Thresholding Projection Estimators in Functional Linear Models.

HERVÉ CARDOT* JAN JOHANNES†

December 17, 2008

Abstract

We consider the problem of estimating the regression function in functional linear regression models by proposing a new type of projection estimators which combine dimension reduction and thresholding. The introduction of a threshold rule allows to get consistency under broad assumptions as well as minimax rates of convergence under additional regularity hypotheses. We also consider the particular case of Sobolev spaces generated by the trigonometric basis which permits to get easily mean squared error of prediction as well as estimators of the derivatives of the regression function. We prove these estimators are minimax and rates of convergence are given for some particular cases.

Keywords: Derivatives estimation, Galerkin method, Linear inverse problem, Mean squared error of prediction, Optimal rate of convergence, Hilbert scale, Sobolev Space.

AMS 2000 subject classifications: Primary 62J05; secondary 62G20, 62G08.

1 Introduction

Functional data analysis (Ramsay and Silverman (2005), Ferraty and Vieu (2006)) is a topic of growing interest in statistics and many applications in chemometrics (Frank and Friedman (1993)), finance (Preda and Saporta (2005)), biometry or climatology (Besse et al. (2000)) are now dealing with the functional linear model. This model is useful to estimate or predict a scalar random variable, say $Y \in \mathbb{R}$, thanks to a random function denoted by $X$. We assume in the following that $Y$ and $X$ are centered random variables and, without loss of generality, that the random function $X$ takes values in $L^2(0,1)$, the space of square integrable functions defined on $[0,1]$ endowed with its usual inner product $\langle f,g \rangle = \int_0^1 f(t)g(t)dt$ and associated norm $\|f\| = (\langle f,f \rangle)^{1/2}$, $f,g \in L^2[0,1]$. The functional linear model is then defined by

$$Y = \int_0^1 \beta(t)X(t)dt + \sigma \epsilon, \quad \sigma > 0,$$

where the function $\beta(t)$ is called the regression or slope function and the error term $\epsilon$ is supposed to be centered $\mathbb{E}(\epsilon) = 0$ and not correlated with $X$; $\forall \ t \in [0,1], \ \mathbb{E}(X(t)\epsilon) = 0$.

*Université de Bourgogne, Institut de Mathématiques de Bourgogne, 9 Av. Alain Savary, 21078 Dijon Cedex, France, e-mail: herve.cardot@u-bourgogne.fr
†Universität Heidelberg, Institut für Angewandte Mathematik, Im Neuenheimer Feld, 294, D-69120 Heidelberg, Germany, e-mail: johannes@statlab.uni-heidelberg.de
Assuming that \( X \) has a finite second moment, i.e. \( \mathbb{E}\|X\|^2 = \int_0^1 \mathbb{E}|X(t)|^2\,dt < \infty \), one can define the covariance operator of \( X \), say \( \Gamma \). This operator is defined on \( L^2[0,1] \) as follows: for any function \( f \in L^2[0,1] \),

\[
\Gamma f(s) = \int_0^1 \text{cov}(X(t), X(s))f(t)\,dt, \quad \forall s \in [0,1].
\]  

(1.2)

It is well known (see e.g. Cardot et al. (1999)) that the regression function \( \beta \) satisfies the following moment equation

\[
g(s) := \mathbb{E}[YX(s)] = [\Gamma \beta](s), \quad s \in [0,1],
\]  

(1.3)

where \( g \) belongs to \( L^2[0,1] \). Since \( \Gamma \) is a non-negative nuclear operator (Dauxois et al. (1982)) a continuous generalized inverse of \( \Gamma \) does not exist as long as the range of the operator \( \Gamma \) is an infinite dimensional subspace of \( L^2[0,1] \). Consequently inverting equation (1.3) to recover \( \beta \) can be seen as an ill-posed inverse problem. Cardot et al. (2003) provides a necessary and sufficient condition for the existence of a unique solution of equation (1.3).

**Assumption 1.1.** The covariance operator \( \Gamma \) of the random function \( X \) is injective and the function \( g = \mathbb{E}[YX] \) belongs to the range \( \mathcal{R}(\Gamma) \) of \( \Gamma \).

