ADAPTIVE ESTIMATION FOR AFFINE STOCHASTIC DELAY DIFFERENTIAL EQUATIONS

MARKUS REISS

Abstract. We consider the stochastic delay differential equation

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

with $r, \sigma > 0$, $\gamma_0, \gamma_r \in \mathbb{R}$ and a weight function $g \in L^1([-r,0])$. For stationary solutions of this equation we consider the problem of nonparametric inference for the weight function g and for γ_0, γ_r from the continuous observation of one trajectory $(X(t),\, t \in [0,T])$ up to time $T > 0$.

For weight functions in the scale of Besov spaces $B^s_{p,1}$ and L^ρ-type loss functions convergence rates are established for long time asymptotics. The estimation problem is transformed into an ill-posed inverse problem with error in the data and the operator. The degree one of ill-posedness explains the rate $(T/\log T)^{-\frac{s}{2s+3}}$ obtained under certain restrictions on p and ρ. This rate is shown to be optimal in a minimax sense for the estimation problem. Our adaptive estimator is based on a suitable wavelet thresholding algorithm for the ill-posed problem involved.

1. Introduction

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) or in laser technology (Garcia-Ojalvo and Roy 1996), 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, e.g. Jeantheau (2001), Reiß (2001). 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

1991 *Mathematics Subject Classification.* Primary 62M09, 62G20; Secondary 47A52, 60G15.

Key words and phrases. wavelet thresholding, ill-posed inverse problem, Besov space, spatial adaptivity, minimax rates, nonparametric estimation.

This work was partially supported by the Sonderforschungsbereich 373 of the Deutsche Forschungsgemeinschaft.

affine SDDE is of the form

$$dX(t) = \left(\int_{-r}^{0} X(t+u)\,da(u)\right) dt + \sigma\,dW(t), \quad t \geq 0, \tag{1.1}$$

$$X(u) = F(u), \quad u \in [-r, 0]. \tag{1.2}$$

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 σ 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 $\alpha\delta_0$ with $\alpha \in \mathbb{R}$.

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). For this purpose, we assume that the weight measure has a Lebesgue density. As it turns out, it is rather natural from a mathematical point of view that we also allow the measure to have additional point masses in the interval endpoints $-r$ and 0.

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). By solving a related ill-posed inverse problem by the Galerkin projection method 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^ρ-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 ρ 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 which have been proposed for solving ill-posed problems (Donoho 1995, Abramovich and Silverman 1998). In fact, the latter paper presents the main idea: we first threshold the wavelet coefficients and then invert the equation. 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 (2002) for details and mathematical results.

Mathematically speaking, we denoise the data in a certain Sobolev space and follow the lines of the abstract results obtained by Kerkyacharian and Picard (2000) for heteroscedastic noise. Then in a second step, an integral equation with the empirical covariance function as kernel and with the data derived from the wavelet estimator will be solved. For the right choice of the thresholding level and for theoretical purposes the regularity of the covariance function and the mapping properties of the covariance operator have to be thoroughly investigated. A numerical simulation study is beyond the scope of this paper and we refer to Cohen, Hoffmann, and Reiß (2002) where an algorithm is presented that can be adapted to determine our estimator in practice.

Except for the afore-mentioned paper (Reiß 2002), statistical inference for the weight measure in affine SDDEs has so far only be considered for parametric models, e.g. by Kutoyants, Mourid, and Bosq (1992), Gushchin and Küchler (1999) and Gushchin and Küchler (2001), 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 deterioration is due to an ill-posed inverse problem involving the covariance operator of the solution process of the SDDE. Ill-posed problems with stochastic error in the data have attracted increasing attention recently and exact minimax rates have been obtained for idealized settings, see Donoho (1995), Nussbaum and Pereverzev (1999), Cavalier and Tsybakov (2002) or Kalifa and Mallat (2003) and the references therein. These results provide a good guideline for the estimation technique applied in this paper, but their assumptions like known operator, exact mapping between Sobolev scales or Gaussian noise are violated. As far as we know, inverse problems with approximately known operator kernel have only been considered in an abstract deterministic setting by Hämarik (1983) and coauthors.

In Section 2 we introduce the theory of affine SDDEs and their stationarity behaviour and present results on the regularity of their covariance function and mapping properties of their covariance operator. Section 3 is devoted to the construction of the estimator and the statement of the main theorems. In Section 4 we assess the optimality of our estimator by the minimax approach. The proofs of the statements are delayed 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. $\mathbb{P}_a$ and $\mathbb{E}_a$ denote the probability measure and the expectation operator depending on the parameter a. 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 $M(I)$ and equipped with the total variation norm $\|\bullet\|_{TV}$. δ_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 $\mu \in M(I)$ 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 monographs by Hale and Verduyn Lunel (1993) and Diekmann, van Gils, Verduyn Lunel, and Walther (1995), whereas fundamental results on stochastic delay equations can be found in the monographs by Mohammed (1984) and Mao (1997). If we put $\sigma = 0$ in (1.1), we obtain the deterministic linear delay equation

$$\dot{x}(t) = \int_{-r}^{0} x(t+u)\, da(u), \quad t \geq 0. \tag{2.1}$$

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 defined 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

$$v_0(a) := \sup\{\operatorname{Re}(\lambda) \mid \chi_a(\lambda) = 0\}.$$

Without loss of generality we shall henceforth assume $\sigma = 1$; otherwise we rescale X and consider $\tilde{X}(t) = \sigma^{-1} X(t)$ instead. Küchler and Mensch (1992) 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*

$$q_a'(t) = \int_{-r}^{0} q_a(t+u)\, da(u) \qquad \textit{for all } t \geq 0. \tag{2.2}$$

Its spectral density is given by

$$\hat{q}_a(\xi) := \int_{-\infty}^{\infty} q_a(t) e^{i\xi t}\, dt = \frac{1}{|\chi_a(i\xi)|^2}, \quad \xi \in \mathbb{R}. \tag{2.3}$$

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 μ_X of the solution process X on the interval $[0, T]$ is mutually absolutely continuous to the law μ_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*

$$\begin{aligned}
b_T(u) &:= \int_0^T X(t+u)\,dX(t) && u \in [-r, 0],\\
q_T(u,v) &:= \int_0^T X(t+u)X(t+v)\,dt && u, v \in [-r, 0],\\
Q_T\mu(u) &:= \int_{-r}^0 q_T(u,v)\,d\mu(v) && u \in [-r, 0],\ \mu \in M([-r, 0]),\\
Q_a\mu(u) &:= \int_{-r}^0 q_a(u-v)\,d\mu(v) && u \in [-r, 0],\ \mu \in M([-r, 0]).
\end{aligned}$$

The operator Q_a is the covariance operator of the stationary process X on $[-r, 0]$, regarded as element of $C([-r, 0])$, which maps the dual $M([-r, 0])$ to $C([-r, 0])$. The operator $\frac{1}{T}Q_T$ is referred to as empirical covariance operator, since $\frac{1}{T}q_T$ is the empirical covariance function.

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:

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\int_0^T X(t+u_1)\,dW(t) - \int_0^T X(t+u_2)\,dW(t)\Big)^4\Big]\\
&\lesssim \mathbb{E}\Big[\Big(\int_0^T (X(t+u_1) - X(t+u_2))^2\,dt\Big)^2\Big] \qquad (2.4)\\
&\lesssim T^2(u_1 - u_2)^2,
\end{aligned}$$

which follows from the Burkholder-Davis-Gundy inequality and the uniform Lipschitz continuity of the covariance function q_a, see Proposition 2.8.

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

$$\begin{aligned}
\Lambda_T(X, X(0)+W) &:= \frac{d\mu_X}{d\mu_W}\\
&= \exp\Big(\int_0^T \int_{-r}^0 X(t+u)\,da(u)\,dX(t) - \frac{1}{2}\int_0^T \Big(\int_{-r}^0 X(t+u)\,da(u)\Big)^2 dt\Big)\\
&= \exp\Big(\langle b_T, a\rangle - \tfrac{1}{2}\langle Q_T a, a\rangle\Big).
\end{aligned}$$

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 $\frac{1}{T}Q_T \to Q_a$ and $\frac{1}{T}b_T \to Q_a a$ for $T \to \infty$, a consequence of Theorem 3.2 and Corollary 3.3, 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 $B^s_{p,\alpha}$ of functions with L^p-regularity s and fine-tuning parameter α we refer to the appendix. Just recall the identity $B^s_{2,2} = W^{s,2}$ 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 $B^s_{p,\alpha}$ are mapped to $B^{s+2}_{p,\alpha}$, cf. Theorem 2.9. Moreover, the point measures δ_{-r} and δ_0 at the boundary of $[-r, 0]$ are even mapped to $B^{s+3}_{p,\alpha}$, which is not true for point measures in $(-r, 0)$ that "see" the irregularity of q_a at zero. In anticipation of these 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*

$$\mathcal{B}^s_{p,\alpha} := \Big\{\gamma_0\delta_0 + \gamma_r\delta_{-r} + g\circ\lambda \,|\, \gamma_0,\gamma_r \in \mathbb{R},\, g \in B^s_{p,\alpha}([-r,0])\Big\},$$

$$\mathcal{B}^s_{p,\alpha}(v) := \Big\{a \in \mathcal{B}^s_{p,\alpha} \,|\, v_0(a) \le -v\Big\}.$$

On $\mathcal{B}^s_{p,\alpha}$ we introduce the norm

$$\|\gamma_0\delta_0 + \gamma_r\delta_{-r} + g\circ\lambda\|_{s,p,\alpha} := |\gamma_0| + |\gamma_r| + \|g\|_{B^s_{p,\alpha}([-r,0])}.$$

- *(L^p scale) For $p \in (1,\infty)$, $v < 0$ set*

$$\mathcal{L}^p := \Big\{\gamma_0\delta_0 + \gamma_r\delta_{-r} + g\circ\lambda \,|\, \gamma_0,\gamma_r \in \mathbb{R},\, g \in L^p([-r,0])\Big\},$$

$$\mathcal{L}^p(v) := \Big\{a \in \mathcal{L}^p \,|\, v_0(a) \le -v\Big\}.$$

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

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

The set $\mathcal{B}^s_{p,\alpha}(v)$ is a closed subset of $\mathcal{B}^s_{p,\alpha}$, 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 $\mathcal{L}^p$. 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 $\mathcal{L}^p$ 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*

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

The choice $\alpha = 1$ in the definition will be discussed in Section 3.3. We shall need a very precise regularity and tail behaviour result for the covariance function q_a depending on properties of a. Roughly speaking, the covariance function is three times more regular than the weight itself and decreases exponentially fast.

