N^{-v}, & \text{if } (t+2q)/u > v. \end{cases}
\]

Moreover, for every \( c > 0 \), the sum on the left is asymptotically equivalent to the same sum restricted to the terms \( i \leq cN^{1/u} \) if and only if \( (t+2q)/u \geq v \).

**PROOF.** As in the proof of the preceding lemma, we split the infinite series in the sum over the terms \( i \leq N^{1/u} \) and \( i > N^{1/u} \). For the first part of the series

\[
\sum_{i \leq N^{1/u}} \frac{\xi_i^{2i-t}}{(1 + Ni^{-u})^v} \leq \sum_{i \leq N^{1/u}} S^2(i) / i.
\]

If \( uv-t-2q > 0 \) [i.e., \( (t+2q)/u < v \)], the right-hand side is of the order \( N^{-(t+2q)/u} S^2(N^{1/u}) \), by Theorem 1(b) on page 281 in [33], while if \( uv-t-2q < 0 \), it is of the order \( N^{-v} \) by Lemma on page 280 in [33]. Finally, if \( uv-t-2q = 0 \), then the right-hand side is identical to \( N^{-v} \sum_{i \leq N^{1/u}} S^2(i) / i \).

The other part of the infinite series satisfies, by Theorem 1(a) on page 281 in [33],

\[
\sum_{i > N^{1/u}} \frac{\xi_i^{2i-t}}{(1 + Ni^{-u})^v} \leq \sum_{i > N^{1/u}} S(i)^2 i^{-t-2q-1} \leq N^{-\frac{t+2q}{u}} S^2(N^{1/u}).
\]

This is never bigger than the contribution of the first part of the sum, and of equal order if \( (t+2q)/u < v \). If \( (t+2q)/u > v \), then the leading polynomial term is strictly smaller than
$N^{-v}$. If $(t + 2q)/u = v$, then the leading term is equal to $N^{-v}$, but the slowly varying part satisfies $S^2(N^{1/u}) \ll \sum_{i \leq N^{1/u}} S^2(i)/i$, by Theorem 1(b) on page 281 in [33]. Therefore, in both cases the preceding display is negligible relative to the first part of the sum. This proves the final assertion of the lemma for $c = 1$. The proof for general $c > 0$ is analogous. □

By the Cauchy–Schwarz inequality, for any $\mu \in S^{t/2}$,

$$\left| \sum_{i=1}^{\infty} \frac{\xi_i \mu_i}{1 + Ni^{-u}} \right|^2 \leq \|\mu\|_{t/2}^2 \sum_{i=1}^{\infty} \frac{\xi_i^2 i^{-t}}{(1 + Ni^{-u})^2}.$$ 

The preceding lemma gives the exact order of the right-hand side. The application of the Cauchy–Schwarz inequality is sharp, in that there is equality for some $\mu \in S^{t/2}$. However, this $\mu$ depends on $N$. For fixed $\mu \in S^{t/2}$ the left-hand side is strictly smaller than the right-hand side.

**Lemma A.4.** For any $t, u \geq 0$, $\mu \in S^{t/2}$ and $(\xi_i)$ such that $|\xi_i| = i^{-q-1/2}S(i)$ for $0 < t + 2q < 2u$ and a slowly varying function $S: (0, \infty) \to (0, \infty)$, as $N \to \infty$,

$$\sum_{i=1}^{\infty} \frac{|\xi_i \mu_i|}{1 + Ni^{-u}} \ll N^{-\frac{t+2q}{2u}} S(N^{1/u}).$$

**Proof.** We split the series in two parts, and bound the denominator $1 + Ni^{-u}$ by $Ni^{-u}$ or 1. By the Cauchy–Schwarz inequality, for any $r > 0$,

$$\left| \sum_{i \leq N^{1/u}} \frac{|\xi_i \mu_i|}{Ni^{-u}} \right|^2 \leq \frac{1}{N^2} \sum_{i \leq N^{1/u}} \frac{S^2(i) i^r}{i} \sum_{i \leq N^{1/u}} \mu_i^2 i^{2u-2q-r}$$

$$\ll \frac{1}{N^2} S^2 \left( N^{\frac{1}{u}} \right) N^{\frac{r}{u}}$$

$$\times \sum_{i \leq N^{1/u}} \mu_i^2 \left( \frac{i}{N^{1/u}} \right)^{2u-2q-r-t} N^{\frac{2u-2q-r-t}{s}},$$

$$\left| \sum_{i > N^{1/u}} \frac{|\xi_i \mu_i|}{1} \right|^2 \leq \sum_{i > N^{1/u}} \frac{S^2(i) i^{-2q}}{i} \sum_{i > N^{1/u}} \mu_i^2 \ll S^2 \left( N^{\frac{1}{u}} \right) N^{-\frac{2q}{u}} \sum_{i > N^{1/u}} \mu_i^2.$$

