ADAPTIVE ESTIMATION FOR AFFINE STOCHASTIC
DELAY DIFFERENTIAL EQUATIONS

MARKUS REISS
Weierstraß Institute (WIAS)
of Applied Analysis and Stochastics
Mohrenstraße 39
D-10117 Berlin (Germany)
mreiss@wias-berlin.de

Abstract. For stationary solutions of the affine stochastic delay differential equation
\[ dX(t) = \left( \gamma_0 X(t) + \gamma_r X(t - r) + \int_{-r}^{0} X(t + u)g(u)\,du \right) dt + \sigma \, dW(t) \]
we consider the problem of nonparametric inference for the weight function \( g \) and for \( \gamma_0, \gamma_r \) from the continuous observation of one trajectory up to time \( T > 0 \). For weight functions in the scale of Besov spaces \( B_{p,1}^s \) and \( L^p \)-type loss functions convergence rates are established for long time asymptotics. The estimation problem is equivalent to an ill-posed inverse problem with error in the data and unknown operator. We propose a wavelet thresholding estimator that achieves the rate \( (T/\log T)^{-s/(2s+3)} \) under certain restrictions on \( p \) and \( \rho \). This rate is shown to be optimal in a minimax sense.

Mathematics subject classification (2000): 62M09, 62G20, 60G15
Key words: wavelet thresholding, ill-posed inverse problem, Besov space, spatial adaptivity, minimax rates

1. Introduction

1.1. The model. Stochastic delay differential equations (SDDEs for short) appear naturally in the description of many processes, e.g. in population dynamics with a time lag due to an age-dependent birth rate (Scheutzow 1984), in economics where a certain “time to build” is needed (Kydland and Prescott 1982), in finance (Hobson and Rogers 1998) and in many engineering applications, see Kolmanovskii and Myshkis (1992) for an overview. They
are also obtained as continuous-time limits of time series models. Among the huge variety of types of equations, the so-called affine stochastic delay differential equations form the fundamental class. They generalize the Langevin equation leading to the Ornstein-Uhlenbeck process and appear as continuous-time limits of linear autoregressive schemes. A general scalar affine SDDE is of the form

\[
\begin{align*}
\frac{dX(t)}{dt} &= \left( \int_{-r}^{0} X(t + u) \, da(u) \right) \, dt + \sigma \, dW(t), \quad t \geq 0, \\
X(u) &= F(u), \quad u \in [-r, 0].
\end{align*}
\]

The drift coefficient depends linearly on the past trajectory \((X(u), u \in [t - r, t])\) by means of an integration with respect to the finite signed Borel measure \(a\) on \([-r, 0]\). The values \(r\) and \(\sigma\) are supposed to be positive and \((W(t), t \geq 0)\) denotes a standard Wiener process adapted to a filtration \((\mathcal{F}_t)_{t \geq 0}\), satisfying the usual conditions. In order to ensure well-posedness of the differential equation, a whole initial function \(F\) independent of the Wiener process is prescribed. The Langevin equation without memory effect is obtained if \(a\) is taken to be a point measure in zero.

The asymptotic properties and the existence of stationary solutions for affine SDDEs, even with more general driving processes, have been studied in detail by Mohammed, Scheutzow, and von Weizsaecker (1986) and Gushchin and Küchler (2000). Our goal here is to estimate the weight measure \(a\) nonparametrically from the observation \((X(t), t \in [-r, T])\) of one realization of a stationary solution to (1.1). In many applications the drift term consists of an instantaneous feedback and a distributed delayed feedback. Having nonparametric estimation and these kind of models in mind, our parameter class should thus contain weight measures \(a\) with a Lebesgue density except for a possible point mass in zero. Our estimation procedure will more generally cope with weight measures having a Lebesgue density in the interior and possible point masses in both interval endpoints \(-r\) and 0.
1.2. **The estimation procedure.** The main idea for the construction of the estimator is that the weight measure \( a \) solves the integral equation

\[
Q_a a(t) := \int_{-r}^{0} q_a(t-u) \, da(u) = q'_a(-t), \quad t \geq 0, \tag{1.3}
\]

where the function \( q_a \) denotes the covariance function of the process \( X \).

From observations up to time \( T \) we can construct estimators of the kernel of the operator \( Q_a \) and of the data \( q'_a \) and solve the integral equation for \( a \). We shall show that \( Q_a \) is smoothing like twofold integration independently of the weight \( a \) such that the solution of equation (1.3) does not depend continuously on the data. We face an ill-posed inverse problem of degree two with errors occurring in the operator and in the right hand side.

A linear estimation technique for \( L^2 \)-risk and weight functions in \( L^2 \)-Sobolev balls of regularity \( s > \frac{1}{2} \) has been presented in Reiß (2002), which is close in spirit to the work by Efroimovich and Koltchinskii (2001). The ill-posed inverse problem has been solved by the Galerkin projection method and a minimax risk of order \( T^{-s/(2s+3)} \) for observation times \( T \to \infty \) has been established. Here, we strive for adaptive estimation, that is we do not suppose the regularity of the unknown weight function to be known and we automatically adapt to spatial inhomogeneity of the function. Moreover, we allow for more general \( L^\rho \)-loss, \( \rho \in (1, \infty) \). As usually in adaptive estimation theory, we are lead to consider density functions in Besov spaces \( B^s_{p,\alpha}([-r,0]) \) and to use nonlinear approximation techniques. Under suitable conditions on \( p \) and \( \rho \) we shall find for our adaptive estimator an asymptotic risk of order \( (T/ \log T)^{-s/(2s+3)} \), which will be shown to be minimax with respect to the Besov classes considered.

For the construction of the estimator wavelet thresholding techniques in a suitable image domain are used. Our approach is related to the wavelet-vaguelette decomposition and vaguelette-wavelet decomposition methods proposed by Donoho (1995) and Abramovich and Silverman (1998). In fact, the latter paper presents the main idea: we first threshold the wavelet coefficients and then invert the approximation of equation (1.3). Since our operator is not exactly known (and for each observation different) the inversion should not be performed by calculating the corresponding vaguelettes.
exactly, but rather by applying a numerical inversion algorithm. For this we can allow for numerical errors up to the order of the statistical error in the first step and even rely on adaptive procedures, see Cohen, Hoffmann, and Reiß (2004) for a suitable algorithm.

Mathematically speaking, we denoise the data in a certain Sobolev space along the lines of the abstract results obtained by Kerkyacharian and Picard (2000) for heteroscedastic noise. For the right choice of the thresholding level and for theoretical purposes the mapping properties of the covariance operator have to be studied in detail.

1.3. Related approaches. Except for the afore-mentioned paper (Reiß 2002), statistical inference for the weight measure in affine SDDEs has so far focused on parametric models, e.g. Kutoyants, Mourid, and Bosq (1992), Gushchin and Küchler (1999) and Gushchin and Küchler (2003), where for sufficiently smooth parametrisations of the weight measure $a = a_\theta$ the classical LAN-property with rate $T^{-1/2}$ holds under stationarity assumptions. On the other hand, nonparametric and even adaptive estimation of the drift coefficient $b$ in ergodic diffusions

$$dX(t) = b(X(t)) \, dt + \sigma \, dW(t), \quad t \in [0, T],$$

is well established (Hoffmann 1999, Dalalyan 2001). Albeit a similar structure of the estimation problem, under recurrency conditions the minimax rate for estimating drift functions $b$ of regularity $s$ is $T^{-s/(2s+1)}$, indicating a close relationship with classical regression estimation. In our SDDE case the worse, because smaller, exponent $s/(2s+3)$ can be explained intuitively by the presence of an integration in the drift term, which leads to additional smoothing of the observation and thus makes the inference more difficult.

More correctly, the interpretation via the ill-posed problem (1.3) and the framework developed by Nussbaum and Pereverzev (1999) allows a deeper understanding of this rate. In a rough summary, the model consists of a linear operator $K$ on $L^2([0, 1])$ that is regularizing of degree $\tau$ and a centred Gaussian noise $\Gamma$ on $L^2([0, 1])$ with regularity $\gamma$ ($\gamma = 0$ corresponds to white noise, $\gamma = 1$ to Brownian motion and so forth). Under suitable conditions
their main result states that functions $f$ from a Sobolev ball of regularity $s$ can be estimated from the observation $Y_{\varepsilon} = Kf + \varepsilon \Gamma$ with an $L^2$-risk of order $\varepsilon^{2s/(2s+2(\tau-\gamma)+1)}$. The values $\tau = \gamma = 0$ correspond to the classical white noise model for direct observations, which should still be coined statistically ill-posed because convergence rates can only be ensured under compactness assumptions on the function class. Translated to our situation, we set formally $\varepsilon = T^{-1/2}$, $K = Q_a$ and $\tau = 2$ and the noise process has regularity $\gamma = 1$ (Theorem 3.2) yielding the announced exponent $s/(2s+2(2-1)+1)$, though it should be emphasized that neither the conditions like exact knowledge of the operator and the Gaussian noise structure nor the methods of proof using singular value decomposition can be adapted suitably.

Ill-posed problems with stochastic error structure have attracted increasing attention recently and the case of weight estimation in SDDEs furnishes a good example from statistical inference with the special feature of a merely approximately known operator and dependent non-Gaussian noise structure. The archetype of inverse problems in statistics is density deconvolution, where ill-posedness arises due to the smoothness of the contaminating noise density. Our problem might be compared to the case of an error in this density, see Butucea and Matias (2003) for a study in the case of an unknown scaling parameter. For idealized situations, like the Gaussian model sketched above, adaptive estimation procedures based on wavelet thresholding or general oracle inequalities have been obtained, see for instance Donoho (1995), Kalifa and Mallat (2003), Cavalier and Tsybakov (2002) or Cavalier, Golubev, Picard, and Tsybakov (2002).

Inverse problems with approximately known operator have been considered in a deterministic setting already by Hämärik (1983) and recently in a stochastic framework by Efroimovich and Koltchinskii (2001) and Cavalier and Hengartner (2002). The latter, however, suppose all unknown operators to have the same eigenfunctions and can thus transfer their problem to a sequence space model. Although the operators $Q_a$ are of convolution-type, their kernels are not periodic and their eigenfunctions are not given by the Fourier basis, but depend on the weight $a$ such that this approach is not
feasible in our case. Most closely, the general procedure of Efro-movich and Koltchinskii (2001) resembles the way we proceed. They also use suitable basis functions and establish the same risk bounds as in the case of known operators whenever the level of the noise in the operator is not larger than that of the noise in the data. We also recover the same minimax rates although we do not dispose of a training sample and the two noise sources are correlated. It is likely that we loose in terms of constants, at least in the proof the two errors just sum up. The main difference in the estimation procedure is that we use a nonlinear projection method and can thus adapt to spatial inhomogeneity whereas Efro-movich and Koltchinskii (2001) employ a linear Lepski method.

1.4. Structure and notation. In Section 2 we introduce the theory of affine SDDEs and their stationarity behaviour and present results on the mapping properties of the covariance operator. Section 3 is devoted to the construction of the estimator and a discussion of its properties. In Section 4 we assess the optimality of our estimator by the minimax approach. The proofs of the statements are deferred to Sections 5 to 7, Section $n$ providing the proofs for Section $n−3$. In the appendix we have collected some essentials on function spaces and wavelet bases.

Let us fix some notation. $\mathbb{P}_a(\cdot)$ and $\mathbb{E}_a[\cdot]$ denote the probability measure and the expectation operator depending on the parameter $a$. By $\mathcal{F}_T^X$ we denote the $\sigma$-field generated by the process $X$ up to time $T$. The space of continuous (resp. $p$-integrable) functions on the interval $I$ is denoted by $C(I)$ (resp. $L^p(I)$). The space of finite signed Borel measures on $I$ is written as $\mathcal{M}(I)$ and equipped with the total variation norm $\|\cdot\|_{TV}$. $\delta_x$ is the Dirac measure in $x$ and $g \circ \lambda$ denotes the measure with Lebesgue density $g \in L^1$. Usually, the density $g$ is identified with the measure $g \circ \lambda$ and thus operators acting on measures are considered to act on the densities itself. For $f \in C(I)$ and $\mu \in \mathcal{M}(I)$ we introduce the dual pairing $\langle f, \mu \rangle := \int_I f \, d\mu$. The cardinality of a set $M$ is denoted by $|M|$. Finally, the symbol $A(T) \lesssim B(T)$ means that $A(T)$ is bounded by a multiple of $B(T)$ independently of $T$, that
is $A(T) = O(B(T))$ in the $O$-notation. Equally, $A(T) \gtrsim B(T)$ stands for $B(T) \lesssim A(T)$ and $A(T) \sim B(T)$ for $A(T) \lesssim B(T)$ as well as $A(T) \gtrsim B(T)$.

2. Affine SDDE

For the theory of deterministic delay equations we refer to the monograph by Hale and Verduyn Lunel (1993), whereas fundamental results on stochastic delay equations can be found in Mohammed (1984) and Mao (1997). If we put $\sigma = 0$ in (1.1), we obtain the deterministic linear delay equation

$$
\dot{x}(t) = \int_{-r}^{0} x(t + u) \, da(u), \quad t \geq 0.
$$

(2.1)

As for linear ODEs the ansatz $x(t) = e^{\lambda t}$ gives rise to a characteristic function the zeros of which determine the long-time behaviour of general solutions $x$.

**Definition 2.1.** The characteristic function associated to (1.1) is given by

$$
\chi_{a}(\lambda) := \lambda - \int_{-r}^{0} e^{\lambda u} \, da(u), \quad \lambda \in \mathbb{C}.
$$

The maximal real part of its zeros is denoted by

$$
v_{0}(a) := \sup \{ \text{Re}(\lambda) | \chi_{a}(\lambda) = 0 \}.
$$

Without loss of generality we shall henceforth assume $\sigma = 1$; otherwise we rescale $X$ and consider $\tilde{X}(t) = \sigma^{-1}X(t)$ instead. Gushchin and Küchler (2003) then prove the following result:

**Theorem 2.2.** A stationary solution of the affine SDDE (1.1) exists if and only if $v_{0}(a) < 0$ holds. In this case the stationary solution $X$ is unique. It is a centered Gaussian process with (auto)covariance function $q_{a}(t) := \mathbb{E}_{a}[X(0)X(|t|)]$, $t \in \mathbb{R}$, satisfying

$$
q_{a}'(t) = \int_{-r}^{0} q_{a}(t + u) \, da(u) \quad \text{for all } t \geq 0.
$$

(2.2)

Its spectral density is given by

$$
\hat{q}_{a}(\xi) := \int_{-\infty}^{\infty} q_{a}(t) e^{i\xi t} \, dt = \frac{1}{|\chi_{a}(i\xi)|^{2}}, \quad \xi \in \mathbb{R}.
$$

(2.3)
Example 2.3. For the point measure $a = \alpha \delta_0$ equation (1.1) reduces to a stochastic ordinary differential equation with the Ornstein-Uhlenbeck process as solution. We obtain $\chi_a(\lambda) = \lambda - \alpha$ and $v_0(a) = \alpha$. For $\alpha < 0$ a stationary solution exists with covariance function $q_a(t) = \frac{1}{2|\alpha|} e^{-|\alpha|t}$ and spectral density $\hat{q}_a(\xi) = (\xi^2 + \alpha^2)^{-1}$.

