Bayesian inverse problems with unknown operators

Mathias Trabs

Universität Hamburg, Hamburg, Germany

E-mail: mathias.trabs@uni-hamburg.de

Received 30 January 2018, revised 6 April 2018
Accepted for publication 10 May 2018
Published 5 June 2018

Abstract
We consider the Bayesian approach to linear inverse problems when the underlying operator depends on an unknown parameter. Allowing for finite dimensional as well as infinite dimensional parameters, the theory covers several models with different levels of uncertainty in the operator. Using product priors, we prove contraction rates for the posterior distribution which coincide with the optimal convergence rates up to logarithmic factors. In order to adapt to the unknown smoothness, an empirical Bayes procedure is constructed based on Lepski’s method. The procedure is illustrated in numerical examples.

Keywords: rate of contraction, posterior distribution, product priors, ill-posed linear inverse problems, empirical Bayes, non-parametric estimation

(Some figures may appear in colour only in the online journal)

1. Introduction
Bayesian procedures to solve inverse problems became increasingly popular in the last few years, see Stuart [31]. In the inverse problem literature the underlying operator of the forward problem is typically assumed to be known. In practice, there might however be some uncertainty in the operator which has to be taken into account by the procedure. While there are some frequentist approaches in the statistical literature to solve inverse problems with an unknown operator, the Bayesian point of view has not yet been analysed. The aim of this work is to fill this gap.

Let \( f \in L^2(D) \) be a function on a domain \( D \subseteq \mathbb{R}^d \) and \( K_\theta : L^2(D) \to L^2(Q), Q \subseteq \mathbb{R}^q \), be an injective, continuous linear operator depending on some parameter \( \theta \in \Theta \). We consider the linear inverse problem

\[
Y = K_\theta f + \varepsilon Z.
\]
where $Z$ Gaussian white noise in $L^2(Q)$ and $\varepsilon > 0$ is the noise level which converges to zero asymptotically. If the operator $K_\theta$ is known, the inverse problem to recover $f$ non-parametrically, i.e. as element of the infinite dimensional space $L^2(D)$, from the observation $Y$ is well studied, see for instance Cavalier [5]. The Bayesian approach has been analysed by Knapik et al [22] with Gaussian priors, by Ray [30] non-conjugate priors and many subsequent articles including [1, 2, 21, 23]. Also non-linear inverse problems have been successfully solved via Bayesian methods, for example, [3, 9, 27–29, 35].

Focussing on linear inverse problems, we will extend the Bayesian methodology to unknown operators. To this end, the unknown parameter $\theta \in \Theta$ is introduced in (1.1) where $K_\theta$ may depend non-linearly on $\theta$. Unknown operators are relevant in numerous applications. Examples include semi-blind and blind deconvolution for image analysis. Therein, the operator is given by $K_{\theta} f = \int g_{\theta}(\cdot - y)f(y)dy$ with some unknown convolution kernel $g_{\theta}$ [4, 20, 32]. More general integral operators such as singular layer potential operators appear in the context of partial differential equations, see examples in [7, 15]. If the coefficients of the underlying PDE are unknown, then the operator itself is only partially known. A typical example of this type is the backwards heat equation where the solution $u$ of the PDE $\frac{\partial}{\partial t} u = \theta \Delta u$ (with Dirichlet boundary conditions) is observed at some time $t$ and the aim is to estimate the initial value function $f = u(0, \cdot)$. Here, we take into account an unknown diffusivity parameter $\theta > 0$. The solution $u(t, \cdot)$ depends linearly on $f$ and the resulting operator admits a singular value decomposition (SVD) with respect to the sine basis and with $\theta$ dependent singular values $\rho_{\theta, k} = e^{-\theta \pi^2 k^2 t}$, $k \geq 1$, see section 6. In particular, the resulting inverse problem is severely ill-posed.

Even without measurement errors the target function $f$ is in general not identifiable any more for unknown operators, i.e. there may be several solutions $(\theta, f)$ to the equation $Y = K_{\theta} f$. For instance, if $K_{\theta}$ admits a SVD $K_{\theta} \varphi_k = \rho_k \psi_k$ for an orthonormal systems $(\varphi_k)_{k \geq 1}$, $(\psi_k)_{k \geq 1}$ and unknown singular values $\theta = (\rho_k)_{k \geq 1}$, then we have $K_{\theta} f = K_{\theta/ad} (a f)$ for any function $f \in L^2(D)$ and any scalar $a > 0$. We thus require some extra information. There are different approaches in the inverse problem literature to deal with this identifiability problem, particularly in the context of semi-blind or blind deconvolution. One approach is to find the so called minimum norm solution which has a minimal distance to some $a$ priori estimates for $\theta$ and $f$, see [4, 20]. Another idea is to assume that some approximation of the unknown operator is available for the reconstruction of $f$, see [18, 32]. Similarly, we may assume to have some noisy observation of the unknown parameter $\theta$ which then allows to construct an estimator for $K_{\theta}$.

In this paper we will study this last setting. More precisely, we suppose that the parameter set $\Theta$ (a subset of) some Hilbert space and we consider the additional sample

$$T = \theta + \delta W$$

(1.2)

where $W$ is white noise on $\Theta$, independent of $Z$, and $\delta > 0$ is some noise level. Thereby, $\theta$ is considered as a nuisance parameter and we will not impose any regularity assumptions on $\theta$. Our aim is the estimation of $f$ from the observations (1.1) and (1.2). This setting includes several models with different levels of uncertainty in the operator $K_{\theta}$:

A If $\Theta \subseteq \mathbb{R}^p$, we have a parametric characterization of the operator $K_{\theta}$ and $T$ can be understood as an independent estimator for $\theta$.

B Cavalier and Hengartner [6] have studied the case where the eigenfunctions of $K_{\theta}$ are known, but only noisy observations of the singular values $(\rho_k)_{k \geq 1}$ are observed: $T_k = \rho_k + \delta W_k$, $k \geq 1$, with i.i.d. standard normal $(W_k)_{k}$. In this case $\Theta = \ell^2$, supposing $K_{\theta}$ is Hilbert–Schmidt, and $\theta = (\rho_k)_{k}$ is the sequences of singular values of $K_{\theta}$.
C Efroymovich and Kol’tchinskii [11], Hoffmann and Reiß [16] as well as Marteau [25] have assumed the operator as completely unknown and considered additional observations of the form

\[ L = K + \delta W \]

where the operator \( L \) is blurred by some independent white noise \( W \) on the space of linear operators from \( L^2(D) \) to \( L^2(Q) \) with some noise level \( \delta \). Fixing basis \((e_k)\) and \((h_l)\) of \( L^2(D) \) and \( L^2(Q) \), respectively, \( K \) is characterised by the infinite matrix \( \vartheta = ((K_{e_k}, h_l))_{k,l \geq 1} \in \mathbb{R}^{K \times L} \) and \( W \) can be identified with the random matrix \((\langle W_{e_k}, g_l \rangle)_{k,l \geq 1}\) consisting of i.i.d. standard Gaussian entries.

In contrast to the just mentioned articles [6, 11, 16, 25], we will investigate the Bayesian approach. We thus put a prior distribution \( \Pi \) on \((f, \vartheta) \in L^2(D) \times \Theta\). Denoting the probability density of \((Y, T)\) under the parameters \((f, \vartheta)\) with respect to some reference measure by \( p_{f,\vartheta} \), the posterior distribution given the observations \((Y, T)\) is given by Bayes’ theorem:

\[ \Pi(B|Y, T) = \frac{\int_B p_{f,\vartheta}(Y, T) d\Pi(f, \vartheta)}{\int_{L^2(D) \times \Theta} p_{f,\vartheta}(Y, T) d\Pi(f, \vartheta)} \quad (1.3) \]

for measurable subsets \( B \subseteq L^2(D) \times \Theta \). Due to the white noise model, the density \( p_{f,\vartheta} \) inherits the nice structure from the normal distribution, see section 2. Although we cannot hope for nice conjugate pairs of prior and posterior distribution due to the non-linear structure of \((f, \vartheta) \mapsto K_{\vartheta} f\), there are efficient Markov chain Monte Carlo algorithms to draw from \( \Pi(f|Y, T) \), see [34].

To analyse the behaviour of the posterior distribution, we will take a frequentist point of view and assume the observations are generated under some true, but unknown \( f_0 \in L^2(D) \) and \( \vartheta_0 \in \Theta \). In a first step we will identify general conditions on a prior \( \Pi \) under which the posterior \( \Pi(f \in \cdot|Y, T) \) for \( f \) concentrates in a neighbourhood of \( f_0 \) with a certain rate of contraction \( \xi_{\varepsilon, \delta} \). We show for some constant \( D > 0 \) the convergence

\[ \Pi(f) \in L^2(D) : \|f - f_0\| > D\xi_{\varepsilon, \delta} \rightarrow Y, T \rightarrow 0 \quad (1.4) \]

in \( \mathbb{P}_{f_0, \vartheta_0} \)-probability as \( \varepsilon \) and \( \delta \) go to zero. This contraction result verifies that whole probability mass the posterior distribution is asymptotically located in a small ball around \( f_0 \) with radius of order \( \xi_{\varepsilon, \delta} \downarrow 0 \). Hence, draws from the posterior distribution will be close to the unknown function \( f_0 \) with high probability. This especially implies that the posterior mean and the posterior median are consistent estimators of the unknown function \( f_0 \). Interestingly, the difficulty to recover \( f \) from \((Y, T)\) is same in all three above mentioned models.

The proof of the contraction result follows general principles developed by Ghosal et al [12]. The analysis of the posterior distribution requires to control both, the numerator in (1.3) and the normalising constant. To find a lower bound for the latter, a so-called small ball probability condition is imposed ensuring that the prior puts some minimal weight in a neighbourhood of the truth. Given this bound, the contraction theorem can be shown by constructing sufficiently powerful tests for the hypothesis \( H_0 : f = f_0 \) against the alternative \( H_1 : \|f - f_0\| > D\xi_{\varepsilon, \delta} \) for the constant \( D > 0 \) from (1.4). To find the test, we follow Giné and Nickl [13] and use a plug-in test based on a frequentist estimator. This estimator obtained by the Galerkin projection method, as proposed in [11, 16] for the Model C.

The main difficulty is that without structural assumptions on \( \Theta \), e.g. if \( \Theta = \ell_2 \), an infinite dimensional nuisance parameter \( \vartheta \) cannot be consistently estimated. We thus cannot expect a concentration of \( \Pi(\vartheta \in \cdot|Y, T) \). Why should then \( \Pi(K_{\vartheta} f \in \cdot|Y, T) \) concentrate around
the truth? Fortunately, \( K_{\delta_0} f_0 \) is regular, such that a finite dimensional projection suffices to reconstruct \( f_0 \) with high accuracy. Under the reasonable assumption that the projection of \( K_{\delta_0} \) depends only on a finite dimensional projection \( P \delta_0 \) of \( \delta_0 \), we can indeed estimate \( f_0 \) without estimating the full \( \delta_0 \). Similarly, we show in the Bayesian setting that a concentration of this finite dimensional projection \( P \delta \) is sufficient resulting in a small ball probability condition depending only on \( f \) and \( P \delta \).

The conditions of the general result are verified in the mildly ill-posed case and in the severely ill-posed case, assuming some Sobolev regularity of \( f_0 \). We use a truncated product prior of \( f \) and a product prior on \( \delta \). Choosing the truncation level \( J \) of \( \delta \) in an optimal way, the resulting contraction rates coincide with the minimax optimal rates which are known in several models up to logarithmic factors. These rates are indeed the same as for the known parameter case, see Ray [30], if \( \delta = O(\varepsilon) \).

Since the optimal level \( J \) depends on the unknown regularity \( s \) of \( f_0 \), a data-driven procedure to select \( J \) is desirable. There are basically two ways to tackle this problem. Setting a hyper prior on \( s \), a hierarchical Bayes procedure could be considered. Alternatively, although not purely Bayesian, we can try to select some \( J \) empirically using Lepski’s method [24] which yields an easy to implement procedure (note that [21] used a maximum likelihood approach to estimate \( s \)). We prove that the final adaptive procedure attains the same rate as the optimized non-adaptive method.

This paper is organised as follows: the posterior distribution is derived in section 2. The general contraction theorem is presented in section 3. In section 4 specific rates for Sobolev regular functions \( f \) in the mildly and the severely case are determined using a truncated product prior. An adaptive choice of the truncation level is constructed in section 5. In section 6 we discuss the implementation of the Bayes method using a Markov chain Monte Carlo algorithm and illustrate the method in two numerical examples. All proofs are postponed to section 7.

2. Setting and posterior distribution

Let us fix some notation: \( \langle \cdot, \cdot \rangle \) and \( \| \cdot \| \) denote the scalar product and the norm of \( L^2(\Omega) \) or \( \Theta \). We write \( x \lesssim y \) if there is some universal constant \( C > 0 \) such that \( x \leq Cy \). If \( x \lesssim y \) and \( y \lesssim x \) we write \( x \asymp y \). We recall that noise process \( Z \) in (1.1) is the standard iso-normal process, i.e. \( \langle g, Z \rangle \) is \( \mathcal{N}(0, \| g \|^2) \)-distributed for any \( g \in L^2(\Omega) \) and covariances are given by

\[
\mathbb{E}[\langle Z, g_1 \rangle \langle Z, g_2 \rangle] = \langle g_1, g_2 \rangle \quad \text{for all } g_1, g_2 \in L^2(\Omega).
\]

We write \( Z \sim \mathcal{N}(0, \text{Id}) \). Note that \( Z \) cannot be realised as an element of \( L^2(\Omega) \), but only as a Gaussian process \( g \mapsto \langle g, Z \rangle \).