Proposition 2.8. *Let E_δ be the multiplication operator with the exponential $E_\delta(f)(t) := f(t)e^{\delta t}$. Then for $a \in \mathcal{B}^s_{p,\alpha}(v)$ with $v < 0$, $s > 0$ and $p \in (1,\infty)$, $\alpha \in [1,\infty]$ the covariance function q_a has for any $\delta \in (0,|v|)$ the property $E_\delta q_a \in B^{s+3}_{p,\alpha}([0,\infty))$.*

In Reiß (2001) and Gushchin and Küchler (2001) 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 $W^{2,2}([-r,0])$. Here, we need the mapping properties along the scale of Besov spaces. Using Proposition 2.8 we show that the covariance operator is for a certain range of Besov spaces also smoothing of order two. The inclusion of point measures at the interval boundary makes this mapping isomorphic:

Theorem 2.9. *For weight measures a in $\mathcal{B}^s_{p,\alpha}(v)$ and the parameters as before the covariance operator is a bijective bounded linear operator on the appropriate spaces:*

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

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

Theorem 2.10. *If (a_n) is a sequence of $\mathcal{B}^s_{p,\alpha}(v)$-weights that converges in $\mathcal{B}^\sigma_{p,\alpha}$-norm to the $\mathcal{B}^s_{p,\alpha}$-weight a for some $\sigma > s - 2 + (1 \vee \frac{2}{p})$ and s, p, α, v as before, then the covariance operators converge in operator norm:*

$$\lim_{n\to\infty} \|Q_{a_n} - Q_a\|_{\mathcal{B}^s_{p,\alpha} \to B^{s+2}_{p,\alpha}} = 0.$$

Consequently, for $s, S, \delta > 0$, $p \in (1,\infty)$ the operator norms are uniformly bounded:

$$\sup_{a\in M(s,p,S,\delta)} \|Q_a\|_{\mathcal{B}^s_{p,\alpha}\to B^{s+2}_{p,\alpha}} < \infty, \qquad \sup_{a\in M(s,p,S,\delta)} \|Q_a^{-1}\|_{B^{s+2}_{p,\alpha}\to \mathcal{B}^s_{p,\alpha}} < \infty.$$

The statements remain true if $\mathcal{B}^s_{p,\alpha}$ is everywhere replaced by $\mathcal{L}^p$ and $B^{s+2}_{p,\alpha}$ by $W^{2,p}$ and under the condition that (a_n) converges in $\mathcal{L}^p$-norm to a.

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 (ψ_λ) of an s-regular wavelet basis on $[-r,0]$ we refer to the appendix. The thresholding techniques used are similar in spirit to the methods developed by Donoho (1995), Abramovich and Silverman (1998) and Kerkyacharian and Picard (2000).

First, we need to clarify the functional nature of the noise in the estimate $\frac{1}{T}b_T$ of $Q_a a$. A good intuition gives the decomposition (recall $\sigma = 1$)

$$\begin{aligned} b_T(u) &= \int_0^T \Big(X(t+u)\int_{-r}^0 X(t+v)\,da(v)\Big)\,dt + \int_0^T X(t+u)\,dW(t) \\ &= (Q_T a)(u) + \int_0^T X(t+u)\,dW(t). \qquad (3.1) \end{aligned}$$

Suppose for a moment that $\frac{1}{T}Q_T$ equals Q_a exactly (it is in fact the less important estimation error). 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

$$Y(t) = f(t) + \varepsilon W(t), \quad t \in [0,1].$$

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 $Q_a^{-1} f = a$. Taking account of Theorem 2.9 we are going to minimise the expected L^ρ-loss in estimating the signal $Q_a a$ with minimal $W^{2,\rho}$-loss. Although our noise process is more regular than white noise, we encounter an ill-posed problem because the noise is far less regular than the norm we want the signal to be estimated in; the regularity differs roughly by $\frac{3}{2}$, since the $W^{2,\rho}$-norm postulates regularity 2 while Brownian motion is of regularity $\frac{1}{2}$ only. A general result in an idealised Gaussian and Hilbert scale setting for this kind of ill-posed problems has been developed by Nussbaum and Pereverzev (1999). They obtain our minimax rate $T^{-s/(2(s+3/2))}$ in our case of (known) operators of smoothing order 2 and noise regularity $\frac{1}{2}$.

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 $s_{max} > 2$ be fixed. With b_T from Definition 2.4 introduce for any multi-index λ the wavelet coefficient*

$$\beta_{\lambda,T} := \langle \tfrac{1}{T} b_T, \psi_\lambda \rangle,$$

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

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

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 $T^{-1}\langle Q_a\psi_\lambda, \psi_\lambda\rangle \sim T^{-1}2^{-2|\lambda|}$ by Lemma 8.8. In Section 3.3 we comment on the choice for a specific sample, here however, we only strive for asymptotically optimal threshold values $\kappa_\lambda(T)$, $T\to\infty$, that obey uniform exponential tail estimates. To this end we study the convergence of exponential-type moments of $\frac{1}{T}Q_T$.

Theorem 3.2. *Let $(\psi_\lambda)_\lambda$ be a compactly supported 2-regular wavelet basis of $L^2([-r,0])$ and let $\delta, R > 0$ be given. Then there are constants $K, T_0 > 0$ such that for all weight measures a with $v_0(a) \le -\delta$, $\|a\|_{TV} \le R$, all multi-indices λ, all measures $\mu \in M([-r,0])$, all $T \ge T_0$ and all $\alpha \in [0, T^{1/2}(K\|\mu\|_{TV})^{-1})$ the following moment bound holds true:*

$$\mathbb{E}_a\big[\cosh\big(\alpha T^{1/2} 2^{3|\lambda|/2}\langle(\tfrac{1}{T}Q_T - Q_a)\mu, \psi_\lambda\rangle\big)\big] \le \exp\big(K\|\mu\|_{TV}^2\alpha^2\big).$$

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

$$\mathbb{E}_a[\langle(\tfrac{1}{T}Q_T - Q_a)\mu, \psi_\lambda\rangle^{2m}]^{1/2m} \lesssim T^{-1/2}2^{-3|\lambda|/2}\|\mu\|_{TV}. \tag{3.2}$$

We note that the overall noise level is $T^{-1/2}$ as expected and that the noise in $\frac{1}{T}q_T(x,\bullet) - q_a(x-\bullet)$ is α-Hölder continuous for all $\alpha < 1$ (put $\mu = \delta_x$). This is intuitively clear by the definition of q_T as a convolution-type integral. For b_T however, the noise regularity will only be like Brownian motion of order $\frac{1}{2}$, which can be established by using the decomposition 3.1 and martingale inequalities in combination with the result for q_T. Even more, we obtain Gaussian tail estimates.

Corollary 3.3. *Let R, δ, $\rho > 0$ and fix a wavelet basis (ψ_λ) as before. Then there is a universal bound $\kappa^* > 0$ such that uniformly for all weight measures a with $v_0(a) < -\delta$, $\|a\|_{TV} \le R$, all multi-indices λ and all T sufficiently large the following large deviation bound holds:*

$$\mathbb{P}_a\Big(2^{|\lambda|}T^{1/2}|\langle \tfrac{1}{T}b_T - Q_a a, \psi_\lambda\rangle| \ge \tfrac{\kappa}{2}\sqrt{\log T}\Big) \lesssim T^{-3\rho} \quad \forall\, \kappa \ge \kappa^*. \tag{3.3}$$

We shall therefore set $\kappa_\lambda(T) = 2^{-|\lambda|}T^{-1/2}\sqrt{\log T}\kappa$. 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 3.4. *Let $s \in (0, s_{max} - 2]$, $S > 0$, $\rho, p \in (1, \infty)$ and $\delta > 0$ be given satisfying*

$$\frac{1}{p} - \frac{1}{\rho} \leq \frac{2}{\rho}\frac{s}{3} \tag{3.4}$$

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

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

In the next step we construct an operator $\widehat{Q}_T$ 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 3.2, but it is even simpler to use the relationship $q_a'(t) = Q_a a(-t)$ for $t \in (0, r]$ deduced from (2.2):

$$q_a'(t) = \int_{-r}^0 q_a(t+s)\,da(s) = \int_{-r}^0 q_a(-t-s)\,da(s) = Q_a a(-t).$$

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. This is exactly the construction method of $\widehat{Q}_T$ we shall adopt. We thus avoid further time consuming calculations.

Theorem 3.5. *Let the parameters s, S, p and δ be as before. Introduce the integral operator $\widehat{Q}_T$ 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. $\widehat{Q}_T \mu(t) := \int_{-r}^0 \hat{q}_T(t-u)\,d\mu(u)$ for $t \in [-r, 0]$, $\mu \in M([-r, 0])$.

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

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

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

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

3.3. **Discussion.** Our method differs from the classical wavelet thresholding algorithm for density estimation or regression due to the ill-posedness involved. Our threshold κ_λ 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 that the weight lies in $W^{\frac{s}{2s+3},\rho}$ because the restriction (3.4) is much stronger than in the classical setting. 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 (6.2). It is not known whether this value is the maximal possible.

Which rate of convergence do we obtain for the L^ρ-risk of the weight function $g = -\mathbf{1}_{[-\frac{1}{2},0]}$ with delay $r = 1$? This might be seen as a toy example

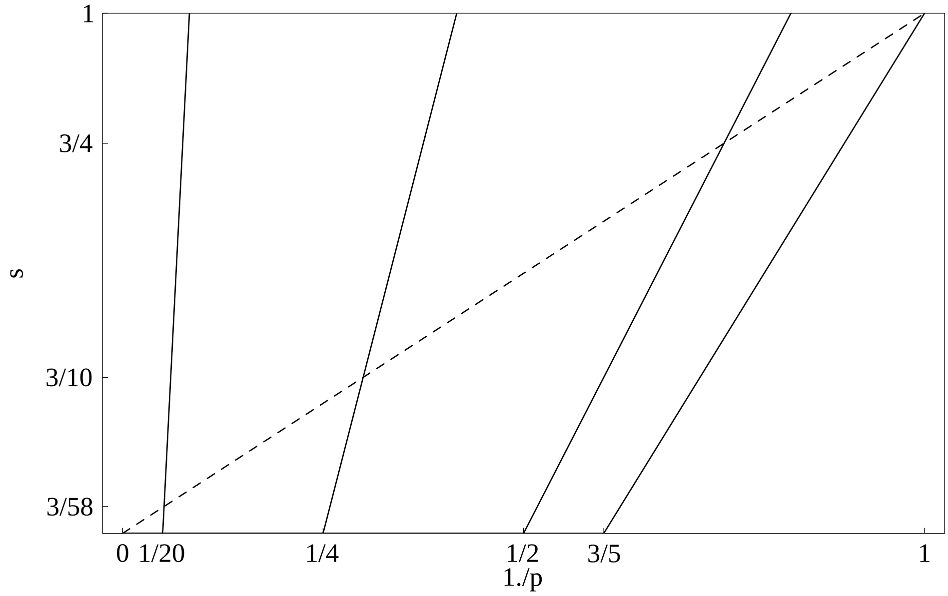

FIGURE 1. Restriction (3.4) for adaptive estimation of $\mathcal{B}^s_{p,1}$-weights in L^ρ-loss, $\rho = 5/3;\ 2;\ 4;\ 20$. For any space above the lines the rate is $(T/\log T)^{-s/(2s+3)}$, for those below we have to use embeddings. The dashed line shows the embedding of a piecewise constant weight function.

for estimating a change point or the maximal delay time. The function g lies in $B^1_{1,\infty}([-1,0])$ and thus by embedding in $W^{\sigma,\rho}$ and $B^\sigma_{p,1}$ for $\sigma < \frac{1}{\rho}$. This shows that linear methods as in (Reiß 2002) cannot converge faster than with rate $T^{-1/(2+3\rho)}$ whereas 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, see also Figure 1. 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 $g \in B^{1/p}_{p,\infty}$ for $p < 1$ would yield (almost) the L^1-rate $T^{-1/3}$.