The law $\mu_X$ of the solution process $X$ on the interval $[0, T]$ is mutually absolutely continuous to the law $\mu_W$ of Brownian motion starting in $X(0)$ in the canonical space $C([0, T])$. We express the likelihood ratio by certain sufficient statistics $b_T$ and $Q_T$ that will be of major importance subsequently.

Definition 2.4. For the solution process $X$ of (1.1) define
\[
\begin{align*}
    b_T(u) &:= \int_0^T X(t + u) \, dX(t) \quad u \in [-r, 0], \\
    q_T(u, v) &:= \int_0^T X(t + u) X(t + v) \, dt \quad u, v \in [-r, 0], \\
    Q_T\mu(u) &:= \int_{-r}^0 q_T(u, v) \, d\mu(v) \quad u \in [-r, 0], \mu \in M([-r, 0]), \\
    Q_a\mu(u) &:= \int_{-r}^0 q_a(u - v) \, d\mu(v) \quad u \in [-r, 0], \mu \in M([-r, 0]).
\end{align*}
\]

The operator $Q_a$ is the covariance operator of the stationary process $X$ on $[-r, 0]$, regarded as an element of $C([-r, 0])$.

It is understood that for $b_T$ a continuous version in $u \in [-r, 0]$ is chosen, which is possible since the Kolmogorov continuity theorem applies due to the moment bound:
\[
\begin{align*}
    \mathbb{E}\left[ \left( \int_0^T X(t + u_1) \, dW(t) - \int_0^T X(t + u_2) \, dW(t) \right)^4 \right] \\
    \lesssim \mathbb{E}\left[ \left( \int_0^T (X(t + u_1) - X(t + u_2))^2 \, dt \right)^2 \right] \\
    \lesssim T^2(u_1 - u_2)^2,
\end{align*}
\]  

which follows from the Burkholder-Davis-Gundy inequality and the uniform Lipschitz continuity of the covariance function $q_a$, see Proposition 5.1.

Theorem 2.5. For a deterministic initial function $F$ in (1.1) the Radon-Nikodym derivative $\Lambda_T(X, X(0) + W)$ of $\mu_X$ with respect to $\mu_W$ is given
by

\[ \Lambda_T(X, X(0) + W) := \frac{d\mu_X}{d\mu_W} \]

\[ = \exp\left( \int_0^T \int_{-r}^0 X(t + u) da(u) \, dX(t) - \frac{1}{2} \int_0^T \left( \int_{-r}^0 X(t + u) da(u) \right)^2 \, dt \right) \]

\[ = \exp\left( \langle b_T, a \rangle - \frac{1}{2} \langle Q_T a, a \rangle \right). \]

This result is the basis for the maximum-likelihood theory developed by Gushchin and Küchler (1999). Its proof is derived from the Girsanov theorem for diffusion-type processes and the stochastic Fubini theorem, see Liptser and Shiryaev (2001) or Küchler and Sørensen (1997).

The first impulse to define a nonparametric estimator \( \hat{a}_T \) of \( a \) would thus be to maximise the likelihood function which amounts to solving the infinite-dimensional equation \( Q_T \hat{a}_T = b_T \). However, the empirical covariance operator \( Q_T \) need not be invertible, and although the covariance operator \( Q_a \), obtained in the limit, is invertible, its inverse \( Q_a^{-1} \) is an unbounded operator, as will be shown later. Hence, we are in a classical nonparametric situation and smoothing methods need to be employed. Our basic idea is to smooth first and then to solve the maximum likelihood equation in terms of the smoothed quantities. The convergences \( \frac{1}{T} Q_T \rightarrow Q_a \) and \( \frac{1}{T} b_T \rightarrow Q_a a \) for \( T \rightarrow \infty \), a consequence of Theorem 6.1 and Theorem 3.2, show that in the limit of an infinitely long observation period the weight measure \( a \) is always identifiable. Having adapted an asymptotic viewpoint, we proceed by analysing the covariance operator \( Q_a \) in detail. From this analysis and the exact convergence properties all subsequent results will be derived.

For the notion of Besov spaces \( B_{p,\alpha}^s \) of functions with \( L^p \)-regularity \( s \) and fine-tuning parameter \( \alpha \) we refer to the appendix. Just recall the identity \( B_{2,2}^s = W^{s,2} \) so that the subsequent results are in particular valid for the scale of \( L^2 \)-Sobolev spaces \( W^{s,2} \). Küchler and Mensch (1992) show that the covariance function \( q_a \) is twice differentiable on \( \mathbb{R} \setminus \{0\} \), but its derivative has a jump in zero which implies that, roughly speaking, the covariance operator \( Q_a \) is smoothing of order two, that is measures with density \( B_{p,\alpha}^s \) are mapped to \( B_{p,\alpha}^{s+2} \), cf. Theorem 2.8. Now let us consider the images of
point masses $\delta_\rho$, $\rho \in [-r,0]$: $$Q_a\delta_\rho(u) = q_a(u - \rho) = q_a(|u - \rho|), \quad u \in [-r,0].$$

This shows that for values of $\rho$ in the interior $(-r,0)$ the image $Q_a\delta_\rho \in C([-r,0])$ has a jump in its first derivative at $u = \rho$, whereas at the boundary $\rho \in \{-r,0\}$ the image $Q_a\delta_\rho$ is as regular as the covariance function $q_a$ on $[-r,0]$. As a consequence, the inclusion of the point measures at the boundary will not complicate the estimator and even imply that the covariance operator is onto. In anticipation of the precise mapping properties we introduce suitable spaces of weight measures, just recall that $g \circ \lambda$ denotes the measure with Lebesgue density $g$ on $[-r,0]$.

**Definition 2.6.**

- (Besov scale) For $s > 0$, $p \in (1,\infty)$, $\alpha \in [1,\infty)$, $v < 0$ set
  $$\mathcal{B}_{p,\alpha}^s := \left\{ \gamma_0 \delta_0 + \gamma_r \delta_{-r} + g \circ \lambda \mid \gamma_0, \gamma_r \in \mathbb{R}, g \in \mathcal{B}_{p,\alpha}^s([-r,0]) \right\},$$
  $$\mathcal{B}_{p,\alpha}^s(v) := \left\{ a \in \mathcal{B}_{p,\alpha}^s \mid v_0(a) \leq v \right\}.$$

  On $\mathcal{B}_{p,\alpha}^s$ we introduce the norm
  $$\| \gamma_0 \delta_0 + \gamma_r \delta_{-r} + g \circ \lambda \|_{s,p,\alpha} := |\gamma_0| + |\gamma_r| + \|g\|_{B_{p,\alpha}^s([-r,0])}.$$

- (L$^p$ scale) For $p \in (1,\infty)$, $v < 0$ set
  $$\mathcal{L}^p := \left\{ \gamma_0 \delta_0 + \gamma_r \delta_{-r} + g \circ \lambda \mid \gamma_0, \gamma_r \in \mathbb{R}, g \in \mathcal{L}^p([-r,0]) \right\},$$
  $$\mathcal{L}^p(v) := \left\{ a \in \mathcal{L}^p \mid v_0(a) \leq v \right\}.$$

  On $\mathcal{L}^p$ we introduce the norm
  $$\| \gamma_0 \delta_0 + \gamma_r \delta_{-r} + g \circ \lambda \|_p := |\gamma_0| + |\gamma_r| + \|g\|_{L^p([-r,0])}.$$

The space $\mathcal{B}_{p,\alpha}^s$ is isomorphic to the tensor product $\mathbb{R}^2 \otimes \mathcal{B}_{p,\alpha}^s$ and thus $(\mathcal{B}_{p,\alpha}^s, \|\cdot\|_{s,p,\alpha})$ is a Banach space. The SDDE (1.1) with weight $a \in \mathcal{B}_{p,\alpha}^s$ typically reads like

$$dX(t) = \left( \gamma_0 X(t) + \gamma_r X(t-r) + \int_{-r}^0 X(t+u)g(u) \, du \right) \, dt + dW(t), \quad t \geq 0.$$ 

The set $\mathcal{B}_{p,\alpha}^s(v)$ is a closed subset of $\mathcal{B}_{p,\alpha}^s$, due to $v < 0$ consisting of weights with a uniform mixing behaviour. This follows from a result in Reiß (2002).
adapted to more general weight measures. By the same arguments these properties also hold for $L^p$. The Besov-type weights form the nonparametric class $M(s, p, S, \delta)$ for which our estimator will be shown to be rate-optimal, whereas the space $L^p$ will merely occur in the context of mapping properties of the covariance operator.

**Definition 2.7.** For $s > 0$, $S > 0$, $p \in [1, \infty]$ and $\delta > 0$ set

$$M(s, p, S, \delta) := \{a \in B_{p,1}^s(-\delta) \mid \|a\|_{s,p,1} \leq S\}.$$ 

The choice $\alpha = 1$ in the definition will be discussed in Section 3.3. In Reiß (2001) and Gushchin and Küchler (2003) it was shown that the covariance operator is always one-to-one on $M([-r, 0])$ and maps densities in $L^2([-r, 0])$ to $W^{2,2}([-r, 0])$. We show that the covariance operator is for a certain range of Besov spaces also smoothing of order two.

**Theorem 2.8.** For weight measures $a$ in $B_{p,\alpha}^s(v)$ and the parameters as before the covariance operator is a bijective bounded linear operator on the appropriate spaces:

$$Q_a : B_{p,\alpha}^s \to B_{p,\alpha}^{s+2}([-r, 0]) \text{ and } Q_a : L^p \to W^{2,p}.$$ 

In order to obtain upper bounds in a minimax sense for our estimator, we shall need uniform norm bounds in the preceding statement.

**Theorem 2.9.** For $s, S, \delta > 0$, $p \in (1, \infty)$ the following operator norms are uniformly bounded:

$$\sup_{a \in M(s,p,S,\delta)} \|Q_a\|_{B_{p,\alpha}^s \to B_{p,\alpha}^{s+2}} < \infty, \quad \sup_{a \in M(s,p,S,\delta)} \|Q_a^{-1}\|_{B_{p,\alpha}^{s+2} \to B_{p,\alpha}^s} < \infty,$$

$$\sup_{a \in M(s,p,S,\delta)} \|Q_a\|_{L^p \to W^{2,p}} < \infty, \quad \sup_{a \in M(s,p,S,\delta)} \|Q_a^{-1}\|_{W^{2,p} \to L^p} < \infty.$$ 

3. **Construction of the estimator**

3.1. **The general idea.** We start by smoothing the statistic $b_T$ adaptively. To this end kernel and wavelet methods are equally applicable, but since an integral equation with this estimator as data has to be solved later, wavelet techniques avoid a second numerical discretisation step. For the notation $(\psi_s)$ of an $s$-regular wavelet basis on $[-r, 0]$ we refer to the appendix.
First, we need to clarify the functional nature of the noise in the estimate $b_T$ of $Q_a a$. A good intuition gives the decomposition (recall $\sigma = 1$)

$$b_T(u) = \int_0^T \left( X(t + u) \int_{-r}^0 X(t + v) \, da(v) \right) \, dt + \int_0^T X(t + u) \, dW(t)$$

$$= (Q_T a)(u) + \int_0^T X(t + u) \, dW(t). \tag{3.1}$$

Suppose for a moment that $\frac{1}{T} Q_T$ equals $Q_a$ exactly (its kernel is in fact easier to estimate). The error term is then due to the stochastic integral term which is as regular with respect to $u$ as Brownian motion due to the Kolmogorov continuity theorem. Thus, we do not face the classical “signal+white noise”-model, but rather an integrated form involving the signal and a perturbation by Brownian motion; one may think of recovering the function $f$ from the noisy observation

$$Y(u) = f(u) + \varepsilon W(u), \quad u \in [-r, 0].$$

However, in our setting the noise is not Gaussian and we are not interested in the signal $f = Q_a a$ itself, but rather in $Q_a^{-1} f = a$. Nevertheless the analogue ill-posed problem in white noise would be

$$dY(u) = DQ_a a(u) \, dt + \varepsilon \, dW(u), \quad u \in [-r, 0],$$

where $D$ denotes differentiation. The operator $DQ_a$ has only degree one of ill-posedness and this explains the exponent $s/(2s + 3)$ in the minimax rate. This formal argument could be substantiated by regarding the stochastic differential $db_T(u)$ instead of $b_T(u)$. The main drawback of regarding differentials in our situation is that $DQ_a$ is no longer one-to-one.

Taking account of Theorem 2.8 we are going to minimise the expected $L^p$-loss in estimating the signal $Q_a a$ with minimal $W^{2,p}$-loss. Although our noise process is more regular than white noise, the noise is still less regular than the norm we want the signal to be estimated in.

For resolving our problem in practice the abstract wavelet thresholding results by Kerkyacharian and Picard (2000) can be well adapted. Our estimator $\hat{b}_T$ is obtained by expanding $b_T$ in a wavelet basis up to a certain level and only keeping the significant coefficients (hard thresholding).
Definition 3.1. Let \( s_{\text{max}} > 2 \) be fixed. With \( b_T \) from Definition 2.4 introduce for any multi-index \( \lambda \) the wavelet coefficient

\[
\beta_{\lambda,T} := \langle \frac{1}{T} b_T, \psi_\lambda \rangle,
\]

where \( (\psi_\lambda)_\lambda \) is a compactly supported \( s_{\text{max}} \)-regular wavelet basis in \( L^2([-r,0]) \) (see Appendix 8.2). Define the thresholding estimator

\[
\hat{b}_T := \hat{b}_{T,J(T),\kappa(T)} := \sum_{|\lambda| \leq J(T)} (\beta_{\lambda,T} \mathbf{1}_{|\beta_{\lambda,T}| > \kappa_\lambda(T)}) \psi_\lambda
\]

for a certain resolution level \( J(T) \) and thresholds \( (\kappa_\lambda(T))_\lambda \).