The observation scheme (1.1) is equivalent to observing

\[
\langle Y, g \rangle = \langle K_{\delta_0} f, g \rangle + \varepsilon \langle Z, g \rangle \quad \text{for all } g \in L^2(\Omega).
\]

Choosing an orthonormal basis \( \{ \varphi_k \}_{k \geq 1} \) of \( L^2(\Omega) \) with respect to the standard \( L^2 \)-scalar product, we obtain the series representation

\[
Y_k := \langle Y, \varphi_k \rangle = \langle K_{\delta_0} f, \varphi_k \rangle + \varepsilon Z_k
\]
for i.i.d. random variables \( Z_k \sim \mathcal{N}(0, 1), k \geq 1 \). Note that the distribution of \( (Z_k) \) does not depend on \( \vartheta \). If \( K_\vartheta \) is compact, it might be tempting to choose \( (\varphi_k) \) from the singular value decomposition of \( K_\vartheta \) simplifying \( K_{\vartheta f} \varphi_k \). However, such a basis of eigenfunctions will in general depend on the unknown \( \vartheta \) and thus cannot be used. Since \( (Z_k)_{k \geq 1} \) are i.i.d., the distribution of the vector \( (Y_k)_{k \geq 1} \) is given by

\[
\mathbb{P}^{Y}_{\vartheta f} = \bigotimes_{k \geq 1} \mathcal{N}((K_{\vartheta f}, \varphi_k), \varepsilon^2).
\]

By Kakutani’s theorem, see Da Prato [8, theorem 2.7], \( \mathbb{P}^{Y}_{\vartheta f} \) is equivalent to the law \( \mathbb{P}^{\varepsilon}_{\vartheta} = \bigotimes_{k \geq 1} \mathcal{N}(0, \varepsilon^2) \) of the white noise \( \varepsilon Z \). Writing \( (K_{\vartheta f}, Z) := \sum_{k \geq 1} (K_{\vartheta f}, \varphi_k) Z_k \) with some abuse of notation, since \( Z \) is not in \( L^2(\mathcal{Q}) \), we obtain the density

\[
\frac{\text{d}\mathbb{P}^{\varepsilon}_{\vartheta f}}{\text{d}\mathbb{P}^{\varepsilon}_{\vartheta}} = \exp \left( \frac{1}{\varepsilon^2} \langle K_{\vartheta f}, Z \rangle - \frac{1}{2\varepsilon^2} \sum_{k \geq 1} |(K_{\vartheta f}, \varphi_k)|^2 \right) = \exp \left( \frac{1}{\varepsilon^2} \langle K_{\vartheta f}, Y \rangle - \frac{1}{2\varepsilon^2} \|K_{\vartheta f}\|^2 \right),
\]

where we have used \( Y_k = \varepsilon Z_k \) under \( \mathbb{P}^{\varepsilon}_{\vartheta} \) for the second equality.

Since any continuous operator \( K_\vartheta \) can be described by the infinite matrix \( (\langle K_\vartheta \varphi_j, \varphi_k \rangle)_{j,k \geq 1} \), we may assume without loss of generality that \( \Theta \subseteq \ell^2 \). The distribution of \( T \) is then similarly given by \( \mathbb{P}^{\vartheta} = \bigotimes_{k \geq 1} \mathcal{N}(\delta_{\vartheta_k}, \delta^2) \) being equivalent to \( \mathbb{P}^{\vartheta} = \bigotimes_{k \geq 1} \mathcal{N}(0, \delta^2) \). Writing \( T = (T_k)_{k \geq 1} \) and \( (\vartheta, T) = \sum_{k \geq 1} \vartheta_k T_k \), we obtain the density

\[
\frac{\text{d}\mathbb{P}^{T}_{\vartheta}}{\text{d}\mathbb{P}^{\varepsilon}_{\vartheta}} = \exp \left( \frac{1}{\delta^2} \langle \vartheta, T \rangle - \frac{1}{2\delta^2} \|\vartheta\|^2 \right).
\]

Therefore, the likelihood of the observations \( (Y, T) \) with respect to \( \mathbb{P}^{\varepsilon}_{\vartheta} \otimes \mathbb{P}^{\varepsilon}_{\vartheta} \) is given by

\[
\frac{\text{d}\mathbb{P}^{Y}_{\vartheta f} \otimes \mathbb{P}^{T}_{\vartheta}}{\text{d}\mathbb{P}^{\varepsilon}_{\vartheta} \otimes \mathbb{P}^{\varepsilon}_{\vartheta}} = \exp \left( \frac{1}{\varepsilon^2} \langle K_{\vartheta f}, Y \rangle - \frac{1}{2\varepsilon^2} \|K_{\vartheta f}\|^2 + \frac{1}{\delta^2} \langle \vartheta, T \rangle - \frac{1}{2\delta^2} \|\vartheta\|^2 \right).
\]

(2.1)

Applying a prior \( \Pi \) on the parameter \( (f, \vartheta) \in L^2(\mathcal{D}) \times \Theta \), we obtain the posterior distribution

\[
\Pi(B|Y, T) = \frac{\int_{\mathcal{D} \times \Theta} e^{-(\langle f, Y \rangle - (\vartheta, T)\|\vartheta\|^2)} \|K_{\vartheta f}\|^2} \Pi(f, \vartheta)}{\int_{\mathcal{D} \times \Theta} e^{-(\langle f, Y \rangle - (\vartheta, T)\|\vartheta\|^2)} \|K_{\vartheta f}\|^2} \Pi(f, \vartheta) \text{d}\Pi(f, \vartheta), \quad B \in \mathcal{B},
\]

(2.2)

with the Borel-\( \sigma \)-algebra \( \mathcal{B} \) on \( L^2(\mathcal{Q}) \times \Theta \). Under the frequentist assumption that \( Y \) and \( T \) are generated under some \( f_0 \) and \( \vartheta_0 \), we obtain the representation

\[
\Pi(B|Y, T) = \frac{\int_{\mathcal{D} \times \Theta} \varphi_f(z, w) \Pi(f, \vartheta)}{\int_{\mathcal{D} \times \Theta} \varphi_f(z, w) \Pi(f, \vartheta) \text{d}\Pi(f, \vartheta)}, \quad B \in \mathcal{B},
\]

(2.3)

for

\[
\varphi_f(z, w) := \exp \left( \frac{1}{\varepsilon^2} \langle K_{\vartheta f} - K_{\vartheta f_0}, z \rangle - \frac{1}{2\varepsilon^2} \|K_{\vartheta f} - K_{\vartheta f_0}\|^2 + \frac{1}{\delta^2} \langle \vartheta - \vartheta_0, w \rangle - \frac{1}{2\delta^2} \|\vartheta - \vartheta_0\|^2 \right)
\]

corresponding to the density of \( \mathbb{P}^{Y}_{\vartheta f} \otimes \mathbb{P}^{T}_{\vartheta} \) with respect to \( \mathbb{P}^{Y}_{\vartheta f_0} \otimes \mathbb{P}^{T}_{\vartheta_0} \).

Note that even if a Gaussian prior is chosen, the posterior distribution is in general not Gaussian, since \( \vartheta \mapsto K_\vartheta \) might be non-linear. Hence, the posterior distribution cannot be explicitly calculated in most cases, but has to be approximated by an MCMC algorithm, see for instance Tierney [34] and section 6.
3. Contraction rates

For simplicity we throughout suppose $\mathcal{D} = \mathcal{Q}$ such that $L^2 := L^2(\mathcal{D}) = L^2(\mathcal{Q})$ and assume $K_0$ to be self-adjoint. The general case is discussed in remark 7.

Taking a frequentist point of view, we assume that the observations (1.1) and (1.2) are generated by some fixed unknown $f_0 \in L^2$ and $\vartheta_0 \in \Theta$. As a first main result the following theorem gives general conditions on the prior which ensure a contraction rate for the posterior distribution (2.3) around the true $f_0$.

Let $(\varphi_{j,l}) : j \in \mathcal{I}, l \in \mathcal{Z}_l)$ be an orthonormal basis of $L^2$. We use here the double-index notation which is especially common for wavelet bases, but also the single-indexed notation is included if $\mathcal{Z}_l$ contains only one element. For any index $k = (j, l)$ we write $|k| := j$. Let moreover $V_j = \text{span}\{\varphi_k : |k| \leq j\}$ be a sequence of approximation spaces with dimensions $d_j \in \mathbb{N}$ associated to $(\varphi_k)$. We impose the following compatibility assumption on $(\varphi_k)$:

**Assumption 1.** There is some $m \in \mathbb{N}$ such that $(K_0 \varphi_i, \varphi_k) = 0$ for any $\vartheta \in \Theta$ if $|l| - |k| > m$.

If $K_0$ is compact and admits an orthonormal basis of eigenfunction $(e_k)_{k \geq 1}$ being independent of $\vartheta$, then this assumption is trivially satisfied for $(\varphi_k) = (e_k)$ and $m = 0$. On the other hand this assumption allows for more flexibility for the considered approximation spaces and can be compared to Condition 1 by [30]. As a typical example, the possibly $\vartheta$ depended eigenfunctions $(e_{\vartheta,k})$ of $K_0$ may be the trigonometric basis of $L^2$ while $V_j$ are generated by band-limited wavelets.

Having $(\varphi_k)$ and thus $V_j$ fixed, we write $\|A\|_{V_j \to V_j} := \sup_{v \in V_j, \|v\| = 1} \|Av\|$ for the operator norm for any bounded linear operator $A : V_j \to V_j$ where $V_j$ is equipped with the $L^2$-norm. We denote by $P_j$ the orthogonal projection of $L^2$ onto $V_j$ and define the operator

$$K_{\vartheta,j} := P_j K_0 |_{V_j}$$

as restriction of $K_0$ to an operator from $V_j$ to $V_j$. Note that $K_{\vartheta,j}$ is given by the finite dimensional matrix $((K_{\vartheta,j} \varphi_i, \varphi_l))_{|i|, |l| \leq j} \in \mathbb{R}^{d_j \times d_j}$.

**Assumption 2.** Let $K_{\vartheta,j}$ depend only on a finite dimensional projection $P_j \vartheta := (\vartheta_1, \ldots, \vartheta_{l_j})$ of $\vartheta \in \Theta$ for some integer $1 \leq l_j \leq d_j^2$, $j \in \mathcal{I}$. Moreover, let $K_{\vartheta,j}$ be Lipschitz continuous with respect to $\vartheta$ in the following sense:

$$\|K_{\vartheta,j} - K_{\vartheta',j}\|_{V_j \to V_j} \leq L \|P_j (\vartheta - \vartheta')\|_j$$

for all $\vartheta, \vartheta' \in \Theta$.

where $L > 0$ is a constant being independent of $j, \vartheta, \vartheta'$ and where $\|\cdot\|_j$ is a norm on $P_j \Theta$. We suppose that the norm $\|\cdot\|_j$ satisfies $\mathbb{P}_\vartheta(\|P_j W\|_j > C(\kappa + \sqrt{j})) \leq \exp(-c \kappa^2)$.

Although projections on $L^2$ and on $\Theta$ are both denoted by $P_j$, it will be always clear from the context which is used such that this abuse of notation is quite convenient. Since $K_{\vartheta,j}$ is fully described by a $d_j \times d_j$ matrix, we naturally have the upper bound $l_j \leq d_j^2$. Let us illustrate the previous assumptions in the models A,B and C from the introduction:

**Examples 3.**

1. In Model A we have a finite dimensional parameter space $\Theta \subseteq \mathbb{R}^p$ with fixed $p \in \mathbb{N}$. Assumption 1 is, for instance, satisfied if $K_0 f = g_0 * f$ is a convolution operator with a kernel $g_0$ whose Fourier transform has compact support and if we choose a band-limited wavelet basis. Note that in this case we do not have to know the SVD of $K_0$. For assumption 2 we may choose $P_j = I_d$ and $\|\cdot\|_j = |\cdot|$ as the Euclidean distance on $\mathbb{R}^p$ leading to the
Lipschitz condition \( \|K_{\theta} - K_{\theta'}\|_{V_j \to V_j} \leq L|\theta - \theta'| \). Then, \( P_\theta(\|W\| > \sqrt{RN}) \leq 2pe^{-\kappa^2/2} \) follows from the Gaussian concentration of \( W \).

2. In Model B let \( K_\theta \) be compact and let \((e_i)_{i \geq 1}\) be an orthonormal basis consisting of eigenfunctions with corresponding eigenvectors \((\rho_{\theta,i})_{i \geq 1}\) and let \((\varphi_k)\) be a wavelet basis fulfilling \( d_j \simeq 2^j \). Then assumption 1 is satisfied if there is some \( C > 0 \) such that \( \langle e_i, \varphi_k \rangle \neq 0 \) only if \( C^{-1}2^{dk} \leq i \leq C2^{dk} \). Since then \( \langle e_k, v \rangle = 0 \) for any \( v \in V_j \) if \( k \geq C2^d \), we moreover have for any \( v \in V_j \)

\[
\| (K_{\theta} - K_{\theta'})v \|^2 = \| P_j \sum_{i \geq 1} (\rho_{\theta,i} - \rho_{\theta',i})(e_i, v)e_i \|^2 \\
\leq \sup_{i \leq 2^d} |\rho_{\theta,i} - \rho_{\theta',i}|^2 \sum_{i \leq 2^d} (e_i, v)^2 \leq \sup_{i \leq 2^d} |\rho_{\theta,i} - \rho_{\theta',i}|^2 \| v \|^2.
\]

We thus choose \( l_j \simeq d_j \) and \( \| \cdot \| \) as the supremum norm on \( P_j \Theta \). Since \( W_k \) are i.i.d. Gaussian, we have for some \( c > 0 \)

\[
P\left( \sup_{k \leq 2^d} |W_k| > \kappa + \sqrt{c \log d_j} \right) \leq 2C2^d e^{-\kappa^2/2} \leq 2Ce^{-\kappa^2/2}.
\]

Therefore, assumption 2 is satisfied.

3. In Model C the projected operators \( K_{\theta,j} \) are given by \( \mathbb{R}^{d \times d} \) matrices. Assumption 1 is satisfied if and only if all \( K_{\theta,j} \) are band matrices with some fixed bandwidth \( n \) independent from \( f \) and \( \theta \). To verify assumption 2, \( \| \cdot \| \) can be chosen as the operator norm or spectral norm of these matrices. The Lipschitz condition is then obviously satisfied. Moreover \( P_j WP_j \) is a \( \mathbb{R}^{d \times d} \) random matrix where all entries are i.i.d. \( \mathcal{N}(0, 1) \) random variables. A standard result for i.i.d. random matrices is the bound \( \mathbb{E}[\|P_j WP_j\|_{V_j \to V_j}] \lesssim \sqrt{d_j} \) for the operator norm, see [33, corollary 2.3.5]. Together with the Borell–Sudakov–Tsirelson concentration inequality for Gaussian processes, see [14, theorem 2.5.8], we immediately obtain the concentration inequality in assumption 2.