In the mathematical results we have focussed 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_{max}), some bound on the size (S) and some uniform mixing behaviour (δ). The resulting minimal asymptotically optimal threshold κ^* depends in a complicated way on these quantities. So clearly the question is how to choose κ_λ for a specific observation up to time T.

First of all note that $T\beta_{\lambda,T} - \langle Q_T a, \psi_\lambda\rangle$ 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

$$\eta_{\lambda,T} := T^{1/2}\frac{\beta_{\lambda,T} - \langle \frac{1}{T}Q_T a, \psi_\lambda\rangle}{\sigma_{\lambda,T}} \quad (\text{with } \sigma^2_{\lambda,T} := \langle \tfrac{1}{T}Q_T\psi_\lambda, \psi_\lambda\rangle \sim 2^{-2|\lambda|})$$

is therefore $N(0,1)$-distributed by the martingale central limit theorem. In other words, we observe the coefficient $\langle \frac{1}{T}Q_T a, \psi_\lambda\rangle$ under the noise

$T^{-1/2}\sigma_{\lambda,T}\eta_{\lambda,T}$. Because $\sigma^2_{\lambda,T}$ converges stochastically, we conclude that the noise is approximately normal distributed with variance $T^{-1}\sigma^2_{\lambda,T}$, 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 $W^{2,2}$-norm used so that we have to be a little bit more conservative and should choose in the Hilbertian case $\rho = 2$

$$\kappa_j = T^{-1/2}\sigma_{\lambda,T}\sqrt{6\log(T\sigma^{-2}_{\lambda,T})},$$

provided the isomorphy constants in Theorem 2.9 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 $J^{-1}2^{-J} \approx T^{-1}\max_\lambda \sigma^2_{\lambda,T}$, which is an estimate of the squared noise level.

If only discrete observations (X_{t_i}) are available with $0 = t_0 \le t_1 \le \cdots \le t_N = T$, 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 measures, i.e. the weight space $B^s_{p,1}$ instead of $\mathcal{B}^s_{p,1}$. For this one can project the estimator $\hat{a}_T$ onto $L^\rho([-r,0])$ by neglecting the point measure part. The asymptotic risk bound remains the same.

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 $Q_T\hat{A} \approx b_T$ to determine an estimator $\hat{A}$. A mathematical analysis of an adaptive version of $\hat{A}$ seems feasible 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^ρ-risk functions, in the sense that one cannot improve on the restriction (3.4) in order to obtain the speed of convergence $(T/\log T)^{-\frac{s}{2s+3}}$ for weights in $\mathcal{B}^s_{p,1}$. For smaller values of p the rate of convergence is indeed worse and is obtained by embedding $\mathcal{B}^s_{p,1}$ to $\mathcal{B}^\sigma_{\pi,1}$ with some properly chosen $\sigma < s$ and $\pi > p$, see Figure 1 for an illustration. In the sequel, we merely assume $s + \frac{1}{\rho} - \frac{1}{p} \ge 0$ in order to have the embedding $\mathcal{B}^s_{p,1} \subset \mathcal{L}^\rho$ 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 $s+\frac{1}{\rho}-\frac{1}{p}\geq 0$ and such that $M(s,p,S,\delta)$ has nonempty interior in $\mathcal{B}^s_{p,1}$. Then the following asymptotic minimax lower bound holds for $T\to\infty$:*

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

where the infimum is taken over all $\sigma(X(t),\,-r\leq t\leq T)$-measurable estimators $\hat{a}_T$.

We obtain a fairly complete picture of the minimax rates for the L^2-risk of certain Besov regularity classes $M(s,p,S,\delta)$.

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 $\mathcal{B}^s_{p,1}$. In what follows the infima are taken over all $\sigma(X(t),\,-r\leq t\leq T)$-measurable estimators $\hat{a}_T$.*

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

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

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

(2) *(regular case) For $\frac{1}{p}-\frac{1}{\rho}<\frac{2}{\rho}\frac{s}{3}$ we have*

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

It should be noted that the actual minimax rates in the regular case are of course $T^{-s/(2s+3)}$, which can be attained by estimators taking the regularity s of the unknown parameter for granted. A Lepski-type adaptive estimation rule could be applied, too. The transfer of the estimation problem to an ill-posed inverse problem and the mathematical tools developed allow to apply further nonparametric inference techniques, e.g. change point analysis to detect the maximal delay time or hypothesis testing with nonparametric alternatives.

5. Proofs for the covariance function and operator

5.1. **Proof of Proposition 2.8.**

First we establish the exponential decay property and the regularity result separately.

Lemma 5.1. *The covariance functions decrease uniformly, in the sense that for all $S>0$ and $v>\delta>0$*

$$\sup_{\|a\|_{TV}\leq S,\,v_0(a)\leq -v}\|E_\delta q_a\|_\infty<\infty.$$

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 $\{z\in\mathbb{C}\mid|\text{Im}(z)|<v_0(a)\}$ (Katznelson 1976, Section VI.7.1) and satisfies

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

The assumptions guarantee $|\chi_a(\pm i\xi \mp \delta)| \geq |i\xi+\delta| - Se^{\delta r}$. 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 $\{\chi_a \,|\, a \in U\}$ 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

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

where the supremum is taken over all measures a as in the statement of the lemma. We conclude

$$\begin{aligned}\sup_a \|E_\delta q_a\|_\infty &\leq \sup_a \int_{-\infty}^{\infty} \Big|\chi_a(i\xi-\delta)^{-1}\chi_a(-i\xi+\delta)^{-1}\Big|\,d\xi \\ &\leq S(K_+ + K_-) + \int_{|\xi|>2S} (|i\xi+\delta| - Se^{\delta r})^{-2}\,d\xi < \infty.\end{aligned}$$

□

Lemma 5.2. *For $s \geq 0$, $p \in (1,\infty)$, $\alpha \in [1,\infty]$ and $v < 0$ we have*

$$a \in \mathcal{B}^s_{p,\alpha}(v) \Rightarrow q_a \in B^{s+3}_{p,\alpha}([0,r]),$$
$$a \in \mathcal{L}^p(v) \Rightarrow q_a \in W^{3,p}([0,r]).$$

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 $q_a'(0+) = -\frac{1}{2}$ and $q_a \in C^2([0,\infty))$ from Küchler and Mensch (1992) imply for $t \in (0,r)$

$$\begin{aligned} q_a'''(t) &= \frac{d}{dt}\Big(\int_{-r}^0 q_a'(|t+u|)\operatorname{sgn}(t+u)g(u)\,du + \alpha q_a'(t) - \beta q_a'(r-t)\Big) \\ &= \int_{-r}^0 q_a''(|t+u|)g(u)\,du + 2q_a'(0+)g(-t) + \alpha q_a''(t) + \beta q_a''(r-t).\end{aligned}$$

This shows that the third derivative q_a''' is as regular in a Besov space sense as g and q_a''. For the first term note that the integral can be split up according to $\int_{-r}^0 = \int_{-r}^{-t} + \int_{-t}^0$ such that the regularity result from Lemma 8.3 with $k = g$, $f = q_a''$ and with obvious modifications of the interval boundaries applies. Since q_a'' is always more regular than q_a''' the result follows. Formally, one proceeds by putting $\sigma := \sup\{s \geq 2 \,|\, q_a \in B^s_{p,\alpha}\}$ and noting that the right hand side is an element of $B^{(\sigma-2-\varepsilon)\wedge s}_{p,\alpha}$ for any $\varepsilon > 0$, hence q_a is in $B^{(\sigma+1-\varepsilon)\wedge(s+3)}_{p,\alpha}$ and $\sigma = s+3$ follows. For the $\mathcal{L}^p$-scale the proof is the same except that Lemma 8.4 can be immediately applied to the expression for q_a'. □

Proof. (of Proposition 2.8) Again, we make use of identity (2.2). We infer that q_a becomes increasingly regular as the time evolves so that q_a is in

$B^{s+3}_{p,\alpha}([0,t])$ for all $t \geq 0$ by Lemma 5.2. Moreover, applying (2.2) repeatedly, we conclude

$$|q_a^{(n)}(t)| \leq \|a\|_{TV}^n \|q_a\|_{C([t-nr,t])} \text{ for } t \geq nr.$$

By Lemma 5.1 the choice $n = [s+3]+1$ then implies that $q_a^{(n)}$ decays with exponential order δ. Expanding the n-fold derivative of $e^{\delta t} q_a(t)$ then proves the assertion $\|E_\delta q_a\|_{B^{s+3}_{p,\alpha}([0,\infty))} < \infty$. □

5.2. **Proof of Theorem 2.9.**

Proof. First, let us see how Q_a acts on functions $f \in B^s_{p,\alpha}$. 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 $q_a \in B^{s+3}_{p,\alpha}([0,r])$ (Proposition 2.8) we obtain for $t \in [-r,0]$ like in the proof of Lemma 5.2

$$(Q_a f)''(t) = \int_{-r}^0 q_a''(t-s) f(s)\,ds - f(t).$$

From Lemma 8.3 we infer as before the estimate

$$\|(Q_a f)''\|_{s,p,\alpha} \lesssim \|f\|_{s,p,\alpha} + \|f\|_{s-1,p,\alpha} \|q_a''\|_{s,p',\alpha'} \lesssim (1 + \|q_a\|_{s+3,p,\alpha}) \|f\|_{s,p,\alpha},$$

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

$$\|(D^2 Q_a + \mathrm{Id}) f\|_{s+\varepsilon} \lesssim \|f\|_{s+\varepsilon-1,p,\alpha} \|q_a''\|_{s+\varepsilon,p',\alpha'} \lesssim \|q_a\|_{s+3,p,\alpha} \|f\|_{s,p,\alpha}.$$

Hence, $D^2 Q_a + \mathrm{Id}$ is a compact operator on $B^s_{p,\alpha}$ ($B^{s+\varepsilon}_{p,\alpha}([-r,0]) \subset B^s_{p,\alpha}([-r,0])$ compactly).