Under this assumption, the covariance operator \( \Gamma \) admits a discrete spectral decomposition given by a sequence \( (\lambda_j)_{j \in \mathbb{N}} \) of strictly positive eigenvalues and a sequence of corresponding orthonormal eigenfunctions \( \{\phi_j\}_{j \in \mathbb{N}} \). Then, the normal equation (1.3) can be rewritten as follows

\[
\beta = \sum_{j \in \mathbb{N}} \frac{g_j}{\lambda_j} \cdot \phi_j \quad \text{with} \quad g_j := \langle g, \phi_j \rangle, \quad j \in \mathbb{N}.
\]  

(1.4)

It is well-known that, even in case of a-priori known eigenvalues \( \{\lambda_j\} \) and eigenfunctions \( \{\phi_j\} \), replacing in (1.4) the unknown function \( g \) by a consistent estimator \( \hat{g} \) does in general not lead to a consistent estimator of \( \beta \). To be more precise, since the sequence \( (\lambda_j)_{j \in \mathbb{N}} \) tends to zero, \( \mathbb{E}\|\hat{g} - g\|^2 = o(1) \) does generally not imply \( \sum_{j \in \mathbb{N}} |\lambda_j|^{-2} \cdot \mathbb{E}\|\hat{g} - g, \phi_j\|^2 = o(1) \). Consequently, the estimation in functional linear model is called ill-posed and additional regularity assumptions on the regression function \( \beta \) are necessary in order to obtain a uniform rate of convergence (c.f. Engl et al. (2000)).

The objective is to estimate the regression function \( \beta \), as well as its derivatives, when observing a sample \( (Y_i, X_i) \) of \( n \) i.i.d realizations of \( (Y,X) \). We can define the empirical estimators of \( g \) and \( \Gamma \) respectively as follows

\[
\hat{g} := \frac{1}{n} \sum_{i=1}^n Y_i X_i \quad \text{and} \quad \hat{\Gamma} := \frac{1}{n} \sum_{i=1}^n \langle X_i, \cdot \rangle X_i.
\]  

(1.5)

The main class of estimation procedures studied in the statistical literature are based on principal components regression and consist in reducing the dimension by inverting equation (1.3) in the finite dimension space generated by the eigenfunctions of \( \hat{\Gamma} \) associated to the largest eigenvalues (see e.g. Bosq (2000), Franklin and Friedman (1993), Cardot et al. (1999), Cardot et al. (2007) or Müller and Stadtmüller (2003) in the context of generalized linear models).

The second important class of estimators relies on minimizing a penalized least squares criterion which can be seen as generalization of the ridge regression. Marx and Eilers (1998)
and Cardot et al. (2003) proposed B-splines expansion of the regression function with a penalty dealing with the squared norm of a fixed order derivative of the estimators. More recently [Crambes et al. (2008)] proposed a spline smoothing decomposition with the same type of penalty and proved the optimality of their estimators according to a criterion that can be interpreted as a squared error of prediction. Note that this question has given rise to numerous publications in the machine learning community with similar ideas based on reproducing kernel Hilbert spaces (RKHS) and Tikhonov regularization (see e.g. Smale and Zhou (2007), Bauer et al. (2007) and references therein).

Borrowing ideas from the inverse problems community [Efrodimovich and Koltchinskii (2001)] and [Hoffmann and Reiβ (2008)] we propose in this article a new class of estimators which rely on dimension reduction by projecting the data onto some basis of orthonormal functions and threshold techniques that allow to control the accuracy of the estimator. More precisely, let us consider a set of orthonormal functions such as wavelet or trigonometric basis denoted by \( \{\psi_1, \ldots, \psi_m, \ldots \} \) which forms a basis of \( L^2[0,1] \). Given a dimension \( m \geq 1 \), we denote by \( \hat{\Gamma}^m \) the \( m \times m \) matrix with generic elements \( \langle \hat{\Gamma} \psi_j, \psi_j \rangle \), \( j, \ell = 1, \ldots, m \) and by \( [g]^m \) the \( m \) vector with elements \( \langle \hat{g}, \psi_j \rangle \), \( \ell = 1, \ldots, m \). We can first remark, that the least squares estimator of \( \beta \) obtained with the projections of the \( X_i \) onto \( \Psi_m \), the subspace of \( L^2[0,1] \) spanned by the functions \( \{\psi_1, \ldots, \psi_m\} \), is simply given, when \( \hat{\Gamma}^m \) is non singular, by \( ([\hat{\Gamma}^m]^{-1}[g]^m)^t[\psi]^m(\cdot) \) where \( [\psi]^m(\cdot) = (\psi_1(\cdot), \ldots, \psi_m(\cdot))^t \). Our estimator, in its simplest form, consists in thresholding this projection estimator when, roughly speaking, the norm of the inverse of the matrix \( \hat{\Gamma}^m \) is too large. More precisely, introducing a threshold value \( \gamma \) which will depend on \( m \) and \( n \) we propose to estimate \( \beta \) as follows

\[
\hat{\beta}(t) = \sum_{\ell=1}^m \hat{\beta}_\ell \cdot \mathbb{1}\{||[\hat{\Gamma}]^{-1}_m|| \leq \gamma \} \cdot \psi_\ell(t), \quad t \in [0,1], \tag{1.6}
\]

where the \( \hat{\beta}_\ell \) are the generic elements of the vector of coordinates obtained by least squares projection and \( \mathbb{1} \) is the indicator function. This new thresholding step can be seen as an improvement of the estimator proposed by Ramsay and Dalzell (1991) which was built by projecting the data onto finite dimensional basis of functions. From an inverse problems perspective this approach is similar to the linear Galerkin procedure (Natterer (1997) or Engl et al. (2000)) defined as follows, \( \beta^m \in \Psi_m \) denotes a Galerkin solution of the operator equation \( g = \Gamma \beta \) when

\[
||g - \Gamma \beta^m|| = ||g - \Gamma \hat{\beta}||, \quad \forall \hat{\beta} \in \Psi_m. \tag{1.7}
\]