Finally, the degree of ill-posedness of \( K_\theta \) can be quantified by the smoothing effect of the operator:

**Assumption 4.** For a decreasing sequence \( (\sigma_j) \subseteq (0, \infty) \) and some constant \( Q > 0 \) let the operator \( K_\theta \) satisfy \( Q^{-1} \sum_k \sigma_k |(f, \varphi_k)|^2 \leq (K_\theta f, f) \leq Q \sum_k |(f, \varphi_k)|^2 \) for all \( f \in L^2 \) and \( \vartheta \in \Theta \).

Note that assumptions 1 and 4 with \( \sigma_j \downarrow 0 \) imply that \( K_\theta \) is compact, because it can be approximated by the operator sequence \( K_{\theta,j}P_j \) having finite dimensional ranges. The rate of the decay of \( \sigma_j \) will determine the degree of ill-posedness of the inverse problem. If \( \sigma_j \) decays polynomially or exponentially, we obtain a mildly or severely illposed problem, respectively.

Recall that the nuisance parameter \( \vartheta \) cannot be consistently estimated without additional assumptions. Therefore, we study the contraction rate of the marginal posterior distribution \( P(f \mid Y, T) \). While we allow for a general prior \( P_f \) on \( L^2 \) for \( f \), we will use a product prior on \( \vartheta \). For densities \( \beta_k \) on \( \mathbb{R} \) we thus consider prior distributions of the form

\[
dP(f, \vartheta) = dP_f(f) \otimes \bigotimes_{k \geq 1} \beta_k(\vartheta_k) d\vartheta_k.
\]

**Theorem 5.** Consider the model (1.1) and (1.2) generated by some \( f_0 \in L^2 \) and \( \vartheta_0 \in \Theta \) with \( \varepsilon = \varepsilon_n \to 0 \) and \( \delta = \delta_n \to 0 \) for \( n \to \infty \), respectively, and let assumptions 1, 2 and 4 be
satisfied. Let \( \Pi_n \) be a sequence of prior distributions of the form (3.1) on the Borel-\( \sigma \)-algebra on \( L^2 \times \Theta \). Let \((\kappa_n), (\xi_n)\) two positive sequences converging to zero and \((j_n)\) a sequence of integers with \( j_n \to \infty \). Suppose \( \kappa_n/(\xi_n \lor \delta_n) \to \infty \) as \( n \to \infty \) as well as

\[
d_{j_n} \leq c_1 \frac{\kappa_n}{(\xi_n \lor \delta_n)^2}, \quad \frac{\kappa_n}{\sigma_{j_n}} \leq c_2 \xi_n \quad \text{and} \quad \frac{\delta_n}{\sigma_{j_n}} \sqrt{d_{j_n}} \to 0
\]

for constants \( c_1, c_2 > 0 \) and all \( n \geq 0 \). Suppose \( f_0 \) satisfies \( \|f_0\| \leq R \) and \( \|f_0 - P_{j_n}(f_0)\| \leq C_0 \xi_n \) for some \( R, C_0 > 0 \). Let \( \mathcal{F}_n \subseteq \{ f \in L^2 : \|f - P_{j_n}f\| \leq C_0 \xi_n \} \) be a sequence and \( C_1 > 0 \) such that

\[
\Pi_n(L^2 \setminus \mathcal{F}_n) \leq e^{-(C_1 + 4) \kappa_n^2 / (\xi_n \lor \delta_n)^2}. \tag{3.2}
\]

Moreover assume for sufficiently large \( n \)

\[
\Pi_n((f, \vartheta) \in V_{j_n} \times \Theta : \frac{\|P_{j_n+m}(K_{\vartheta}f - K_{\vartheta}f_0)\|}{\varepsilon_n^2} + \frac{\|P_{j_n+m}(\vartheta - \vartheta_0)\|}{\delta_n^2} \leq \frac{\kappa_n}{(\xi_n \lor \delta_n)^2} ) \geq e^{-(C_1 + 4) \kappa_n^2 / (\xi_n \lor \delta_n)^2}. \tag{3.3}
\]

Then there exists a finite constant \( D > 0 \) such that the posterior distribution from (2.3) satisfies

\[
\Pi_n(f \in V_{j_n} : \|f - f_0\| > D \xi_n | Y, T) \to 0 \tag{3.4}
\]

as \( n \to \infty \) in \( \mathbb{P}_{j_n, \vartheta_0, \delta_0} \)-probability.

Theorem 5 states that the posterior distribution \( \Pi(f \in : Y, T) \) is consistent and concentrates asymptotically its whole probability mass in a ball around the true \( f_0 \) with decaying radius \( D \xi_n \downarrow 0 \), that is, the posterior ‘contracts to \( f_0 \)’ with the rate \( \xi_n \). This result is similarly to Ray [30, theorem 2.1] who has proven a corresponding theorem for known operators. However, the contraction rate is now determined by the maximum \( \varepsilon \lor \delta \) instead of \( \varepsilon \), which is natural in view of the results by Hoffmann and Reiß [16] who have included the case \( \delta > \varepsilon \) in their frequentist analysis.

To gain some intuition on the interplay between \( \kappa_n \) and the noise level \( \varepsilon_n \lor \delta_n \), let us set for simplicity \( m = 0 \) in assumption 1 and \( \xi_n = \delta_n \). Using assumption 4 (with lemma 14) and assumption 2, we then can decompose

\[
\|f - f_0\| \leq \|P_{j_n}f_0 - f_0\| + \|f - P_{j_n}f_0\|
\]

\[
\leq \|P_{j_n}f_0 - f_0\| + \sigma_{j_n}^{-1} \|K_{\vartheta}f - K_{\vartheta}f_0\|
\]

\[
\leq \|P_{j_n}f_0 - f_0\| + \sigma_{j_n}^{-1} \|K_{\vartheta}f - K_{\vartheta}f_0\| + \sigma_{j_n}^{-1} L \|P_{j_n}(\vartheta - \vartheta_0)\| \|f_0\|.
\]

The first term in the last line is the approximation error being bounded by \( \xi_n \). It corresponds to the classical bias. Indeed, the prior sequence \( \Pi_n \) is concentrated on a subset of \( \{ f : \|f - P_{j_n}f\| \leq C_0 \xi_n \} \) due to (3.2) such that the projection of \( f \) to the level \( j_n \) serves as reference measure for the prior and the deterministic error remains bounded by \( \xi_n \). The last two terms in the previous display correspond to the stochastic errors in \( f \) and \( \vartheta \) and are of the order \( \kappa_n / \sigma_{j_n} \), owing the the minimal spread of \( \Pi_n \) imposed by the small ball probability condition (3.3). In particular, we recover the ill-posedness of the inverse problem due to \( \sigma_{j_n} \to 0 \) in
the denominator. To obtain the best possible contraction rate, we need to choose $j_n$ in way that ensures that $\xi_n$ is close to $\kappa_n/\sigma_n$, i.e. we will balance the deterministic and the stochastic error. The conditions on the dimension $d_n$ are mild technical assumptions.

The crucial small ball probability assumption (3.3) ensures that the prior sequence $\Pi_n$ has some minimal mass in a neighbourhood of the underlying $f_0$ and $\vartheta_0$. The distance from $(f_0, \vartheta_0)$ is measured in a (semi-)metric which reflects the structure of our inverse problem. If $\varepsilon_n = \delta_n$, it would be sufficient if $\|K_\delta f - K_\vartheta f_0\|$ and $\|\vartheta - \vartheta_0\|$ are smaller than $\kappa_n$. However, condition (3.3) is more subtle. Firstly, the maximum of $\varepsilon$ and $\delta$ on the right-hand side within the probability introduces some difficulties. The prior has to weight a smaller neighbourhood of $K_\delta f_0$ or $\vartheta_0$, respectively, depending on whether $\varepsilon$ is smaller than $\delta$ or the other way around. If, for instance, $\varepsilon < \delta$ the contraction rate is determined by $\delta$ but the prior has to put enough probability to the smaller $\varepsilon$-ball around $K_\delta f_0$. We see such effects also in the construction of lower bounds, see [16], where we may have in the extreme case a $\delta$ distance between $f$ and $f_0$ while $K_\delta f_0 = K_\delta f$. Secondly, (3.3) depends only on finite dimensional projections of both $K_\delta f$ and $\vartheta$. This is particularly important as we do not assume any regularity conditions on $\vartheta$ such that we cannot expect the projection remainder ($\text{Id} - P_{f+n}$)$\vartheta$ to be small.

To allow for this relaxed small ball probability condition, the contraction rate is restricted to the set $V_j$. The result can be extended to $L^2$ by appropriate constructions of the prior, in particular, if the support of $\Pi_n$ is contained in $V_j$ we can immediately replace $V_j$ by $L^2$ in (3.4). Another possibility are general product prior if the basis is chosen according to the singular value decomposition of $K_\delta$.

To prove theorem 5, we use the techniques by Ghosal et al. [12, theorem 2.1], see also [14, theorem 7.3.5]. A main step is the construction of tests for the testing problem

$$H_0 : f = f_0 \quad \text{versus} \quad H_1 : f \in \mathcal{F}_n, \|f - f_0\| \geq D\xi_n.$$  

To this end, we first study a frequentist estimator of $f$ which then allows to construct a plug in test as proposed by Giné and Nickl [13].

The natural estimator for $\vartheta$ is $T$ itself. In order to estimate $f$, we use a linear Galerkin method based on the perturbed operator $K_T$ similar to the approaches in [11, 16]. We thus aim for a solution $\hat f_{e,\delta} \in V_j$ to

$$\langle K_T \hat f_{e,\delta}, v \rangle = \langle Y, v \rangle \quad \text{for all } v \in V_j.$$  

Choosing $v \in \{\varphi_k : |k| \leq j\}$, we obtain a system of linear equations depending only on the projected operator $K_{T_j}$. There is a unique solution if $K_{T_j}$ is invertible. Noting that for the unperturbed operator $K_{\delta_j}$ assumption 4 implies $\|K_{\delta_j}^{-1}\|_{V_j \rightarrow V_j} \leq Q\sigma_j^{-1}$ (see lemma 14 below), we set

$$\hat f_j := \begin{cases} K_{T_j}^{-1}P_jY, & \text{if } \|K_{T_j}\|_{V_j \rightarrow V_j} \leq \tau/\sigma_j, \\ 0, & \text{otherwise}, \end{cases}$$

for a projection level $j$ and a cut-off parameter $\tau > 0$. Adopting ideas from [13, 16], we obtain the following non-asymptotic concentration result for the estimator $\hat f_j$.

**Proposition 6.** Let $j \in \mathbb{N}$, $\kappa > 0$ such that $d_j \leq C_1 \kappa^2/(\varepsilon \vee \delta)^2$ for some $C_1 > 0$. Under assumptions 2 and 4 there are constants $c, C > 0$ such that, if $\delta \sigma_j^{-1}(\kappa + \sqrt{d_j}) \leq c^{-2}Q$ and $\tau > Q$, then $\hat f_j$ from (3.6) fulfills

$$\mathbb{P}_{f,\vartheta} \left(\|\hat f_j - f\| \geq C\sigma_j^{-1}(\|f\| \vee 1)\kappa + \|f - P_j f\|\right) \leq 3e^{-\kappa^2/(\varepsilon \vee \delta)^2}.$$
Note that some care will be needed to analyse the above mentioned tests since also the stochastic error term \( \sigma_j^{-1}(\|f\| \vee 1) \) depends on the unknown function \( f \) and, for instance, a Gaussian prior on \( f \) will not sufficiently concentrate on a fixed ball \( \{ f \in L^2 : \|f\| \leq R \} \).

**Remark 7.** While the assumption that \( K_\vartheta \) is self-adjoint simplifies the analysis and the presentation of our approach, the methodology can be generalised to general compact operators \( K_\vartheta \). In this case assumption 4 should be replaced by the assumption \( \|K_\vartheta f\|^2 \approx \sum_k \sigma_k^2 \langle f, \varphi_k \rangle^2 \) which is consistent with the original condition, see remark 15. The Galerkin projection method (3.5) can then be generalised to solve
\[
\langle K^*_\vartheta K_\vartheta \hat{f}, \varphi \rangle = \langle Y, K_\vartheta \varphi \rangle \quad \text{for all } \varphi \in V_f,
\]
see Cohen et al [7, appendix A]. This modified estimator should have a similar behaviour as above such that we can construct the tests which we needed to prove theorem 5. The rest of the proof of the contraction theorem and the subsequent results would remain as before.

4. A truncated product prior and the resulting rates

For the ease of clarity we fix a (\( S \)-regular) wavelet basis \( \{\varphi_k\}_{k \in \{-1,0,1,\ldots\} \times \mathbb{Z}} \) of \( L^2 \) with the associated approximation spaces \( V_j = \text{span}\{\varphi_k : |k| \leq j\} \). We write \(|k| = |(j, l)| = j \) as before. Investigating a bounded domain \( D \subseteq \mathbb{R}^d \), we have in particular \( d_j \approx 2^jd \). The regularity of \( f \) will be measured in the Sobolev balls
\[
H^s(R) := \left\{ f \in L^2([0, 1]) : \|f\|^2_{H^s} := \sum_{j=1}^{\infty} 2^{2js} \sum_l |(f, \varphi_{j,l})|^2 \leq R^2 \right\}, \quad s \in \mathbb{R}.
\]
(4.1)

We will use Jackson’s inequality and the Bernstein inequality: For \(-S < s \leq t < S\) and \( f \in H^s \), \( g \in V_j \) we have
\[
\|(\text{Id} - P_j)f\|_{H^s} \lesssim 2^{-j(t-s)} \|f\|_{H^t} \quad \text{and} \quad \|g\|_{H^s} \lesssim 2^{j(t-s)} \|g\|_{H^t}.
\]
(4.2)

**Remark 8.** The subsequent analysis applies also to the trigonometric as well as the sine basis in the case of periodic functions. Considering more specifically \( L^2_{\text{per}}([0, 1]) = \{ f \in L^2([0, 1]) : f(0) = f(1) = 0 \} \), we may set \( \varphi_k = \sqrt{2} \sin(\pi k \cdot) \) for \( k \in \mathbb{N} \). Since \( \|f\|^2_{H^s} \approx \sum_{k \geq 1} J^2(f, \varphi_k)^2 \) holds for any \( f \in L^2_{\text{per}}([0, 1]) \), it is then easy to see that the inequalities (4.2) are satisfied for \( V_j = \text{span}\{\varphi_1, \ldots, \varphi_j\} \) if \( 2^j \) is replaced by \( j \). Alternatively we may set \( V_j = \text{span}\{\varphi_1, \ldots, \varphi_j\} \) which gives exactly (4.2).