3.2. Results. How should we choose the threshold values \( \kappa_\lambda(T) \)? The second term in the decomposition (3.1) gives for \( \beta_{\lambda,T} \) the variance estimate

\[
T^{-1} \langle Q_a \psi_\lambda, \psi_\lambda \rangle \sim T^{-1} \mathbf{2}^{-2|\lambda|}
\]

by Lemma 8.8. In Section 3.3 we comment on the choice for a specific sample, here however, we strive for asymptotically optimal threshold values \( \kappa_\lambda(T) \), \( T \to \infty \). We note that the overall noise level is \( T^{-1/2} \) as expected and that level-wise the noise intensity decays like \( \mathbf{2}^{-|\lambda|} \), which indicates that the noise is by one degree smoother than white noise. A detailed analysis confirms this picture and even yields uniform Gaussian tail estimates, which cannot be obtained by standard large deviation techniques.

Theorem 3.2. Let \( (\psi_\lambda)_\lambda \) be a compactly supported \( 2 \)-regular wavelet basis of \( L^2([-r,0]) \) and let \( R, \delta, \rho > 0 \). Then there is a universal bound \( \kappa^* > 0 \) such that we obtain uniformly for all weight measures \( a \) with \( v_0(a) \leq -\delta \), \( \|a\|_{TV} \leq R \), all multi-indices \( \lambda \) and all sufficiently large \( T \)

\[
\mathbb{P}_a \left( 2^{|\lambda|} T^{1/2} |\frac{1}{T} b_T - Q_a \psi_\lambda| \geq \frac{\sqrt{T}}{2} \sqrt{T} \right) \lesssim T^{-3\rho} \quad \forall \kappa \geq \kappa^*.
\]

We shall therefore set \( \kappa_\lambda(T) = \kappa \mathbf{2}^{-|\lambda|} T^{-1/2} \sqrt{T} \). As is classical in wavelet methods, we choose the maximal frequency level \( J(T) \) such that \( J(T) 2^{J(T)} \) is anti-proportional to the variance level \( T^{-1} \).

Proposition 3.3. Let \( s \in (0, s_{\text{max}} - 2] \), \( S > 0 \), \( \rho, p \in (1, \infty) \) and \( \delta > 0 \) be given satisfying

\[
\frac{1}{p} - \frac{1}{\rho} \leq \frac{2}{\rho} \frac{s}{3}
\]

(3.3)
Set $2^j(T) \sim T/\log T$ and $\kappa_\lambda(T) = \kappa 2^{-|\lambda|} T^{-1/2} \log(T)$ with $\kappa$ chosen as in Theorem 3.2. Then we obtain the following asymptotic estimate for the estimator $\hat{b}_T$ from Definition 3.1:

$$\sup_{a \in M(s,p,S,\delta)} E_a[\|\hat{b}_T - Q_a a\|_{W^{2,\rho}}] \lesssim \left( \frac{T}{\log T} \right)^{-\frac{s}{2s+3}}.$$

In the next step we construct an operator $\hat{Q}_T$ from the observations up to time $T$, which is close to the true covariance operator. We could, of course, use the results for $Q_T$ from Theorem 6.1, but it is even simpler to use (1.3), that is $q_a'(t) = Q_a a(-t)$ obtained from (2.2). Writing $q_a(t) = q_a(0) + \int_0^t q_a'(u) du$, we can thus determine $q_a$ from $q_a(0)$ and $Q_a a$ and derive an estimator from estimators for these two quantities. The convolution kernel of $\hat{Q}_T$ will have a jump in the first derivative at zero, but will be regular elsewhere. This explains why the mapping properties of $\hat{Q}_T$ resemble those of $Q_a$. In particular, $\hat{Q}_T$ is with high probability an isomorphism between $L^\rho$ and $W^{2,\rho}$. This gives our main result.

**Theorem 3.4.** Let the parameters $s$, $S$, $p$ and $\delta$ be as before. Introduce the integral operator $\hat{Q}_T : L^\rho \to W^{2,\rho}$ with convolution kernel

$$\hat{q}_T(u) := \frac{1}{T} \int_0^T X(t)^2 dt + \int_{-|u|}^0 \hat{b}_T(v) dv, \quad u \in [-r, r],$$

i.e. $\hat{Q}_T \mu(t) := \int_{-r}^0 \hat{q}_T(t - u) d\mu(u)$ for $t \in [-r, 0]$, $\mu \in L^\rho$.

Define the estimator $\hat{a}_T$ by

$$\hat{a}_T := \begin{cases} \min \left( S \| \hat{Q}_T^{-1} \hat{b}_T \|_{L^\rho}^{-1}, 1 \right) \hat{Q}_T^{-1} \hat{b}_T, & \text{if } \hat{Q}_T : L^\rho \to W^{2,\rho} \text{ is invertible}, \\ 0, & \text{otherwise}. \end{cases}$$

Then the following asymptotic upper bound holds for $T \to \infty$:

$$\sup_{a \in M(s,p,S,\delta)} E_a[\|\hat{a}_T - a\|_{L^\rho}] \lesssim \left( \frac{T}{\log T} \right)^{-\frac{s}{2s+3}}.$$

**3.3. Discussion.** Our method differs from the classical wavelet thresholding algorithm due to the ill-posedness involved. Our threshold $\kappa_\lambda$ depends on the resolution level $|\lambda|$, because the intensity of the noise coefficients is of order $2^{-|\lambda|} T^{-1/2}$. Furthermore, it is not necessary to suppose additionally
the linear constraint that the weight lies in \( W^{\frac{s}{p} + \frac{3}{\rho}} \) because the restriction (3.3) is much stronger than in the classical setting. If we call \( \beta = s/3 \) the effective smoothness (like in multidimensional smoothing) in our model, then (3.3) in terms of \( \beta \) is well known (Härdle, Kerkyacharian, Picard, and Tsybakov 1998, Section 10.4). We have chosen the Besov scale \( (B^s_{p,\alpha}) \) with \( \alpha = 1 \) for simpler embedding relations. In fact, \( \alpha = p/\rho \) would do as can be seen from (6.3). It is not known whether this value is the maximal possible.

**Example 3.5.** Which rate of convergence do we obtain for the \( L^\rho \)-risk of the weight function \( g = -1_{[-\frac{1}{2},0]} \) with delay \( r = 1 \)? This might be seen as a toy example for estimating a change point or the maximal delay time. The function \( g \) lies in \( B^{s}_{1,\infty}([-1,0]) \) and thus by embedding in \( W^{s,\rho}([-1,0]) \) for \( s < \frac{1}{p} \). This shows that linear methods cannot converge faster than with rate \( T^{-1/(2+3\rho)} \) whereas due to \( g \in B^{s}_{p,1} \) for any \( s < \frac{1}{p} \), relationship (3.3) is satisfied for all \( s < \frac{3}{3p-2} \). Owing to \( s \leq 1 \) our wavelet thresholding estimator achieves (almost) the rate \( T^{-1/5} \) for \( \rho < 5/3 \) and the rate \( T^{-1/(3\rho)} \) for \( \rho \geq 5/3 \), which is a significant gain. If our results could be generalized to cover the case of the quasi-Banach spaces \( L^p \) for \( p < 1 \) (which is to be expected), then \( g \in B^{s}_{1/s,\infty} \) would yield (almost) the \( L^1 \)-rate \( T^{-1/3} \). The latter is in fact the minimax rate for nonparametric change point detection, cf. Reiß (2004).

In the mathematical results we have focused on the spatial adaptivity of our estimator, but the construction is clearly independent of major a priori knowledge of the unknown parameter. However, we had to assume some maximal domain of regularity \( (s_{\text{max}}) \), some bound on the size \( (S) \) and some uniform mixing behaviour \( (\delta) \). The resulting minimal asymptotically optimal threshold \( \kappa^* \) depends in a complicated way on these quantities. So clearly the question arises how to choose \( \kappa \) for a specific observation.

First of all note that \( T^{1/2} \beta_{\lambda,T} - \langle Q_T a, \psi_\lambda \rangle \) is a martingale with respect to \( T \) with quadratic variation \( \langle Q_T \psi_\lambda, \psi_\lambda \rangle \). Asymptotically for \( T \to \infty \) the random variable

\[
\eta_{\lambda,T} := T^{1/2} \beta_{\lambda,T} - \frac{\langle \frac{1}{T} Q_T a, \psi_\lambda \rangle}{\sigma_{\lambda,T}} \quad \text{(with } \sigma_{\lambda,T}^2 := \langle \frac{1}{T} Q_T \psi_\lambda, \psi_\lambda \rangle \sim 2^{-2|\lambda|})
\]
is therefore $N(0, 1)$-distributed by the martingale central limit theorem. In other words, we observe the coefficient $\langle \frac{1}{\sqrt{T}} Q^T a, \psi_\lambda \rangle$ under the noise $T^{-1/2} \sigma_{\lambda, T} \eta_{\lambda, T}$. Because $\sigma_{\lambda, T}^2$ converges stochastically, we conclude that the noise is approximately normal distributed with variance $T^{-1} \sigma_{\lambda, T}^2$, which is observable. Thus, we are lead to apply the usual threshold rules in the Gaussian shift setting, see Donoho and Johnstone (1994) for a detailed discussion and Neumann and von Sachs (1995) for the case of only asymptotically Gaussian noise. It only remains to take account of the $W^{2, \rho}$-norm used so that we have to be a little bit more conservative and should choose in the Hilbertian case $\rho = 2$ the thresholds $\kappa_j = T^{-1/2} \sigma_{\lambda, T} \sqrt{6 \log(T \sigma_{\lambda, T}^{-2})}$, provided the isomorphy constants in Theorem 2.8 are close to one when measured in the corresponding wavelet coefficient norms, cf. equation (17) of Abramovich and Silverman (1998). The maximal frequency $J$ should be chosen such that $J^{-1/2} \approx T^{-1} \max_{\lambda} \sigma_{\lambda, T}^2$, which is an estimate of the squared noise level.

The cut-off in the definition of $\hat{a}_T$ avoids uncontrolled errors in the inversion and is rather natural, cf. Efromovich and Koltchinskii (2001). If only discrete observations $(X_{t_i})$ are available with $0 = t_0 \leq t_1 \leq \cdots \leq t_N = T$, then it can be shown that the error in approximating the stochastic integral $b_T$ does not increase the asymptotics as long as $\Delta := \max_i (t_{i+1} - t_i)$ satisfies $\Delta \lesssim T^{-1/2}$. For low-frequency observations, that is $\Delta > 0$ fixed and $N \to \infty$, it is an open question whether a consistent estimator exists at all.

One might want to consider the submodel in which the weights do not include any point masses, that is the weight space $B_{p, 1}^s([0, r])$ instead of $B_{p, 1}^s$. For this purpose one can project the estimator $\hat{a}_T$ onto $L^p([0, r])$ by neglecting the point measure part. The asymptotic risk bound remains the same. Conversely, adding finitely many point masses with known location to our model does not increase the asymptotic rates for the risk. It must only be taken into account that the image functions of the weight measures under the covariance operator are less regular at these locations.

Finally, note that the approach can be extended naturally to multi-dimensional affine SDDEs where a matrix $A$ of weight measures is to be estimated. In this case, we use the matrix-valued statistics $b_T$ and $q_T$ formed
by applying the one-dimensional definition to all cross terms and we are lead to the analogous inverse problem $Q_T \hat{A}_T \approx b_T$ to determine an estimator $\hat{A}_T$.

A mathematical analysis of an adaptive version of $\hat{A}_T$ will be accomplished by similar methods and a wide range of applications could be addressed, the model being the counterpart of vector autoregressive processes in time series analysis.

4. Optimality of the estimator

We show that the adaptive wavelet thresholding estimator is rate-optimal with respect to $L^\rho$-risk functions, in the sense that one cannot improve on the restriction (3.3) in order to obtain the speed of convergence $(T/ \log T)^{-\frac{s-1}{2s+3}}$ for weights in $B^s_{p,1}$. For smaller values of $p$ the rate of convergence is indeed worse and is obtained by embedding $B^s_{p,1}$ to $B^u_{\pi,1}$ with some properly chosen $u < s$ and $\pi > p$. In the sequel, we merely assume $s + \frac{1}{p} - \frac{1}{p} \geq 0$ in order to have the embedding $B^s_{p,1} \subset L^\rho$ and thus a well-defined risk. For the sake of simplicity we do not present the proofs for the stationary case, but for fixed deterministic initial functions in (1.1). Due to ergodicity the initial segment is not significant for asymptotic statements, but the proofs for stochastic initial conditions are lengthy and tedious, see Reiß (2001).

**Theorem 4.1.** Let $s > 0$, $p > 0$, $S > 0$ and $\delta > 0$ be given with $s + \frac{1}{p} - \frac{1}{p} \geq 0$ and such that $M(s, p, S, \delta)$ has nonempty interior in $B^s_{p,1}$. Then the following asymptotic minimax lower bound holds for $T \to \infty$:

$$\inf_{\hat{a}_T} \sup_{a \in M(s, p, S, \delta)} \mathbb{E}_n[\|\hat{a}_T - a\|_{L^\rho}] \gtrsim \left( \frac{T}{\log T} \right)^{-\frac{s+\frac{1}{p} - \frac{1}{p}}{2s+3-\frac{2}{p}}} ,$$

where the infimum is taken over all $\mathcal{F}_T^X$-measurable estimators $\hat{a}_T$.

We obtain a fairly complete picture of the minimax rates for the $L^\rho$-risk and certain Besov regularity classes $M(s, p, S, \delta)$. Again replacing $\frac{s}{3}$ by the efficient regularity $\beta$ we obtain the usual picture (Härdle, Kerkyacharian, Picard, and Tsybakov 1998, Section 10.4).
Corollary 4.2. Assume that $s > 0$, $p \in (1, \infty)$, $S > 0$ and $\delta > 0$ are given such that $M(s, p, S, \delta)$ has nonempty interior in $\mathcal{B}_p$. In what follows the infima are taken over all $\mathcal{F}_T$-measurable estimators $\hat{a}_T$.

1. (sparse case) For $\frac{1}{p} - \frac{1}{p} \geq \frac{2s}{p} + \frac{3}{2}$ the risk lower and upper bound match, that is our estimator is rate-optimal in a minimax sense. We find

$$\inf_{\hat{a}_T} \sup_{a \in M(s, p, S, \delta)} \mathbb{E}_a[\|\hat{a}_T - a\|_{\mathcal{L}_p}] \sim \left(\frac{T}{\log T}\right)^{-\frac{s}{2s+3}}$$

with $u = \frac{3p}{5p-2} (s + \frac{1}{p} - \frac{1}{p}) \leq s$.

2. (regular case) For $\frac{1}{p} - \frac{1}{p} < \frac{2s}{5}$ we have

$$\sup_{\hat{a}_T} \inf_{a \in M(s, p, S, \delta)} \mathbb{E}_a[\|\hat{a}_T - a\|_{\mathcal{L}_p}] \lesssim \left(\frac{T}{\log T}\right)^{-\frac{s}{2s+3}}$$

It should be noted that the actual minimax rates in the regular case are of course $T^{s/(2s+3)}$, which are attained by linear estimators.