Since \( \Gamma \) is strictly positive it follows that \( \beta^m = [\beta^m]^m_m [\psi]^m(\cdot) \) with \( [\beta^m]^m_m = [\Gamma]^{-1}_m [g]^m_m \) is the unique Galerkin solution satisfying \( \Gamma(\beta - \beta^m)\big|_m = 0 \). It has the advantage compared to principal components regression that it does not necessitate to estimate the eigenfunctions of the empirical covariance operator.

We will consider a large class of weighted norms to evaluate the asymptotic rates of convergence of the thresholded projection estimators. For \( f \in L^2[0,1] \), we define

\[
||f||_{\omega}^2 = \sum_{j=1}^{\infty} \omega_j \langle f, \psi_j \rangle^2 \tag{1.8}
\]

for some strictly positive sequence of weights \( (\omega_j)_{j \in \mathbb{N}} \). Then, the performance of the estimator \( \hat{\beta} \) of \( \beta \) is evaluated according to the risk \( \mathbb{E}||\hat{\beta} - \beta||_{\omega}^2 \), called \( \mathcal{W}_\omega \)-risk in the following,
which is simply the $L^2[0,1]$-risk when $\omega_j = 1$ for all $j \in \mathbb{N}$. This general framework allows us with appropriate choices of the weight sequence $\omega$ to cover the estimation of derivatives of $\beta$ as well as the optimal estimation with respect to the mean squared prediction error. Indeed, the prediction error of a new value of $Y$ given any random function $X_{n+1}$ possessing the same distribution as $X$ and being independent of $X_1, \ldots, X_n$ can be evaluated as follows (see for example Cardot et al. (2003) or Crambes et al. (2008) for similar setups)

$$E\left[\left|\int_0^1 \hat{\beta}(s)X_{n+1}(s)ds - \int_0^1 \beta(s)X_{n+1}(s)ds\right|^2 \parallel \hat{\beta}\right] = \langle \Gamma(\hat{\beta} - \beta), (\hat{\beta} - \beta)\rangle.$$ 

Consequently, if we suppose, now for sake of simplicity, that the functions $\psi_j$ are also the eigenfunctions $\phi_j$ of operator $\Gamma$ then it is clear that choosing $\omega_j = \lambda_j$ leads to evaluate, according to the $\omega$-norm, the mean squared prediction error of the estimator.

The paper is organized as follows. In section 2, we fix notations and we first derive consistency of the estimator in the general case under broad moment assumptions and then prove minimax results under some additional regularity assumptions based on a link condition between the operator $\Gamma$ and the basis $\{\psi_j\}$. Section 3 is devoted to the particular case of trigonometric basis and focuses on finitely and infinitely smoothing operator $\Gamma$ as well as different regularity conditions for the function $\beta$. We first consider the case of mean squared prediction error and get asymptotic rates of convergence which are comparable to those of Crambes et al. (2008) in the polynomial case. One remarkable result is that for the exponential case, one can attain the parametric rates up to a power of a log $n$ factor. Rates of convergence for the function itself and its derivatives are also given. They are similar to those obtained by Hall and Horowitz (2007) in the case of the estimation of the function itself. Finally, a brief section 4 presents the concluding remarks and some perspectives. The proofs are gathered in the Appendix.

2 Asymptotic properties, the general case

2.1 Notations and assumptions.

We assume from now on that the regression function $\beta$ belongs to some ellipsoid $\mathcal{W}_b^\rho$, $\rho > 0$, defined as follows

$$\mathcal{W}_b^\rho := \{ f \in L^2[0,1] : \sum_{j=1}^\infty b_j|\langle f, \psi_j\rangle|^2 =: \|f\|_b^2 \leq \rho\}, \tag{2.1}$$

where $\{\psi_j, j \in \mathbb{N}\}$ is as before some orthonormal basis in $L^2[0,1]$ not necessarily corresponding to the eigenfunctions of $\Gamma$, and the sequence of weights $(b_j)_{j \in \mathbb{N}}$ is non-decreasing. Here $\mathcal{W}_b^\rho$ captures all the prior information (such as the smoothness) about the unknown slope function $\beta$.