For \( \vartheta \) we use the product prior as in (3.1) with a fixed density \( \beta_k = \beta \). For \( f \) we also apply a product prior. More precisely, we take a prior \( \Pi_f \) determined by the random series
\[
f(x) = \sum_{|k| \leq J} \tau_{|k|} \Phi_k \varphi_k(x), \quad x \in [0, 1],
\]
for a sequence \( \tau_j_{j \geq 1} \), i.i.d. random coefficients \( \Phi_k \) (independent of \( \varphi_k \)) distributed according to a density \( \alpha \) and a cut-off \( J \in \mathbb{N} \). Hence,
\[
d\Pi(\vartheta, f) = \prod_{|k| \leq J} \tau_{|k|}^{-d} \alpha(\tau_{|k|}^{-1} f_k) \, df_k \cdot \prod_{k \geq 1} \beta(\varphi_k) \, d\vartheta_k.
\]
(4.3)
Under appropriate conditions on the distributions $\alpha, \beta$ and on $J$ we will verify the conditions of theorem 5.

**Assumption 9.** There are constants $\gamma, \Gamma > 0$ such that the densities $\alpha$ and $\beta$ satisfy

$$\alpha(x) \wedge \beta(x) \geq \Gamma e^{-\gamma|x|^2} \quad \text{for all } x \in \mathbb{R}. $$

Assumption 9 is very weak and is satisfied for many distributions with unbounded support, for example, Gaussian, Cauchy, Laplace distributions or Student’s $t$-distribution. Also uninformative priors where $\alpha$ or $\beta$ are constant are included. A consequence of the previous assumption is that any random variable $\Phi$ with probability density $\alpha$ (or $\beta$) satisfies

$$\Pr(|\Phi - x| \leq \kappa) \geq \Gamma \int_{|y| \leq \kappa} e^{-\gamma(|y|+\kappa)^2} dy \geq 2\Gamma \kappa e^{-\gamma(|x|+\kappa)^2} \quad \text{for all } \kappa > 0, x \in \mathbb{R}. $$

(4.4)

This lower bound will be helpful to verify the small ball probabilities (3.3).

To apply theorem 5, we choose $J = j_n$ to ensure that the support of $\Pi_f$ lies in \{ $f \in \mathcal{F}$ : $\|P_{jn}(f) - f\| \leq C_j \xi_n$\}. Note that the optimal $j_n$ is not known in practice. We will discuss the a data-driven choice of $J$ in section 5. Alternatively to truncating the random series for $f$, the small bias condition could be satisfied if ($\tau$) decays sufficiently fast and $\alpha$ has bounded support, as it is the case for uniform wavelet priors.

We start with the mildly ill-posed case imposing $\sigma_j = 2^{-j\beta}$ for some $\beta > 0$ in assumption 4. In this case the operators $K_\theta$ are naturally adapted to Sobolev scale, since then $K_\theta : L^2 \rightarrow H^s$ is continuous with $\|K_\theta f\| \lesssim \|f\|_{H^s}$, see remark 15.

**Theorem 10.** Let $\epsilon^{\eta} \lesssim \delta \lesssim \epsilon$ for some $\eta > 1$ and let assumptions 1, 2 with $l_j \leq 2^j$, assumption 4 with $\sigma_j = 2^{-j\beta}$ for some $\beta > 0$ as well as assumption 9 be fulfilled. Then the posterior distribution from (2.3) with prior given by (4.3) where $J$ is chosen such that $2^J = (\epsilon \log(1/\epsilon))^{-2/(2s+2d)}$ and $2^{-j(2n+d)} \leq \tau_j^2 \leq 2^C j$ for constants $c, C > 0$ and some $0 < s_0 < s$ satisfies for any $f_0 \in H^s(R)$ and $\theta_0 \in \Theta_0$

$$\Pi_\alpha \left( f \in L^2 : \|f - f_0\| > D \epsilon \log(1/\epsilon)^{2s/(2s+2d)} \right) \rightarrow 0$$

with some constant $D > 0$ and in $\Pr_{f_0, \alpha_\epsilon}$-probability.

**Remark 11.** This theorem is restricted to the case $\epsilon \gtrsim \delta$. However, its proof reveals that in the special case where $m = 0$, for instance, if $(\varphi_k)$ are eigenfunctions, the condition $\epsilon^\eta \lesssim \delta \lesssim \epsilon$ can be weakened to $\log \delta \simeq \log \epsilon$, which also allows for $\epsilon < \delta$. The second restriction is $l_j \leq 2^j$ which is especially satisfied in the model $B$ of unknown eigenvalues in the singular value decomposition of $K_\theta$. Larger $l_j$ could be incorporated if we put some structure on $\Theta$ which allows for applying a different prior on $\theta$ with better concentration of $P_f \theta$.

The contraction rate coincides with the minimax optimal convergence rate, as determined in [6, 16] for specific settings of $\theta \mapsto K_\theta$, up to the logarithmic term. The conditions on $\tau_j$ are very weak and allow for a large flexibility in the choice of prior, particularly, a constant $\tau_j = 1$ for all $j$ is included. In contrast, the choice of the cut-off parameter $J$ is crucial and depends on the regularity $s$ of $f_0$ and the ill-posedness $t$ of the operator.

In the severely ill-posed case the contraction rates deteriorates to a logarithmic dependence on $\sqrt[\vee]{\delta}$ and coincide again with the minimax optimal rate.

**Theorem 12.** Let $\log \epsilon \simeq \log \delta$ and let assumptions 1, 2 and 4 with $\sigma_j = \exp(-r2^j)$ for some $r, t > 0$ as well as assumption 9 be fulfilled. Then the posterior distribution from
(2.3) with prior given by (4.3) where \( J \) is chosen such that \( 2^j = \left( -\frac{1}{2^j} \log(\varepsilon \vee \delta) \right)^{1/t} \) and \( 2^{-j(2n+t+d)} \leq \tau_j^2 \leq \exp(C2^j) \) for a constant \( C > 0 \) satisfies for any \( f_0 \in \mathcal{H}^t(\mathcal{R}) \) and \( \vartheta_0 \in \Theta_0 \)

\[
\Pi_n \left( f \in L^2 : \|f - f_0\| > D \left( \log(\varepsilon \vee \delta)^{-1} \right)^{-1/t} \right) \rightarrow 0
\]

with some constant \( D > 0 \) and in \( \mathbb{P}_{f_0, \vartheta_0} \)-probability.

5. Adaptation via empirical Bayes

We saw above that the choice of the projection level \( J \) of the prior depends on the unknown regularity \( s \) (and the ill-posedness \( t \)) in order to achieve the optimal rate. We will now discuss how \( J \) can be chosen purely data-driven resulting in an empirical Bayes procedure that adapts on \( s \). Noting that choice of \( J \) in theorem 12 is already independent of \( s \), we focus on the mildly ill-posed case and \( \delta \lesssim \varepsilon \).

The method is based on the observation that all conditions on the level \( j_n \) in theorem 5 are monotone (in the sense that they are also satisfied for all \( j \) smaller than the optimal \( j_n \)) except for the bias condition on \( \|f_0 - P_j f_0\| \lesssim \xi_n \). Given the optimal \( J_n = j_n \), the so-called oracle, the result in theorem 10 continues to hold for any, empirically chosen \( J \) satisfying

\[
\hat{J} \leq J_n \quad \text{and} \quad \|f_0 - P_{\hat{J}} f_0\| \lesssim \xi_n.
\]

To find \( J \), we use Lepski’s method [24] which is generally known for these two properties.

In proposition 6 we saw that the variance of the estimator \( \hat{f}_j \) from (3.6) is of the order \( \varepsilon^2 j / \sigma_j^2 = \varepsilon^2 2^{2t + jd} \). For some fixed lower bound \( s_0 \) on the regularity \( s \) of \( f_0 \in \mathcal{H}^t \) let

\[
J_{\varepsilon} = \left\lfloor \frac{\log \varepsilon^{-1}}{(s_0 + t + d/2) \log 2} \right\rfloor
\]

where \( \lfloor x \rfloor \) denotes be the largest integer smaller than \( x \). The choice of \( J_{\varepsilon} \) allows for applying the concentration inequality from proposition 6 to all \( \hat{f}_j \) with \( j \leq J_{\varepsilon} \). We then choose

\[
\hat{J} := \min \left\{ j \in \{1, \ldots, J_{\varepsilon}\} : \|\hat{f}_j - \hat{f}_j\| \leq \Delta \varepsilon (\log \varepsilon^{-1})^{2^{(t+d)/2}} e^{-2^{(t+d)/2}} \right\}
\]

for a constant \( \Delta \in (0, 1) \) which can be chosen by the practitioner. The idea of the choice \( \hat{J} \) is as follows: Starting with large \( j \) the projection estimator \( \hat{f}_j \) has a small bias, but a standard deviation of order \( \varepsilon 2^{j(t+d/2)} \). Decreasing \( j \) reduces the variance while the bias increases. At the point where there is some \( i > j \) such that \( \|\hat{f}_i - \hat{f}_j\| + \|\hat{f}_j - \hat{f}_0\| \geq \|\hat{f}_i - \hat{f}_0\| \) is larger than the order of the variance the bias starts dominating the estimation error. At this point we stop lowering \( j \) and select \( \hat{J} \).

Theorem 13. Let \( \varepsilon^2 \lesssim \delta \lesssim \varepsilon \) for some \( \eta > 1 \) and let assumptions 1 and 2 with \( l_j \leq 2^{jd} \), assumption 4 with \( \sigma_j = 2^{-j} \) for some \( t > 0 \) as well as assumption 9 be fulfilled. Then the posterior distribution from (2.3) with prior given by (4.3) with \( \hat{J} \) instead of \( J \) and \( \varepsilon^2 2^{-j(2n+t+d)} \leq \tau_j^2 \leq 2^C \) for constants \( c, C > 0 \) and some \( 0 < s_0 < s \) satisfies for any \( f_0 \in \mathcal{H}^t(\mathcal{R}) \) and \( \vartheta_0 \in \Theta_0 \)

\[
\Pi_n \left( f \in L^2 : \|f - f_0\| > D (\log \varepsilon^{-1})^\chi e^{2^s/((2s + 2t + d))} \right) \rightarrow 0
\]

with some constant \( D > 0, \chi = (4s + 2t + d)/(2s + 2t + d) \) and in \( \mathbb{P}_{f_0, \vartheta_0} \)-probability.
Note that the empirical Bayes procedure is adaptive with respect to $s$ and the Sobolev radius $R$. Compared to theorem 10 where the oracle choice for $J$ is used, we only lose a logarithmic factor for adaptivity.

6. Examples and simulations

6.1. Heat equation with unknown diffusivity parameter

To illustrate the previous theory, we consider the heat equation

$$\frac{\partial}{\partial t} u(x, t) = \alpha \frac{\partial^2 u(x, t)}{\partial x^2}, \quad u(\cdot, 0) = f, \quad u(0, t) = u(1, t) = 0 \quad (6.1)$$

with Dirichlet boundary condition at $x = 0$ and $x = 1$ and some initial value function $f \in L^2([0, 1])$ satisfying $f(0) = f(1) = 0$. Different to [23, 30] we take an unknown diffusivity parameter $\vartheta > 0$ into account. A solution to (6.1) is observed at some time $t > 0$

$$Y = u(\cdot, t) + \varepsilon Z \quad (6.2)$$

with white noise $Z$ on $L^2([0,1])$. The aim is to recover $f$ from $Y$.

The solution $u(\cdot, t)$ depends linearly on $f$ via an operator $K_\vartheta$ which is diagonalised by the sine basis $e_k = \sqrt{2} \sin(\pi k \cdot)$, $k \geq 1$, of $L^2_{\text{per}}([0, 1])$ building a system of eigenfunctions of the Laplace operator. The corresponding eigenvalues of $K_\vartheta$ are given by $\rho_{\vartheta,k} := e^{-\vartheta \pi^2 k^2}$, $k \geq 1$, and we obtain the singular value decomposition

$$K_{\text{of}} = \sum_{k \geq 1} \langle f, e_k \rangle \rho_{\vartheta,k} e_k = \sum_{k \geq 1} \langle f, e_k \rangle e^{-\vartheta \pi^2 k^2 \sqrt{2} \sin(\pi k \cdot)}.$$

Note that $K_{\vartheta}$ depends on $\vartheta$ only via its eigenvalues $\rho_{\vartheta,k}$ while the eigenfunctions and thus the considered basis is independent of $\vartheta$. Moreover the dependence of $\rho_{\vartheta,k}$ on $\vartheta$ is non-linear. From the decay of the eigenvalues we see that the resulting inverse problem is severely ill-posed with $s_\vartheta = \exp(-\vartheta \pi^2 \vartheta^2)$. Since we can easily construct pairs $(\vartheta, f)$ and $(\vartheta', f')$ with $K_{\text{of}} = K_{\vartheta} f'$, the function $f$ is indeed not identifiable only based on the observation $Y$ and we need the additional observation $T = \vartheta + \delta W$ for some $W \sim \mathcal{N}(0, 1)$.

Since the eigenfunctions are independent of $\vartheta$, we can choose the basis $\varphi_k = e_k$ thanks to remark 8. We moreover apply the truncated product prior (4.3) with centered normal densities $\alpha$ and $\beta$ and fixed variances $\tau^2$ and $\sigma^2$. In our numerical example we set $t = 0.1,

$$f_0(x) = 4x(1 - x)(8x - 5) \quad \text{and} \quad \vartheta_0 = 1 \quad (6.3)$$

reproducing the same setting as considered in [23], but taking the unknown $\vartheta$ into account.