The terms in the remaining series in the right-hand side of the first inequality are bounded by $\mu_i^2 i^t$ and tend to zero pointwise as $N \to \infty$ if $2u - 2q - r - t > 0$. If $t + 2q < 2u$, then there exists $r > 0$ such that the latter is true, and for this $r$ the sum tends to zero by the dominated convergence theorem. The other terms collect to $N^{-(t+2q)/(u)} S^2(N^{1/u})$. The sum in the right-hand side of the second inequality is bounded by $\sum_{i > N^{1/u}} \mu_i^2 i^t N^{-t/u} = o(N^{-t/u})$. □

**Lemma A.5.** For any $t, u, v \geq 0$, $p > 0$, and $0 \leq r < vp$, as $N \to \infty$,

$$\sum_{i=1}^{\infty} \frac{i^{-t} e^{-ri^2}}{(1 + Ni^{-u} e^{-ri^2})^v} \ll \begin{cases} N^{-\frac{r}{p}} (\log N)^{-\frac{1}{2} + \frac{vp}{2p}}, & \text{if } r \neq 0, \\ (\log N)^{-\frac{1}{2} + \frac{1}{p}}, & \text{if } r = 0. \end{cases}$$
PROOF. As in the preceding proof we split the infinite series in the sum over the terms \( i \leq I_N \) and \( i \geq I_N \). For the first part of the sum we get

\[
\sum_{i \leq I_N} \frac{i^{-t} e^{-ri^2}}{1 + Ni^u e^{-pi^2}} \leq \sum_{i \leq I_N} \frac{i^{uv-t} e^{(vp-r)i^2}}{N^v}.
\]

Most certainly

\[
N^v \cdot I_N^t e^{-rI_N^2} = I_N^{uv-t} e^{(vp-r)I_N^2} \leq \sum_{i \leq I_N} i^{uv-t} e^{(vp-r)i^2}.
\]

If \( i^{uv-t} e^{(vp-r)i^2} \) as a function of \( i \) is strictly increasing, then the sum is upper bounded by the integral in the same range, and the value at the right end-point. Otherwise \( i^{uv-t} e^{(vp-r)i^2} \) first decreases, and then increases, and, therefore, the sum is upper bounded by the integral, and values at both endpoints:

\[
\sum_{i \geq I_N} i^{uv-t} e^{(vp-r)i^2} \leq \int_1^{I_N} x^{uv-t} e^{(vp-r)x^2} dx + e^{vp-r} + I_N^{uv-t} e^{(vp-r)I_N^2}
= \frac{1}{2(vp-r)} I_N^{uv-t-1} e^{(vp-r)I_N^2} (1 + o(1)) + e^{vp-r} + I_N^{uv-t} e^{(vp-r)I_N^2}
\leq I_N^{uv-t} e^{(vp-r)I_N^2} (1 + o(1)),
\]

by Lemma A.8. Therefore, by Lemma A.7,

\[
\sum_{i \leq I_N} \frac{i^{uv-t} e^{(vp-r)i^2}}{N^v} \times I_N^t e^{-rI_N^2} = N^{-\frac{v}{p}} I_N^{-t + \frac{uv}{p}} \times N^{-\frac{v}{p}} (\log N)^{-\frac{1}{2} + \frac{uv}{p}}.
\]

The other part of the sum satisfies

\[
\sum_{i \geq I_N} \frac{i^{-t} e^{-ri^2}}{1 + Ni^u e^{-pi^2}} \leq \sum_{i \geq I_N} i^{-t} e^{-ri^2}.
\]

Suppose \( r > 0 \). Again, the latter sum is bounded from below by

\[
I_N^{-t} e^{-rI_N^2} \times N^{-\frac{v}{p}} (\log N)^{-\frac{1}{2} + \frac{uv}{p}}.
\]