Matrix and operator notations. Given $m \geq 1$, $\Psi_m$ denotes the subspace of $L^2[0,1]$ spanned by the functions $\{\psi_1, \ldots, \psi_m\}$. $\Pi_m$ and $\Pi_m^\perp$ denote the orthogonal projections on $\Psi_m$ and its orthogonal complement $\Psi_m^\perp$, respectively. Given an operator (matrix) $K$, $\|K\|_\omega$ denotes its operator $\mathcal{W}_b^\omega$-norm, i.e. $\|K\|_\omega := \sup_{\|f\|_\omega = 1} \|Kf\|_\omega$. The inverse operator (matrix) of $K$ is denoted by $K^{-1}$, the adjoint (transposed) operator (matrix) of $K$ by $K^t$. The identity operator (matrix) is denoted by $I$. For a vector $v$ and a matrix $K$, the upper
Let \( m \) subvector and \( m \times m \) sub-matrix is denoted by \( [v]_m \) and \([K]_m\) and its entries by \( v_i \) and \( K_{i,j} \) respectively. The diagonal matrix with entries \( v \) is denoted by \( \text{Diag}(v) \). \([f]_m\) and \([K]_m\) denote the (infinite) vector and matrix of the function \( f \) and the operator \( K \) with the entries \([f]_i = \langle f, \psi_i \rangle \) and \([K]_{i,j} = \langle K \psi_j, \psi_i \rangle\) respectively. Clearly, \( [\Pi_m f]_m = [f]_m \) and if we restrict \( \Pi_m K \Pi_m \) to an operator from \( \Psi_m \) into itself, then it has the matrix \([K]_m\). Moreover, \( \Pi_m f = [f]_m \psi_m(\cdot) \) and \( \Pi_m K \Pi_m f = [f]_m \psi_m(\cdot) \) with \([\psi]_m = (\psi_1(\cdot), \ldots, \psi_m(\cdot))^t\).

Consider the covariance operator \( \Gamma \). We assume throughout the paper that \( \Gamma \) is strictly positive definite and hence the matrix \([\Gamma]_m^{-1}\) is nonsingular for all \( m \in \mathbb{N} \), so that \([\Gamma]_m^{-1}\) always exists. Under this assumption the notation \( \Gamma^{-1}_m \) is used for the operator from \( L^2[0,1] \) to itself, whose matrix in the basis \( \{\psi_j\} \) has the entries \(([\Gamma]_m^{-1})_{i,j}\) for \( 1 \leq i, j \leq m \) and zeroes otherwise.

**Moment assumptions.** The results derived below involve additional conditions on the moments of the random function \( X \), which we formalize now. Here and subsequently, we denote by \( \mathcal{X} \) the set of all centered random functions \( X \) with finite second moment, i.e., \( \mathbb{E} \|X\|^2 < \infty \), and strictly positive covariance operator. Given \( X \in \mathcal{X} \) consider the random vector \([X]_m\), then its entries \([X]_j = \langle X, \psi_j \rangle\) have mean zero and variance \([\Gamma]_{j,j} = \langle \psi_j, \psi_j \rangle\), but they are not uncorrelated. In fact, \([\Gamma]_m \) is the covariance matrix of \([X]_m\). Since \( \Gamma \) is strictly positive definite it follows that \([\Gamma]_m \) is non-singular. Therefore, the random vector \([\Gamma]^{-1/2}_m [X]_m\) has mean zero and identity \( I_m \) as covariance matrix. Then we denote by \( \mathcal{X}^k_\eta \), \( k \in \mathbb{N}, \eta \geq 1 \), the subset of \( \mathcal{X} \) containing only random functions \( X \) with uniformly bounded \( k \)-th moment of the corresponding random variables \([X]_j / [\Gamma]_{j,j}^{1/2}, j \in \mathbb{N} \), and \([\Gamma]^{-1/2}_m [X]_m j_j, 1 \leq j \leq m, m \in \mathbb{N} \), that is

\[
\mathcal{X}^k_\eta := \left\{ X \in \mathcal{X} \text{ such that } \sup_{j \in \mathbb{N}} \mathbb{E} \left| [X]_j / [\Gamma]_{j,j}^{1/2} \right|^k \leq \eta \right\}
\]

and

\[
\sup_{m \in \mathbb{N}} \sup_{1 \leq j \leq m} \mathbb{E} \left| ([\Gamma]^{-1/2}_m [X]_m)_{j, j} \right|^k \leq \eta.
\]

(2.2)

It is worth noting that in case \( X \in \mathcal{X} \) is a Gaussian random function the corresponding random variables \([X]_j / [\Gamma]_{j,j}^{1/2}, j \in \mathbb{N} \) and \([\Gamma]^{-1/2}_m [X]_m j_j, 1 \leq j \leq m, m \in \mathbb{N} \), are Gaussian with mean zero and variance one. Hence, for each \( k \in \mathbb{N} \) there exists \( \eta \) such that any Gaussian random function \( X \in \mathcal{X} \) belongs also to \( \mathcal{X}^k_\eta \). Furthermore, in what follows, \( \mathcal{E}^k_\eta \) stands for the set of all centered error terms \( \epsilon \) with variance one and finite \( k \)-th moment, i.e., \( \mathbb{E} |\epsilon|^k \leq \eta \).