The Fourier coefficients of $f_0$ with respect to the sine series $\varphi_k$ are given by

$$f_{0,k} = \langle f_0, \varphi_k \rangle = \frac{8\sqrt{2}(13 + 11(-1)^k)}{\pi^3 k^3}, \quad k \geq 1.$$

By the decay of the coefficients, we have $f_{0,k} \in H^s$ for every $s < 5/2$.

To implement our Bayes procedure, we need to sample from the posterior distribution which is not explicitly accessible. Fortunately, using independent normal $\mathcal{N}(0, \tau^2)$ priors on the coefficients $f_k = \langle f, \varphi_k \rangle$, we see from (2.2) that at least the conditional posterior distribution of $f$ given $\vartheta, Y, T$ can be explicitly computed as
\[ \Pi(f \in \cdot | \vartheta, Y, T) = \bigotimes_{k \leq J} \mathcal{N}\left( \frac{\varepsilon^{-2} \rho_{\vartheta,k}^{-1} \varepsilon^{-2} + \rho_{\vartheta,k}^{-2} \tau^{-2} - 2}{\varepsilon^{-2} + \rho_{\vartheta,k}^{-2} \tau^{-2}}, \frac{\rho_{\vartheta,k}^{-2}}{\varepsilon^{-2} + \rho_{\vartheta,k}^{-2} \tau^{-2}} \right). \] (6.4)

Profiting from this known conditional posterior distribution, we use a Gibbs sampler to draw (approximately) from the unconditional posterior distribution of \( f \) given \( Y, T \), see [34]. Given some initial \( \vartheta^{(0)} \), the algorithm alternates between draws from \( f^{(i+1)} | \vartheta = \vartheta^{(i)}, Y, T \) and \( \vartheta^{(i+1)} | f = f^{(i+1)}, Y, T \) for \( i \in \mathbb{N} \). The second conditional distribution is not explicitly given, due to the non-linear dependence of \( \rho_{\vartheta, \vartheta} \) from \( \vartheta \). We apply a standard Metropolis–Hastings algorithm to approximate the distribution of \( \vartheta | f, Y, T \) using a random walk with \( \mathcal{N}(0, \nu^2) \) increments as proposal chain. A similar Metropolis-within-Gibbs method has been used in [21] in a comparable simulation task. Using the sequence \( (\vartheta^{(i)})_i \) from this algorithm, the final Markov chain Monte Carlo (MCMC) approximation of \( \Pi(f \in \cdot | Y, T) \) is then given by an average

\[
\frac{1}{M} \sum_{m=1}^{M} \Pi(f \in \cdot | \vartheta = \vartheta^{(B+m-l)}), Y, T)
\]

for sufficiently large \( B, M, l \in \mathbb{N} \), where we again profit from the explicitly given conditional posterior distribution (6.4).

Figure 1 shows the typical posterior mean and 20 draws from the posterior distribution in a simulation using \( \varepsilon = \delta = 10^{-6} \) and \( 10^{-8} \). In both cases the projection level is chosen as \( J = 4 \simeq \sqrt{-\log(\varepsilon)} \). Especially for the smaller noise level, the common intersections of all sampled functions are conspicuous. They reflect a quite low variance of the posterior distribution in the first coefficients compared to a relatively large variance already for \( f_4 \) due to the severe ill-posedness, see (6.4).

As a reference estimator the Galerkin projector \( \hat{f}_J \) from (3.6) is plotted, too. We see that for \( \varepsilon = 10^{-6} \) the posterior mean is much closer to the true function indicating an efficiency gain of the Bayesian procedure compared to the projection estimator. For \( \varepsilon = 10^{-8} \) both estimators coincide almost perfectly. As shown by the theory, the figure illustrates that the posterior
distribution concentrates around the truth for smaller noise levels. Monte Carlo simulations based on 500 iterations yield a root mean integrated squared error (RMISE) 0.3353 and 0.0512 for ε = 10^{-6} and ε = 10^{-8}, respectively. For the posterior mean of ɣ we observe a root mean squared error of approximately 1.0 · 10^{-6} and 9.7 · 10^{-9}, respectively. Additionally, Table 1 reports the RMISE for several different combinations of the noise levels ε and δ.

### 6.2. Deconvolution with unknown kernel

Another example is the deconvolution problem occurring for instance in image processing, see Johnstone et al [19]. The aim is to recover some unknown 1-periodic function f from the observations

\[ Y = K_0 f + \varepsilon Z \quad \text{with} \quad K_0 f := g_\theta * f := \int_0^1 f(t-x)g_\theta(x)dx \]

where \( g_\theta \in L^2_{\text{per}} \) is some 1-periodic convolution kernel (more general it might be a signed measure). Since the convolution operator \( K_0 \) is smoothing, the inverse problem is ill-posed. If the kernel \( g_\theta \) is unknown, the problem is called blind deconvolution occurring in many applications [4, 20, 32]. In a density estimation setting this problem as already been intensively investigated, see [10, 17, 18, 26] among others. However, the Bayesian perspective on this problem seem not thoroughly studied.

We consider the trigonometric basis

\[ \varphi_0 = 1, \quad \varphi_{j,0} = \sqrt{2} \sin(2\pi j \cdot), \quad \varphi_{j,1} = \sqrt{2} \cos(2\pi j \cdot), \quad j \in \mathbb{N}, \]

with the corresponding approximation spaces \( V_J = \text{span}(\varphi_{j,l} : j \leq J, l \in \{0,1\}) \). Assuming \( g_\theta \) is symmetric, we have \( \langle g_\theta, \varphi_{j,0} \rangle = 0 \) and

\[ K_0 \varphi_{j,0} = \langle g_\theta, \varphi_{j,0} \rangle \varphi_0, \quad K_0 \varphi_{j,l} = \sum_m \langle g_\theta, \varphi_{m,l} \rangle \varphi_{j,m,l} = \frac{1}{\sqrt{2}} \langle g_\theta, \varphi_{j,0} \rangle \varphi_{j,1}, \quad j \in \mathbb{N}, l \in \{0,1\} \]

by the angle sum identities (for non-symmetric kernels \( K_0 \) could be diagonalised by the complex valued Fourier basis). We thus obtain the singular value decomposition \( K_0 f = \sum_k \rho_{\theta,k} f_k \varphi_k \) again in multi-index notation \( k = (j,l), j \in \mathbb{N}, l \in \{0,1\} \), where \( \rho_{\theta,k} = \langle g_\theta, \varphi_{k,j,l} \rangle / \sqrt{2} \) and \( f_k = \langle f, \varphi_k \rangle \). Depending on the regularity of \( g \) and thus the decay of \( \langle g_\theta, \varphi_{j,l} \rangle \) the problem is mildly or severely ill-posed.

If the convolution kernel is fully unknown, we parametrise \( g_\theta = \vartheta \) by all (symmetric) 1-periodic kernels \( \vartheta \). Due to the SVD, we then can identify \( g_\theta \) with the singular values, that is, we set \( \vartheta = (\rho_{\theta,k})_k \). The sample \( T \) can be understood as training data, where the convolution experiment is applied to all basis functions \( f \in \{ \varphi_{j,l} \} \). In this scenario we obtain \( \varepsilon = \delta \).

In our simulation \( \vartheta_0 \) is given by the periodic Laplace kernel \( g_{\vartheta_0}(x) = \frac{1}{\pi} e^{-|x|/h} \mathbb{1}_{[-1/2,1/2]}(x) \) with normalisation constant \( C_0 = 2h(1 + e^{-1/(2h)}) \) and fixed bandwidth \( h = 0.1 \). Hence, we have for \( k \in \mathbb{N} \times \{0,1\} \)

### Table 1. RMISE for different values of ε and δ.

| \( \varepsilon \) | 10^{-4} | 10^{-6} | 10^{-8} |
|------------------|--------|--------|--------|
| 10^{-5}          | 0.5728 | 0.3173 | 0.5656 |
| 10^{-6}          | 0.5515 | 0.3353 | 0.0545 |
| 10^{-8}          | 0.5548 | 0.3269 | 0.0512 |
\[ \rho_{\theta_0,0} = 1, \quad \rho_{\theta_0,k} = \frac{1}{C_h(4\pi^2 |k|^2 + h^{-2})} \left( 1 - e^{-1/(2h)} \cos(\pi |k|) + e^{-1/(2h)} 2\pi |k|h \sin(\pi |k|) \right). \]

In particular, we have two degree of illposedness. We moreover use \( f_0 \) from (6.3).

To implement the empirical Bayes procedure with the trigonometric basis and corresponding approximation spaces \( V_j = \text{span}(\varphi_k : |k| \leq j) \), we need to replace \( 2j \) by \( 2j \) as mentioned in remark 8. Choosing some \( b > 1 \) and setting \( J_b = \left\lfloor \frac{\log \varepsilon^{-1}}{(s_0 + s + d/2) \log b} \right\rfloor \) for some lower bound \( s_0 \leq s \), the selection rule then reads as

\[ \hat{J} := \min \left\{ j \in \{1, b, b^2, \ldots, b^{J_b} \} : \| \hat{f}_j - \hat{f} \| \leq \Delta \varepsilon (\log \varepsilon^{-1})^{3/2} \gamma i > j \right\}. \]

Using the again Gaussian product priors for \( \varphi \) and \( \vartheta \), the posterior distribution can be similarly approximated as described in section 6.1. However, the nuisance parameter \( \vartheta \) is now infinite dimensional. Here, we can profit from the truncated product structure of the prior which implies that the posterior distribution only depends on the \( J \)-dimensional projection \( P_J \vartheta \) (note that assumption 1 is satisfied with \( m = 0 \)). More precisely, we only have to draw from the posterior given by

\[ \Pi(B|Y, T) = \frac{1}{C} \int_B \exp \left( \frac{1}{2\varepsilon^2} \| P_J K_0 f, Y \| - \frac{1}{2\varepsilon^2} \| P_J K_0 f \|^2 + \frac{1}{\lambda^2} \| P_J \vartheta, T \| - \frac{1}{2\delta^2} \| P_J \vartheta \|^2 \right) \] 

with normalisation constant \( C > 0 \) and for all Borel sets \( B \subseteq L^2 \times P_J \Theta \), see proof of theorem 5. Therefore, a Gibbs sampler can be used to draw successively the coordinates of \( P_J \vartheta \) with a Metropolis–Hastings algorithm and iterate as above with draws of \( f \). This simulation approach is not restricted to this particular example, but applies generally. Note that in the specific deconvolution setting, the map \( \vartheta \mapsto K_0 f \) is linear for fixed \( f \), such that \( \vartheta | f, Y, T \) can be directly sampled from a Gaussian distribution.

For \( \varepsilon = \delta = 10^{-2} \) and \( \varepsilon = \delta = 10^{-3} \) a typical trajectory of the posterior mean and 20 draws from the posterior are presented in figure 2 where the Lepski rule has chosen \( J = 3 \) (i.e. \( 7 \) basis functions) and \( J = 5 \) (\( 11 \) basis functions), respectively. For the larger noise level, the posterior mean slightly improves the Galerkin projector, while for the smaller noise level both estimators basically coincide. We see a much better concentration of the posterior distribution than in the severely ill-posed case discussed previously. In a Monte Carlo simulation for \( \varepsilon = \delta = 10^{-2} \) based on 500 iterations in this setting the posterior mean for \( f \) achieved a RMISE of 0.1142 which is approximately 8.6\% of \( \| f_0 \| \). The Lepski method has chosen \( J \in \{2, 3\} \) with relative frequency 0.97. For \( \varepsilon = \delta = 10^{-2} \) the simulation yields a RMISE of 0.0174, which is 1.3\% of \( \| f_0 \| \), and projections levels \( J \) in \( \{4, 5\} \) in 0.82\% of the Monte Carlo iterations.

7. Proofs

We first study some smoothing properties of the operator \( K_\vartheta \).

**Lemma 14.** Under assumption 4 we have \( \| K_{a,j}^{-1} \|_{V_j \to V_j} \leq Q \sigma_j^{-1} \) for all \( \vartheta \in \Theta \).

**Proof.** For \( g \in V_j \) the function \( h = K_{a,j}^{-1} g \in V_j \) is given by the unique solution to the linear system

\[ \rho_{\theta_0,j} = 1, \quad \rho_{\theta_0,k} = \frac{1}{C_h(4\pi^2 |k|^2 + h^{-2})} \left( 1 - e^{-1/(2h)} \cos(\pi |k|) + e^{-1/(2h)} 2\pi |k|h \sin(\pi |k|) \right). \]
\langle K_{\alpha} h, v \rangle = \langle g, v \rangle, \text{ for all } v \in V_j.