Since \( i^{-t} e^{-ri^2} \) is decreasing, we get the following upper bound

\[
\sum_{i \geq I_N} i^{-t} e^{-ri^2} \leq I_N^{-t} e^{-rI_N^2} + \int_{I_N}^{\infty} x^{-t} e^{-rx^2} dx \leq I_N^{-t} e^{-rI_N^2} + \frac{1}{2r} I_N^{-t-1} e^{-rI_N^2}
\leq I_N^{-t} e^{-rI_N^2} (1 + o(1)) \times N^{-\frac{v}{p}} (\log N)^{-\frac{1}{2} + \frac{uv}{p}},
\]

where the upper bound for the integral follows from Lemma A.8. In case \( r = 0 \), we get \( \sum_{i > I_N} i^{-t} \asymp I_N^{-t+1} \asymp (\log N)^{-(t+1)/2} \) (see Lemma A.3). \( \square \)
Lemma A.6. For any $t \geq 0$, $u, p > 0$, $\mu \in S^{t/2}$, and $q > -t/2$, as $N \to \infty$
\[ \sum_{i=1}^{\infty} \frac{|\mu_i i^{-q - 1/2}|}{1 + N i^{-u} e^{-p i^2}} \ll (\log N)^{-t/2 - q}. \]

PROOF. We split the series in two parts, and bound the denominator $1 + N i^{-u} e^{-p i^2}$ by $N i^{-u} e^{-p i^2}$ or 1. By the Cauchy–Schwarz inequality, for any $r > 0$,
\[ \left| \sum_{i \leq I_N} |\mu_i i^{-q - 1/2}|^2 \right| \leq \frac{1}{N^2} \sum_{i \leq I_N} \frac{i^r}{i} \sum_{i \leq I_N} \mu_i^2 i^{-2u - 2q - r} e^{2p i^2} \]
\[ \leq \frac{1}{N^2} I_N^2 \sum_{i \leq I_N} \mu_i^2 i^{-2} \frac{i^{2u - 2q - r} e^{2p i^2}}{I_N^2} \sum_{i \leq I_N} \frac{i^{2} e^{2p i^2}}{I_N^2} \]
\[ = I_N^{-2q} \sum_{i \leq I_N} \mu_i^2 i^{-2} \frac{i^{2u - 2q - r} e^{2p i^2}}{I_N^2} \sum_{i \leq I_N} \frac{i^{2} e^{2p i^2}}{I_N^2} \]

The terms in the remaining series in the right side are bounded by a constant times $\mu_i^2 i^{-2}$ for large enough $N$ and all $i$ bigger than a fixed number, and tend to zero pointwise as $N \to \infty$, and the sum tends to zero by the dominated convergence theorem. Therefore, the first part of the sum in the assertion is $o(I_N^{-2q - t})$. As for the other part we have
\[ \left| \sum_{i > I_N} |\mu_i i^{-q - 1/2}|^2 \right| \leq \sum_{i > I_N} i^{-2q - 1} \sum_{i > I_N} \mu_i^2 \leq I_N^{-t - 2q} \sum_{i > I_N} \mu_i^2 i^{-2}, \]
which completes the proof upon noting that $\mu \in S^{t/2}$, and $I_N^{-t - 2q} \ll (\log N)^{-t/2 - q}$ by Lemma A.7. 

Lemma A.7. Let $I_N$ be the solution for $1 = N i^{-u} e^{-p i^2}$, for $u \geq 0$ and $p > 0$. Then
\[ I_N \sim \sqrt{\frac{1}{p} \log N}. \]

PROOF. If $u = 0$ the assertion is obvious. Consider $u > 0$. The Lambert function $W$ satisfies the following identity $z = W(z) \exp W(z)$. The equation $1 = N i^{-u} e^{-p i^2}$ can be rewritten as
\[ \frac{2p}{u} N^{2/u} = \exp \left( \frac{2p}{u} i^2 \right) \frac{2p}{u} i^2 \]
and, therefore, by definition of $W(z)$
\[ I_N = \sqrt{\frac{u}{2p} W \left( N^{2/u} \frac{2p}{u} \right)}. \]
By [25] $W(x) \sim \log(x)$, which completes the proof. 

Lemma A.8. 1. For $\gamma \in \mathbb{R}$, $\zeta > 0$ we have, as $K \to \infty$,
\[ \int_{1}^{K} e^{\zeta x^2} x^\gamma dx \sim \frac{1}{2\zeta} e^{\zeta K^2} K^{\gamma - 1}. \]
2. For $K > 0$, $\gamma > 0$, $\zeta > 0$ we have

$$\int_K^\infty e^{-\zeta x^2} x^{-\gamma} \, dx \leq \frac{1}{2\zeta} e^{-\zeta K^2} K^{-\gamma-1}.$$ 

PROOF. First integrating by substitution $y = x^2$ and then by parts proves the lemma, with the help of the dominated convergence theorem in case 1. 

Lemma A.9. Let $c > 0$ and $r \geq 1 + c$.