### 2.2 Consistency.

The \( \mathcal{W}_\omega \)-risk of \( \hat{\beta} \) is essentially determined by the deviation of the estimators of \([g]_m \) and \([\Gamma]_m \), and by the regularization error due to the projection. The next assertion summarizes then minimal conditions to ensure consistency of \( \hat{\beta} \) proposed in (1.1).

**Proposition 2.1.** Assume an \( n \)-sample of \((Y, X)\) satisfying (1.1) with \( \sigma > 0 \). Let \( \beta \in \mathcal{W}_\omega \), \( X \in \mathcal{X}^k_\eta \) and \( \epsilon \in \mathcal{E}^k_\eta \), \( \eta \geq 1 \). Consider the estimator \( \hat{\beta} \) with parameter \( m = m(n) \) and threshold \( \gamma := \gamma(n) \) are chosen such that \( \gamma \geq 2 \left\| \Gamma^{-1}_m \right\| \) and suppose, as \( n \to \infty \), that

\[
1/m = o(1), \quad (m/n) \sup_{1 \leq j \leq m} \{\omega_j\} = o(1), \quad (m^2/n) = o(1) \quad \text{and} \quad \gamma^2 (m^3/n^{1+1/2}) = O(1).
\]

If in addition \( \sup_{m \in \mathbb{N}} \| \Gamma^{-1}_m \Pi_m \Pi_m \| \omega < \infty \), then \( \mathbb{E} \| \hat{\beta} - \beta \|^2 = o(1) \) as \( n \to \infty \).
Remark 2.1. The last result covers the case $\omega \equiv 1$, i.e., the estimator of $\beta$ is consistent without an additional assumption on $\beta$. However, consistency is only obtained under the condition $\sup_{m \in \mathbb{N}} \| \Gamma_m^{-1} \Pi_n \Gamma_{m} \|_{\omega} < \infty$, which is known to be sufficient to ensure convergence in the $\mathcal{W}_{\omega}$-norm as $m \to \infty$ of the Galerkin solution $\beta_m = \left[ \beta_m^{[f]}(\omega) \right]_{m}$ with $[\beta_m]_{m} = [\Gamma_m^{-1} g]_{m}$ to the slope parameter $\beta$. Furthermore, if $\omega$ is increasing, as in case of a Sobolev norm, then $\hat{\beta}$ is obviously a consistent estimator only if $\beta \in \mathcal{W}_{\omega}$. Moreover, in the last assertion we may replace the condition $\beta \in \mathcal{W}_{\omega}$ by the assumption $\beta \in \mathcal{W}_{b}$ and $(\omega_j/b_j)$ is non-increasing. In this situation we have $\mathcal{W}_{b} \subset \mathcal{W}_{\omega}$ and thus the result still holds true. Roughly speaking this corresponds to the condition that at least $p \geq s$ derivatives exist in case we want to estimate the $s$-th derivative. 

Link condition. In the last assertion the choice of the smoothing parameter $m$ and $\gamma$, i.e. $\gamma \geq 2 \| \Gamma \|^1_{m}$, depends on the relation between the covariance operator $\Gamma$ associated to the regressor $X$ and the basis $\{ \psi_j \}$ used for the projection, which we formalize next. Consider the sequence $(\| \Gamma \psi_j \|_{j \geq 1})$, which is summable and hence converges to zero since $\Gamma$ is nuclear. In what follows we impose restriction on the decay of this sequence. Therefore, consider a strictly positive, monotonically decreasing and summable sequence of weights $\nu := (\nu_j)_{j \in \mathbb{N}}$ with $\nu_1 = 1$. Then for $s \in \mathbb{R}$ denote by $\| . \|_{\nu s}$ the associated weighted norm given by $\| f \|_{\nu s}^2 := \sum_{j=1}^{\infty} \nu_j^s |\langle f, \psi_j \rangle|^2$. Let $\mathcal{N}$ be the set of all self-adjoint nuclear operator defined on $L^2[0,1]$. Then for $d \geq 1$ define the subset $\mathcal{N}^d$ of $\mathcal{N}$ by

$$
\mathcal{N}^d := \left\{ \Gamma \in \mathcal{N} : \| f \|_{\nu d}^2 / d^2 \leq \| \Gamma f \|^2 \leq d^2 \| f \|_{\nu d}^2, \quad \forall f \in L^2[0,1] \right\}.
$$