5. Proofs for the Covariance Operator

5.1. Proof of Theorem 2.8. We first need precise regularity and tail behaviour results for the covariance function $q_a$. Roughly speaking, the covariance function is three times more regular than the weight itself and decreases exponentially fast.

Proposition 5.1. For $a \in \mathcal{B}_p^{s,v}(v)$ with $v < 0$, $s > 0$ and $p \in (1, \infty)$, $\alpha \in [1, \infty]$ the covariance function $q_a$ lies in $\mathcal{B}_p^{s+3,\alpha}(0, r]$.

Proof. Let us write $a = g + \alpha \delta_0 + \beta \delta_{-r}$, as in Definition 2.6. The covariance function satisfies (2.2) such that for $t \in (0, r)$

$$q'_a(t) = \int_{-r}^0 q_a(|t + u|) \, da(u) = \int_{-r}^0 q_a(|t + u|) g(u) \, du + \alpha q_a(t) + \beta q_a(r - t)$$

holds. The properties $q_a(-t) = q_a(t)$, $q_0' (0+) = -\frac{1}{2}$ and $q_a \in C^2([0, \infty))$ from Küchler and Mensch (1992) imply for $t \in (0, r)$ that $q''_a(t)$ equals

$$= \frac{d}{dt} \left( \int_{-r}^0 q'_a(|t + u|) \mathrm{sgn}(t + u) g(u) \, du + \alpha q'_a(t) - \beta q'_a(r - t) \right)$$

$$= \int_{-r}^0 q'_a(|t + u|) g(u) \, du + 2q'_a(0+) g(-t) + \alpha q''_a(t) + \beta q''_a(r - t)$$

$$= \int_{t-r}^0 q''_a(v) g(v - t) \, dv + \int_0^1 q''_a(v) g(v - t) \, dv - g(-t) + \alpha q''_a(t) + \beta q''_a(r - t).$$
This shows that the third derivative $q''_a$ is at least as regular in a Besov space sense as the most irregular term among the terms on the right hand side. For the two convolution terms note that the regularity result from Lemma 8.3 with $f = g$, $k = q''_a$ and with obvious modifications of the interval boundaries applies: $g \in B^s_{p,\alpha}$ and $q_a \in B^w_{p,\alpha}$ for some $w \geq 1$ imply by embedding $q''_a \in B^{\tilde{w} - 2}_{p,\alpha}$ for all $\tilde{w} < w - 1 + 2(p - 1)/p$ and thus by Lemma 8.3 that the convolutions lie in $B^{\tilde{w} - 2}_{p,\alpha}$. Formally, we now proceed by putting $w := \sup\{s \geq 2 \mid q_a \in B^s_{p,\alpha}\}$ and by deducing that the convolution terms are in $B^{\tilde{w} - 2}_{p,\alpha}$ for some $\tilde{w} > w - 1$ while $g$ is in $B^s_{p,\alpha}$ and $q''_a$ in $B^{w - 2 - \varepsilon}_{p,\alpha}$ for all $\varepsilon > 0$. Consequently, the right hand side is an element of $B^{(\tilde{w} + 1)\wedge (s + 3)}_{p,\alpha}$, hence $q_a$ is in $B^{\tilde{w} - 1}_{p,\alpha}$ and therefore $q_a \in B^s_{p,\alpha}$ follows. For the $L^p$-scale the proof is the same except that Lemma 8.4 can be applied immediately to the expression for $q''_a$.

The following lemma extends well-known non-uniform results, e.g. Hale and Verduyn Lunel (1993), Küchler and Sørensen (1997).

**Lemma 5.2.** For all $S > 0$ and $v > \delta > 0$ we have the uniform bound

$$\sup_{\|a\|_{TV} \leq S, v(0) \leq -v} \|E_\delta q_a\|_{C^1([0,\infty))} < \infty.$$  

**Proof.** We consider the formula $\hat{q}_a(\xi) = |\chi_a(i\xi)|^{-2}$ from (2.3). Due to $|\chi_a(i\xi)|^2 = \chi_a(i\xi)\chi_a(-i\xi)$ the Fourier transform $\hat{q}$ can be extended to a holomorphic function on the strip $\{z \in \mathbb{C} \mid |\text{Im}(z)| < v(0)\}$ (Katznelson 1976, Section VI.7.1) and satisfies

$$\hat{E}_\delta(q_a)(\xi) = \hat{q}_a(\xi + i\delta) = \chi_a(i\xi - \delta)^{-1}\chi_a(-i\xi + \delta)^{-1}.$$  

The assumptions guarantee $|\chi_a(\pm i\xi + \delta)| \geq |i\xi + \delta| - Se^{\delta r}$. Since subsets $U \subset M([-r, 0])$ that are bounded and closed in total variation norm are compact in the weak*-topology of $M([-r, 0])$ by the Banach-Alaoglu Theorem, the set of characteristic functions $\{\chi_a \mid a \in U\}$ is compact in the space of entire functions equipped with the convergence on compact sets. Consequently, the classical result from calculus about the convergence of maxima on compact
sets yields for the respective choice of signs

\[ K_\pm := \sup_a \max_{|\xi| \leq 2Se^{\delta r}} |\chi_a(\pm i\xi \mp \delta)|^{-1} < \infty, \]

where the supremum is taken over all measures \( a \) as in the statement of the lemma. We compare with the covariance function \( q_{OU}(t) = \frac{1}{2e^{-t|\xi|}} \) of the Ornstein-Uhlenbeck process with parameter \( -\nu \)

and Lipschitz continuous.

\[ \text{Proof.} \]

Because of \( \| E_\delta(q_{OU}) \|_{C^1(\mathbb{R})} < \infty \) the result follows. \( \square \)

**Proposition 5.3.** Let \( E_\delta \) be the multiplication operator with the exponential:

\[ E_\delta(f)(t) := f(t)e^{\delta t}. \]

Then for \( S, \nu > 0 \) and \( p \geq 1 \) the covariance function \( q_a \) has for any \( \delta \in (0, \nu) \) the property

\[ \sup_{a \in L^p(-\nu): \|a\|_p \leq S} \| E_\delta(q_a) \|_{C^{2,1}([0, \infty))} < \infty. \]

In particular, the second derivative of \( E_\delta(q_a) \) is uniformly bounded on \([0, \infty)\) and Lipschitz continuous.

**Proof.** The solution property (2.2) of the covariance function \( q_a \) implies for \( t \geq r \)

\[ |q_a''(t)| \leq \| q_a'' \|_{C([t-r, t])} \| a \|_{TV} \leq \| q_a \|_{C([t-2r, t])} \| a \|_{TV}^2. \]

By Lemma 5.2 and the product rule of differentiation we conclude that \( \sup_{t \geq r} |E_\delta(q_a)''(t)| \) is uniformly bounded over the weights \( a \). Similarly, the bound \( |q_a''(t+v) - q_a''(t)| \) \( \leq \| q_a(\bullet + v) - q_a \|_{C([t-2r, t])} \| a \|_{TV}^2 \) for \( t \geq r \) and \( v \geq 0 \) yields the existence of a uniform Lipschitz on \( q_a'' \) on \([r, \infty)\) by Lemma 5.2.
It remains to apply Proposition 5.1, the uniformity result of Proposition 5.4 and Sobolev embeddings to obtain the same results on the interval \([0, r]\).

Proof. (of Theorem 2.8) First, let us see how \(Q_a\) acts on functions \(f \in B_{p,\alpha}^s([−r,0])\). Since \(Q_a\) maps \(M([−r,0])\) continuously to \(C([−r,0])\) by general properties of covariance operators (Vakhaniya, Tarieladze, and Chobanyan 1987, Thm. III.2.2), we only need to estimate \(\|Q_a f''\|_{s,p,\alpha}\). By symmetry of \(q_a\) and by the regularity result \(q_a \in B_{p,\alpha}^{s+3}(0, r)\) (Proposition 5.1) we obtain for \(t \in [−r, 0]\) like in the proof of Lemma 5.1 \((Q_a f)''(t) = \int_0^t q_a''(t − s)f(s)ds − f(t)\). From Lemma 8.3 we infer further

\[
\|Q_a f''\|_{s,p,\alpha} \lesssim \|f\|_{s,p,\alpha} + \|f\|_{s−1,p,\alpha}\|q_a''\|_{s',p',\alpha'} \lesssim (1 + \|q_a\|_{s+3,p,\alpha})\|f\|_{s,p,\alpha},
\]

which shows that \(Q_a\) maps \(B_{p,\alpha}^s\) continuously to \(B_{p,\alpha}^{s+2}\). Writing the derivative operator as \(D\), we find for any \(\varepsilon \in (0, 2 − (\frac{3}{p} ∨ 1))\) by Lemma 8.3 with \(k = q_a''\) and by the embedding \(B_{p,\alpha}^{s+1} \subset B_{p',\alpha'}^{s+\varepsilon}\),

\[
\|(D^2 Q_a + \text{Id}) f\|_{s+\varepsilon,p,\alpha} \lesssim \|f\|_{s+\varepsilon−1,p,\alpha}\|q_a''\|_{s+\varepsilon,p',\alpha'} \lesssim \|q_a\|_{s+3,p,\alpha}\|f\|_{s,p,\alpha}.
\]

Hence, \((D^2 Q_a + \text{Id})\) is a compact operator on \(B_{p,\alpha}^s\) \((B_{p,\alpha}^{s+\varepsilon}([−r,0]) \subset B_{p,\alpha}^s([-r,0])\) compactly).

Let \(V \subset B_{p,\alpha}^s\) denote the kernel of \(D^2 Q_a\) and let \(V^c\) be a complementing subspace of \(V\). By Fredholm theory (Rudin 1991) the range of \(D^2 Q_a\) is closed and its codimension equals the finite dimension of \(V\). Therefore there exists a complementing subspace \(U\) of \(D^2 Q_a(B_{p,\alpha}^s)\) with \(\dim U = \dim V\). The situation is illustrated by the following diagram:

\[
\begin{array}{ccc}
B_{p,\alpha}^s & = & V^c \oplus V \\
& & \downarrow Q_a \\
B_{p,\alpha}^{s+2} & = & Q_a(B_{p,\alpha}^s) \oplus (D^2)^{-1}(U) \\
& & \downarrow D^2 \\
B_{p,\alpha}^s & = & D^2 Q_a(B_{p,\alpha}^s) \oplus U
\end{array}
\]

While the decomposition in the first and in the third line hold by definition, the representation of \(B_{p,\alpha}^{s+2}\) in the second line follows from the third line due to \((D^2)^{-1}(B_{p,\alpha}^s) = (D^2)^{-1}(D^2 Q_a(B_{p,\alpha}^s) \oplus U) \subset Q_a(B_{p,\alpha}^s) + (D^2)^{-1}(U)\). The
fact that \( Q_a(V) \) is contained in the kernel of \( D^2 \) implies that the operators \( Q_a \) and \( D^2 \) each map the vertically corresponding subspaces into each other.

This argument shows that \( Q_a(B_{p,\alpha}^s) \) is a closed subspace of \( B_{p,\alpha}^{s+2} \) of codimension not larger than two. Due to \( Q_a(\delta_{-r}) = q_a(\bullet + r) \) and \( Q_a\delta_0 = q_a \) we have \( Q_a(\text{span}(\delta_{-r}, \delta_0)) \subset B_{p,\alpha}^{s+3} \subset B_{p,\alpha}^{s+2} \) by Proposition 5.1. The injectivity of \( Q_a \) on \( M([-r,0]) \) implies that \( Q_a(\text{span}(\delta_{-r}, \delta_0)) \) is a two-dimensional subspace of \( B_{p,\alpha}^{s+2} \) in the complement of \( Q_a(B_{p,\alpha}^s) \). Owing to codim \( Q_a(B_{p,\alpha}^s) \leq 2 \) this codimension must equal two and \( Q_a : B_{p,\alpha}^s \to B_{p,\alpha}^{s+2} \) is onto, hence bijective. Because \( Q_a \) is separately continuous on these two subspaces, it is continuous on its span \( B_{p,\alpha}^s \) and by the open mapping theorem it is an isomorphism.

Exactly the same reasoning applies for \( L^p \) and \( L^p \) instead of \( B_{p,\alpha}^s \) and \( B_{p,\alpha}^s \), just apply Lemma 8.4 to \( (Q_a f)' \).

\[ \square \]

5.2. Proof of Theorem 2.9. This will be a consequence of the following continuity properties.

**Proposition 5.4.** Suppose \( s > 0, 1 < p < \infty, \alpha \in [1,\infty] \) and \( v < 0 \) are given. If \( (a_n) \) is a sequence in \( B_{p,\alpha}^s(v) \) that converges in \( B_{p,\alpha}^s \)-norm to the \( B_{p,\alpha}^s(v) \)-weight \( a \), then \( \|q_{a_n} - q_a\|_{B_{p,\alpha}^{s+2}} \to 0 \) follows.