Assumption 4 then yields
\[ \sigma_j \|h\|^2 = \sigma_j \sum_{|k| \leq j} \langle h, \varphi_k \rangle^2 \leq \sum_{|k| \leq j} \sigma_{|k|} \langle h, \varphi_k \rangle^2 = Q \|K_{\alpha}h, h\| \sup_{v \in V_j : \|v\| \leq 1} \langle K_{\alpha}h, v \rangle = Q \|h\| \|g\| . \]

Therefore, \( \sigma_j \|K_{\alpha,j}^{-1} g\| \leq Q \|g\| \) holds true for all \( g \in V_j \). \(\square\)

Remark 15. As soon as \( (\sigma_j) \) decays at least geometrically, assumptions 1 and 4 also yield \( \|K_{af}\|^2 \lesssim \sum_{k} \sigma_{|k|}^2 \langle f, \varphi_k \rangle^2 \). Indeed, we have for any \( f \in L^2 \) such that the right-hand side is finite:
\[
\|K_{af}\|^2 = \sum_k \|K_{af}P_{|k|+m} f, \varphi_k \|^2 \leq \sum_k \|K_{f}^{1/2}P_{|k|+m} f\|^2 \|K_{o}^{1/2} \varphi_k\|^2 \lesssim \sum_k \sigma_{|k|} \sigma_{|l|} \langle f, \varphi_l \rangle^2 \leq \sum_{|k| \geq (|l|+m) \vee 0} \sigma_{|k|} \sum_{|l|} \sigma_{|l|} \langle f, \varphi_l \rangle^2 .
\]

7.1. Proof of proposition 6

To simplify the notation, we abbreviate \( \mathbb{P} = \mathbb{P}_{f, \theta} \) in the sequel and define the operator \( \Delta_{T,j} := K_{\alpha,j} - K_{\theta,j} \). Set for \( \gamma \in (0, 1 - Q/\tau) \)
\[
\Omega_{T,j} := \{ \|K_{\theta,j}^{-1} \Delta_{T,j} \|_{V_j \to V_j} \leq \gamma \} .
\]

Lemma 14 yields
Under assumption 2 we have due to the condition $\delta \sigma_j^{-1}(\kappa + \sqrt{d_j}) \leq \gamma/(CQL)$
\[
\mathbb{P}(\Omega_{T,j}) \leq \mathbb{P}(\|K_{T,j}^{-1}\|_{V_j \rightarrow V_j} \Delta_{T,j} \|_{V_j \rightarrow V_j} > \gamma) \leq \mathbb{P}(\|\Delta_{T,j}\|_{V_j \rightarrow V_j} > \gamma \sigma_j/Q).
\]

We thus may restrict on $\Omega_{T,j}$ on which the operator $K_{T,j} = K_{\theta,j}(\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j})$ is invertible satisfying
\[
\|K_{T,j}^{-1}\|_{V_j \rightarrow V_j} \leq \|\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j})^{-1}\|_{V_j \rightarrow V_j} \|K_{T,j}^{-1}\|_{V_j \rightarrow V_j} \leq \frac{1}{1 - \gamma} \|K_{\theta,j}^{-1}\|_{V_j \rightarrow V_j} \leq \frac{Q}{(1 - \gamma)\sigma_j}
\]
where we used lemma 14 in the last step. Hence, for $\gamma \leq 1 - Q/\tau$ we have $\Omega_{T,j} \subseteq \{\|K_{T,j}^{-1}\|_{V_j \rightarrow V_j} \leq \tau \sigma_j^{-1}\}$. Therefore, we can decompose on $\Omega_{T,j}$
\[
\|\hat{f} - f\|^2 = \|P_j f - f\|^2 + \|\hat{f} - P_j f\|^2
\]
\[
\leq \|P_j f - f\|^2 + \|K_{T,j}^{-1} P_j Y - P_j f\|^2. \tag{7.1}
\]

The first term is the usual bias. For the second term in (7.1) we write on $\Omega_{T,j}$
\[
K_{T,j}^{-1} P_j Y - P_j f = (\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j})^{-1}(\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j}) P_j f + \varepsilon(\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j})^{-1} K_{\theta,j}^{-1} P_j Z
\]
\[
= (\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j})^{-1} K_{\theta,j}^{-1} \Delta_{T,j} P_j f + \varepsilon(\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j})^{-1} K_{\theta,j}^{-1} P_j Z.
\]

Since $\|\text{Id} - K_{\theta,j}^{-1} \Delta_{T,j})^{-1}\|_{V_j \rightarrow V_j} \leq 1/(1 - \gamma)$ on $\Omega_{T,j}$, we obtain
\[
\|K_{T,j}^{-1} P_j Y - P_j f\| \leq \frac{1}{1 - \gamma} \|K_{\theta,j}^{-1}\|_{V_j \rightarrow V_j} \|\Delta_{T,j}\|_{V_j \rightarrow V_j} \|P_j f\| + \frac{\varepsilon}{1 - \gamma} \|K_{\theta,j}^{-1}\|_{V_j \rightarrow V_j} \|P_j Z\|
\]
\[
\leq \frac{Q}{(1 - \gamma)\sigma_j} (\|\|\Delta_{T,j}\|_{V_j \rightarrow V_j} + \varepsilon\|P_j Z\|).
\tag{7.2}
\]

To deduce a concentration inequality for $\|P_j Z\|$, we proceed as proposed in [13]: For a countable dense subset $B$ of the unit ball in $L^2$, we have $\|P_j Z\| = \sup_{f \in B} \|P_j Z(f)\|$. The Borell–Sudakov–Tsirelson inequality [16, theorem 2.5.8] yields for any $\kappa > 0$
\[
\mathbb{P}(\|P_j Z\| \geq \kappa + \varepsilon \sup_{f \in B} \|P_j Z(f)\|) \leq \mathbb{P}(\sup_{f \in B} \|P_j Z(f)\| - \mathbb{E}[\sup_{f \in B} \|P_j Z(f)\|] \geq \kappa) \leq 2^{-\kappa^2/(2\sigma^2)}
\]
with $\sigma^2 = \sup_{f \in B} \text{Var}(P_j Z(f)) \leq \|f\|^2 \leq 1$. Since
\[
\mathbb{E}[\|P_j Z\|] \leq \mathbb{E}[\|P_j Z\|^2]^{1/2} = \left(\sum_{k \leq j} \mathbb{E}[Z_k^2]\right)^{1/2} = d_j^{1/2}
\]
and $d_j \leq \varepsilon^{-2}\kappa^2$, we find for some constant $C > 0$
\[
\mathbb{P}(\varepsilon \|P_j Z\| \geq C\kappa) \leq \mathbb{P}(\varepsilon \|P_j Z\| \geq \kappa + \varepsilon d_j^{1/2}) \leq e^{-\kappa^2/(2\epsilon^2)}.
\]

Under assumption 2 and due to $d_j \leq \delta^{-2}\kappa^2$, we analogously obtain
\[ \mathbb{P}( \| \Delta r \|_{V_f \to V_i} \geq C \kappa ) \leq e^{-\kappa^2 / (2\delta^2)}. \]

In combination with (7.2), the asserted concentration inequality is proven. \(\square\)

### 7.2. Proof of theorem 5

We proof the theorem in two steps.

**Step 1:** We construct tests \( \Psi_n = \Psi_n(Y, T) \) such that

\[ \mathbb{E}_{\theta, \alpha}[\Psi_n] \to 0 \quad \text{and} \quad \sup_{f \in \mathcal{F}, \theta \in \Theta} \mathbb{E}_{\theta, \alpha}[1 - \Psi_n] \leq 3e^{-(C_1 + 4)\kappa_\gamma^2 / (\varepsilon \sqrt{\delta^2})^2}. \]

(7.3)

Based on the estimator \( \hat{f}_n \) from (3.6), we set

\[ \Psi_n := \mathbb{I}_{\{\|\hat{f}_n - f_0\| \geq D_1 \varepsilon_n\}} \]

for \( D_1 = 2C_2 \sqrt{C_1} + 4R + 2C_0 \varepsilon_n \) with the constant \( C \) from proposition 6. Due to proposition 6 and \( \kappa_\gamma / \sigma_{\kappa_\gamma} \leq \varepsilon_2 \varepsilon_n \), we then have

\[ \mathbb{E}_{\theta, \alpha}[\Psi_n] = \mathbb{P}_{\theta, \alpha}\left(\|\hat{f}_n - f_0\| \geq 2C_2 \varepsilon_n \right) \leq \mathbb{P}_{\theta, \alpha}\left(\|\hat{f}_n - f_0\| \geq C_2 \varepsilon_n \right) \leq 3e^{-(C_1 + 4)\kappa_\gamma^2 / (\varepsilon \sqrt{\delta^2})^2} \]

converging to 0.

On the alternative we set \( D = D_2(1 + R) \) for \( D_2 = 2\max(C_0 + D_1, C\sqrt{C_1} + 4 / \varepsilon_2) \). For any \( \Psi \in \Theta \) and any \( f \in \mathcal{F} \), with \( \|f - f_0\| \geq D_2(1 + R) \varepsilon_n \), we have \( (2 - D_2 \varepsilon_n) \|f - f_0\| \geq D_2(1 + R) \varepsilon_n \) for sufficiently small \( \varepsilon_n \). Therefore,

\[ \|f - f_0\| \geq \frac{D_2}{2} (1 + R + \|f - f_0\|) \leq \frac{D_2}{2} (1 + \|f_0\| + \|f - f_0\|) \varepsilon_n \]

\[ \leq \frac{D_2}{2} (1 + \|f\|) \varepsilon_n \geq C_\kappa_\gamma^{-1} \sqrt{C_1 + 4 \kappa_\gamma \|f\| + (C_0 + D_1) \varepsilon_n}, \]

(7.4)

where the last inequality holds by the choice of \( D_2 \). We obtain

\[ \mathbb{E}_{\theta, \alpha}[1 - \Psi_n] = \mathbb{P}_{\theta, \alpha}\left(\|\hat{f}_n - f_0\| < D_1 \varepsilon_n\right) \]

\[ \leq \mathbb{P}_{\theta, \alpha}\left(\|\hat{f}_n - f\| > \|f - f_0\| - D_1 \varepsilon_n\right) \]

\[ \leq \mathbb{P}_{\theta, \alpha}\left(\|\hat{f}_n - f\| > C_\kappa_\gamma^{-1} \sqrt{C_1 + 4 \kappa_\gamma \|f\| + (C_0 + D_1) \varepsilon_n}\right). \]

Proposition 6 yields again \( \mathbb{E}_{\theta, \alpha}[1 - \Psi_n] \leq 3e^{-(C_1 + 4)\kappa_\gamma^2 / (\varepsilon \sqrt{\delta^2})^2} \).

**Step 2:** Since \( \mathbb{E}_{\theta, \alpha}[\Psi_n] \to 0 \), it suffices to prove that

\[ \Pi_n(f \in V_f, \|f - f_0\| > D \varepsilon_n | Y, T) (1 - \Psi_n) \]

\[ = \frac{\mathbb{E}_{\theta, \alpha}[\Psi_n]}{\mathbb{E}_{\theta, \alpha}[\Psi_n]} \rightarrow 0 \quad \text{in} \quad P_{\theta, \alpha} \text{-probability.} \]

Due to assumption 1, we have \( K_{\alpha} P_{\alpha} = P_{\alpha} K_{\alpha} P_{\alpha} = K_{\alpha} P_{\alpha} + m P_{\alpha} \). Hence, restricted on \( f \in V_f \), we obtain
\[
\begin{align*}
  p_{f,\vartheta}(z, w) &= \exp \left( \frac{1}{\varepsilon} \left( (K_{\vartheta,\lambda}f - K_{\vartheta,\lambda}f_0, z) \right) - \frac{1}{2\varepsilon^2} \| K_{\vartheta,\lambda}f - K_{\vartheta,\lambda}f_0 \|_2^2 \\
  &\quad + \frac{1}{\delta} (\vartheta - \vartheta_0, w) - \frac{1}{2\delta^2} \| \vartheta - \vartheta_0 \|_2^2 \right) \\
  &= \exp \left( \frac{1}{\varepsilon} \left( (K_{\vartheta,\lambda}f - K_{\vartheta,\lambda}f_0, z) \right) - \frac{1}{2\varepsilon^2} \| K_{\vartheta,\lambda}f - P_{\vartheta,\lambda}mK_{\vartheta,\lambda}f_0 \|_2^2 \\
  &\quad + \frac{1}{\delta^2} \| (\mathrm{Id} - P_{\vartheta,\lambda}m)K_{\vartheta,\lambda}f_0 \|_2^2 + \frac{1}{\delta} (\vartheta - \vartheta_0, w) - \frac{1}{2\delta^2} \| \vartheta - \vartheta_0 \|_2^2 \right).
\end{align*}
\]

Since we assume that \( K_{\vartheta,\lambda}f_0 \) depends only on \( P_{\vartheta,\lambda}m \vartheta = (\vartheta_1, \ldots, \vartheta_{\lambda+m}) \) and \( \Pi \) is a product prior in \( (\vartheta_\lambda) \), we may rewrite
\[
\begin{align*}
  \Pi_n(f \in V_n : \| f - f_0 \| > D \Sigma_\vartheta(Y, T) (1 - \Psi_n)) \\
  &\leq \frac{\int_{f \in \mathcal{F} \cap V_n, \| f - f_0 \| > D \Sigma_\vartheta, \vartheta \in \Theta} p_{f,\vartheta}^{(h)}(Z, W) \mathrm{d}\Pi_n(f, \vartheta) (1 - \Psi_n)}{\int_{f \in \mathcal{F} \cap V_n, \vartheta \in \Theta} p_{f,\vartheta}^{(h)}(Z, W) \mathrm{d}\Pi_n(f, \vartheta)} \\
  &\leq \frac{p_{f,\vartheta}^{(h)}(Z, W) \mathrm{d}\Pi_n(f, \vartheta) \geq e^{-\left( C_1 + 2 \kappa_2 \sqrt{\varepsilon \delta \lambda} \right)} \geq 1 - \frac{\varepsilon_\lambda + \delta_\lambda}{\kappa_2}}{\Pi_n(B_n)},
\end{align*}
\]

we obtain
\[
\begin{align*}
  \mathbb{P}_{f,\vartheta}(\int_{f \in \mathcal{F} \cap V_n, \vartheta \in \Theta} p_{f,\vartheta}^{(h)}(Z, W) \mathrm{d}\Pi_n(f, \vartheta) \geq e^{-\left( C_1 + 2 \kappa_2 \sqrt{\varepsilon \delta \lambda} \right)} \geq 1 - \frac{\varepsilon_\lambda + \delta_\lambda}{\kappa_2}) \\
  &\leq \mathbb{P}_{f,\vartheta}(\int_{f, \vartheta : \| f - f_0 \| > D \Sigma_\vartheta} p_{f,\vartheta}^{(h)}(Z, W) \mathrm{d}\Pi_n(f, \vartheta) \geq e^{-\left( C_1 + 2 \kappa_2 \sqrt{\varepsilon \delta \lambda} \right)} \geq 1 - \frac{\varepsilon_\lambda + \delta_\lambda}{\kappa_2}) \\
  &\leq \mathbb{P}_{f,\vartheta}(\int_{f, \vartheta : \| f - f_0 \| > D \Sigma_\vartheta} p_{f,\vartheta}^{(h)}(Z, W) \mathrm{d}\Pi_n(f, \vartheta) \geq e^{-\left( C_1 + 2 \kappa_2 \sqrt{\varepsilon \delta \lambda} \right)} \geq 1 - \frac{\varepsilon_\lambda + \delta_\lambda}{\kappa_2}) \\
  &\leq \frac{\varepsilon_\lambda + \delta_\lambda}{\kappa_2}.
\end{align*}
\]
Note that $p_{f,j}^{(m)}$ corresponds to the density of the law of $(Y', T')$ where

$$Y' = P_{\pi, m} K_{\theta} f + (\text{Id} - P_{\pi, m}) K_{\partial_f} f + \varepsilon Z \quad \text{and} \quad T' = P_{\pi, m} \vartheta + (\text{Id} - P_{\pi, m}) \vartheta_0 + \delta W$$

with respect to $\mathbb{P}_{\partial_f} \otimes \mathbb{P}_{\partial_\vartheta}$ and we have $\Psi_n(Y, T) = \Psi_n(Y', T')$ by construction. Therefore, we can apply Step 1 to bound the previous display and conclude

$$\mathbb{P}_{\partial_f} \left( \Pi_n \left( f \in V_n : \|f - f_0\| > D \varepsilon_n, \|Y, T\| \geq r \right) \right) \lesssim \frac{1}{r} e^{-2n^2 / \left( \varepsilon \delta_n \right)^2} + \left( \frac{\varepsilon_n + \delta_n}{\kappa_n} \right)^2 + \mathbb{E}_{\partial_f} \mathbb{E}_{\partial_\vartheta} \left[ \Psi_n \right].$$

(7.6)

It remains to note that for any $r > 0$ the right-hand side converges to zero as $n \to \infty$. □

7.3. Proof of theorem 10

For the sake of brevity we omit the subscript $n$ in the proof. $c_1, c_2, \ldots$ will denote positive, universal constants. We will choose $\kappa, \varepsilon$ and $j = J$ according to

$$\varepsilon \asymp \left( \varepsilon \delta \right)^{2/(2+2+2d)}, \kappa \asymp \left( \varepsilon \delta \right)^{(2+1)/(2+2+2d)}, 2^j = \kappa^{-1/(2+1)},$$

(7.7)

It is not difficult to see that these choices satisfy the requirements of theorem 5 and $\|f_0 - P_j f_0\| \lesssim \varepsilon \delta$ holds by (4.2). Moreover, the support of $\Pi_{j, \delta}$ lies in $V_j$ such that (3.2) is trivially satisfied for $J = \{ f : \|f - P_j f\| \leq C_\delta \}$. It only remains to verify the small ball probability (3.3).