Let $V \subset B^s_{p,\alpha}$ 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 $D^2 Q_a(B^s_{p,\alpha})$ with $\dim U = \dim V$. The situation is illustrated by the following diagram:

$$\begin{array}{ccccc}
B^s_{p,\alpha} & = & V^c & \oplus & V \\
 & & & \big\downarrow{\scriptstyle Q_a} & \\
B^{s+2}_{p,\alpha} & = & Q_a(B^s_{p,\alpha}) & + & (D^2)^{-1}(U) \\
 & & & \big\downarrow{\scriptstyle D^2} & \\
B^s_{p,\alpha} & = & D^2 Q_a(B^s_{p,\alpha}) & \oplus & U
\end{array}$$

While the decomposition in the first and in the third line hold by definition, the representation of $B^{s+2}_{p,\alpha}$ in the second line follows from the third line due to $(D^2)^{-1}(B^s_{p,\alpha}) = (D^2)^{-1}(D^2 Q_a(B^s_{p,\alpha}) \oplus U) \subset Q_a(B^s_{p,\alpha}) + (D^2)^{-1}(U)$. 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 $Q_a(B^s_{p,\alpha})$ is a closed subspace of $B^{s+2}_{p,\alpha}$ 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 $Q_a(\operatorname{span}(\delta_{-r},\delta_0)) \subset B^{s+3}_{p,\alpha} \subset B^{s+2}_{p,\alpha}$ by Proposition 2.8. The injectivity of Q_a on $M([-r,0])$ implies that $Q_a(\operatorname{span}(\delta_{-r},\delta_0))$ is a two-dimensional subspace of $B^{s+2}_{p,\alpha}$ in the complement of $Q_a(B^s_{p,\alpha})$. Owing to $\operatorname{codim} Q_a(B^s_{p,\alpha}) \le 2$ this codimension must equal two and $Q_a : \mathcal{B}^s_{p,\alpha} \to B^{s+2}_{p,\alpha}$ is onto, hence bijective. Because Q_a is separately continuous on these two subspaces, it is continuous on its span $\mathcal{B}^s_{p,\alpha}$ and by the open mapping theorem it is an isomorphism.

Exactly the same reasoning applies for L^p and $\mathcal{L}^p$ instead of $B^s_{p,\alpha}$ and $\mathcal{B}^s_{p,\alpha}$, just apply Lemma 8.4 to $(Q_af)'$. □

5.3. **Proof of Theorem 2.10.** This will be a consequence of the following continuity property.

Proposition 5.3. *Suppose $s>0$, $1<p<\infty$, $\alpha\in[1,\infty]$ and $v<0$ are given. If (a_n) is a sequence in $\mathcal{B}^s_{p,\alpha}(v)$ that converges in $\mathcal{B}^s_{p,\alpha}$-norm to the $\mathcal{B}^s_{p,\alpha}(v)$-weight a, then $\|E_\delta(q_{a_n}-q_a)\|_{B^{s+3}_{p,\alpha}} \to 0$ follows for all $\delta<|v|$.*

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 $t\in(0,r)$:

$$\begin{aligned}
f_n''(t) &= \Big(\int_{-r}^0 q_{a_n}(\bullet+u)\,da_n(u) - \int_{-r}^0 q_a(\bullet+u)\,da(u)\Big)'(t)\\
&= \Big(\int_{-r}^0 f_n(\bullet+u)\,da_n(u)\Big)'(t) + (Q_a(a_n-a)(-\bullet))'(t)\\
&= -\int_{-r}^{-t} f_n'(-t-u)g_n(u)\,du + \int_{-t}^0 f_n'(t+u)g_n(u)\,du\\
&\quad -\gamma_{r,n}f_n'(r-t)+\gamma_{0,n}f_n'(t) - (Q_a(a_n-a))'(-t)\\
&= -\int_0^{r-t} f_n'(u)g_n(-u-t)\,du + \int_{-t}^0 f_n'(u)g_n(u-t)\,du\\
&\quad -\gamma_{r,n}f_n'(r-t)+\gamma_{0,n}f_n'(t) - (Q_a(a_n-a))'(-t).
\end{aligned}$$

By Lemma 8.3 we obtain for all $\sigma>0$ (allowing the value ∞) the estimate

$$\begin{aligned}
\|f_n''\|_{\sigma,p,\alpha} &\lesssim \|f_n'\|_{\sigma,p',\alpha'}\|g_n\|_{\sigma-1,p,\alpha} + (|\gamma_{1,n}|+|\gamma_{2,n}|)\|f_n'\|_{\sigma,p,\alpha}\\
&\quad + \|Q_a\|_{B^{(\sigma-1)\vee 0}_{p,\alpha}\to B^{\sigma+1}_{p,\alpha}}\|a_n-a\|_{(\sigma-1)\vee 0,p,\alpha}.
\end{aligned} \tag{5.1}$$

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

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

Starting with $\sigma_0=\varepsilon$, we can iterate this estimate ($\sigma_{n+1}:=\min(\sigma_n+\varepsilon,s+1)$). 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 $\|f_n\|_{s+3,p,\alpha}\to 0$. □

Proof. (of Theorem 2.10) By linearity we have for $f \in B^s_{p,\alpha}$ 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 $\|\bullet\|_{s+2,p',\alpha'} \lesssim \|\bullet\|_{\sigma+3,p,\alpha}$ and $\|\bullet\|_{L^{p'}} \lesssim \|\bullet\|_{\sigma+3,p,\alpha}$ we infer the bound

$$\begin{aligned} \|(Q_{a_n} - Q_a)f\|_{s+2,p,\alpha} &\lesssim \|(Q_{a_n} - Q_a)f\|_{L^\infty} + \|f\|_{s-1,p,\alpha}\|(q_{a_n} - q_a)''\|_{s,p',\alpha'} \\ &\lesssim \big(\|q_{a_n} - q_a\|_{L^{p'}} + \|q_{a_n} - q_a\|_{\sigma+3,p,\alpha}\big)\|f\|_{s,p,\alpha} \\ &\lesssim \|q_{a_n} - q_a\|_{\sigma+3,p,\alpha}\|f\|_{s,p,\alpha}. \end{aligned}$$

Since $M(s,p,S,\delta)$ is bounded in $\mathcal{B}^s_{p,\alpha}$, it is relatively compact in any $\mathcal{B}^\sigma_{p,\alpha}$ for $\sigma < s$. Since the operator norm of Q_a depends continuously on a in $\mathcal{B}^\sigma_{p,\alpha}$-norm for some $\sigma < s$, the supremum of $\|Q_a\|$ is attained and finite. The norm continuity of the mapping $Q_a \mapsto Q_a^{-1}$ (Rudin 1991, Thm. 10.11) yields the second bound.

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

6. Proof of the upper bound

6.1. Proof of Theorem 3.2.

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

$$\begin{aligned} &\mathbb{E}_a[\langle(\tfrac{1}{T}Q_T - Q_a)\mu, \psi_\lambda\rangle^{2m}] \\ &= \mathbb{E}_a[\langle\mu, (\tfrac{1}{T}Q_T - Q_a)\psi_\lambda\rangle^{2m}] \\ &= \int_{[-r,0]^{2m}} \mathbb{E}_a\Big[\prod_{i=1}^{2m}(\tfrac{1}{T}Q_T - Q_a)\psi(u_i)\Big] d\mu(u_{2m})\dots d\mu(u_1) \\ &\le \|\mu\|^{2m}_{TV} \sup_{u_j} \int_{[-r,0]^{2m}} \mathbb{E}_a\Big[\prod_{i=1}^{2m}(\tfrac{1}{T}q_T(u_i,v_i) - q_a(u_i,v_i))\Big] \prod_{i=1}^{2m}\psi_\lambda(v_i)\,dv_{2m}\dots dv_1 \\ &= \|\mu\|^{2m}_{TV} T^{-2m} \sup_{u_j} \int_{[-r,0]^{2m}} dv_{2m}\dots dv_1 \int_{[0,T]^{2m}} dt_{2m}\dots dt_1 \\ &\quad \mathbb{E}_a\Big[\prod_{i=1}^{2m}(X(t_i+u_i)X(t_i+v_i) - q_a(u_i - v_i))\Big] \prod_{i=1}^{2m}\psi_\lambda(v_i). \end{aligned}$$

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

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

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):

$$\mathbb{E}_a\Big[\prod_{i=1}^{2m}(A_i B_i - \alpha)\Big] = \sum_{\substack{\Gamma\in P_2(4m)\\ \forall i\colon \{2i-1,2i\}\notin\Gamma}} \prod_{(k,l)\in\Gamma} \mathbb{E}_a[N_k N_l]. \tag{6.1}$$

In our case the expected value of the product equals

$$\sum_{\substack{\Gamma\in P_2(4m)\\ \forall i:\{2i-1,2i\}\notin\Gamma}} \prod_{(k,l)\in\Gamma} q_a(z_k - z_l)$$

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, $i = 1, \dots, 2m$. 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:

$$\begin{aligned} &q_a(t_i + u_i - t_j - u_j), && \text{(I)},\\ &\int_{-r}^0 q_a(t_i + u_i - t_j - v_j)\psi_\lambda(v_j)\,dv_j && \text{(II)},\\ \int_{-r}^0 &\int_{-r}^0 q_a(t_i + v_i - t_j - v_j)\psi_\lambda(v_i)\psi_\lambda(v_j)\,dv_i\,dv_j && \text{(III)}. \end{aligned}$$

For the factor (I) we shall use $|q_a(t_i + u_i - t_j - u_j)| \le C_1 e^{-\delta|t_i - t_j|}$ derived from Proposition 2.8 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 2.8, 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 ψ and distinguish the cases (1) $|t_i - t_j| > 2^{-|\lambda|}S$ and (2) $|t_i - t_j| \le 2^{-|\lambda|}S$. A substitution gives

$$\begin{aligned} &\int_{-r}^0 \int_{-r}^0 q_a(t_i + v_i - t_j - v_j)\psi_\lambda(v_i)\psi_\lambda(v_j)\,dv_i\,dv_j \\ &\qquad = \iint_{|v_i - v_j| \le S} q_a(t_i - t_j + 2^{-|\lambda|}(v_i - v_j))2^{-|\lambda|}\psi(v_i)\psi(v_j)\,dv_i\,dv_j, \end{aligned}$$

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 2.8) 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 $C := \max_j C_j$

$$\int_{[-r,0]^{2m}} dv_{2m} \dots dv_1 \, \mathbb{E}_a \left[\prod_{i=1}^{2m} (X(t_i+u_i)X(t_i+v_i) - q_a(u_i-v_i)) \right] \prod_{i=1}^{2m} \psi_\lambda(v_i)$$
$$\leq \sum_\Gamma 2^{-3|\lambda|m} C^{2m} \prod_{(k,l)\in\Gamma} \big(1+2^{|\lambda|}\mathbf{1}_{\{k,l \text{ even}, |t_{k/2}-t_{l/2}|\leq S2^{-|\lambda|}\}}\big) e^{-\delta|t_{\lceil k/2\rceil}-t_{\lceil l/2\rceil}|}.$$