**Proof.** Put \( f_n := q_{a_n} - q_a \) and \( a_n = g_n + \gamma_{r,n} \delta_{-r} + \gamma_{0,n} \delta_0 \). As before the following identities hold for \( t \in (0,r) \):

\[
\begin{align*}
\frac{d^2}{dt^2} f_n'(t) &= \left( \int_{-r}^0 q_{a_n}(\bullet + u) \, da_n(u) - \int_{-r}^0 q_a(\bullet + u) \, da(u) \right)'(t) \\
&= \left( \int_{-r}^0 f_n(\bullet + u) \, da_n(u) \right)'(t) + (Q_a(a_n - a)(-\bullet))'(t) \\
&= - \int_{-r}^t f_n'(-t - u)g_n(u) \, du + \int_{-t}^0 f_n'(t + u)g_n(u) \, du \\
&\quad - \gamma_{r,n} f_n'(r - t) + \gamma_{0,n} f_n'(t) - (Q_a(a_n - a))'(-t) \\
&= - \int_0^{r-t} f_n'(u)g_n(-u - t) \, du + \int_{-t}^0 f_n'(u)g_n(u - t) \, du \\
&\quad - \gamma_{r,n} f_n'(r - t) + \gamma_{0,n} f_n'(t) - (Q_a(a_n - a))'(-t).
\end{align*}
\]
By Lemma 8.3 we obtain for all \( w > 0 \) (allowing the value \( \infty \)) the estimate

\[
\| f_n'' \|_{w,p,\alpha} \lesssim \| f_n'' \|_{w + 2, p, \alpha} + \| \gamma_{1,n} \|_{w - 1, p, \alpha} + \| \gamma_{2,n} \|_{w - 1, p, \alpha}
\]

\[+ \| Q_a \|_{B_{p,\alpha}^{(w-1)\vee 0}} \| a_n - a \|_{(w-1)\vee 0, p, \alpha}.
\]

(5.1)

For \( a_n \to a \) weakly the covariance functions converge in \( W^{p,2}([0, r]) \) for all \( \rho < \frac{5}{2} \), which has been established in Reiß (2002) by spectral methods. Hence, \( \| f_n \|_{w,p,\alpha} \to 0 \) holds for all \( w < 2 + \frac{1}{p} \). In particular, the convergence \( \| f_n \|_{L^p} \to 0 \) follows. The right hand side of estimate (5.1) is thus finite for all \( w \in (0, \frac{1}{p}) \). Once again using \( B_{p,\alpha}^w \subset B_{p',\alpha'}^{w-1+\varepsilon} \) for any \( \varepsilon \in (0, 2 - (\frac{2}{p} \lor 1)) \), we obtain for all \( w \leq s + 1 \)

\[
\| f_n \|_{w + 2, p, \alpha} \lesssim \| f_n \|_{L^p} + \| a_n \|_{s, p, \alpha} \| f_n \|_{w + 2 - \varepsilon, p, \alpha} + \| Q_a \| \| a_n - a \|_{s, p, \alpha}.
\]

Starting with \( w_0 = \varepsilon \), we can iterate this estimate \( (w_{n+1} := (w_n + \varepsilon) \wedge (s + 1)) \). Hence \( \| f_n \|_{s + 3, p, \alpha} \) is bounded by a multiple of \( \| f_n \|_{L^p} + \| f_n \|_{2, p, \alpha} + \| a_n - a \|_{s, p, \alpha} \), which tends to zero for \( n \to \infty \). This proves \( \| f_n \|_{s + 3, p, \alpha} \to 0 \). \( \square \)

**Proposition 5.5.** If \((a_n)\) is a sequence of \( B_{p,\alpha}^w(v)\)-weights that converges in \( B_{p,\alpha}^w\)-norm to the \( B_{p,\alpha}^s\)-weight \( a \) for some \( w > s - 2 + (1 \lor \frac{2}{p}) \) and \( s > 0 \), \( p \in (1, \infty) \), \( \alpha \in [1, \infty] \), \( v < 0 \), then the covariance operators converge in operator norm:

\[
\lim_{n \to \infty} \| Q_{a_n} - Q_a \|_{B_{p,\alpha}^s \to B_{p,\alpha}^{s+2}} = 0.
\]

The statement remains true if \( B_{p,\alpha}^s \) is replaced by \( L^p \) and \( B_{p,\alpha}^{s+2} \) by \( W^{2,p} \) and under the condition that \( (a_n) \) converges in \( L^p \)-norm to \( a \).

**Proof.** By linearity we have for \( f \in B_{p,\alpha}^s \) and \( t \in [-r, 0] \)

\[
((Q_{a_n} - Q_a)f)'(t) = \int_0^{r + t} f(t - u)(q_{a_n} - q_a)'(u) \, du + \int_0^{-t} f(t + u)(q_{a_n} - q_a)'(u) \, du.
\]
By Lemma 8.3 and by the norm estimates \( \| \cdot \|_{s+2,p',\alpha'} \lesssim \| \cdot \|_{w+3,p,\alpha} \) and \( \| \cdot \|_{L^p'} \lesssim \| \cdot \|_{w+3,p,\alpha} \) we infer the bound

\[
\|(Q_a - Q_a)f\|_{s+2,p,\alpha} \lesssim \|(Q_a - Q_a)f\|_{L^\infty} + \|f\|_{s-1,p,\alpha} \|(q_a - q_a)''\|_{s,p',\alpha'} \\
\lesssim \|q_a - q_a\|_{L^{p'}} + \|q_a - q_a\|_{w+3,p,\alpha} \|f\|_{s,p,\alpha}.
\]

The proof in the \( L^p \)-case is performed in a completely analogous way. □

**Proof.** (of Theorem 2.9) The following argument works for the Besov scale as well as for the \( L^p \)-scale. Since the set \( M(s, p, S, \delta) \) is bounded in \( B_{s,p,\alpha} \), it is relatively compact in any \( B_{w,p,\alpha} \) for \( w < s \). By the preceding result the operator norm of \( Q_a \) depends continuously on \( a \) in \( B_{w,p,\alpha} \)-norm for some \( w < s \) such that the supremum of \( \|Q_a\| \) over \( M(s, p, S, \delta) \) is attained and finite. The norm continuity of the mapping \( Q_a \mapsto Q_a^{-1} \) (Rudin 1991, Thm. 10.11) yields the bound for the inverse. □

6. **Proof of the upper bound**

6.1. **Proof of Theorem 3.2.** This result follows from an exponential moment bound on \( \frac{1}{T} Q_T \), which for small exponents \( \alpha \) yields a Gaussian moment behaviour.

**Theorem 6.1.** Let \( \delta, R > 0 \) be given. Then there are constants \( K, T_0 > 0 \) such that for all weight measures \( a \) with \( v_0(a) \leq -\delta, \|a\|_{TV} \leq R \), all multi-indices \( \lambda \), all measures \( \mu \in M([-r,0]) \), all \( T \geq T_0 \) and all \( \alpha \in [0, T^{1/2}(K\|\mu\|_{TV})^{-1}] \) the following moment bound holds true:

\[
\mathbb{E}_a \left[ \cosh\left( \alpha T^{1/2} 2^{3|\lambda|/2} ((\frac{1}{T} Q_T - Q_a)\mu, \psi_\lambda) \right) \right] \leq \exp(K\|\mu\|_{TV}^2 \alpha^2).
\]

In particular, using \( x^{2m} \lesssim \cosh(x) \) we obtain

\[
\mathbb{E}_a \left[ ((\frac{1}{T} Q_T - Q_a)\mu, \psi_\lambda)^{2m} \right]^{1/2m} \lesssim T^{-1/2} 2^{-3|\lambda|/2} \|\mu\|_{TV}.
\]

**Proof.** Due to \( \cosh(x) = \sum_{m} \frac{x^{2m}}{(2m)!} \) we shall estimate polynomial moments. Using the finiteness of \( \mathbb{E}[\|X\|_{C([-r,T])}^{2m}] \) by the Fernique theorem on \( C([-r,T]) \)
as requirement for the Fubini Theorem we obtain:

\[
\mathbb{E}_a[(\frac{1}{T} Q_T - Q_a)\mu, \psi_\lambda)^{2m}] = \mathbb{E}_a[(\mu, (\frac{1}{T} Q_T - Q_a)\psi_\lambda)^{2m}]
\]

\[
= \int_{[r_0]} \mathbb{E}_a \left[ \prod_{i=1}^{2m} (\frac{1}{T} Q_T - Q_a) \psi(u_i) \right] d\mu(u_{2m}) \ldots d\mu(u_1)
\]

\[
\leq \|\mu\|^{2m}_{TV} \sup_{u_1, \ldots, u_{2m}} \left| \int_{[r_0]} \prod_{i=1}^{2m} (\frac{1}{T} Q_T(u_i) - q_a(u_i, v_i)) \right| \prod_{i=1}^{2m} \psi_\lambda(v_i)
\]

\[
= \|\mu\|^{2m}_{TV} \sup_{u_1, \ldots, u_{2m}} \left| \int_{[r_0]} \prod_{i=1}^{2m} (X(t_i + u_i)X(t_i + v_i) - q_a(u_i - v_i)) \right| \prod_{i=1}^{2m} \psi_\lambda(v_i)
\]

In order to evaluate the expected value of the product, let us introduce the set \( P_2(2n) \) of all partitions of the set \( \{1, \ldots, 2n\} \) into subsets with two elements. An easy argument based on the characteristic function shows that for a centered Gaussian random vector \((N_1, \ldots, N_{2n})\) the formula

\[
\mathbb{E} \left[ \prod_{i=1}^{2n} N_i \right] = \sum_{\Gamma \in P_2(2n)} \prod_{(k,l) \in \Gamma} \mathbb{E}[N_k N_l]
\]

is valid. Let us set \( n = 2m \), \( A_i = N_{2i-1} \), \( B_i = N_{2i} \), and \( \alpha = \mathbb{E}[A_i B_i] \).

Then we obtain the following formula because terms involving neighbouring random variables \( N_{2i-1}, N_{2i} \) cancel (proof by induction over \( n \)):

\[
\mathbb{E}_a \left[ \prod_{i=1}^{2m} (A_i B_i - \alpha) \right] = \sum_{\Gamma \in P_2(4m)} \prod_{(k,l) \in \Gamma} \mathbb{E}_a[N_k N_l]. \quad (6.2)
\]

In our case the expected value of the product equals

\[
\sum_{\Gamma \in P_2(4m)} \prod_{(k,l) \in \Gamma} q_a(z_k - z_l)
\]

with \( z_{2i-1} = t_i + u_i \) and \( z_{2i} = t_i + v_i \). Changing the order of integration, we start with the integration over \( v_i, i = 1, \ldots, 2m \). Since any \( v_i \) appears only
once in the product, we have to deal with products over terms which have one of the following three forms:

\[
q_a(t_i + u_i - t_j - u_j), \quad (I),
\]

\[
\int_{-r}^0 q_a(t_i + u_i - t_j - v_j)\psi_\lambda(v_j) \, dv_j \quad (II),
\]

\[
\int_{-r}^0 \int_{-r}^0 q_a(t_i + v_i - t_j - v_j)\psi_\lambda(v_i)\psi_\lambda(v_j) \, dv_i \, dv_j \quad (III).
\]

For the factor (I) we shall use \( |q_a(t_i + u_i - t_j - u_j)| \leq C_1 e^{-\delta|t_i - t_j|} \) derived from Proposition 5.1 for \( \delta < -v_0(a) \).

The Lipschitz constant of \( q_a(t_i + u_i - t_j - \bullet) \) on \([-r, 0]\) is of order \( e^{-\delta(|t_i - t_j| - r)} \) by Proposition 5.1, which implies the existence of a constant \( C_2 \) such that the modulus of the integral (II) is smaller than \( C_2 2^{-3|\lambda|/2}e^{-\delta|t_i - t_j|} \) (Lemma 8.9).

For the estimation of the integral (III) we let \( S \) denote the length of the minimal interval supporting \( \psi \) and distinguish the cases (1) \( |t_i - t_j| > 2^{-|\lambda|}S \) and (2) \( |t_i - t_j| \leq 2^{-|\lambda|}S \). A substitution gives

\[
\int_{-r}^0 \int_{-r}^0 q_a(t_i + v_i - t_j - v_j)\psi_\lambda(v_i)\psi_\lambda(v_j) \, dv_i \, dv_j
\]

\[
= \int \int_{|v_i - v_j| \leq S} q_a(t_i - t_j + 2^{-|\lambda|}(v_i - v_j))2^{-|\lambda|}\psi(v_i)\psi(v_j) \, dv_i \, dv_j,
\]

which shows that in case (1) \( q_a \) needs only to be evaluated at either positive arguments or at negative ones. Due to the Lipschitz continuity of \( q'_a \) with exponentially decaying norm (Proposition 5.1) the estimate in Lemma 8.9 shows that the modulus of (III) is in case (1) smaller than \( C_3 2^{-3|\lambda|}e^{-\delta|t_i - t_j|} \), \( C_3 > 0 \) a constant. In case (2) \( q_a \) is at least Lipschitz continuous and the modulus of (III) is by the same arguments smaller than \( C_4 2^{-2|\lambda|}e^{-\delta|t_i - t_j|} \), \( C_4 > 0 \) a constant.

Finally note that each \( u_i \) and \( v_i \) appears exactly once in the product and that each \( t_i \) appears twice so that with \( C := \max_j C_j \)

\[
\left| \int_{[-r, 0]^{2m}} dv_{2m} \ldots dv_1 \sum_{a} \left\langle \prod_{i=1}^{2m} (X(t_i + u_i)X(t_i + v_i) - q_a(u_i - v_i)) \prod_{i=1}^{2m} \psi_\lambda(v_i) \right\rangle \right| ^m
\]

\[
\leq \sum_{\Gamma} 2^{-3|\lambda|m} C^{2m} \prod_{(k,l) \in \Gamma} (1 + 2^{|\lambda|}1_{k \text{ even}, |(k/2 - l/2)| \leq S |\lambda|}) e^{-\delta|t_{(k/2)} - t_{(l/2)}|}. 
\]
The partitions $\Gamma$ can also be described by fixed point-free permutations. Let us denote $2k - 1$ and $2k$ by the same symbol $s(k)$. The idea is to start with one pair $\{k_0, k_1\} \in \Gamma$, to look for $\{k'_0, k_2\} \in \Gamma$ with $s(k'_0) = s(k_1)$, then for $\{k'_2, k_3\}$ with $s(k'_2) = s(k_2)$ and so forth until $s(k_1)$ equals $s(k_0)$. This describes a cyclic permutation of $\{s(k_0), \ldots, s(k_{l-1})\}$. Proceeding in the same manner for the remaining elements of $\Gamma$ and identifying $s(k)$ with $[k/2]$ a fixed point-free permutation $\pi = \pi(\Gamma)$ of $\{1, \ldots, 2m\}$ is defined. To clarify the construction look at the following example ($m = 6$):

\[
\Gamma = \{\{1, 3\}, \{2, 11\}, \{4, 7\}, \{5, 10\}, \{6, 9\}, \{8, 12\}\}
\]
\[
\Rightarrow s(1) \mapsto s(2) \mapsto s(4) \mapsto s(6); \quad s(3) \mapsto s(5) \Rightarrow \pi(G) = (1\, 2\, 4\, 6)\, (3\, 5).
\]

Let us denote by $C(\pi)$ the set of cycles in $\pi$ and by $|\tau|$ the length of a cycle $\tau$. Then we can easily evaluate the integral over the product:

\[
\int_{[0,T]^{2m}} \prod_{(k,l) \in \Gamma} \left(1 + 2^{|\lambda|} \mathbf{1}_{\{k,l\ \text{even}; |k/2 - \ell/2| \leq S2^{-|\lambda|}\}}\right) e^{-\delta|t_{k/2} - t_{\ell/2}|} dt_1 \ldots dt_{2m}
\]
\[
= \prod_{\tau \in C(\pi(\Gamma))} \int_{[0,T]^{|\tau|}} \prod_{k=1}^{|\tau|} \left(1 + 2^{|\lambda|} \mathbf{1}_{|s_{k+1} - s_k| \leq 2^{-|\lambda|}}\right) e^{-\delta|s_{k+1} - s_k|} ds_1 \ldots ds_{|\tau|}
\]
\[
\leq \prod_{\tau \in C(\pi(\Gamma))} \int_{0}^{T} ds_1 \int_{-T}^{T} dt_1 \prod_{k=1}^{|\tau| - 1} \left(1 + 2^{|\lambda|} \mathbf{1}_{|u_k| \leq 2^{-|\lambda|}}\right) e^{-\delta|u_k|} du_1 \ldots, du_{|\tau| - 1}
\]
\[
\leq \prod_{\tau \in C(\pi(\Gamma))} \left(T(2\delta^{-1} + 2S)^{|\tau|-1}\right)
\]
\[
\leq T^{|C(\pi(\Gamma))|}(2\delta^{-1} + 2S)^{2m}.
\]

So far we have shown

\[
\mathbb{E}_u[(\frac{1}{2} Q_T - Q_0)\mu, \psi_\lambda)^{2m}] \leq \|\mu\|^{2m} T^{-2m} \left(2C(\delta^{-1} + S)\right)^{2m} 2^{-3|m|} \sum_{\Gamma} T^{\|C(\pi(\Gamma))\|}.
\]

It remains to solve the combinatorial problem to determine the number $a_{n,k}$ of fixed point-free permutations of $\{1, \ldots, n\}$ with exactly $k$ cycles. We claim that the following recursive relation is true for all $n \geq 3, k \geq 1$:

\[
a_{n,k} = (n - 1)a_{n-1,k} + (n - 1)a_{n-2,k-1}, \quad a_{n,1} = a_{n,0} = 1, \quad a_{1,k} = 0.
\]

We classify with regard to the element $n$. If in a permutation $n$ is in a cycle of length at least three, then by leaving $n$ away, we obtain a fixed-point free
permutation of \{1, \ldots, n-1\} with \(k\) cycles. Since \(n\) can stand in front of every other element, there are exactly \(n-1\) possibilities to generate from a valid \((n-1)\)-permutation such an \(n\)-permutation. This explains the first term, the second stems from the permutations where \(n\) lies in a cycle of length 2. By removing this 2-cycle we obtain a fixed point-free permutation of \(n-2\) elements with \(k-1\) cycles. Since the other element of the cycle involving \(n\) can be chosen from all other elements, we find the second summand.