Owing to $P_j K_\beta = P_j K_\beta P_{j+m} = P_j K_\beta P_{j+m}$ (4.2) and $\|K_\beta f\| \lesssim \|f\|_{H^{-\delta}}$, we can estimate for any $f \in V_j$

$$\|P_{j+m} (K_{\beta} f - K_{\beta} f_0)\| \lesssim \|P_{j+m} (K_{\beta} - K_{\beta} f_0) f\| + \|P_{j+m} K_{\beta} f_{j+m} (f - f_0)\|
\lesssim \|P_{j+m} (K_{\beta} - K_{\beta} f_0) f\| + \|P_{j+m} f_{j+m} (f - f_0)\|
\lesssim 2 \|K_{\beta} f_{j+m} - K_{\beta} f_{j+m} f_{j+m} + (f - P_{j+m} f_0)\|_{H^{-\delta}}$$

(7.8)

The last term is bounded by $\| (\text{Id} - P_j) f_0 \|_{H^{-\delta}} \lesssim 2^{-j/(2+1)} \|f_0\|_{H^{-\delta}}$ being of the order $O(\kappa \varepsilon / (\varepsilon \delta))$ due to $\varepsilon \geq \delta$ and the choice $j$ as in (7.7). We obtain

$$\Pi \left( f \in \mathcal{F} \cap V_j, \vartheta \in \Theta : \text{e}^{-2} \|P_{j+m} (K_{\beta} f - K_{\beta} f_0)\|^2 + \delta^{-2} \|P_{j+m} (\vartheta - \vartheta_0)\|^2 \lesssim \kappa^2 / (\varepsilon \delta)^2 \right)$$

$$\geq \Pi \left( f \in \mathcal{F} \cap V_j : \|f - P_j f_0\|_{H^{-\delta}} \leq 2^{-j/(2+1)} \|f_0\|_{H^{-\delta}} \right)$$

(4.4)

where the last line follows from independence of $f$ and $\vartheta$ under $\Pi$. The first term can be bounded using the product structure and the estimate (4.4). Setting $\tilde{\kappa} = \frac{\kappa^2}{\varepsilon \delta}$ and taking $\log \tau_j \leq j$ into account, we obtain

$$\ldots $$

21
\[ \Pi_{\nu} \left( f \in \mathcal{F} \cap V_{j}: \| f - P_{j} f \|_{H^{-1}} \leq \frac{c_{1} \epsilon_{K}}{\sqrt{2(\epsilon \vee \delta)}} \right) \\
= \Pi_{\nu} \left( f \in \mathcal{F} \cap V_{j}: \sum_{|k| \leq j} 2^{-2|k|} (f_{k} - f_{0,k})^{2} \leq \frac{\epsilon^{2} \kappa^{2}}{2} \right) \\
\geq \prod_{|k| \leq j} \Pi_{\nu} \left( |f_{k} - f_{0,k}| \leq c_{2} \epsilon^{2} (\delta / |k|^{2}) \right) \\
\geq \exp \left( c_{d} \epsilon^{2d} \log(2\Gamma) + c_{\gamma} \sum_{|k| \leq j} (2\epsilon - d|k|) \log \tau_{|k|} + \log \kappa \right) - 2\gamma \sum_{|k| \leq j} \tau_{|k|}^{-2} (|f_{0,k}|^{2} + c_{3} \epsilon^{2} 2^{(2\epsilon - d)|k|}) \\
\geq \exp \left( c_{d} \epsilon^{2d} (\log \kappa - \log 2) - 2\gamma \max_{|k| \leq j} (2\epsilon - 2d|k| - 2\log(\epsilon \vee \delta)) \right).
\]

Since \( \kappa \simeq 2^{-j(\epsilon + \delta)} \), we have \( \log \kappa^{-1} \lesssim j + \log \frac{\epsilon \vee \delta}{\epsilon} \lesssim j \). From the assumptions on \( \tau_{j} \) we thus deduce
\[ \Pi_{\nu} \left( f \in \mathcal{F} \cap V_{j}: \| f - P_{j} f \|_{H^{-1}} \leq \frac{c_{1} \epsilon_{K}}{\sqrt{2(\epsilon \vee \delta)}} \right) \geq e^{c_{d} \epsilon^{2d} (\log \kappa - j)} \geq e^{-c_{d} \epsilon^{2d} j}. \]  

(7.9)

By the the Lipschitz continuity \( \| K_{\delta f} - K_{\delta \theta} \|_{V_{j} \rightarrow V_{j}} \lesssim \| P_{j} (\theta - \theta_{0}) \|^{2} \), the second term in (7.8) is bounded by
\[ \Pi_{\nu} \left( \left( \frac{1}{\epsilon} + \frac{1}{\delta} \right) \| P_{j+2m} (\theta - \theta_{0}) \|^{2} \leq \frac{c_{7} \epsilon^{2}}{(\epsilon \vee \delta)^{2}} \right) \geq \Pi_{\nu} \left( \| P_{j+2m} (\theta - \theta_{0}) \| \leq \frac{c_{7} \epsilon \vee \delta}{\epsilon \vee \delta} \right).
\]

Due to assumption 9 and using again (4.4), we can estimate for \( \tilde{\kappa} = \frac{(\epsilon \vee \delta) \kappa}{\epsilon \vee \delta} \),
\[ \Pi_{\nu} \left( \| P_{j+2m} (\theta - \theta_{0}) \| \leq c_{7} \tilde{\kappa} \right) \\
\geq \prod_{|k| \leq j+2m} \Pi_{\nu} \left( |\theta_{k} - \theta_{0,k}| \leq c_{7} \tilde{\kappa} / \sqrt{j+2m} \right) \\
\geq \exp \left( c_{d} j+2m \log(2\Gamma) + c_{\gamma} j+2m \log \kappa - \frac{c_{\gamma}}{2} j+2m \log j+2m - \gamma \sum_{|k| \leq j+2m} (|\theta_{0,k}| + c_{7} \tilde{\kappa} / \sqrt{j+2m})^{2} \right) \\
\geq \exp \left( c_{d} j+2m \log \kappa - c_{d} j+2m \log j+2m - c_{\gamma} \| \theta_{0} \|^{2} - c_{9} \kappa^{2} \right) \\
\geq \exp \left( - c_{10} j+2m \log(\kappa^{-1}) + \log j+2m \right),
\]

where we have used in the last step that \( \kappa \leq \kappa \rightarrow 0 \). Because \( \epsilon \theta \lesssim \delta \) implies \( \log \kappa^{-1} \lesssim j + \log \frac{\epsilon \vee \delta}{\epsilon} \lesssim j \), we find in combination with \( \log j+2m \lesssim j \) that
\[ \Pi_{\nu} \left( \| P_{j+2m} (\theta - \theta_{0}) \| \leq c_{7} \kappa \right) \geq e^{-c_{10} j+2m} \geq e^{-c_{12} \epsilon^{2d} j}. \]  

(7.10)

Therefore, (3.3) follows from \( j2^{d} \lesssim \kappa^{2} (\epsilon \vee \delta)^{-2} \), which is satisfied due to (7.7), in combination with (7.8)–(7.10).

\[ \Box \]

7.4. Proof of theorem 12

The proof is similar to the previous one. The choices of \( \kappa, \xi \) and \( j \) given by
\[ \xi \simeq \left( \log \frac{1}{\epsilon \vee \delta} \right)^{-s/\gamma}, \quad \kappa \simeq (\epsilon \vee \delta)^{1/2}, \quad 2^{j} = \left( - \frac{1}{2r} \log \left( \frac{\kappa \xi}{\epsilon \vee \delta} \right) \right)^{1/r} \]  

(7.11)
satisfy the conditions of theorem 5. Especially, we have \[ \|f_0 - P_j f_0\| \lesssim 2^{-j\varepsilon}\|f_0\| \lesssim \xi \] and \[ 2^j = 2^{j(1 + o(1))} \] because of \( \log \varepsilon \simeq \log \delta \). Since \( \|K_0 f\|^2 \lesssim \sum_k e^{-2\varepsilon}f_k^2 \), we estimate for any \( f \in \mathcal{F} \cap V_j \)

\[
\|P_{j+m}(K_0 f - K_0 f_0)\|^2 \\
\leq 2\|\langle \theta, \phi_0 \rangle\| + 2\|K_0 f - f_0\|^2 + \sum_k e^{-2\delta\varepsilon}e^{2\delta\varepsilon}\|f_k - f_0, \phi_k\|^2 \\
\leq 2\|\langle \theta, \phi_0 \rangle\| + 2\|K_0 f - f_0\|^2 + 4\sum_{|k| \leq j} e^{-2\delta\varepsilon}f_k^2 + 4\sum_{|k| > j} e^{-2\delta\varepsilon}f_k^2.
\]

Using

\[
\sum_{|k| > j} e^{-2\delta\varepsilon}f_k^2 \leq 2^{-2\delta\varepsilon}e^{-2\delta\varepsilon}\|f\|^2 \lesssim e^{-2\delta\varepsilon},
\]

together with the choice \( j \) from (7.11), the last term in (7.12) is \( \mathcal{O}(\kappa^2/e) \). Analogously to (7.8) we obtain for some \( c_1 > 0 \)

\[
\Pi \left( f \in \mathcal{F} \cap V_j : \left| \frac{1}{e^\varepsilon} \right| \|P_{j+m}(K_0 f - K_0 f_0)\|^2 + \frac{1}{e^\varepsilon} \|P_{j+m}(\theta - \phi_0)\|^2 \leq \frac{\kappa^2}{(e^\varepsilon \delta)} \right) \\
\geq \prod_{|k| \leq j} \Pi\left( f \in \mathcal{F} \cap V_j : e^{-2\delta\varepsilon}f_k^2 \leq c_1 \frac{\kappa^2}{(e^\varepsilon \delta)} \right) \\
\times \prod_{|k| > j} \left( \frac{1}{e^\varepsilon} \|K_0 f - f_0\| \leq c_2 \frac{\kappa^2}{(e^\varepsilon \delta)} \right) \\
\geq e^{-c_2 \kappa^2}.
\]

The second factor is the same as in the proof of theorem 10. Taking into account that \( \log \delta \simeq \log \varepsilon \) and (7.11) imply \( -\log (e^\varepsilon/\delta) \kappa = -\log \kappa = -\log (e^\varepsilon/\delta) \simeq 2^\theta \), we find

\[
\Pi\left( \frac{1}{e^\varepsilon} \|K_0 f - f_0\| \leq c_2 \frac{\kappa^2}{(e^\varepsilon \delta)} \right) \\
\geq e^{-c_2 \kappa^2}.
\]

Setting \( \bar{\kappa} = \frac{e^\varepsilon}{e^\varepsilon \delta} \) and applying (4.4), we obtain for the first term

\[
\Pi\left( f \in \mathcal{F} \cap V_j : \sum_{|k| \leq j} e^{-2\delta\varepsilon}f_k^2 \leq \frac{c_1 e^\varepsilon \delta}{(e^\varepsilon \delta)} \right) \\
\geq \prod_{|k| \leq j} \Pi\left( |f_k - f_0| \leq c_3 \bar{\kappa} 2^{-d|k|/2} e^{2\delta\varepsilon} \right) \\
\geq \exp \left( c_4 2^{j\varepsilon} \log(2\Gamma) + c_4 \sum_{|k| \leq j} (r^2|k| - d|k| - \log \tau|k| + \log \bar{\kappa}) \\
- 2 \gamma \sum_{|k| \leq j} \tau^{-2}|k| (|f_0, \phi_k|^2 + c_3 \bar{\kappa}^2 2^{-d|k|} e^{2\delta\varepsilon}) \right) \\
\geq \exp \left( c_5 2^{j\varepsilon} (\log \bar{\kappa} - 2^\theta) - c_5 \max_{l \in l_\ell} (2^{-2d^\theta \gamma - 2} \|f_0\|^2_{P_l} - c_5 \bar{\kappa}^2 \tau^{-2} e^{2\delta\varepsilon}) \right).
\]
From the assumptions on \( \tau_j \) and \(- \log \delta \lesssim 2^t\) we thus deduce
\[
\Pi_f \left( f \in \mathcal{F} \cap V_j : \sum_{|k| \leq j} e^{-2^j t} |f - f_0, \varphi_k| \leq \frac{\delta_k}{\delta} \right) \geq c e^{2^j (\log \delta - 2^t)} \geq c e^{-2^j (\log \delta)}.
\]

Combining (7.14) and (7.15) yields
\[
\Pi \left( f \in \mathcal{F} \cap V_j, \vartheta \in \Theta : \frac{1}{\vartheta} \left| \|f_{j+1} - f_{j}, \vartheta\|_2 \right|^2 + \frac{1}{\vartheta} \left| \|f_{j+1} - \vartheta\|_2 \right|^2 \leq \frac{\kappa}{(e \vee \delta)^2} \right) \geq e^{-2^j (\log \delta)^2}.
\]

Therefore, (3.3) follows from \( \epsilon 2^{(t^2 + q)} \lesssim \log(\kappa^{-1}) \lesssim \kappa^2 (e \vee \delta)^{-2} \) by the choice of \( \kappa \) from (7.11). \( \square \)

75. Proof of theorem 13

Let us introduce the oracle which balances the bias and the variance term:
\[
J_o := \min \{ j \leq J : R 2^{-j} \leq CR \log(1/\epsilon) \epsilon 2^{(t^2 + q/2)} \}
\]

where \( C \) is the constant from proposition 6 and \( R \) is the radius of the Hölder ball. As \( \epsilon \to 0 \) we see that
\[
2^{J_0} \lesssim (\epsilon (\log \epsilon^{-1}))^{-2(t^2 + q)}
\]

which coincides with the choice of \( J \) in the proof of theorem 10. The rest of the proof is divided into three steps.