The partitions Γ 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 $\{k_1',k_2\}\in\Gamma$ with $s(k_1')=s(k_1)$, then for $\{k_2',k_3\}$ with $s(k_2')=s(k_2)$ and so forth until $s(k_l)$ equals $s(k_0)$. This describes a cyclic permutation of $\{s(k_0),\dots,s(k_{l-1})\}$. Proceeding in the same manner for the remaining elements of Γ and identifying $s(k)$ with $\lceil k/2\rceil$ a fixed point-free permutation $\pi=\pi(\Gamma)$ of $\{1,\dots,2m\}$ is defined. To clarify the construction look at the following example ($m=6$):

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

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

$$\int_{[0,T]^{2m}} \prod_{(k,l)\in\Gamma} \big(1+2^{|\lambda|}\mathbf{1}_{\{k,l \text{ even}, |t_{k/2}-t_{l/2}|\leq S2^{-|\lambda|}\}}\big) e^{-\delta|t_{\lceil k/2\rceil}-t_{\lceil l/2\rceil}|}\, dt_1\dots dt_{2m}$$
$$= \prod_{\sigma\in C(\pi(\Gamma))} \int_{[0,T]^{|\sigma|}} \prod_{k=1}^{|\sigma|} \big(1+2^{|\lambda|}\mathbf{1}_{|s_{k+1}-s_k|\leq S2^{-|\lambda|}\}}\big) e^{-\delta|s_{k+1}-s_k|}\, ds_1\dots,ds_{|\sigma|}$$
$$\leq \prod_{\sigma\in C(\pi(\Gamma))} \int_0^T ds_1 \int_{[-T,T]^{|\sigma|-1}} \prod_{k=1}^{|\sigma|-1} \big(1+2^{|\lambda|}\mathbf{1}_{|u_k|\leq S2^{-|\lambda|}\}}\big) e^{-\delta|u_k|}\, du_1\dots,du_{|\sigma|-1}$$
$$\leq \prod_{\sigma\in C(\pi(\Gamma))} \Big(T\big(2\delta^{-1}+2S\big)^{|\sigma|-1}\Big)$$
$$\leq T^{|C(\pi(\Gamma))|}\big(2\delta^{-1}+2S\big)^{2m}.$$

So far we have shown

$$\mathbb{E}_a[\langle(\tfrac{1}{T}Q_T - Q_a)\mu,\psi_\lambda\rangle^{2m}] \leq \|\mu\|_{TV}^{2m} T^{-2m}\big(2C(\delta^{-1}+S)\big)^{2m} 2^{-3|\lambda|m} \sum_\Gamma T^{|C(\pi(\Gamma))|}.$$

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

$$a_{n,k} = (n-1)a_{n-1,k} + (n-1)a_{n-2,k-1}, \qquad a_{n,1}=a_{n,0}=1,\ a_{1,k}=0.$$

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 $\{1,\dots,n-1\}$ 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

$$\sum_{k=1}^{2m} a_{2m,k} x^k \le \frac{(2m)!}{m!}(x+1)\cdots(x+m), \quad x \ge 0.$$

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

$$\begin{aligned}
&\mathbb{E}_a[\cosh(\alpha T^{1/2} 2^{3|\lambda|/2}\langle(\tfrac{1}{T}Q_T - Q_a)\mu, \psi_\lambda\rangle)] \\
&= \sum_{m=0}^{\infty} \frac{(\alpha T^{1/2} 2^{3|\lambda|/2}\langle(\tfrac{1}{T}Q_T - Q_a)\mu, \psi_\lambda\rangle)^{2m}}{(2m)!} \\
&\le \sum_{m=0}^{\infty} \frac{\|\mu\|_{TV}^{2m} T^{-m}\big(2C(\delta^{-1}+S)\big)^{2m}\alpha^{2m}}{(2m)!} \sum_{k=1}^{2m} a_{2m,k} T^k \\
&\le \sum_{m=0}^{\infty} \frac{\big(\|\mu\|_{TV}^2 T^{-1}\big(2C(\delta^{-1}+S)\big)^2\alpha^2\big)^m}{m!}(T+1)\cdots(T+m) \\
&= \sum_{m=0}^{\infty} \big(\|\mu\|_{TV} T^{-1/2}\big(2C(\delta^{-1}+S)\big)\alpha\big)^{2m}\binom{T+m}{m} \\
&= \big(1 - \big(\|\mu\|_{TV}\big(2C(\delta^{-1}+S)\big)\alpha\big)^2 T^{-1}\big)^{-(T+1)} \\
&\le \exp\big(K\|\mu\|_{TV}^2\alpha^2\big)
\end{aligned}$$

where $K := 2\big(2C(\delta^{-1}+S)\big)^2$ and $T \ge T_0$ large enough. The estimates of the covariance function relied only on Proposition 2.8 whence by Proposition 5.3 the uniformity of the constant follows. □

Proof. (of Corollary 3.4) The moment inequality in Proposition 3.2 yields

$$\mathbb{P}_a\big(T^{1/2}2^{3|\lambda|/2}|\langle(\tfrac{1}{T}Q_T - Q_a)\mu, \psi_\lambda\rangle| \ge \tfrac{\kappa}{2}\sqrt{\log T}\big) \le \frac{\exp\big(K\|\mu\|_{TV}^2\alpha^2\big)}{\cosh(\alpha\frac{\kappa}{2}\sqrt{\log T})}$$

for any suitable α. 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

$$|\beta_{\lambda,T} - \langle Q_a a, \psi_\lambda\rangle| \le |\langle(\tfrac{1}{T}Q_T - Q_a)a, \psi_\lambda\rangle| + \Big|\frac{1}{T}\int_0^T \langle X(t+\bullet), \psi_\lambda\rangle\, dW(t)\Big|.$$

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

$$\begin{aligned}
&\mathbb{P}_a\Big(\frac{2^{|\lambda|}}{T}\Big|\int_0^T \langle X(t+\bullet),\psi_\lambda\rangle\,dW(t)\Big| > \tfrac{\kappa}{2}\sqrt{2^{2|\lambda|}\langle Q_a\psi_\lambda,\psi_\lambda\rangle T\log T}\Big)\\
&\le 4\sqrt{e}\kappa(\log T)T^{-\kappa^2/(8+2\kappa(\log T)^{1/2}T^{-1/2})}\\
&\quad + \mathbb{P}_a\big(2^{2|\lambda|}|\langle(\tfrac{1}{T}Q_T - Q_a)\psi_\lambda,\psi_\lambda\rangle| > \tfrac{\kappa}{2}\sqrt{T^{-1}\log T}\big)\\
&\lesssim T^{-\kappa^2/9} + T^{-\kappa^2/(16K\|\psi\|_{L^1}^2)}.
\end{aligned}$$

By Lemma 8.8 the expression $2^{2|\lambda|}\langle Q_a\psi_\lambda,\psi_\lambda\rangle$ is uniformly bounded from below by some $m > 0$ and we obtain the uniform estimate

$$\begin{aligned}
&\mathbb{P}_a\big(2^{|\lambda|}T^{1/2}|\beta_{\lambda,T} - \langle Q_a a,\psi_\lambda\rangle| > \tfrac{\kappa}{2}\sqrt{\log T}\big)\\
&\qquad\lesssim T^{-\kappa^2 2^{|\lambda|}/(16K\|a\|_{TV}^2)} + T^{-\kappa^2/(9m^2)} + T^{-\kappa^2/(16K\|\psi\|_{L^1}^2 m^2)}.
\end{aligned}$$

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

6.2. **Proof of Proposition 3.4.**

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 $(e_{j,k})$

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

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

$$\mathbb{E}_a[\|\hat{b}_T - Q_a a\|_{W^{2,\rho}}] \le \mathbb{E}_a[\|\hat{b}_T - P_J Q_a a\|_{W^{2,\rho}}] + \|(\mathrm{Id} - P_J)Q_a a\|_{W^{2,\rho}}.$$

The second (bias) term can be dealt with by linear approximation theory. The Besov space embeddings (8.1) yield under the restriction (3.4) that $B^{s+2}_{p,1} \subset W^{\frac{s}{2s+3}+2,\rho}$. By Jackson's inequality (8.2) in $W^{2,\rho}([-r,0])$ we thus find

$$\|(\mathrm{Id} - P_J)Q_a a\|_{W^{2,\rho}} \lesssim 2^{-J\frac{s}{2s+3}}\|Q_a a\|_{W^{\frac{s}{2s+3}+2,\rho}} \lesssim (T/\log T)^{-\frac{s}{2s+3}}\|Q_a a\|_{B^{s+2}_{p,1}}.$$

Due to $Q_a : \mathcal{B}^s_{p,1} \to B^{s+2}_{p,1}$ isomorphically (Theorem 2.9) with uniform constants (Theorem 2.10), this second term is of order $(T/\log T)^{-\frac{s}{2s+3}}$ uniformly over $M(s,p,S,\delta)$.

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

$$\begin{aligned}
\mathbb{E}_a[\|\hat{b}_T - P_J Q_a a\|_{W^{2,\rho}}] &\lesssim \mathbb{E}_a\Big[\sum_{j\le J}2^{j(\frac{5}{2}-\frac{1}{\rho})}\Big(\sum_k \Big|\beta_{j,k}\mathbf{1}_{|\beta_{j,k}|>\kappa_j} - b_{j,k}\Big|^\rho\Big)^{1/\rho}\Big]\\
&\le \sum_{j\le J}2^{j(\frac{5}{2}-\frac{1}{\rho})}\Big(\sum_k \mathbb{E}_a\Big[\Big|\beta_{j,k}\mathbf{1}_{|\beta_{j,k}|>\kappa_j} - b_{j,k}\Big|^\rho\Big]\Big)^{1/\rho}.
\end{aligned}$$

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

$$
\begin{aligned}
&|\beta_{j,k}-b_{j,k}|^\rho \mathbf{1}_{|\beta_{j,k}|>\kappa_j}+|b_{j,k}|^\rho \mathbf{1}_{|\beta_{j,k}|\le\kappa_j}\\
&=|e_{j,k}|^\rho \mathbf{1}_{\substack{|\beta_{j,k}|>\kappa_j\\ |b_{j,k}|\le\kappa_j/2}}+|e_{j,k}|^\rho \mathbf{1}_{\substack{|\beta_{j,k}|>\kappa_j\\ |b_{j,k}|>\kappa_j/2}}+|b_{j,k}|^\rho \mathbf{1}_{\substack{|\beta_{j,k}|\le\kappa_j\\ |b_{j,k}|>2\kappa_j}}+|b_{j,k}|^\rho \mathbf{1}_{\substack{|\beta_{j,k}|\le\kappa_j\\ |b_{j,k}|\le2\kappa_j}}\\
&\le |e_{j,k}|^\rho \mathbf{1}_{|e_{j,k}|>\kappa_j/2}+|e_{j,k}|^\rho \mathbf{1}_{|b_{j,k}|>\kappa_j/2}+|b_{j,k}|^\rho \mathbf{1}_{|e_{j,k}|>\kappa_j}+|b_{j,k}|^\rho \mathbf{1}_{|b_{j,k}|\le2\kappa_j}\\
&=: S_1(j,k)+S_2(j,k)+S_3(j,k)+S_4(j,k).
\end{aligned}
$$