A similar condition, but in a different context, can be found, for example, in Nair et al. (2005) and Chen and Reiß (2008). Note, for all $\Gamma \in \mathcal{N}^d$ by using the inequality of Heinz (1951) it follows that $\| \Gamma \psi_j \| \asymp \nu_j$. Hence, the sequence $(\nu_j)_{j \in \mathbb{N}}$ has to be summable, i.e., $\sum_j \nu_j < \infty$, since $\Gamma$ is nuclear. We first consider this general class of operator. However, we illustrate condition (2.3) in Section 3 by considering the particular cases of a sequence $\nu$ with polynomial or exponential decay which are naturally linked to polynomial or exponential decreasing rates for the eigenvalues of $\Gamma$. To be more precise, if the eigenvalue decomposition of $\Gamma \in \mathcal{N}$ is given by $\{ \lambda_j, \psi_j, j \in \mathbb{N} \}$ then $\Gamma \in \mathcal{N}^d$ if and only if $\lambda_j \asymp \nu_j$ for all $j \in \mathbb{N}$. All the results below are derived under the following basic regularity assumption.

Assumption 2.1. Let $\omega := (\omega_j)_{j \geq 1}$, $b := (b_j)_{j \geq 1}$ and $\nu := (\nu_j)_{j \geq 1}$ be strictly positive sequences of weights with $\omega_1 = 1$, $b_1 = 1$ and $\nu_1 = 1$ such that $b$ and $(b_j/\omega_j)_{j \geq 1}$ are non-decreasing and $\nu$ and $(\nu_j^2/\omega_j)_{j \geq 1}$ are non-increasing with $\Lambda := \sum_j \nu_j < \infty$.

Note that under Assumption 2.1 i.e., $(b_j/\omega_j)_{j \geq 1}$ is non-decreasing, the ellipsoid $\mathcal{W}^d_{\omega}$ is a subset of $\mathcal{W}^d_{b}$. Roughly speaking, if $\mathcal{W}^d_{b}$ describes $p$-times differentiable functions, then the Assumption 2.1 ensures that the $\mathcal{W}^d_{\omega}$-risk involves maximal $s \leq p$ derivatives. On the other hand if the sequence $\omega$ is decreasing, i.e., the $\mathcal{W}^d_{\omega}$-norm is roughly speaking smoothing, the Assumption 2.1 excludes cases in which $\omega$ decreases faster than the sequence $\nu^2$. However, in case $\omega \equiv \nu^2$ we show below that the obtainable optimal-rate is parametric, and hence, whenever $(\omega_j/\nu_j^2) = o(1)$ it is parametric too.

The next assertion summarizes now minimal conditions to ensure consistency of the estimator $\beta$ given in (1.6) when the covariance operator satisfies a link condition.

Corollary 2.2. Assume an $n$-sample of $(Y, X)$ satisfying (1.1) with $\sigma > 0$ and associated covariance operator $\Gamma \in \mathcal{N}^d$, $d \geq 1$. Let $\beta \in \mathcal{W}_{b}$, $X \in \mathcal{X}^d_{\eta}$ and $\epsilon \in \mathcal{E}^d_{\eta}$, $\eta \geq 1$. Consider

\footnote{We write $a \asymp d$ if $d^{-1} \leq b/a \leq d$.}
the estimator \( \hat{\beta} \) with threshold \( \gamma = 8d^3/v_m \) and parameter \( m := m(n) \) chosen such that 
\[ 1/m = o(1), \ (m/n) \sup_{1 \leq j \leq m} \{ \omega_j / v_j \} = o(1), \ (m^2 / n) = o(1) \] 
and \( m^3 / (v_m^2 n^{1/2}) = O(1) \) as \( n \to \infty \). If in addition Assumption 2.1 is satisfied, then \( E\| \hat{\beta} - \beta \|_2^2 = o(1) \) as \( n \to \infty \).

It is worth noting that the link condition \( \Gamma \in N_0^d \) used in the last assertion implies 
\( \sup_{m \in \mathbb{N}} \| \Gamma_m^1 \Pi_m \Gamma_m^2 \|_\infty \leq \infty \) and hence ensures automatically the consistency in the \( W_\omega \)-norm of the Galerkin solution \( \beta^m \) as \( m \to \infty \). However, in order to obtain a rate of convergence it is necessary to impose additional regularity assumption on the slope parameter \( \beta \). First we derive a lower bound for any estimator when these regularity assumptions are formalized by the condition that \( \beta \) belongs to the ellipsoid \( W_\omega^\rho \).

### 2.3 The lower bound.

It is well-known that in general the hardest one-dimensional subproblem does not capture the full difficulty in estimating the solution of an inverse problem even in case of a known operator (for details see e.g. the proof in Mair and Ruymgaart (1994)). In other words, there does not exist two sequences of slope functions \( \beta_{1,n}, \beta_{2,n} \in W_\omega^\rho \), which are statistically not consistently distinguishable and which satisfy \( \|\beta_{1,n} - \beta_{2,n}\|_\omega \geq C\delta_n^\ast \), where \( \delta_n^\ast \) is the optimal rate of convergence. Therefore we need to consider subsets of \( W_\omega^\rho \) with growing number of elements in order to get the optimal lower bound. More precisely, we obtain the following lower bound by applying Assouad’s cube technique (see e.g. Korostolev and Tsybakov (1993) or Chen and Reiß (2008)). Moreover, the following lower bound is obtained under the additional assumption that distribution of the error term \( \epsilon \) is Gaussian with mean zero and variance one, i.e., \( \epsilon \sim \mathcal{N}(0,1) \).