From this recursive relationship we infer by an easy induction argument that the generating function satisfies

\[
\sum_{m=0}^{\infty} \frac{\|\mu\|^2 TV T^{-m} (2C(\delta^{-1} + S))^2 \alpha^m}{(2m)!} \sum_{k=1}^{2m} a_{2m,k} T^k \leq \frac{2m}{m!} (x+1) \cdots (x+m), \quad x \geq 0.
\]

Now we are in the position to prove the assertion of the theorem:

\[
\mathbb{E}_a[\cosh(\alpha T^{1/2} q \langle \mathbf{1} | Q_T - Q_a | \mathbf{1} \rangle \mu, \psi_{\lambda})] = \sum_{m=0}^{\infty} \frac{\|\mu\|^2 TV T^{-m} (2C(\delta^{-1} + S))^2 \alpha^m}{(2m)!} \sum_{k=1}^{2m} a_{2m,k} T^k 
\]

where \(K := 2(2C(\delta^{-1} + S))^2\) and \(T \geq T_0\) large enough. The estimates of the covariance function relied only on Proposition 5.1 whence by Proposition 5.4 the uniformity of the constant follows. \(\square\)

**Proof.** (of Theorem 3.2) The moment inequality in Proposition 6.1 yields

\[
\mathbb{P}_a(T^{1/2} q \langle \mathbf{1} | Q_T - Q_a | \mathbf{1} \rangle \mu, \psi_{\lambda}) \geq \frac{x}{2} \sqrt{T} \leq \frac{\exp(K\|\mu\|^2 TV^2 \alpha^2)}{\cosh(\alpha^2 \sqrt{T})}
\]
for any suitable $\alpha$. The choice $\alpha = \frac{\kappa}{2}\sqrt{\log T/(2K\|\mu\|_{TV}^2)}$ yields the bound $2T^{-\kappa^2/(16K\|\mu\|_{TV}^2)}$. From the decomposition (3.1) it follows that

$$|\beta_{\lambda,T} - \langle Q_\lambda a, \psi_\lambda \rangle| \leq \left| \left( \langle \frac{1}{T} Q_T - Q_\lambda a, \psi_\lambda \rangle \right) \right| + \frac{1}{T} \int_0^T \langle X(t + \bullet), \psi_\lambda \rangle \, dW(t).$$

The stochastic integral has quadratic variation $\langle Q_T \psi_\lambda, \psi_\lambda \rangle$ and by the exact deviation probability bound found by Liptser and Spokoiny (2000) we infer for any $\kappa > 0$ and large $T$

$$\mathbb{P}_a\left( \frac{2|\lambda|}{T} \left| \int_0^T \langle X(t + \bullet), \psi_\lambda \rangle \, dW(t) \right| > \frac{\kappa}{2} \sqrt{2^2|\lambda|(Q_\lambda \psi_\lambda, \psi_\lambda) T \log T} \right)$$

$$\leq 4\sqrt{\kappa}(\log T)T^{-\kappa^2/(8+2\kappa(\log T)^{1/2}T^{1/2})} + \mathbb{P}_a\left( \frac{2|\lambda|}{T} \left| \left( \langle \frac{1}{T} Q_T - Q_\lambda a \psi_\lambda, \psi_\lambda \rangle \right) \right| > \frac{\kappa}{2} \sqrt{T^{-1} \log T} \right)$$

$$\lesssim T^{-\kappa^2/9} + T^{-\kappa^2/(16K\|\psi\|_{L^1}^2)}.$$ 

By Lemma 8.8 the expression $2^2|\lambda|(Q_\lambda \psi_\lambda, \psi_\lambda)$ is uniformly bounded from below by some $m > 0$ and we obtain the uniform estimate

$$\mathbb{P}_a\left( \frac{2|\lambda|}{T} \left| \beta_{\lambda,T} - \langle Q_\lambda a, \psi_\lambda \rangle \right| > \frac{\kappa}{2} \sqrt{\log T} \right)$$

$$\lesssim T^{-\kappa^2/2|\lambda|/(16K\|a\|_{TV}^2)} + T^{-\kappa^2/(9m^2)} + T^{-\kappa^2/(16K\|\psi\|_{L^1}^2 m^2)}.$$ 

If we choose $\kappa^2 \geq \max(48KR^2, 48m^2\|\psi\|_{L^1}^2, 27m^2)\rho =: \kappa^*$, then the right hand side is of maximal order $T^{-3\rho}$.

## 6.2. Proof of Proposition 3.3.

**Proof.** Without loss of generality we assume $p < \rho$ and we omit the $T$-dependence of the quantities. Let us introduce the true coefficients $(b_{j,k})$ and error coefficients $(e_{j,k})$

$$b_{j,k} := \langle Q_\lambda a, \psi_{j,k} \rangle, \quad e_{j,k} := \beta_{j,k} - b_{j,k} = \langle \frac{1}{T} b_T - Q_\lambda a, \psi_{j,k} \rangle.$$ 

We split the risk according to the usual bias-variance decomposition:

$$\mathbb{E}_a[\|\hat{b}_T - Q_\lambda a\|_{W^{2,p}}] \leq \mathbb{E}_a[\|\hat{b}_T - P_J Q_\lambda a\|_{W^{2,p}}] + \|(\text{Id} - P_J) Q_\lambda a\|_{W^{2,p}}.$$ 

The second (bias) term can be dealt with by linear approximation theory. The Besov space embeddings (8.1) yield under the restriction (3.3) that
$B_{p,1}^{s+2} \subset W^{2+2,\rho}$ together with Jackson’s inequality (8.2) in $W^{2,\rho}([-r,0])$

$$
|| (\mathrm{Id} - P_J) Q_a ||_{W^{2,\rho}} \lesssim 2^{-J} \frac{\rho}{\pi \beta} || Q_a ||_{W^{2,\rho}} \lesssim (T/\log T)^{-\frac{\rho}{\pi \beta}} || Q_a ||_{B_{p,1}^{s+2}}.
$$

Due to $Q_a : B_{p,1}^s \to B_{p,1}^{s+2}$ isomorphically (Theorem 2.8) with uniform constants (Theorem 2.9), this second term is of order $(T/\log T)^{-\frac{\rho}{\pi \beta}}$ uniformly over $M(s,p,S,\delta)$.

The first term can be estimated using the imbedding $B_{p,1}^2([-r,0]) \subset W^{2,\rho}([-r,0])$, the characterisation of $B_{p,1}^2$ by 2-regular wavelets (Appendix 8.2) and Jensen’s inequality:

$$
\mathbb{E}_a[|| \hat{b}_T - P_J Q_a ||_{W^{2,\rho}}] \lesssim \mathbb{E}_a \left[ \sum_{j \leq J} 2^{j(\frac{3}{2} - \frac{1}{\rho})} \left( \sum_k |\beta_{j,k}^1|_{|\beta_{j,k}| > \kappa_j} - |b_{j,k}|^{\rho} \right)^{1/\rho} \right]
\lesssim \sum_{j \leq J} 2^{j(\frac{3}{2} - \frac{1}{\rho})} \left( \sum_k \mathbb{E}_a \left[ |\beta_{j,k}^1|_{|\beta_{j,k}| > \kappa_j} - |b_{j,k}|^{\rho} \right] \right)^{1/\rho}.
$$

The term $|\beta_{j,k}^1|_{|\beta_{j,k}| > \kappa_j} - |b_{j,k}|^{\rho}$ can be split according to whether thresholding takes place or not and whether the true coefficient is large or not. It equals

$$
|\beta_{j,k} - b_{j,k}|^{\rho} 1_{|\beta_{j,k}| > \kappa_j} + |b_{j,k}|^{\rho} 1_{|\beta_{j,k}| \leq \kappa_j}
= |e_{j,k}|^{\rho} 1_{|e_{j,k}| > \kappa_j/\beta_{j,k}} + |b_{j,k}|^{\rho} 1_{|\beta_{j,k}| \leq \kappa_j/\beta_{j,k}} + |b_{j,k}|^{\rho} 1_{|\beta_{j,k}| > 2\kappa_j/\beta_{j,k}} + |b_{j,k}|^{\rho} 1_{|\beta_{j,k}| \leq 2\kappa_j/\beta_{j,k}}
\leq |e_{j,k}|^{\rho} 1_{|e_{j,k}| > \kappa_j/2} + |e_{j,k}|^{\rho} 1_{|\beta_{j,k}| > \kappa_j/2} + |b_{j,k}|^{\rho} 1_{|\beta_{j,k}| > \kappa_j} + |b_{j,k}|^{\rho} 1_{|\beta_{j,k}| \leq 2\kappa_j/\beta_{j,k}}
=: S_1(j,k) + S_2(j,k) + S_3(j,k) + S_4(j,k).
$$

By the Cauchy-Schwarz inequality, the large deviation bound (3.2) on $e_{j,k}$ and by (6.1) we obtain a fast decay for the sum involving $S_1(j,k)$:

$$
\sum_{j \leq J} 2^{j(\frac{3}{2} - \frac{1}{\rho})} \left( \sum_k \mathbb{E}_a |S_1(j,k)| \right)^{1/\rho}
\leq \sum_{j \leq J} 2^{j(\frac{3}{2} - \frac{1}{\rho})} \left( \sum_k \mathbb{E}_a |e_{j,k}|^{\rho} 1_{|e_{j,k}| > \kappa_j/2} \right)^{1/\rho}
\leq \sum_{j \leq J} 2^{j(\frac{3}{2} - \frac{1}{\rho})} \left( \sum_k T^{-3\rho/2} 2^{-j\rho} T^{-\rho/2} \right)^{1/\rho}
\sim T^{-2-3J/2} \lesssim T^{-1/2}.
$$
The large deviation estimate also bounds the sum over $S_3(j, k)$:

$$\sum_{j \leq J} 2^{j(\frac{s}{p} - \frac{1}{p})} \left( \sum_k \mathbb{P}_a[S_3(j, k)] \right)^{1/p} \leq \sum_{j \leq J} 2^{j(\frac{s}{p} - \frac{1}{p})} \left( \sum_k \mathbb{P}_a(|e_{j,k}| > \frac{\kappa_j}{2})|b_{j,k}|^p \right)^{1/p} \lesssim \|Q_a a\|_{2, p, 1} T^{-3} \lesssim T^{-3}\|a\|_{B_{p, 1}^s}.$$

The remaining estimates rely on nonlinear approximation theory. Using the characterisation by $(s + 2)$-regular wavelets (Appendix 8.2)

$$\|Q_a a\|_{B_{p, p}^{s+2}} \sim \left( \sum_{j \geq 0} 2^{j(p+(s+\frac{5}{2})/p)} \left( \sum_k |b_{j,k}|^p \right)^{1/p} \right)^{p/p},$$

we infer for all $j \in \mathbb{N}_0$ and $\tau_j > 0$ by a Chebyshev inequality-type argument the following bound on the cardinality of large wavelet coefficients:

$$\sum_{j \geq 0} 2^{j(p+(s+\frac{5}{2})/p)\rho/\rho_j} |\{k \mid |b_{j,k}| \geq \tau_j\}|^{1/p} \rho_j^{p/p} \lesssim \|Q_a a\|_{B_{p, p}^{s+2}} \leq \|Q_a a\|_{B_{p, p}^{s+2}}. \quad (6.3)$$

The sum involving $S_2(j, k)$ can be bounded by separate estimates, where $j_0$ is such that $2^{j_0} \sim T^{\frac{1}{2\tau_0 + \frac{s}{p}}}$:

$$\sum_{j \leq J} 2^{j(\frac{s}{p} - \frac{1}{p})} \left( \sum_k \mathbb{P}_a[S_2(j, k)] \right)^{1/p} = \sum_{j \leq J} 2^{j(\frac{s}{p} - \frac{1}{p})} \left( \sum_k \mathbb{P}_a[|e_{j,k}|^p 1_{|b_{j,k}| > \kappa_j/2}] \right)^{1/p} \lesssim \sum_{j \geq j_0} 2^{j(\frac{s}{p} - \frac{1}{p})} T^{-1/2} 2^{-j} + \sum_{j > j_0} 2^{j(\frac{s}{p} - \frac{1}{p})} \left( \sum_k T^{-p/2} 2^{-j} 1_{|b_{j,k}| > \kappa_j/2} \right)^{1/p} \lesssim T^{-1/2} 2^{3j_0/2} + T^{-1/2} 2^{3j_0/2} \sum_{j > j_0} 2^{j(\frac{s}{p} - \frac{1}{p})} \left( \frac{\kappa_j}{2} \right)^{p/p} |\{k \mid |b_{j,k}| > \kappa_j/2\}|^{1/p} \lesssim T^{-1/2} 2^{3j_0/2} + T^{-1/2} 2^{-j_0} 2^{-3j_0/2} 2^{-j_0(\frac{3p}{2p} - \frac{3}{2})} \sim T^{-1/2} 2^{3j_0/2} (1 + T^{p/2 p - j_0(s+\frac{5}{2})/p}) \sim T^{-\frac{3}{2\tau_0 + \frac{s}{p}}}.$$

In the fifth line we have used the sparsity estimate (6.3) and the fact that $\frac{ps}{p} + \frac{3p}{2p} - \frac{3}{2} \geq 0$ is non-negative due to assumption (3.3).