By the Cauchy-Schwarz inequality, the large deviation bound on $e_{j,k}$ (3.3) and by (3.2) we obtain a fast decay for the sum involving $S_1(j,k)$:

$$
\begin{aligned}
\sum_{j\le J}2^{j(\frac52-\frac1\rho)}\Big(\sum_k \mathbb{E}_a[S_1(j,k)]\Big)^{1/\rho} &\le \sum_{j\le J}2^{j(\frac52-\frac1\rho)}\Big(\sum_k \mathbb{P}_a(|e_{j,k}|>\tfrac{\kappa_j}{2})^{1/2}\,\mathbb{E}_a[e_{j,k}^{2\rho}]^{1/2}\Big)^{1/\rho}\\
&\le \sum_{j\le J}2^{j(\frac52-\frac1\rho)}\Big(\sum_k T^{-3\rho/2}2^{-j\rho}T^{-\rho/2}\Big)^{1/\rho}\\
&\sim T^{-2}2^{3J/2}\lesssim T^{-1/2}.
\end{aligned}
$$

Even more easily, the large deviation estimate bounds the sum over $S_3(j,k)$:

$$
\begin{aligned}
\sum_{j\le J}2^{j(\frac52-\frac1\rho)}\Big(\sum_k \mathbb{E}_a[S_3(j,k)]\Big)^{1/\rho} &= \sum_{j\le J}2^{j(\frac52-\frac1\rho)}\Big(\sum_k \mathbb{P}_a(|e_{j,k}|>\tfrac{\kappa_j}{2})|b_{j,k}|^\rho\Big)^{1/\rho}\\
&\lesssim \|Q_a a\|_{2,\rho,1}T^{-3}\lesssim T^{-3}\|a\|_{\mathcal{B}^s_{p,1}}.
\end{aligned}
$$

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

$$
\|Q_a a\|_{B^{s+2}_{p,p/\rho}}\sim\Big(\sum_{j\ge0}2^{jp(s+\frac52-\frac1p)/\rho}\Big(\sum_k|b_{jk}|^p\Big)^{1/\rho}\Big)^{\rho/p},
$$

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:

$$
\sum_{j\ge0}2^{jp(s+\frac52-\frac1p)/\rho}|\{k\,|\,|b_{jk}|\ge\tau_j\}|^{1/\rho}\tau_j^{p/\rho}\lesssim\|Q_a a\|^{p/\rho}_{B^{s+2}_{p,p/\rho}}\le\|Q_a a\|^{p/\rho}_{B^{s+2}_{p,1}}. \tag{6.2}
$$

The sum involving $S_2(j,k)$ can be bounded by separate estimates, where j_0 is such that $2^{j_0}\sim T^{\frac{1}{2s+3}}$:

$$
\begin{aligned}
&\sum_{j\le J}2^{j(\frac52-\frac1\rho)}\Big(\sum_k \mathbb{E}_a[S_2(j,k)]\Big)^{1/\rho}\\
&=\sum_{j\le J}2^{j(\frac52-\frac1\rho)}\Big(\sum_k \mathbb{E}_a[|e_{j,k}|^\rho]\mathbf{1}_{|b_{j,k}|>\kappa_j/2}\Big)^{1/\rho}\\
&\lesssim \sum_{j\le j_0}2^{j(\frac52-\frac1\rho)}2^{j/\rho}T^{-1/2}2^{-j}+\sum_{j>j_0}2^{j(\frac52-\frac1\rho)}\Big(\sum_k T^{-\rho/2}2^{-j\rho}\mathbf{1}_{|b_{j,k}|>\kappa_j/2}\Big)^{1/\rho}\\
&\lesssim T^{-1/2}2^{3j_0/2}+T^{-\frac12+\frac{p}{2\rho}}\sum_{j>j_0}2^{j(\frac32-\frac1\rho+\frac p\rho)}\Big(\frac{\kappa_j}{2}\Big)^{p/\rho}|\{k\,|\,|b_{j,k}|>\kappa_j/2\}|^{1/\rho}
\end{aligned}
$$

$$\lesssim T^{-1/2}2^{3j_0/2} + T^{-\frac{1}{2}+\frac{p}{2\rho}}2^{-j_0(\frac{ps}{\rho}+\frac{3p}{2\rho}-\frac{3}{2})}$$
$$\sim T^{-1/2}2^{3j_0/2}(1+T^{p/2\rho}2^{-j_0 p(s+\frac{3}{2})/\rho}) \sim T^{-\frac{s}{2s+3}}.$$

In the fifth line we have used the sparsity estimate (6.2) and the fact that $\frac{ps}{\rho}+\frac{3p}{2\rho}-\frac{3}{2}$ is non-negative due to $\frac{1}{p}-\frac{1}{\rho}\leq\frac{2}{\rho}\frac{s}{3}$.

The slightly extended technique also applies to the estimate of the sum over $S_4(j,k)$. Here, one must choose $2^{j_0}\sim(T/\log T)^{\frac{1}{2s+3}}$ for balancing the two appearing sums:

$$\sum_{j\leq J}2^{j(\frac{5}{2}-\frac{1}{\rho})}\Big(\sum_k \mathbb{E}_a[S_4(j,k)]\Big)^{1/\rho}$$
$$=\sum_{j\leq J}2^{j(\frac{5}{2}-\frac{1}{\rho})}\Big(\sum_k |b_{j,k}|^\rho \mathbf{1}_{|b_{j,k}|\leq 2\kappa_j}\Big)^{1/\rho}$$
$$\leq \sum_{j\leq j_0}2^{j(\frac{5}{2}-\frac{1}{\rho})}2^{j/\rho}2\kappa_j + \sum_{j>j_0}2^{j(\frac{5}{2}-\frac{1}{\rho})}\Big(\sum_k |b_{j,k}|^\rho \sum_{m\geq 0}\mathbf{1}_{2^{-m}\kappa_j<|b_{j,k}|\leq 2^{-m+1}\kappa_j}\Big)^{1/\rho}$$
$$\lesssim 2^{3j_0/2}T^{-1/2}(\log T)^{1/2} + \sum_{j>j_0}2^{j(\frac{5}{2}-\frac{1}{\rho})}\Big(\sum_k\sum_{m\geq 0}2^{(-m+1)\rho}\kappa_j^\rho \mathbf{1}_{|b_{j,k}|>2^{-m}\kappa_j}\Big)^{1/\rho}$$
$$\lesssim \Big(\frac{T}{\log T}\Big)^{-\frac{s}{2s+3}} + \sum_{m\geq 0}\sum_{j>j_0}2^{j(\frac{5}{2}-\frac{1}{\rho})}2^{-m}\kappa_j|\{k\,|\,|b_{j,k}|>2^{-m}\kappa_j\}|^{1/\rho}$$
$$\lesssim \Big(\frac{T}{\log T}\Big)^{-\frac{s}{2s+3}} + \sum_{m\geq 0}2^{-j_0(\frac{ps}{\rho}+\frac{3p}{2\rho}-\frac{3}{2})}\Big(\frac{2^{-m}(\log T)^{1/2}}{T^{1/2}}\Big)^{(\rho-p)/\rho}$$
$$\sim \Big(\frac{T}{\log T}\Big)^{-\frac{s}{2s+3}}.$$

All estimates together yield

$$\mathbb{E}_a[\|\hat{b}_T - Q_a a\|_{W^{2,\rho}}] \lesssim \Big(\frac{T}{\log T}\Big)^{-\frac{s}{2s+3}},$$

where the constant holds uniformly for $a\in M(s,p,S,\delta)$. □

6.3. **Proof of Theorem 3.5.**

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 $\widehat{Q}_T : \mathcal{L}^\rho \to W^{2,\rho}([-r,0])$ follows from Lemma 8.4. Formally, the Neumann series expansion yields for $\widehat{Q}_T^{-1}$

$$\widehat{Q}_T^{-1} = (\mathrm{Id} - Q_a^{-1}(Q_a - \widehat{Q}_T))^{-1}Q_a^{-1} = \sum_{m=0}^\infty (Q_a^{-1}(Q_a-\widehat{Q}_T))^m Q_a^{-1}.$$

Introducing the random set

$$\mathcal{C}_T := \{\|Q_a^{-1}\|_{W^{2,\rho}\to\mathcal{L}^\rho}\|Q_a - \widehat{Q}_T\|_{\mathcal{L}^\rho\to W^{2,\rho}} \leq \tfrac{1}{2}\},$$

the operator $\widehat{Q}_T$ is therefore invertible on $\mathcal{C}_T$ with

$$\|\widehat{Q}_T^{-1}\|_{W^{2,\rho}\to\mathcal{L}^\rho} \leq 2\|Q_a^{-1}\|_{W^{2,\rho}\to\mathcal{L}^\rho},$$
$$\|\widehat{Q}_T^{-1} - Q_a^{-1}\|_{W^{2,\rho}\to\mathcal{L}^\rho} \leq 2\|Q_a^{-1}\|^2_{W^{2,\rho}\to\mathcal{L}^\rho}\|\widehat{Q}_T - Q_a\|_{\mathcal{L}^\rho\to W^{2,\rho}}.$$

In order to bound the probability of $\mathcal{C}_T$ from below, we use the estimate $\|Q_a - \widehat{Q}_T\|_{\mathcal{L}^\rho \to W^{2,\rho}} \lesssim \|q_a - \hat{q}_T\|_{W^{3,\rho}([-r,0])}$, derived from Lemma 8.4 with $k(t) = (q_a - \hat{q}_T)(-t)$ and $W^{3,\rho} \subset W^{2,\rho'}$. From Proposition 3.4 we know

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

Furthermore, we infer from Propositions 2.8 and 5.3

$$\mathbb{E}_a\Big[\Big|q_a(0) - \frac{1}{T}\int_0^T X(u)^2\,du\Big|^2\Big] = \frac{1}{T^2}\int_0^T\int_0^T 2q_a^2(u-v)\,du\,dv \lesssim \frac{1}{T}$$

with uniform constants. We conclude

$$\mathbb{E}_a[\|q_a - \hat{q}_T\|_{W^{3,\rho}}] \lesssim \left(\frac{T}{\log T}\right)^{-\frac{s}{2s+3}}. \tag{6.3}$$

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

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

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

$$\begin{aligned}&\|\hat{a}_T - a\|_{\mathcal{L}^\rho}\\ &\lesssim \|\widehat{Q}_T^{-1}\hat{b}_T - Q_a^{-1}Q_a a\|_{\mathcal{L}^\rho}\\ &\le \|\widehat{Q}_T^{-1}\|_{W^{2,\rho}\to\mathcal{L}^\rho}\|\hat{b}_T - Q_a a\|_{W^{2,\rho}} + \|\widehat{Q}_T^{-1} - Q_a^{-1}\|_{W^{2,\rho}\to\mathcal{L}^\rho}\|Q_a a\|_{W^{2,\rho}}\\ &\le 2\|Q_a^{-1}\|_{W^{2,\rho}\to\mathcal{L}^\rho}\|\hat{b}_T - Q_a a\|_{W^{2,\rho}} + 2\|Q_a^{-1}\|^2_{W^{2,\rho}\to\mathcal{L}^\rho}\|Q_a - \widehat{Q}_T\|_{\mathcal{L}^\rho\to W^{2,\rho}}\|Q_a a\|_{W^{2,\rho}}\\ &\lesssim \|\hat{b}_T - Q_a a\|_{W^{2,\rho}} + \|q_a - \hat{q}_T\|_{W^{3,\rho}}.\end{aligned}$$

By Theorem 2.10 the last estimate holds uniformly for all $a \in M(s,p,S,\delta)$.