**Theorem 2.3.** Assume an \( n \)-sample of \( (Y,X) \) satisfying (1.1) with \( \sigma > 0 \) and associated covariance operator \( \Gamma \in N_0^d \), \( d \geq 1 \). Suppose the error term \( \epsilon \sim \mathcal{N}(0,1) \) is independent of \( X \). Consider \( W_\omega^\rho \), \( \rho > 0 \), as set of slope functions. Let \( m_* := m_*(n) \) and \( \delta_n^\ast := \delta_n^\ast(m_*) \) for some \( \Delta \geq 1 \) be chosen such that

\[
\Delta^{-1} \leq \frac{b_{m_*}}{n \omega_{m_*}} \sum_{j=1}^{m_*} \frac{\omega_j}{v_j} \leq \Delta \quad \text{and} \quad \delta_n^\ast := \omega_{m_*} / b_{m_*}.
\]  

(2.4)

If in addition the Assumption 2.1 is satisfied, then for any estimator \( \tilde{\beta} \) of \( \beta \) we have

\[
\sup_{\beta \in W_\omega^\rho} \left\{ E\| \tilde{\beta} - \beta \|_2^2 \right\} \geq \frac{1}{4\Delta} \cdot \min \left\{ \frac{\sigma^2}{2d}, \rho / \Delta \right\} \cdot \delta_n^\ast.
\]

**Remark 2.2.** The normality and independence assumption on the error term in the last theorem is only used to simplify the calculation of the distance between distributions corresponding to different slope functions. However, below we show an upper bound for the estimator \( \tilde{\beta} \) in case the error term \( \epsilon \in \mathcal{E}_k^\eta \) and the regressor \( X \in \mathcal{X}_k^\eta \) for some \( k \in \mathbb{N} \) and \( \eta \geq 1 \) are only uncorrelated, which includes the particular case of an independent Gaussian error considered in Theorem 2.3 as long as \( \eta \) is sufficiently large. Therefore, by applying Theorem 2.3 an upper bound of order \( \delta_n^\ast \) implies that this rate is optimal and hence the estimator \( \tilde{\beta} \) is minimax-optimal. Note further that if \( \{\omega_j / v_j\} \) is summable then the order \( \delta_n^\ast \) is parametric. This in particular is the case when \( \omega \equiv v^2 \) since \( (v_j) \) is summable. \( \square \)
In case the eigenfunctions of the operator $\Gamma$ are known, the obtainable accuracy of any estimator of $\beta$ is essentially determined by the decay of the eigenvalues $(\lambda_j)_{j \geq 1}$ of $\Gamma$. To be more precise, if for some sequence of weights $v := (v_j)_{j \geq 1}$ we have

$$\exists d \geq 1 : \quad \lambda_j \asymp d^j v_j, \quad j \geq 1,$$

(2.5)

then $v$ determines the obtainable rate of convergence (c.f. Johannes (2008)). If $\{\psi_j\}$ are the eigenfunctions of $\Gamma$, i.e., $\lambda_j = \langle \Gamma \psi_j, \psi_j \rangle$, then the condition (2.5) holds if and only if $\Gamma \in \mathcal{N}_v^d$. In other words, the condition $\Gamma \in \mathcal{N}_v^d$ specifies in this situation the decay of the eigenvalues of $\Gamma$. However, the set $\mathcal{N}_v$ also contains operators whose eigenfunctions are not given by $\{\psi_j\}$. Then the corresponding eigenvalues may decay far slower than the sequence of weights $v$. Hence, for these operators the obtainable rate of convergence may be far slower by using the basis $\{\psi_j\}$ in place of their eigenfunctions. \qed

### 2.4 The upper bound.

In the following theorem we provide an upper bound for the estimator $\hat{\beta}$ defined in (1.6) by assuming sequences $b$, $\omega$ and $v$ with the additional property that

$$\frac{m_{2k}^*}{\delta_n^k n^k} = O(1), \quad \frac{m_*}{\delta_n^k n^k} \sup_{1 \leq j \leq m_*} \left\{ \frac{\omega_j}{v_j} \right\} = O(1) \quad \text{and} \quad \frac{m_{*+k}^*}{n^{k/2-1}} = O(1),$$

(2.6)

where $m_* := m_*(n)$ and $\delta_n^* := \delta_n^*(m_*)$ are given by (2.4). The next theorem states that the rate $\delta_n^*$ of the lower bound given in Theorem 2.3 provides also an upper bound of the estimator $\hat{\beta}$ defined in (1.6).