(i) For $n \geq 1$

$$\sum_{i=1}^\infty \frac{n \log i}{i^r + n} \leq \left(2 + \frac{2}{c} + \frac{2}{c^2 \log 2}\right) \frac{n^{1/r} \log n}{r}.$$ 

(ii) If $r > (\log n) / (\log 2)$, then for $n \geq 1$

$$\sum_{i=1}^\infty \frac{n \log i}{i^r + n} \leq \left(1 + \frac{2}{c} + \frac{2}{c^2 \log 2}\right) (\log 2) n 2^{-r}.$$ 

PROOF. First consider $r \leq (\log n) / (\log 2)$, which implies that $n^{1/r} \geq 2$. We split the series in two parts, and bound the denominator $i^r + n$ by $n$ or $i^r$. Since $\log i$ is increasing, we see that

$$\sum_{i=1}^{\lceil n^{1/r} \rceil} \log i \leq \frac{n^{1/r} \log n}{r}.$$ 

Since $f(x) = x^{-\gamma} \log x$ is decreasing for $x \geq e^{1/\gamma}$, we see that $i^{-r} \log i$ is decreasing on interval $[\lceil n^{1/r} \rceil, \infty)$ for $n \geq e$. Therefore,

$$\sum_{i=\lceil n^{1/r} \rceil}^\infty \frac{n \log i}{i^r} \leq \frac{n \log \lceil n^{1/r} \rceil}{\lceil n^{1/r} \rceil^r} + n \int_{\lceil n^{1/r} \rceil}^\infty \frac{\log x}{x^r} \, dx.$$ 

Since $\lceil x \rceil / x \leq 2$ for $x \geq 1$, and $n^{1/r} \geq 2$,

$$n \frac{\log \lceil n^{1/r} \rceil}{\lceil n^{1/r} \rceil^r} \leq 2 \log n^{1/r} \leq \frac{n^{1/r} \log n}{r}.$$ 

Moreover

$$\int_{\lceil n^{1/r} \rceil}^\infty \frac{\log x}{x^r} \, dx \leq \int_{n^{1/r}}^\infty \frac{\log x}{x^r} \, dx = n^{1-1/r} \frac{(r-1) \log n^{1/r} + 1}{(r-1)^2}.$$ 

Since $r \geq 1 + c$, we have

$$\frac{\log n^{1/r}}{r-1} \leq \frac{1}{c} \cdot \frac{\log n}{r}, \quad \frac{1}{(r-1)^2} \leq \frac{\log n^{1/r}}{(r-1)^2 \log 2} \leq \frac{1}{c^2 \log 2} \cdot \frac{\log n}{r}.$$ 

This proves (i) for the case $r \leq (\log n) / (\log 2)$. 


We now consider \( r > (\log n)/(\log 2) \), which implies that \( n^{1/r} < 2 \). We have

\[
\sum_{i=2}^{\infty} \frac{n \log i}{i^r + n} \leq n \sum_{i=2}^{\infty} \frac{\log i}{i^r} \leq n 2^{-r} \log 2 + n \int_{2}^{\infty} x^{-r} \log x \, dx,
\]

by monotonicity of the function \( f \) defined above (with \( \gamma = r \)). We have

\[
\int_{2}^{\infty} x^{-r} \log x \, dx = 2^{1-r} \left( \frac{r-1}{r} \log 2 + 1 \right),
\]

and since \( r \geq 1 + c \)

\[
\frac{\log 2}{r-1} \leq \frac{\log 2}{c}, \quad \frac{1}{(r-1)^2} \leq \frac{1}{c^2},
\]

which finishes the proof of (ii).

To complete the proof of (i), we consider the function \( f(x) = 2^{-x} x \) and note that it is decreasing for \( x > 1/\log 2 \). Therefore, \( n 2^{-r} = (n 2^{-r}) / (r \log 2) \leq (\log n) / (r \log 2) \), for \( n \geq 3 \). Since \( 1 \leq n^{1/r} \), we get the desired result. \( \square \)

**Lemma A.10.** For any \( m > 0 \), \( l \geq 1 \), \( r_0 > 0 \), \( r \in (0, r_0) \), \( s \in (0, rl - 2) \), and \( n \geq e^{2mr_0} \)

\[
\sum_{i=1}^{\infty} \frac{i^s (\log i)^m}{(i^r + n)^l} \leq 4n \frac{1 + s - r}{r^m} \frac{(\log n)^m}{r^m}.
\]

The same upper bound holds for \( m = 0 \), \( r \in (0, \infty) \), \( s \in (0, rl) \), and \( n \geq 1 \).