Step 1: We will proof that \( \hat{J} \leq J_o \) with probability approaching one. We have for sufficiently small \( \epsilon \)
\[
| \mathbb{P}_{f_0, \|f_0\|} (\hat{J} > J_o) = \mathbb{P}_{f_0, \|f_0\|} (\exists i > J_o : \|\hat{f}_i - \hat{f}_j\| > \Delta \epsilon (\log \epsilon^{-1})^2 2^{(t^2 + q/2)}) \leq \sum_{i > j > J_o} \mathbb{P}_{f_0, \|f_0\|} (\|\hat{f}_i - \hat{f}_j\| > \Delta \epsilon (\log \epsilon^{-1})^2 2^{(t^2 + q/2)}) \leq \sum_{i > j > J_o} \mathbb{P}_{f_0, \|f_0\|} (\|\hat{f}_i - f_0\| + \|\hat{f}_j - f_0\| > \Delta \epsilon (\log \epsilon^{-1})^2 2^{(t^2 + q/2)}) \leq 2J_0 \sum_{i > J_o} \mathbb{P}_{f_0, \|f_0\|} (\|\hat{f}_i - f_0\| > 2CR \epsilon (\log \epsilon^{-1})^2 2^{(t^2 + q/2)}).
\]

By definition of \( J_o \) we have for every \( j \geq J_o \) and \( f_0 \in H^t(R) \) that
\[
\|f_0 - P_{f_0}\| \leq R 2^{-j} \leq CR \log(1/\epsilon) \epsilon 2^{(t^2 + q/2)}. \]

Hence, for \( \epsilon \) sufficiently small we obtain
\[
\mathbb{P}_{f_0, \|f_0\|} (\hat{J} > J_o) \leq 2J_0 \sum_{i > J_o} \mathbb{P}_{f_0, \|f_0\|} (\|\hat{f}_i - f_0\| > C \|f_0\| \epsilon (\log \epsilon^{-1}) 2^{(t^2 + q/2)} + \|f_0 - P_{f_0}\|).
\]

For any \( j \leq J \) we then have \( \epsilon 2^{(t^2 + q/2)} \to 0 \) and the concentration inequality from proposition 6 can be applied to \( \hat{f}_i \) for any \( \kappa \in (C^{-1} 2^{(t^2 + q/2)} , C2^{-t^2/2} \epsilon^{-1}) \) for a certain constant \( C > 0 \). We can choose \( \kappa = 2^{t^2/2} \epsilon (\log \epsilon^{-1}) \) to obtain
\[
\mathbb{P}_{f_0, \|f_0\|} (\hat{J} > J_o) \leq 6J_0^2 \epsilon^{-2(t^2 + q/2) \log \epsilon} \leq 6J_0^2 \epsilon \to 0.
\]
**Step 2:** In order to prove the adaptive contraction rate, we replace the test $\Psi_n$ from the proof of theorem 5 by

$$\hat{\Psi} := 1_{\{\|\hat{f}_j - f_0\| \geq 2\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}\}}$$

requiring to verify (7.3) for $\hat{\Psi}$ and

$$\kappa = (\varepsilon \log (1/\varepsilon))^{2J_o(t+d/2)} \quad \xi = (\log \varepsilon^{-1})^{2J_o(t+d/2)}.$$  

(7.16)

Note that $\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)} \simeq (\log \varepsilon^{-1})^{2J_o(t+d/2)}$ by the choice of the oracle $J_o$.

Thanks to Step 1 we have

$$\mathbb{P}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq 2\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)})$$

$$\leq \mathbb{P}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq 2\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}, J \leq J_o) + 6J_o^2 \varepsilon$$

$$\leq \mathbb{P}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq 2\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}, J \leq J_o) + 6J_o^2 \varepsilon.$$

By construction of $J$ we have $|\hat{f}_j - \hat{f}_0| \leq \varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}$ on the event $\{J \leq J_o\}$. Therefore,

$$\mathbb{E}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq \varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}) \leq 6J_o^2 \varepsilon$$

$$\leq \mathbb{P}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq \varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}) + 6J_o^2 \varepsilon$$

$$\leq 3 \varepsilon + 6J_o^2 \varepsilon \to 0$$

where the last bound follows from proposition 6 exactly as in Step 1. For any $f \in \mathcal{F}_n$ with $\|f - f_0\| \geq C_1 \varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}$ for an sufficiently large constant $C_1$ and $\vartheta \in \Theta$ we obtain on the alternative with an argument as in (7.4)

$$\mathbb{E}_{\vartheta, \vartheta}^f[1 - \hat{\Psi}] = \mathbb{P}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq 2\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)})$$

$$\leq \mathbb{P}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq 2\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}, J \leq J_o) + 6J_o^2 \varepsilon$$

$$\leq \mathbb{P}_{\rho_o, \theta_o}(|\hat{f}_j - f_0| \geq 2\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)}, J \leq J_o) + 6J_o^2 \varepsilon.$$

for some $C_2 \varepsilon > 0$. Since $J_o \vartheta \simeq \log \varepsilon^{-1}$ and

$$\varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)} \simeq \varepsilon (\log \varepsilon^{-1})^{2J_o(t+d/2)} \simeq \kappa^2 \varepsilon^{-2},$$

we indeed have $\mathbb{E}_{\vartheta, \vartheta}^f[1 - \hat{\Psi}] \leq e^{-C \varepsilon^2 \varepsilon^{-2}}$ for some constant $C > 4$.

**Step 3:** With the previous preparations we can now prove the adaptive contraction result. Given $\Psi_n$, we have for any $r > 0$ and $\xi$ from (7.16)

$$\mathbb{P}_{\rho_o, \theta_o}(\Pi_n(f \in \mathcal{F} : \|f - f_0\| > M\xi] (Y, T) > r)$$

$$\leq \mathbb{P}_{\rho_o, \theta_o}(\Pi_n(f \in \mathcal{F} \cap V_j : \|f - f_0\| > M\xi] (1 - \hat{\Psi}) > r, j \leq J_o) + 6J_o^2 \varepsilon$$

$$\leq \sum_{j \leq J_o} \mathbb{P}_{\rho_o, \theta_o}(\Pi_n(f \in \mathcal{F} \cap V_j : \|f - f_0\| > M\xi] (1 - \hat{\Psi}) > r, j = j) + 6J_o^2 \varepsilon.$$

We can now handle each term in the sum exactly as in the proof of theorem 5. It suffices to note that: First, $\hat{\Psi}$ depends only on the $J = j$ projection of $Y$ and $j + m$ projection of $\vartheta$, respectively. Second, if the small ball probability condition (3.3) is satisfied for $\Psi_n$, as verified
in the proof of theorem 10, than by monotonicity it is also satisfied for all \( j \leq J_0 \). We thus conclude from (7.6)

\[
\mathbb{P}_{\theta_0, \theta_0} \left( \Pi_n \left( f \in \mathcal{F} : \| f - f_0 \| > M \xi | Y, T \right) > r \right) \leq \frac{J_0}{r} e^{-2 \kappa^2 \epsilon^2} + J_0 \frac{\epsilon^2}{\kappa^2} + 6J_0^2 \epsilon \to 0.
\]

\[\square\]

**Acknowledgment**

This work is the result of three conferences which I attended in 2017. The first two on Bayesian inverse problems in Cambridge and in Leiden led to my interest for this problem. On the third conference in Luminy on the honour of Oleg Lepski’s and Alexandre B Tsybakov’s 60th birthday, I realised that Lepski’s method can be used to construct an empirical Bayes procedure. I want to thank the organisers of these three conferences. The helpful comments by two anonymous referees are gratefully acknowledged.

**ORCID IDs**

Mathias Trabs © https://orcid.org/0000-0001-8104-4467

**References**

[1] Agapiou S, Larsson S and Stuart A M 2013 Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems *Stoch. Process. Appl.* 123 3828–60

[2] Agapiou S, Stuart A M and Zhang Y X 2014 Bayesian posterior contraction rates for linear severely ill-posed inverse problems *J. Inverse Ill-Posed Probl.* 22 297–321

[3] Bochkina N 2013 Consistency of the posterior distribution in generalized linear inverse problems *Inverse Problems* 29 095010

[4] Burger M and Scherzer O 2001 Regularization methods for blind deconvolution and blind source separation problems *Math. Control Signals Syst.* 14 358–83

[5] Cavalier L 2008 Nonparametric statistical inverse problems *Inverse Problems* 24 034004

[6] Cavalier L and Hengartner N W 2005 Adaptive estimation for inverse problems with noisy operators *Inverse Problems* 21 1345–61

[7] Cohen A, Hoffmann M and Reiβ M 2004 Adaptive wavelet Galerkin methods for linear inverse problems *SIAM J. Numer. Anal.* 42 1479–501

[8] Da Prato G 2006 *An Introduction to Infinite-Dimensional Analysis* (Universitext) (Berlin: Springer) (revised and extended from the 2001 original by Da Prato)

[9] Dashti M, Law K J H, Stuart A M and Voss J 2013 MAP estimators and their consistency in Bayesian nonparametric inverse problems *Inverse Problems* 29 095017

[10] Dattner I, Reiβ M and Trabs M 2016 Adaptive quantile estimation in deconvolution with unknown error distribution *Bernoulli* 22 143–92

[11] Efromovich S and Koltchinskii V 2001 On inverse problems with unknown operators *IEEE Trans. Inform. Theory* 47 2876–94

[12] Ghosal S, Ghosh J K and van der Vaart A W 2000 Convergence rates of posterior distributions *Ann. Stat.* 28 500–31

[13] Giné E and Nickl R 2011 Rates on contraction for posterior distributions in \( L^r \)-metrics, \( 1 \leq r < \infty \) *Ann. Stat.* 39 2883–911

[14] Giné E and Nickl R 2016 *Mathematical Foundations of Infinite-Dimensional Statistical Models* (Cambridge Series in Statistical and Probabilistic Mathematics) (Cambridge: Cambridge University Press)

[15] Gugushvili S, van der Vaart A and Yan D 2018 Bayesian linear inverse problems in regularity scales (arXiv:1802.08992)
[16] Hoffmann M and Reiß M 2008 Nonlinear estimation for linear inverse problems with error in the operator Ann. Stat. 36 310–36
[17] Johannes J 2009 Deconvolution with unknown error distribution Ann. Stat. 37 2301–23
[18] Johannes J, Van Bellegem S and Vanhems A 2011 Convergence rates for ill-posed inverse problems with an unknown operator Econometric Theory 27 522–45
[19] Johnstone I M, Kerkyacharian G, Picard D and Raimondo M 2004 Wavelet deconvolution in a periodic setting J. R. Stat. Soc. B 66 547–73
[20] Justen L and Ramlau R 2006 A non-iterative regularization approach to blind deconvolution Inverse Problems 22 771–800
[21] Knapi B T, Szabó B T, van der Vaart A W and van Zanten J H 2016 Bayes procedures for adaptive inference in inverse problems for the white noise model Probab. Theory Relat. Fields 164 771–813
[22] Knapi B T, van der Vaart A W and van Zanten J H 2011 Bayesian inverse problems with Gaussian priors Ann. Stat. 39 2626–57
[23] Knapi B T, van der Vaart A W and van Zanten J H 2013 Bayesian recovery of the initial condition for the heat equation Commun. Stat. Theory Methods 42 1294–313
[24] Lepski O V 1990 A problem of adaptive estimation in Gaussian white noise Teor. Veroyatn. Primen. 35 459–70
[25] Marteau C 2006 Regularization of inverse problems with unknown operator Math. Methods Stat. 15 415–43
[26] Neumann M H 1997 On the effect of estimating the error density in nonparametric deconvolution J. Nonparameter Stat. 7 307–30
[27] Nickl R 2017 Bernstein-von mises theorems for statistical inverse problems I: Schrödinger equation (arXiv:1707.01764)
[28] Nickl R and Söhl J 2017 Bernstein-von mises theorems for statistical inverse problems II: compound poisson processes (arXiv:1709.07752)
[29] Nickl R and Söhl J 2017 Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions Ann. Stat. 45 1664–93
[30] Ray K 2013 Bayesian inverse problems with non-conjugate priors Electron. J. Stat. 7 2516–49
[31] Stuart A M 2010 Inverse problems: a Bayesian perspective Acta Numer. 19 451–559
[32] Stück R, Burger M and Hohage T 2012 The iteratively regularized Gauss–Newton method with convex constraints and applications in 4Pi microscopy Inverse Problems 28 015012
[33] Tao T 2012 Topics in random matrix theory (Graduate Studies in Mathematics vol 132) (Providence, RI: American Mathematical Society)
[34] Tierney L 1994 Markov chains for exploring posterior distributions Ann. Stat. 22 1701–62 (with discussion and a rejoinder by the author)
[35] Vollmer S J 2013 Posterior consistency for Bayesian inverse problems through stability and regression results Inverse Problems 29 125011