The slightly extended technique also applies to the estimate of the sum over $S_4(j, k)$. Here, one must choose $2^{j_0} \sim (T/\log T)^{\frac{1}{2\tau_0 + \frac{s}{p}}}$ for balancing the
two appearing sums:
\[ \sum_{j \leq J} 2^{j(\frac{\gamma}{p} - \frac{1}{p})} \left( \sum_k |b_{j,k}|^\rho 1_{|b_{j,k}| \leq 2^{\kappa_j}} \right)^{1/\rho} \]
\[ = \sum_{j \leq J} 2^{j(\frac{\gamma}{p} - \frac{1}{p})} \left( \sum_k |b_{j,k}|^\rho \sum_{m \geq 0} 1_{2^{-m} \kappa_j < |b_{j,k}| \leq 2^{-m+1} \kappa_j} \right)^{1/\rho} \]
\[ \leq \sum_{j \leq J_0} 2^{j(\frac{\gamma}{p} - \frac{1}{p})} 2^{j/\rho} 2 \kappa_j + \sum_{j > J_0} 2^{j(\frac{\gamma}{p} - \frac{1}{p})} \left( \sum_k |b_{j,k}|^\rho \sum_{m \geq 0} 1_{2^{-m} \kappa_j < |b_{j,k}| \leq 2^{-m+1} \kappa_j} \right)^{1/\rho} \]
\[ \leq 2^{3j_0/2} T^{-1/2} (\log T)^{1/2} + \sum_{j > J_0} 2^{j(\frac{\gamma}{p} - \frac{1}{p})} \left( \sum_k \sum_{m \geq 0} 2^{(m+1) \rho} \kappa_j^m 1_{|b_{j,k}| > 2^{-m} \kappa_j} \right)^{1/\rho} \]
\[ \leq \left( \frac{T}{\log T} \right)^{-\frac{s}{2 + \gamma}} + \sum_{m \geq 0} \sum_{j > J_0} 2^{j(\frac{\gamma}{p} - \frac{1}{p})} 2^{-m} \kappa_j |\{k \mid |b_{j,k}| > 2^{-m} \kappa_j\}|^{1/\rho} \]
\[ \leq \left( \frac{T}{\log T} \right)^{-\frac{s}{2 + \gamma}} + \sum_{m \geq 0} 2^{-j_0(\frac{\gamma}{p} - \frac{1}{p})} \left( \frac{2^{-m} (\log T)^{1/2}}{T^{1/2}} \right)^{(\rho-p)/\rho} \]
\[ \sim \left( \frac{T}{\log T} \right)^{-\frac{s}{2 + \gamma}} . \]

All estimates together yield \( \mathbb{E}_a[|\|\hat{b}_T - Q_a a\|_{W^2,\rho}|] \lesssim (T/(\log T))^{-\frac{s}{2 + \gamma}} \) with uniform constants. \( \square \)

6.3. **Proof of Theorem 3.4.**

**Proof.** Due to \( \hat{b}_T \in W^{2,\rho}([-r,0]) \) the kernel \( \hat{q}_T|_{[0,r]} \) is an element of \( W^{3,\rho}([0,r]) \) and the continuity of \( \hat{Q}_T : L^\rho \rightarrow W^{2,\rho}([-r,0]) \) follows from Lemma 8.4. Formally, the Neumann series expansion yields for \( \hat{Q}_T^{-1} \)

\[ \hat{Q}_T^{-1} = (\text{Id} - Q_a^{-1}(Q_a - \hat{Q}_T))^{-1} Q_a^{-1} = \sum_{m=0}^{\infty} (Q_a^{-1}(Q_a - \hat{Q}_T))^m Q_a^{-1} . \]

Introducing the random set

\[ \mathcal{C}_T := \{ \|Q_a^{-1}\|_{W^{2,\rho} \rightarrow L^\rho} \|Q_a - \hat{Q}_T\|_{L^\rho \rightarrow W^{2,\rho}} \leq 1/2 \} , \]

the operator \( \hat{Q}_T \) is therefore invertible on \( \mathcal{C}_T \) with

\[ \|\hat{Q}_T^{-1}\|_{W^{2,\rho} \rightarrow L^\rho} \leq 2\|Q_a^{-1}\|_{W^{2,\rho} \rightarrow L^\rho} , \]
\[ \|\hat{Q}_T^{-1} - Q_a^{-1}\|_{W^{2,\rho} \rightarrow L^\rho} \leq 2\|Q_a^{-1}\|_{W^{2,\rho} \rightarrow L^\rho} \|\hat{Q}_T - Q_a\|_{L^\rho \rightarrow W^{2,\rho}} . \]

In order to bound the probability of \( \mathcal{C}_T \) from below, we use the estimate

\[ \|Q_a - \hat{Q}_T\|_{L^\rho \rightarrow W^{2,\rho}} \lesssim \|q_a - \hat{q}_T\|_{W^{3,\rho}([-r,0])} , \]

derived from Lemma 8.4 with
$k(t) = (q_a - \hat{q}_T)(-t)$ and $W^{3,\rho} \subset W^{2,\rho'}$. From Proposition 3.3 we know

$$E_a[\|q_a' - \hat{q}_T'\|_{W^{2,\rho}}] = E_a[\|Q_a a(-\cdot) - \hat{b}_T(-\cdot)\|_{W^{2,\rho}}] \lesssim \left( \frac{T}{\log T} \right)^{-\frac{s}{2s+3}}.$$  

Furthermore, we infer from Propositions 5.1 and 5.4

$$E_a \left[ |q_a(0) - \frac{1}{T} \int_0^T X(u)^2 \, du |^2 \right] = \frac{1}{T^2} \int_0^T \int_0^T 2q_a^2(u - v) \, du \, dv \lesssim \frac{1}{T}$$

with uniform constants. We conclude

$$E_a[\|q_a - \hat{q}_T\|_{W^{3,\rho}}] \lesssim \left( \frac{T}{\log T} \right)^{-\frac{s}{2s+3}}.$$  \hspace{1cm} (6.4)

Finally, Markov’s inequality yields for suitable $c > 0$

$$\sup_{a \in M(s, p, S, \delta)} \mathbb{P}_a(\Omega \setminus C_T) \leq \sup_{a \in M(s, p, S, \delta)} \mathbb{P}_a(\|q_a - \hat{q}_T\|_{W^{3,\rho}} > c) \leq \sup_{a \in M(s, p, S, \delta)} E_a[\|q_a - \hat{q}_T\|_{W^{3,\rho}}] c^{-1} \lesssim \left( \frac{T}{\log T} \right)^{-\frac{s}{2s+3}}.$$

It therefore suffices to work on the set $C_T$, because on its complement the loss is bounded by $2S$. Since our renormalisation uses the a priori knowledge $\|a\|_{L^p} \leq S$, our estimator is on $C_T$ only up to a constant factor worse than the estimator obtained by pure inversion. We obtain on $C_T$

$$\|\hat{a}_T - a\|_{L^p} \lesssim \|\hat{Q}_T^{-1}\hat{b}_T - Q_a^{-1}Q_a a\|_{L^p} \lesssim \|\hat{Q}_T^{-1}\|_{W^{2,\rho} \to L^p} \|\hat{b}_T - Q_a a\|_{W^{2,\rho}}$$

$$+ \|\hat{Q}_T^{-1} - Q_a^{-1}\|_{W^{2,\rho} \to L^p} \|Q_a a\|_{W^{2,\rho}} \leq 2\|Q_a^{-1}\|_{W^{2,\rho} \to L^p} \|\hat{b}_T - Q_a a\|_{W^{2,\rho}}$$

$$+ 2\|Q_a^{-1}\|_{W^{2,\rho} \to L^p} \|Q_a a\|_{W^{2,\rho}} \lesssim \|\hat{b}_T - Q_a a\|_{W^{2,\rho}} + \|q_a - \hat{q}_T\|_{W^{3,\rho}}.$$  

By Theorem 2.9 the last estimate holds uniformly for all $a \in M(s, p, S, \delta)$.

From Proposition 3.3 and the Estimate (6.4) we conclude

$$\sup_{a \in M(s, p, S, \delta)} E_a[\|\hat{a}_T - a\|_{L^p} 1_{C_T}] \lesssim T^{-\frac{s}{2s+3}},$$

which accomplishes the proof of the asymptotic risk upper bound.  \hspace{1cm} \Box
7. Proof of the lower bound

Proof. (of Theorem 4.1) We build from a weight $a_0$ in the interior of $M(s, p, S, \delta)$ a family of local alternatives $(a_{jk})$. Choose a compactly supported $s$-regular wavelet basis in $L^2(\mathbb{R})$ and denote by $R_j$ a maximal set of integers with $\text{supp}(\psi_{jk}) \subset [-r, 0]$ and $\text{supp}(\psi_{jk}) \cap \text{supp}(\psi_{jk'}) = \emptyset$ for all $k, k' \in R_j$, $k \neq k'$. For any $k \in R_j$ we set $a_{jk} := a_0 + \gamma \psi_{jk}$ with $\gamma = \gamma(T) \sim 2^{-j(T)(s+\frac{1}{2}-\frac{1}{p})}$ such that $\|a_{jk}\|_{s, p, 1} \leq S$ and $v_0(a_{jk}) \leq -\delta$ are satisfied, hence $a_{jk} \in B(s, S, p, \delta)$ holds true. We are using a classical lemma on lower bounds in the sparse case: Suppose the likelihood ratio satisfies

$$\mathbb{P}_{a_{jk}}\left(\log(\Lambda_T(X(a_0), X(a_{jk}))) \geq -j\right) \geq p_s > 0$$

uniformly for all $a_{jk}$. Then for $\mathcal{F}^X_T$-measurable estimators $\hat{a}_T$ the following lower bound holds:

$$\inf_{\hat{a}_T} \sup_{a \in M(s, p, S, \delta)} \mathbb{E}_a[\|\hat{a}_T - a\|_{L^p}] \geq \gamma(T)2^{j(T)(\frac{1}{2} - \frac{1}{p})} \sim 2^{-j(T)(s+\frac{1}{2}-\frac{1}{p})}.$$ 

This is an adapted version of (Korostelev and Tsybakov 1993, Theorem 2.5.3). Merely note the relations $M \sim 2^j$ and $s_T \sim \gamma 2^{j(\frac{1}{2} - \frac{1}{p})}$ in their statement, having substituted $n$ by $T$.

We use the likelihood ratio from Theorem 2.5 with some fixed initial condition and apply Lemma 8.8 and Estimate (6.1):

$$\mathbb{E}_{a_{jk}}[\log(\Lambda_T(X(a_0), X(a_{jk}))^2] 
= \mathbb{E}_{a_{jk}}\left[\left(\int_0^T \langle X(t + \bullet), a_{jk} - a_0 \rangle dW(t) - \frac{1}{2}(Q_T(a_{jk} - a_0), a_{jk} - a_0)\right)^2\right]
\leq 2\gamma^2 T\langle Q_{a_{jk}} \psi_{jk}, \psi_{jk} \rangle + \frac{1}{2} \mathbb{E}_{a_{jk}}[\langle Q_T(a_{jk} - a_0), a_{jk} - a_0 \rangle^2]
\leq 2\gamma^2 T\langle Q_{a_{jk}} \psi_{jk}, \psi_{jk} \rangle + \gamma^4 T^2\langle Q_{a_{jk}} \psi_{jk}, \psi_{jk} \rangle^2
+ \gamma^4 \mathbb{E}_{a_{jk}}[\langle (Q_T - TQ_{a_{jk}}) \psi_{jk}, \psi_{jk} \rangle^2]
\lesssim \gamma^2 T 2^{-2j} + \gamma^4 T^2 2^{-4j} + \gamma^4 T 2^{-4j}$$

with a uniform constant for all $a_{jk}$. Thus, by Chebyshev’s inequality the requirements of the lower bound Lemma are satisfied, when we balance the restrictions on $\gamma$ by choosing $2^{(2s+3-\frac{2}{p})j(T)} \sim T/ \log T$ such that

$$\gamma(T)2^{j(T)(4s+6-\frac{4}{p})} \sim (\log T)^2 \sim j(T)^2$$
holds. The lower bound follows. □

Proof. (of Corollary 4.2)

(1) The lower bound is just Theorem 4.1 properly rewritten. For the upper bound use the embedding \( B_s^p \subset B_w^\pi \) with \( \frac{1}{\pi} = w - s + \frac{1}{p} < \frac{1}{p} \). Due to \( \frac{1}{\pi} - \frac{1}{p} = \frac{2w}{p^3} \) we can apply Theorem 3.4 to the class \( M(w, \pi, S', \delta) \), \( S' \) chosen appropriately.

(2) The upper bound is the content of Theorem 3.4, whereas the lower bound follows along the lines of the \( L^2 \)-lower bound proof using Assouad’s cube in Reiß (2002). The details are omitted. □

8. Appendix

8.1. Function spaces. For a more detailed account see Triebel (1983). Let us introduce the scale of Sobolev spaces \( W^{m,p}(I) \), \( m \in \mathbb{N} \), \( p \in [1, \infty) \), \( I \subset \mathbb{R} \) an interval: \( W^{m,p}(I) := \{ f \in L^p(I) \mid f^{(i)} \in L^p(I) \text{ for all } i = 0, \ldots, m \} \), where \( f^{(i)} \) denotes the \( i \)-th derivative of \( f \) in a weak sense. They are Banach spaces with respect to the norm \( ||f||_{m,p} := (\sum_{i=0}^{m} ||f^{(i)}||_{L^p(I)})^{1/p} \).

Besov spaces \( B_{p,\alpha}^s \), measure the regularity \( s \) in an \( L^p \)-sense with an additional fine-tuning parameter \( \alpha \in [1, \infty] \).