**Proof.** We deal with this sum by splitting the sum in the parts \( i \leq n^{1/r} \) and \( i > n^{1/r} \). In the first range we bound the sum by

\[
\sum_{i=1}^{n^{1/r}} n^{-l} i^s (\log i)^m \leq n^{-l} n^{1-s+\frac{s}{r}} \frac{(\log n)^m}{r^m},
\]

by monotonicity of the function \( f(x) = x^s (\log x)^m \).

Suppose that \( m > 0 \). The derivative of the function \( f(x) = x^{-1/2} (\log x)^m \) is \( f'(x) = x^{-3/2} (\log x)^{m-1} (m - (\log x)/2) \), hence it is monotone decreasing for \( x \geq e^{2m} \). Since \( n^{1/r} \geq n^{1/r_0} \) and \( n > e^{2mr_0} \), the function \( f \) is decreasing on interval \([n^{1/r}, \infty)\). Therefore, we bound the sum over the second range by

\[
\sum_{i=n^{1/r}}^{\infty} i^s r \frac{(\log i)^m}{(i^r + n)^l} \leq n^{-l} n^{1-s+\frac{s}{r}} \frac{(\log n)^m}{r^m} \sum_{i=n^{1/r}}^{\infty} i^{1/2+s-r}.
\]

Since \( s \leq rl - 2 \), \( i^{1/2+s-r} \) is decreasing and \( rl - s - 3/2 \geq 1/2 \). We get

\[
\sum_{i=n^{1/r}}^{\infty} i^{1/2+s-r} \leq n^{1/2+s-r} \int_{n^{1/r}}^{\infty} x^{1/2+s-r} \, dx + \int_{n^{1/r}}^{\infty} x^{-3/2-s+rl} \, dx
\]

\[
= n^{1/2+s-r} + \frac{1}{-3/2-s+rl} n^{3/2+s-r}
\]

\[
\leq 3n^{3/2+s-r}.
\]

In the case \( m = 0 \), we use monotonicity of \( i^{s-r} \) for all \( i \geq 1 \). \( \square \)
Lemma A.11. For any $p \geq 0$, $r \in (1, (\log n)/(2 \log(3e/2))]$, and $\gamma > 0$,  

$$\sum_{i=1}^{\infty} \frac{n^\gamma \log i}{(i^r + n)^\gamma} \geq \frac{1}{3 \cdot 2^\gamma r} n^{\frac{1}{r}} \log n.$$  

PROOF. In the range $i \leq n^{1/r}$ we have $i^r + n \leq 2n$, thus  

$$\sum_{i=1}^{\infty} \frac{n^\gamma \log i}{(i^r + n)^\gamma} \geq \frac{1}{2^\gamma} \sum_{i=1}^{\lfloor n^{1/r} \rfloor} \frac{1}{2^\gamma} \int_{1}^{\lfloor n^{1/r} \rfloor} \log x \, dx \geq \frac{1}{2^\gamma} \int_{1}^{\frac{2}{3} n^{1/r}} \log x \, dx,$$

since $n^{1/r} \geq 2$ and $|x| \geq 2x/3$ for $x \geq 2$. Since $\log n \geq 2 \log(3e/2)$ implies that $(\log n)/(2r) \geq \log(3e/2)$, we have  

$$\frac{2}{3} n^{\frac{1}{r}} \left( \log \left( \frac{2}{3} n^{\frac{1}{r}} \right) - 1 \right) = \frac{2}{3} n^{\frac{1}{r}} \left( \frac{1}{r} \log n - \frac{3e}{2} \right) \geq \frac{1}{3r} n^{\frac{1}{r}} \log n.$$

This completes the proof since  

$$\frac{1}{2^\gamma} \int_{1}^{\frac{2}{3} n^{1/r}} \log x \, dx = \frac{2}{3} n^{\frac{1}{r}} \left( \log \left( \frac{2}{3} n^{\frac{1}{r}} \right) - 1 \right) + 1. \quad \Box$$

Lemma A.12. Let $m$, $i$, $r$, and $\xi$ be positive reals. Then for $n \geq e^m$  

$$\frac{ni^r(r \log i)^m}{(i^r + n)^2} \leq (\log n)^m, \quad \text{and} \quad \frac{n^\xi (r \log i)^{\xi m}}{(i^r + n)^{\xi}} \leq (\log n)^{\xi m}.$$  