From Proposition 3.4 and the estimate (6.3) we conclude

$$\sup_{a\in M(s,p,S,\delta)} \mathbb{E}_a[\|\hat{a}_T - a\|_{\mathcal{L}^\rho}\mathbf{1}_{\mathcal{C}_T}] \lesssim T^{-\frac{s}{2s+3}},$$

which accomplishes the proof of the asymptotic risk upper bound. □

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 $\operatorname{supp}(\psi_{jk}) \subset [-r,0]$ and $\operatorname{supp}(\psi_{jk}) \cap \operatorname{supp}_(\psi_{jk'}) = \varnothing$ 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 $\|a_{jk}\|_{s,p,1} \le S$ and $v_0(a_{jk}) \le -\delta$ are satisfied, hence $a_{jk} \in B(s, S, p, \delta)$ holds true. We briefly interrupt the proof for stating a classical lemma on lower bounds in the sparse case.

Lemma 7.1. *Suppose the likelihood ratio satisfies*

$$\mathbb{P}_{a_{jk}}\Big(\log(\Lambda_T(X^{(a_0)}, X^{(a_{jk})})) \geq -j\Big) \geq \pi_0 > 0$$

uniformly for all a_{jk}. Then for $\sigma(X(t), -r \leq t \leq T)$-measurable estimators $\hat{a}_T$ the following lower bound holds:

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

Proof of the Lemma. This is an adapted version of (Härdle, Kerkyacharian, Picard, and Tsybakov 1998, Lemma 10.1). Merely note the relations $K \sim 2^j$, $v_T^K \sim \lambda_T \sim j$ and $\|a_{jk} - a_{jk'}\|_\rho \sim \gamma 2^{j(\frac{1}{2}-\frac{1}{\rho})}$ 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 (3.2):

$$\begin{aligned}
&\mathbb{E}_{a_{jk}}[\log(\Lambda_T(X^{(a_0)}, X^{(a_{jk})}))^2] \\
&= \mathbb{E}_{a_{jk}}\Big[\Big(\int_0^T \langle X(t+\bullet), a_{jk} - a_0\rangle \, dW(t) - \tfrac{1}{2}\langle Q_T(a_{jk} - a_0), a_{jk} - a_0\rangle\Big)^2\Big] \\
&\leq 2\gamma^2 T \langle Q_{a_{jk}} \psi_{jk}, \psi_{jk}\rangle + \tfrac{1}{2}\,\mathbb{E}_{a_{jk}}[\langle Q_T(a_{jk} - a_0), a_{jk} - a_0\rangle^2] \\
&\leq 2\gamma^2 T \langle Q_{a_{jk}} \psi_{jk}, \psi_{jk}\rangle + \gamma^4 T^2 \langle Q_{a_{jk}} \psi_{jk}, \psi_{jk}\rangle^2 + \gamma^4\,\mathbb{E}_{a_{jk}}[\langle (Q_T - T Q_{a_{jk}})\psi_{jk}, \psi_{jk}\rangle^2] \\
&\lesssim \gamma^2 T 2^{-2j} + \gamma^4 T^2 2^{-4j} + \gamma^4 T 2^{-4j}
\end{aligned}$$

with a uniform constant for all a_{jk}. Thus, by Chebyshev's inequality the requirements of Lemma 7.1 are satisfied, when we balance the restrictions on γ by choosing $2^{(2s+3-\frac{2}{p})j(T)} \sim T/\log T$ such that

$$\gamma(T)^4 T^2 2^{-4j(T)} \sim T^2 2^{-j(T)(4s+6-\frac{4}{p})} \sim (\log T)^2 \sim j(T)^2$$

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 embbedding $\mathcal{B}^s_{p,1} \subset \mathcal{B}^\sigma_{\pi,1}$ with $\frac{1}{\pi} := \sigma - s + \frac{1}{p} < \frac{1}{p}$. Due to $\frac{1}{\pi} - \frac{1}{\rho} = \frac{2}{\rho}\frac{\sigma}{3}$ we can apply Theorem 3.5 to the class $M(\sigma, \pi, S', \delta)$, S' chosen appropriately.

(2) The upper bound is the content of Theorem 3.5, 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 for instance 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:

$$W^{m,p}(I) := \{f \in L^p(I) \,|\, f^{(i)} \in L^p(I) \text{ for all } i = 0,\dots,m\},$$

where $f^{(i)}$ denotes the i-th derivative of f in a weak (distributional) sense. These spaces are Banach spaces with respect to the following norm

$$\|f\|_{m,p} := \Big(\sum_{i=0}^{m} \|f^{(i)}\|_{L^p}^p\Big)^{1/p}.$$

An even larger scale of function spaces is given by the Besov spaces $B^s_{p,\alpha}$, measuring 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 $I_h := \{x \in I \,|\, x \pm h \in I\}$. Then the n-th order L^p-modulus of smoothness is defined by*

$$\omega_n(f,\varepsilon)_p := \sup_{|h|\le\varepsilon} \|\Delta_h^n f\|_{L^p(I_{nh})},$$

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

$$\|f\|_{s,p,\alpha} := \|f\|_{L^p(I)} + \bigg(\int_0^1 \Big(\frac{\omega_n(f,t)_p}{t^s}\Big)^\alpha \frac{dt}{t}\bigg)^{1/\alpha}$$

with the usual modification $\sup_t \omega_n(f,t)_p t^{-s}$ for $\alpha = \infty$ and with $n = \lfloor s\rfloor + 1$. The Besov space $B^s_{p,\alpha}(I) := \{f \in L^p(I) \,|\, \|f\|_{s,p,\alpha} < \infty\}$ is a Banach space when equipped with the norm $\|\bullet\|_{s,p,\alpha}$. On a bounded interval I an equivalent norm is given by (n as above)

$$\|f\|_{B^s_{p,\alpha}} \sim \|f\|_{L^p} + \|f^{(n-1)}\|_{s-(n-1),p,\alpha}.$$

Proposition 8.2. *The following embedding relations hold true*

- $B^s_{p,\alpha} \subset B^{s'}_{p,\alpha'}$, *$s > s'$, any α, α';*
- $B^s_{p,\alpha} \subset B^s_{p',\alpha}$, *$p > p'$;*
- $B^s_{p,\alpha} \subset B^s_{p,\alpha'}$, *$\alpha < \alpha'$;*
- *the Sobolev embedding theorem generalizes to*

$$B^s_{p,\alpha} \subset B^{s'}_{p',\alpha} \text{ for } s \ge s' \text{ and } s - \tfrac{1}{p} \ge s' - \tfrac{1}{p'}; \tag{8.1}$$

as a special case $B^s_{p,\alpha} \subset C^{s'}$ 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 α.

In order to apply abstract interpolation theory, we consider the case $s=1$ in a Sobolev scale first:

$$\begin{aligned}\|Tk\|_{W^{1,p}} &\sim \|Tk\|_{L^p}+\|(Tk)'\|_{L^p}\\ &\lesssim \|Tk\|_{L^\infty}+\Big\|\Big(\int_{-\bullet}^0 f(v)k(v+\bullet)\,dv\Big)'\Big\|_{L^p}\\ &\le \|f\|_{L^p}\|k\|_{L^{p'}}+\Big\|f(-\bullet)k(0)+\int_0^\bullet f(u-\bullet)k'(u)\,du\Big\|_{L^p}\\ &\le \|f\|_{L^p}\|k\|_{L^{p'}}+\|f\|_{L^p}\|k\|_\infty+\|T(k')\|_{L^p}\\ &\lesssim \|f\|_{L^p}\|k\|_{W^{1,p'}}.\end{aligned}$$

Due to $\|Tk\|_\infty\le \|f\|_{L^p}\|k\|_{L^{p'}}$ 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 $s\in(0,1)$

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

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

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

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

$$\Big(\int_0^t f(u-t)k^{(s)}(u)\,du\Big)'(t)=f(0)k^{(s)}(t)-\int_0^t f'(u-t)k^{(s)}(u)\,du$$

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

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

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 ρ satisfies $\frac{1}{\rho}+\frac{1}{\rho'}=1$. Then the integral operator*

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

is continuous from $L^\rho([-r,0])$ *to* $W^{2,\rho}([-r,0])$ *with* $\|K\|_{L^\rho\to W^{2,\rho}} \lesssim \|k\|_{W^{2,\rho'}}$.

Proof. First consider the following identities in an almost everywhere-sense for $t\in[-r,0]$ and $f\in L^\rho([-r,0])$

$$\begin{aligned}(Kf)'(t) &= \int_{-r}^0 k'(|t-s|)\operatorname{sgn}(t-s)f(s)\,ds,\\ (Kf)''(t) &= \Big(\int_{-r}^{\bullet} k'(\bullet-s)f(s)\,ds - \int_{\bullet}^0 k'(s-\bullet)f(s)\,ds\Big)'(t)\\ &= \int_{-r}^0 k''(|t-s|)f(s)\,ds + 2k'(0)f(t).\end{aligned}$$

By the Hölder inequality we obtain

$$\begin{aligned}\|Kf\|_{L^\rho} &\le 2\|k\|_{L^{\rho'}}\|f\|_{L^\rho},\\ \|(Kf)''\|_{L^\rho} &\le \rho\|k''\|_{L^{\rho'}}\|f\|_{L^\rho} + \rho\|k'\|_\infty\|f\|_{L^\rho}.\end{aligned}$$

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* $j\in\mathbb{Z}$ *and* $k\in\mathbb{Z}$ *introduce the multi-index* $\lambda=(j,k)$ *and put* $|\lambda| := |(j,k)| := j$. *A* <u>*wavelet basis*</u> $(\psi_\lambda)_\lambda$ *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*

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

Furthermore set V_j *as the closure of* $\operatorname{span}(\psi_\lambda,\ |\lambda|\le j)$. *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 ψ_λ 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 ψ_λ even if they are not directly derived from ψ. 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

$$f=\sum_\lambda\langle f,\psi_\lambda\rangle\psi_\lambda=\sum_{j\ge-1}\sum_k\langle f,\psi_{jk}\rangle\psi_{jk}$$

Note that summation over $|\lambda|\le j_0$ will always mean summation over (j,k) for all $j\le j_0$ and all possible values of k.