Definition 8.1. Let \( I \subset \mathbb{R} \) be an interval, \( \Delta_h f(x) := f(x+h) - f(x) \) and \( I_h := \{ x \in I \mid x \pm h \in I \} \). Then the \( n \)-th order \( L^p \)-modulus of smoothness is defined by

\[
\omega_n(f, \varepsilon)_p := \sup_{|h| \leq \varepsilon} ||\Delta_h^n f||_{L^p(I_h)},
\]

where \( \Delta_h^n \) denoting the \( n \)-fold application of \( \Delta_h \). For \( p, \alpha \in [1, \infty] \) and \( s > 0 \) set

\[
||f||_{s,p,\alpha} := ||f||_{L^p(I)} + \left( \int_0^{1} \left( \frac{\omega_{n}(f, t^s)_p}{t^{s}} \right)^{\alpha} \frac{dt}{t} \right)^{1/\alpha}
\]

with the usual modification \( \sup_{t} \omega_{n}(f, t^s)_p t^{-s} \) for \( \alpha = \infty \) and with \( n = [s] + 1 \). The Besov space \( B_{p,\alpha}^s(I) := \{ f \in L^p(I) \mid ||f||_{s,p,\alpha} < \infty \} \) is a Banach space when equipped with the norm \( ||\cdot||_{s,p,\alpha} \). On a bounded interval an equivalent norm is given by (\( n \) as above) \( ||f||_{B_{p,\alpha}^s} \sim ||f||_{L^p} + ||f^{(n-1)}||_{s-(n-1),p,\alpha} \).
Proposition 8.2. The following embedding relations hold true: \(B^s_{p,\alpha} \subset B^{s'}_{p',\alpha'}, s > s', \text{ any } \alpha, \alpha'; \) \(B^s_{p,\alpha} \subset B^s_{p',\alpha}, p > p'; \) \(B^s_{p,\alpha} \subset B^s_{p',\alpha'}, \alpha < \alpha'; \) the Sobolev embedding theorem generalizes to
\[
B^s_{p,\alpha} \subset B^{s'}_{p',\alpha} \quad \text{for } s \geq s' \text{ and } s - \frac{1}{p} \geq s' - \frac{1}{p'};
\]
(8.1)
as a special case \(B^s_{p,\alpha} \subset C^{s'} \text{ for } s - \frac{1}{p} > s' \) follows. The first embedding is compact for Besov spaces on bounded intervals.

The regularity property of convolutions with variable integral bound seems obvious, but not to be treated in the literature.

Lemma 8.3. For functions \(f \in B^s_{p,\alpha}([-r,0]) \) and \(k \in B^{s+1}_{p',\alpha'}([0,r]), s > 0 \) and \(p, p' \in (1, \infty), \alpha, \alpha' \in [1, \infty] \) with \(\frac{1}{p} + \frac{1}{p'} = \frac{1}{\alpha} + \frac{1}{\alpha'} = 1 \), set
\[
L(f,k)(t) := \int_0^t f(u-t)k(u) du, \quad t \in [0,r].
\]
Then \(L \) is a bilinear mapping from \(B^s_{p,\alpha}([-r,0]) \times B^{s+1}_{p',\alpha'}([0,r]) \) to \(B^{s+1}_{p,\alpha}([0,r]) \) with
\[
\|L(f,k)\|_{s+1,p,\alpha} \lesssim \|f\|_{s,p,\alpha} \|k\|_{s+1,p',\alpha'}.
\]

Proof. First, we show for a fixed function \(f \) in \(L^p([-r,0]) \) that \(Tk := L(f,k) \) maps \(B^s_{p,\alpha}([0,r]) \) to \(B^s_{p,\alpha}([0,r]) \) for \(s \in (0,1) \) and all \(p \) and \(\alpha \).

In order to apply abstract interpolation theory, we consider the case \(s = 1 \) in a Sobolev scale first:
\[
\|Tk\|_{W^{1,p}} \sim \|Tk\|_{L^p} + \|(Tk)\|_{L^p}
\]
\[
\lesssim \|Tk\|_{L^\infty} + \left\| \int_0^t f(v)k(v + \bullet) dv \right\|_{L^p}
\]
\[
\leq \|f\|_{L^p} \|k\|_{L^{p'}} + \left\| f(-\bullet)k(0) + \int_0^\bullet f(u-\bullet)k'(u) du \right\|_{L^p}
\]
\[
\leq \|f\|_{L^p} \|k\|_{L^{p'}} + \|f\|_{L^p} \|k\|_\infty + \|T(k')\|_{L^p}
\]
\[
\lesssim \|f\|_{L^p} \|k\|_{W^{1,p'}}.
\]
Due to \(\|Tk\|_\infty \leq \|f\|_{L^p} \|k\|_{L^{p'}} \) the real interpolation theory ((Triebel 1983, Thm. 3.3.6)) yields for all \(s \in (0,1) \)
\[
\|Tk\|_{s,p,\alpha} \lesssim \|f\|_{L^p} \|k\|_{s,p',\alpha'}.
\]
In a second step, we use an induction argument from $s$ to $s + 1$ for non-integer $s > 0$. Suppose $f \in B^s_{p,\alpha}$ and $k \in B^{s+1}_{p',\alpha'}$. The weak derivative of $L(f, k)$ is given by (see above)

$$L(f, k)'(t) = f(-t)k(0) + L(f, k')(t), \quad t \in [0, r],$$

which yields for $s \in (0, 1)$

$$\|L(f, k)\|_{s,p,\alpha} \leq \|f\|_{s,p,\alpha}\|k\|_{\infty} + \|T(k')\|_{s,p,\alpha} \lesssim \|f\|_{s,p,\alpha}\|k\|_{s+1,p',\alpha'}$$

and a fortiori for $s > 1$, $s \notin \mathbb{N}$ by induction

$$\|L(f, k)\|_{s,p,\alpha} \leq \|f\|_{s,p,\alpha}\|k\|_{\infty} + \|f\|_{s-1,p,\alpha}\|k'\|_{s,p',\alpha'} \lesssim \|f\|_{s,p,\alpha}\|k\|_{s+1,p',\alpha'}.$$  

Since the very first argument provided an estimate for $\|L(f, k)\|_{L^p}$ of the same type, the norm $\|L(f, k)\|_{s+1,p,\alpha}$ is bounded.

Finally, the same induction argument for $s \in \mathbb{N}$ requires an extra estimate for $\|T(k(s))\|_{0,p,\alpha}$. Since $f$ is in $B^s_{p,\alpha} \subset L^\infty$ and $k$ in $B^{s+1}_{p',\alpha'} \subset C^s$, we infer from

$$\left(\int_0^t f(u - t)k(s)(u)\, du\right)'(t) = f(0)k(s)(t) - \int_0^t f'(u - t)k(s)(u)\, du$$

and the convolution estimate (use (Triebel 1983, Thm. 2.11.2, Prop. 3.3.2))

$$\sup_{t \in \mathbb{R}} |(f'1_{\mathbb{R}^-} \ast (k(s)1_{\mathbb{R}^+}))(t)| \lesssim \|f'\|_{0,p,\alpha}\|k(s)\|_{0,p',\alpha'}$$

that $\|T(k(s))\|_{C^1} \lesssim \|f\|_{1,p,\alpha}\|k\|_{s+1,p',\alpha'}$ holds. □

**Lemma 8.4.** Suppose that $k$ is a function in $W^{2,\rho'}([0, r])$, $\rho' \in (1, \infty)$ and $\rho$ satisfies $\frac{1}{p} + \frac{1}{\rho'} = 1$. Then the integral operator

$$Kf(t) := \int_{-r}^0 k(|t - s|)f(s)\, ds, \quad t \in [-r, 0],$$

is continuous from $L^p([-r, 0])$ to $W^{2,\rho}([-r, 0])$ with $\|K\|_{L^p \to W^{2,\rho}} \lesssim \|k\|_{W^{2,\rho'}}$. 
Proof. First consider the following identities in an almost everywhere-sense for \( t \in [-r, 0] \) and \( f \in L^\rho([-r, 0]) \)

\[
(Kf)'(t) = \int_{-r}^0 k'(|t - s|) \text{sgn}(t - s) f(s) \, ds,
\]

\[
(Kf)''(t) = \left( \int_{-r}^0 k'(t - s) f(s) \, ds - \int_{-r}^0 k'(s - t) f(s) \, ds \right)'(t)
\]

\[= \int_{-r}^0 k''(|t - s|) f(s) \, ds + 2k'(0)f(t).
\]

By the Hölder inequality we obtain

\[
\|Kf\|_{L^\rho} \leq 2\|k\|_{L^{\rho'}} \|f\|_{L^\rho}, \quad \|(Kf)''\|_{L^\rho} \leq \rho\|k''\|_{L^{\rho'}} \|f\|_{L^\rho} + \rho\|k'\|_{\infty} \|f\|_{L^\rho}.
\]

The Sobolev embedding \( W^{2,\rho'} \subset C^1 \) proves \( \|Kf\|_{W^{2,\rho'}} \lesssim \|k\|_{W^{2,\rho'}} \|f\|_{L^\rho} \). □

8.2. Wavelets.

Definition 8.5. (We largely follow Cohen (2000).) For \( j, k \in \mathbb{Z} \) introduce the multi-index \( \lambda = (j, k) \) and put \( |\lambda| := |(j, k)| := j \). A wavelet basis \((\psi_\lambda)_\lambda\) is an orthonormal basis of functions in \( L^2(\mathbb{R}) \), derived from one function \( \psi \in L^2(\mathbb{R}) \) by translations and dilations

\[
\psi_\lambda(x) := \psi_{jk}(x) := 2^{j/2} \psi(2^j x - k).
\]

Furthermore set \( V_j \) as the closure of \( \text{span}(\psi_\lambda, |\lambda| \leq j) \). By \( P_j : L^2([-r, 0]) \to V_j \) we denote the orthogonal projection onto \( V_j \).

Cohen, Daubechies, and Vial (1993) constructed orthonormal wavelet bases on a bounded interval \( I \). The basis functions are obtained by restriction. Wavelet functions \( \psi_\lambda \) whose support crosses the boundary of \( I \) are suitably corrected in order to keep the orthogonality and approximation properties. These corrected functions are still denoted by \( \psi_\lambda \) even if they are not directly derived from \( \psi \). A consequence of this construction is that only multi-indices \( \lambda = (j, k) \) with \( |k| \lesssim 2^j \) are used and that the spaces \( V_j \) are finite-dimensional, whence we can start off with a space \( V_{-1} \) and an orthonormal basis \((\psi_{-1,k})_k\) of \( V_{-1} \). Then any function \( f \in L^2(I) \) has the wavelet decomposition

\[
f = \sum_\lambda \langle f, \psi_\lambda \rangle \psi_\lambda = \sum_{j \geq -1} \sum_k \langle f, \psi_{jk} \rangle \psi_{jk}.
\]
Note that summation over $|\lambda| \leq j_0$ will always mean summation over $(j,k)$ for all $j \leq j_0$ and all possible values of $k$.

**Definition 8.6.** A wavelet basis $(\psi_\lambda)$ will be called $s$-regular on the interval $I$ if the following two conditions are satisfied:

1. For all $w \in (0,s]$, $p, \alpha \in [1,\infty]$ the function $f$ is an element of $B^w_{p,\alpha}(I)$ if and only if
   \[ \|P^{-1}f\|_{L^p} + \left( \sum_{j=0}^\infty 2^{\alpha(j+w+\frac{1}{2}-\frac{1}{p})} \left( \sum_k |\langle f, \psi_{jk} \rangle|^p \right)^{\alpha/p} \right)^{1/\alpha} < \infty. \]
   The above expression constitutes a norm equivalent to $\| \cdot \|_{w,p,\alpha}$.

2. For all $k = 0, \ldots, \lfloor s \rfloor$ the vanishing moment property is fulfilled
   \[ \int_{-\infty}^{\infty} x^k \psi(x) \, dx = 0. \]

Sufficiently regular wavelets guarantee that for $m, s > 0$ and $\rho \in [1,\infty]$ the general Jackson inequality
\[ \|f - P_J f\|_{W^{m,\rho}} \lesssim 2^{-Js} \|f\|_{W^{m+s,\rho}} \quad \forall f \in W^{m+s,\rho} \] holds (Cohen 2000). From (Cohen, Daubechies, and Vial 1993) we immediately obtain (only mind the different notion of $s$-regularity there):

**Theorem 8.7.** $s$-regular wavelet bases exist for any $s > 0$. Moreover, they may be chosen to have compact support.

We shall often need estimates on $\langle Q_a \psi_\lambda, \psi_\lambda \rangle$ and similar expressions.

**Lemma 8.8.** For any weight measure $a$ with $\|a\|_{TV} \leq R < \infty$ and $\nu_0(a) \leq -\delta < 0$ and for any multi-index $\lambda$ we have $\langle Q_a \psi_\lambda, \psi_\lambda \rangle \sim 2^{-2|\lambda|}$ uniformly.

**Proof.** Using the formula for the spectral density (2.3), estimates as in the proof of Lemma 5.2 and the spectral characterisation of the space $W^{1,2}$ we obtain
\[ \langle Q_a \psi_\lambda, \psi_\lambda \rangle = \int_{-\infty}^{\infty} \left| \frac{\hat{\psi}_\lambda(\xi)}{\xi - f(u)} \right|^2 \, d\xi \sim \int_{-\infty}^{\infty} (1 + \xi^2)^{-1} |\hat{\psi}_\lambda(\xi)|^2 \, d\xi \]
\[ = \sup_{\|f\|_{L^2} = 1} \left( \int_{-\infty}^{\infty} (1 + \xi^2)^{-1/2} f(\xi) \hat{\psi}_\lambda(\xi) \, d\xi \right)^2 = \sup_{\|h\|_{W^{1,2}} = 1} \langle h, \psi_\lambda \rangle^2. \]

The last expression is clearly of order $2^{-2|\lambda|}$. $\square$
Lemma 8.9. Let $f \in C^{m,1}([-r, r])$, i.e. $f^{(m)}$ is Lipschitz continuous, be given with $m \in \mathbb{N}_0$ and suppose that $(\psi_{\lambda})$ is a compactly supported $(m+1)$-regular wavelet basis of $L^2([-r, 0])$. Then
\[
\left| \int_{-r}^{0} \int_{-r}^{0} f(x-y)\psi_{\lambda}(x)\psi_{\lambda}(y) \, dy \, dx \right| \lesssim \|f\|_{C^{m,1}}2^{-|\lambda|(m+2)}.
\]
holds with a constant independent of $f$ and of the multi-index $\lambda$.

Proof. Note that for $y \in [-r,0]$ the function $f(\bullet - y)|_{[-r,0]}$ lies in $C^{m,1}([-r,0])$. By the $(m+1)$-regularity of $(\psi_{\lambda})$ we find
\[
\left| \int_{-r}^{0} \int_{-r}^{0} f(x-y)\psi_{\lambda}(x)\psi_{\lambda}(y) \, dy \, dx \right|
\leq \sup_{y \in [-r,0]} \|f(\bullet - y), \psi_{\lambda}\|_{L^1} \|\psi_{\lambda}\|_{L^1}
\lesssim \sup_{y \in [-r,0]} \|f(\bullet - y)\|_{C^{m,1}}2^{-|\lambda|(m+1+\frac{1}{2})}2^{-|\lambda|/2}\|\psi\|_{L^1}.
\]
Note that we have used the embedding $C^{m,1} \subset B^{m+1}_{\infty,\infty}$.

References