PROOF. Assume first that $i \leq n^{1/r}$, then the left hand side of the first inequality is bounded above by  

$$\frac{n^2 (r \log n^{1/r})^m}{n^2} = (\log n)^m.$$  

Next assume that $i > n^{1/r}$. The derivative of the function $f(x) = x^{-c} (\log x)^m$ is $f'(x) = x^{-c-1} (\log x)^{m-1} (-c (\log x) + m)$, hence $f(x)$ is monotone decreasing for $x \geq e^{m/c}$. Therefore, the function $(i^r (\log i)^m)$ is monotone decreasing for $i \geq e^{m/r}$ and since by assumption $i > n^{1/r}$, we get that for $n \geq e^m$ the function $f(i) = i^{-r} (\log i)^m$ takes its maximum at $i = n^{1/r}$. Hence, the left hand side of the inequality is bounded above by  

$$n (r \log i)^m i^{-r} \leq nr^m (\log n^{1/r})^m n^{-1} = (\log n)^m.$$  

The second inequality can be proven similarly. \hfill \Box

Lemma A.13. For every differentiable $\eta$ and $\varepsilon > 0$ the following inequalities hold:  

$$\eta(0)\varepsilon - \varepsilon \int_{0}^{\varepsilon} |\eta'(x)| \, dx \leq \int_{0}^{\varepsilon} \eta(x) \, dx \leq \eta(0)\varepsilon + \varepsilon \int_{0}^{\varepsilon} |\eta'(x)| \, dx.$$  

PROOF. Integration by parts yields  

$$\int_{0}^{\varepsilon} \eta(x) \, dx = \eta(0)\varepsilon + \int_{0}^{\varepsilon} (\varepsilon - x)\eta'(x) \, dx.$$  

Since $-\varepsilon |\eta'(x)| \leq (\varepsilon - x)\eta'(x) \leq \varepsilon |\eta'(x)|$ for $x \in [0, \varepsilon]$, the assertion holds. \hfill \Box
Bibliography


Summary

The main goal of statistical estimation is to recover an unknown, fixed parameter of interest from noisy observations, which is achieved by an estimation procedure. In this thesis we consider the Bayesian approach to statistical inference by assigning a prior distribution to the unknown parameter. Next the corresponding posterior distribution serves as a starting point for estimation. If the parameter of interest is infinite-dimensional (nonparametric statistics), the choice of the prior is of significant importance and might dramatically influence the performance of the corresponding posterior.

In Chapter 1 we introduce important notions of Bayesian asymptotics: posterior consistency and posterior contraction, (frequentist) coverage of credible balls, and posterior limits. Even though the existing literature on Bayesian nonparametrics is large, several important aspects of the models considered in this thesis have not been studied so far. We briefly introduce nonparametric inverse problems and show why the general theory of posterior contraction cannot be applied in this setting. We also present the classical Bernstein–von Mises theorem and review the recent developments in the study of posterior limits in semi- and nonparametric statistical problems.

In Chapter 2 we first describe a nonparametric inverse problem in the context of the canonical signal-in-white noise model with the operator acting between two Hilbert spaces, and show its equivalence to the infinite-dimensional normal mean model. The main contribution of this chapter is the study of the asymptotic properties of the posterior in two settings of inverse problems: mildly and extremely ill-posed. The former setting covers, among others, estimation of a derivative of a function, and the latter is presented by a study of the recovery of the initial condition for the heat equation. We consider a certain family of Gaussian prior distributions and show that the rate of contraction depends on the parameters of the prior, characteristics of the inverse problem, and the regularity of the true parameter of interest. These results are compared with the existing frequentist approaches to nonparametric inverse problems. We also discuss frequentist properties of Bayesian credible balls. The results on contraction and credibility are illustrated by simulation examples in both inverse problem settings.

In Chapter 3 we present the first theoretical study of adaptive Bayesian procedures for nonparametric inverse problems. Again, as in Chapter 2, we consider a certain family of Gaussian priors for the parameter of interest. These priors are indexed by a parameter $\alpha$ quantifying the “regularity” of the prior. In Chapter 2 we considered this parameter fixed, and in this chapter we select $\alpha$ using the data. A first approach is fully Bayesian: we endow the parameter $\alpha$ with a prior distribution itself. A second approach we study mixes the Bayesian and the frequentist paradigm: we first “estimate” $\alpha$ from the data in a frequentist manner, and then substitute the estimator $\hat{\alpha}_n$ for $\alpha$ in the posterior distribution obtained in Chapter 2. We show that both methods lead to adaptation and rate-optimality (up to lower order factors) over two families of submodels containing the true parameter of interest, and describing its regularity. We illustrate both methods by the simulation example introduced in Chapter 2 in the mildly ill-posed inverse problem setting.
In Chapter 4 we consider a semiparametric aspect of inverse problems: recovery of linear functionals of the parameter of interest. We consider not only continuous, but also certain discontinuous functionals, belonging to a wider class of prior-measurable linear functionals. The contribution of this chapter is similar to the one of Chapter 2: we study posterior contraction that is not covered by the existing literature on the subject, and we investigate the frequentist coverage of Bayesian credible intervals. The regularity of the linear functional plays an important role in the asymptotic behavior of Bayesian procedures. We show that certain continuous linear functionals cancel the inverse nature of the problem, and put the problem in the regular regime. In this chapter we obtain a semiparametric Bernstein–von Mises theorem, not only with a typical $n^{-1/2}$ rate, but also with a rate slowed down by a slowly varying factor. The results of this chapter are illustrated by the same simulation examples as in Chapter 2.