Wavelets are like tailor-made for the description of Besov spaces.

Definition 8.6. *A wavelet basis* (ψ_λ) *will be called* <u>*s-regular*</u> *on the interval* I *if the following two conditions are satisfied:*

(1) *For all* $\sigma \in (0,s]$, $p,\alpha \in [1,\infty]$ *the function* f *is an element of* $B^{\sigma}_{p,\alpha}(I)$ *if and only if*

$$\|P_{-1}f\|_{L^p} + \Big(\sum_{j=0}^{\infty} 2^{\alpha j(\sigma+\frac{1}{2}-\frac{1}{p})}\Big(\sum_k |\langle f,\psi_{jk}\rangle|^p\Big)^{\alpha/p}\Big)^{1/\alpha} < \infty.$$

The above expression constitutes a norm equivalent to $\|\bullet\|_{\sigma,p,\alpha}$.

(2) *For all* $k = 0,\dots,\lfloor s\rfloor$ *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

$$\|f-P_Jf\|_{W^{m,\rho}} \lesssim 2^{-Js}\|f\|_{W^{m+s,\rho}} \qquad \forall\, f\in W^{m+s,\rho} \tag{8.2}$$

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 $\|a\|_{TV}\le R<\infty$ and $v_0(a)\le -\delta<0$ and for any multi-index λ 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 5.1 and the spctral characterisation of the space $W^{1,2}$ we obtain

$$\langle Q_a\psi_\lambda,\psi_\lambda\rangle = \int_{-\infty}^{\infty}\Big|\frac{\hat\psi_\lambda(\xi)}{i\xi-\int_{-r}^{0}e^{i\xi u}\,da(u)}\Big|^2 d\xi \sim \int_{-\infty}^{\infty}(1+\xi^2)^{-1}|\hat\psi_\lambda(\xi)|^2\,d\xi$$
$$= \sup_{\|f\|_{L^2}=1}\Big|\int_{-\infty}^{\infty}(1+\xi^2)^{-1/2}f(\xi)\hat\psi_\lambda(\xi)\,d\xi\Big|^2 = \sup_{\|h\|_{W^{1,2}}=1}\langle h,\psi_\lambda\rangle^2.$$

The last expression is clearly of order $2^{-2|\lambda|}$. □

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 (ψ_λ) is a compactly supported $(m+1)$-regular wavelet basis of $L^2([-r,0])$. Then*

$$\Big|\int_{-r}^{0}\int_{-r}^{0}f(x-y)\psi_\lambda(x)\psi_\lambda(y)\,dy\,dx\Big| \lesssim \|f\|_{C^{m,1}}2^{-|\lambda|(m+2)}.$$

holds with a constant independent of f and of the multi-index λ.

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 (ψ_λ) we find

$$\Big|\int_{-r}^{0}\int_{-r}^{0}f(x-y)\psi_\lambda(x)\psi_\lambda(y)\,dy\,dx\Big|$$
$$\le \sup_{y\in[-r,0]}|\langle f(\bullet-y),\psi_\lambda\rangle|\,\|\psi_\lambda\|_{L^1}$$
$$\lesssim \sup_{y\in[-r,0]}\|f(\bullet-y)\|_{C^{m,1}}2^{-|\lambda|(m+1+\frac{1}{2})}2^{-|\lambda|/2}\|\psi\|_{L^1}.$$

Note that we have used the embedding $C^{m,1} \subset B^{m+1}_{\infty,\infty}$. □

References

Abramovich, F., and B. Silverman (1998): "Wavelet decomposition approaches to statistical inverse problems.," *Biometrika*, 85(1), 115–129.

Cavalier, L., and A. Tsybakov (2002): "Sharp adaptation for inverse problems with random noise," *Probab Theory Relat Fields*, 123(3), 323–354.

Cohen, A. (2000): "Wavelet methods in numerical analysis.," in *Handbook of numerical analysis. Vol. 7*, ed. by P. G. Ciarlet. North-Holland/ Elsevier, Amsterdam.

Cohen, A., I. Daubechies, and P. Vial (1993): "Wavelets on the interval and fast wavelet transforms.," *Appl. Comput. Harmon. Anal.*, 1(1), 54–81.

Cohen, A., M. Hoffmann, and M. Reiss (2002): "Adaptive wavelet Galerkin methods for linear inverse problems," Preprint 749, Laboratoire de probabilités et modèles aléatoires, Paris (`http://www.proba.jussieu.fr/mathdoc/preprints`).

Dalalyan, A. (2001): "Asymptotically efficient estimation for ergodic diffusion processes," *Ph.D. Thesis, University Le Mans, France.*

Diekmann, O., S. A. van Gils, S. M. Verduyn Lunel, and H.-O. Walther (1995): *Delay equations. Functional-, complex-, and nonlinear analysis.* Springer, New York.

Donoho, D. L. (1995): "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition.," *Appl. Comput. Harmon. Anal.*, 2(2), 101–126.

Donoho, D. L., and I. M. Johnstone (1994): "Ideal spatial adaptation by wavelet shrinkage.," *Biometrika*, 81(3), 425–455.

Garcia-Ojalvo, J., and R. Roy (1996): "Noise amplification in a stochastic Ikeda model.," *Physics Letters A*, 224, 51–56.

Gushchin, A. A., and U. Küchler (1999): "Asymptotic inference for a linear stochastic differential equation with time delay.," *Bernoulli*, 5(6), 1059–1098.

——— (2000): "On stationary solutions of delay differential equations driven by a Lévy process.," *Stochastic Proc. Appl.*, 88, 195–211.

——— (2001): "On parametric statistical models for stationary solutions of affine stochastic delay differential equations.," Discussion paper 91, Sonderforschungsbereich 373, Humbold Universität zu Berlin (`http://sfb.wiwi.hu-berlin.de`).

Hale, J. K., and S. M. Verduyn Lunel (1993): *Introduction to functional differential equations.* Springer, New York.

Hämarik, U. (1983): "Projection methods for regularizing ill-posed problems (in Russian).," *Tr. Vychisl. Tsentra Tartusk*, 50, 69–90.

Härdle, W., G. Kerkyacharian, D. Picard, and A. Tsybakov (1998): *Wavelets, approximation, and statistical applications.* Springer, Berlin.

Hobson, D. G., and L. Rogers (1998): "Complete models with stochastic volatility.," *Math. Finance*, 8(1), 27–48.

Hoffmann, M. (1999): "Adaptive estimation in diffusion processes.," *Stochastic Processes Appl.*, 79(1), 135–163.

Jeantheau, T. (2001): "A link between complete models with stochastic volatility and ARCH models," Preprint, Université de Marne-La-Vallée (France), Equipe d'Analyse et de Mathématiques Appliquées.

Kalifa, J., and J. Mallat (2003): "Thresholding estimators for linear inverse problems.," *Ann. Stat*, 31(1).

Katznelson, Y. (1976): *An introduction to harmonic analysis. 2nd corr. ed.* Dover Publications, New York.

Kerkyacharian, G., and D. Picard (2000): "Thresholding algorithms, maxisets and well-concentrated bases.," *Test*, 9(2), 283–344.

Kolmanovskii, V., and A. Myshkis (1992): *Applied theory of functional differential equations.* Mathematics and Its Applications. Soviet Series. 85. Kluwer, Dordrecht.

Küchler, U., and B. Mensch (1992): "Langevin's stochastic differential equation extended by a time-delayed term.," *Stochastics Stochastics Rep.*, 40(1/2), 23–42.

Küchler, U., and M. Sørensen (1997): *Exponential families of stochastic processes.* Springer, New York.

Kutoyants, Y., T. Mourid, and D. Bosq (1992): "Estimation parametrique d'un processus de diffusion avec retards. (Parametric estimation of a diffusion process with delays).," *Ann. Inst. Henri Poincare, Probab. Stat.*, 28(1), 95–106.

Kydland, F. E., and E. C. Prescott (1982): "Time to build and aggregate fluctuations.," *Econometrica*, 50, 1345–1370.

Liptser, R., and V. Spokoiny (2000): "Deviation probability bound for martingales with applications to statistical estimation.," *Stat. Probab. Lett.*, 46(4), 347–357.

Liptser, R. S., and A. N. Shiryaev (2001): *Statistics of random processes. 1: General theory. 2nd ed.* Springer, Berlin.

Mao, X. (1997): *Stochastic differential equations and their applications.* Horwood Publishing, Chichester.

Mohammed, S.-E. (1984): *Stochastic functional differential equations.* Pitman, Boston.

Mohammed, S.-E. A., M. Scheutzow, and H. von Weizsaecker (1986): "Hyperbolic state space decomposition for a linear stochastic delay equation.," *SIAM J. Control Optimization*, 24, 543–551.

Neumann, M., and R. von Sachs (1995): "Wavelet thresholding: Beyond the Gaussian i.i.d. situation.," in *Wavelets and statistics. Proceedings of the 15th French-Belgian meeting of statisticians, 1994.*, ed. by A. Antoniadis. Springer, New York.

Nussbaum, M., and S. V. Pereverzev (1999): "The degrees of ill-posedness in stochastic and deterministic noise models.," Preprint 509, Weierstraß-Institut Berlin (WIAS).

Reiss, M. (2001): "Nonparametric estimation for stochastic delay differential equations.," Ph.D. thesis, Humboldt Universität zu Berlin, `http://dochost.rz.hu-berlin.de`.

——— (2002): "Minimax rates for nonparametric drift estimation in affine stochastic delay equations.," *J. Stat. Inf. Stoch. Proc.*, 5(2), 131–152.

Rudin, W. (1991): *Functional analysis. 2nd ed.* McGraw-Hill, New York.

Scheutzow, M. (1984): "Qualitative behaviour of stochastic delay equations with a bounded memory.," *Stochastics*, 12, 41–80.

Triebel, H. (1983): *Theory of function spaces.* Mathematik und ihre Anwendungen in Physik und Technik, Bd. 38. Akademische Verlagsgesellschaft, Leipzig.

Vakhaniya, N., V. Tarieladze, and S. Chobanyan (1987): *Probability distributions on Banach spaces.* Reidel, Dordrecht.