In Chapter 5 we first present a simple irregular model, consisting of shifted exponential distributions with scale 1, and consider the resulting posterior limit. Next we introduce local asymptotic exponentiality (LAE), an irregular expansion of the likelihood, presented in the semiparametric setting in which we decompose the parameter as a pair $(\theta, \eta)$, where the parameter of interest $\theta$ lies in an open subset of the real line, and the nuisance parameter $\eta$ is an element of an infinite-dimensional space. We next present the main theorem of the chapter, an irregular version of a semiparametric Bernstein–von Mises theorem. A separate section of the chapter is dedicated to the most demanding condition of the main theorem, namely marginal consistency at $n^{-1}$ rate. Some discussion is provided, followed by a lemma verifying the condition based on a condition on the likelihood ratio. We end the chapter by presenting two semiparametric models exhibiting the LAE property. Both problems are related to the problem of estimation of the boundary point of a distribution. The first is a generalization of the shifted exponential model. The other one generalizes the uniform distribution on the interval $[0, \theta]$. In both settings, based on an i.i.d. sample distributed according to an unknown, but fixed element $(\theta_0, \eta_0)$, we obtain exponential limits for the marginal posterior distributions.
Samenvatting

BAYESIAANSE ASYMPTOTIEK
INVERSE PROBLEMEN EN IRREGULIERE MODELLEN

In de schattingstheorie zoekt men procedures om een vaste, onbekende parameter terug te vinden uit een aantal verstoorde observaties. In dit proefschrift beschouwen we de Bayesiaanse aanpak van statistiek en kennen we een a priori verdeling toe aan de onbekende parameter. Vervolgens dient de corresponderende a posteriori verdeling als beginpunt voor onze schattingen. Als de parameter waarin we geïnteresseerd zijn oneindigdimensionaal is (niet-parametrische statistiek), is de keuze van een a priori verdeling zeer belangrijk en kan deze dramatisch grote invloed uitoefenen op de corresponderende a posteriori verdeling.

In hoofdstuk 1 introduceren we de belangrijkste begrippen binnen de Bayesiaanse asymptotiek: consistentie en contractie van de a posteriori verdeling, (frequentistische) coverage van credible balls, en a posteriori limieten. Hoewel uitgebreide literatuur voorhanden is over niet-parametrische Bayesiaanse statistiek, is een aantal belangrijke aspecten van de modellen die we in dit proefschrift bekijken nog niet eerder bestudeerd. We geven een korte introductie in niet-parametrische inverse problemen en we laten zien waarom de algemene theorie van a posteriori contractie in deze situatie niet van toepassing is. Verder behandelen we de klassieke Bernstein–von Mises stelling en de recente ontwikkelingen in de leer van a posteriori limieten in semi- en niet-parametrische statistische problemen.

In hoofdstuk 2 bekijken we eerst een niet-parametrisch invers probleem in de context van het canonieke signaal-in-ruis model met een operator tussen twee Hilbertruimtes en tonen we aan dat dit equivalent is met het oneindig dimensionele normale mean model. Het belangrijkste onderdeel van dit hoofdstuk is de bestudering van de asymptotische eigenschappen van de a posteriori verdeling voor twee verschillende soorten inverse problemen: mild en extreem ill-posed. De eerste situatie omvat onder andere het schatten van de afgeleide van een functie en de tweede situatie wordt uitgelegd aan de hand van het probleem van het terugvinden van de beginvoorwaarden van de hittevergelijking. We bekijken een zekere familie van Gaussische a priori verdelingen en tonen aan dat de snelheid van contractie afhangt van de parameters van de verdeling, de eigenschappen van het inverse probleem en de regulariteit van de echte parameterwaarde waarin we geïnteresseerd zijn. Deze resultaten worden vergeleken met de bestaande frequentistische aanpak bij niet-parametrische inverse problemen. We bespreken ook de frequentistische eigenschappen van Bayesiaanse credible balls. De resultaten over contractie en credibility worden geïllustreerd aan de hand van simulaties voor beide types inverse problemen.

In hoofdstuk 3 presenteren we het eerste theoretische onderzoek naar adaptieve Bayesiaanse procedures voor niet-parametrische inverse problemen. Net als in hoofdstuk 2 bekijken we een zekere familie van Gaussische a priori verdelingen voor de parameter. Deze verdelingen worden geïndexeerd door een parameter $\alpha$ die de “regulatiteit” van de verdeling kwantificeert. In hoofdstuk 2 fixeren we
Samenvatting

dezelfde parameter en in dit hoofdstuk selecteren we \( \alpha \) met behulp van de data. Een eerste aanpak is geheel Bayesiaans: we nemen voor de parameter \( \alpha \) wederom een a priori verdeling. Een tweede aanpak mengt de Bayesiaanse en de frequentistische methoden: we “schatten” eerst \( \alpha \) uit de data op een frequentistische manier en substitueren dan de verkregen schatter \( \hat{\alpha}_n \) voor \( \alpha \) in de a posteriori verdeling verkregen in hoofdstuk 2. We tonen aan dat beide methoden leiden tot adaptatie en een optimale snelheid (tot op lagere orde factoren) over twee families van submodellen die de echte parameter bevatten en die zijn regulariteit beschrijven. We illustreer beide methoden aan de hand van de simulaties geïntroduceerd in hoofdstuk 2 in de situatie van het mild ill-posed inverse probleem.

In hoofdstuk 4 bekijken we een semiparametrisch aspect van inverse problemen: het terugvinden van lineaire functionalen toegepast op de parameter. We bekijken niet alleen continue, maar ook bepaalde discontinue functionalen, die deel uitmaken van een grotere klasse van lineaire functionalen die meetbaar zijn ten opzichte van de a priori verdeling. De bijdrage van dit hoofdstuk is vergelijkbaar met die van hoofdstuk 2: we bestuderen contractie van de a posteriori verdeling die niet in bestaande literatuur over dit onderwerp is onderzocht en we onderzoeken de frequentistische coverage van Bayesiaanse credible intervals. De regulariteit van de lineaire functional speelt een belangrijke rol in het asymptotische gedrag van Bayesiaanse procedures. We tonen aan dat bepaalde continue lineaire functionalen het inverse karakter van het probleem opheffen en het probleem in een bekend perspectief plaatsen. In dit hoofdstuk verkrijgen we de semiparametrische Bernstein-von Mises stelling, niet alleen met de typische snelheid van \( n^{-1/2} \), maar ook met een snelheid die is afgenomen met een traag varierende factor. De resultaten van dit hoofdstuk worden geïllustreerd aan de hand van dezelfde simulatie als in hoofdstuk 2.

In hoofdstuk 5 presenteren we eerst een simpel irregulier model bestaande uit verschoven exponentiële verdelingen met schaal 1 en bekijken de resulterende a posteriori limiet. Vervolgens introduceren we *lokale asymptotische exponentialiteit (LAE)*, een irreguliere uitbreiding van de likelihood, die we beschouwen in de niet-parametrische setting waarin we de parameter ontbinden in een paar \((\theta, \eta)\), waarbij de parameter \( \theta \) waarin we geïnteresseerd zijn in een open verzameling van de reële lijn ligt en de hinderlijke parameter \( \eta \) een element is van een oneindig-dimensionale ruimte. Vervolgens presenteren we de hoofdstelling van dit hoofdstuk, een irreguliere versie van een semiparametrische Bernstein-von Mises stelling. Een aparte sectie wordt gewijd aan de meest veeleisende conditie van de hoofdstelling, marginale consistentie met snelheid \( n^{-1} \). Dit wordt besproken, gevolgd door een lemma dat de conditie verifieert op basis van een voorwaarde van de likelihood ratio. We sluiten het hoofdstuk af met twee voorbeelden van semiparametrische modellen met de LAE eigenschap. Beide problemen zijn verwant aan het schatten van het randpunt van een verdeling. Het eerste is een generalisatie van het verschoven exponentiële model en het tweede van de uniforme verdeling op het interval \([0, \theta]\). In beide situaties, die gebaseerd zijn op s.o. en identiek verdeelde waarnemingen met onbekende, maar vaste, onderliggende parameter \((\theta_0, \eta_0)\), verkrijgen we exponentiële limieten voor de marginale a posteriori verdelingen.