**Theorem 2.4.** Assume an $n$-sample of $(Y, X)$ satisfying (1.1) with $\sigma > 0$ and associated covariance operator $\Gamma \in \mathcal{N}_v^d$, $d \geq 1$. Consider $\mathcal{W}_b^\rho$, $\rho > 0$ as set of slope functions and suppose that the sequences $b$, $\omega$ and $v$ satisfy the Assumption 2.4. Let $m_* := m_*(n)$ and $\delta_n^* := \delta_n^*(n)$ be given by (2.4) and suppose (2.6) is satisfied for some $k \geq 4$. Consider the estimator $\hat{\beta}$ with parameter $m = m_*$ and threshold $\gamma = n \max(1, 8 d^3 \triangle/b_{m_*})$. If in addition $X \in \mathcal{X}_\eta$ and $\epsilon \in E_{4\eta}^k$, $\eta \geq 1$, then we have

$$\sup_{\beta \in \mathcal{W}_b^\rho} \mathbb{E} \|\hat{\beta} - \beta\|_\omega^2 \leq C \delta_n^* \eta \triangle^2 \left\{ \sigma^2 + \rho \Lambda \right\},$$

where $C$ is a positive constant.

Thus, we have proved that the rate $\delta_n^*$ is optimal and hence the estimator $\hat{\beta}$ is minimax optimal.

**Remark 2.4.** It is worth noting that as long as the sequence $b$ is increasing the condition on the threshold $\gamma$ given in Theorem 2.4 writes $\gamma = n$ for all sufficiently large $n$. Therefore, only the parameter $m$ has to be chosen data-driven in order to build an adaptive estimation procedure. On the other hand, under the assumptions of Theorem 2.4 the parametric rate cannot be obtained. To be more precise, in case that $\sum_j \omega_j/v_j < \infty$, the rate of the lower bound in Theorem 2.4 is given by $\delta_n^* = 1/n$. But in this case the condition $m_*/(\delta_n^*) \sup_{1 \leq j \leq m_*} \left\{ \omega_j/v_j \right\} = O(1)$ is not satisfied and hence we cannot apply Theorem 2.4. However, we conjecture that the proposed estimator attains also the parametric rate under a stronger set of assumptions as, for example, used by Johannes and Schenk (2008) in order to obtain rate optimal estimation of a linear functional of the slope parameter $\beta$. \qed
3 Mean squared prediction error and derivative estimation

In this section we will suppose that the slope function $\beta$ is an element of the Sobolev space of periodic functions $W_p$ for some $p > 0$ given by

$$W_p = \left\{ f \in H_s : f^{(j)}(0) = f^{(j)}(1), \quad j = 0, 1, \ldots, p - 1 \right\},$$

where $H_s := \{ f \in L^2[0,1] : f^{(p-1)} \text{ absolutely continuous}, f^{(p)} \in L^2[0,1] \}$ is a Sobolev space (c.f. Neubauer (1988a,b), Mair and Ruymgaart (1996) or Tsybakov (2004)). Let us first remark that if we consider the sequence of weights $(p_j)_{j \in \mathbb{N}}$ given by

$$b_0^p = 1 \quad \text{and} \quad b_{2j}^p = b_{2j+1}^p = j^{2p}, \quad j \in \mathbb{N},$$

and the trigonometric basis

$$\psi_1(t) = 1, \quad \psi_{2k}(t) = \sqrt{2} \cos(2 \pi k t), \quad \psi_{2k+1}(t) = \sqrt{2} \sin(2 \pi k t), \quad k = 1, 2, \ldots$$

then the Sobolev space of periodic functions is equivalently given by $W_{b^p}$ defined in (3.1). Therefore, let us denote by $W_{b^p} := W_{b^p}$, $\rho > 0$, an ellipsoid in the Sobolev space $W_p$.

**Mean squared prediction error.** We shall first measure the performance of the estimator by considering the mean prediction error (MPE), i.e., $\mathbb{E} \| \tilde{\beta} - \beta \|^2$. In this case, if $\Gamma$ satisfies a link condition, that is $\Gamma \in \mathcal{N}_v^d$, $d \geq 1$, for some weight sequence $v$ (see definition 2.3), then it follows by using the inequality of Heinz (1951) that the MPE is equivalent to the $W_v$-risk, that is $\mathbb{E} \| \tilde{\beta} - \beta \|^2_v$. To illustrate the previous results we assume in the following the sequence $(v_j)_{j \in \mathbb{N}}$ to be either polynomially decreasing, i.e., $v_1 = 1$ and $v_j = |j|^{-2a}$, $j \geq 2$, for some $a > 1/2$, or exponentially decreasing, i.e., $v_1 = 1$ and $v_j = \exp(-|j|^{2a})$, $j \geq 2$, for some $a > 0$. In the polynomial case easy calculus shows that a covariance operator $\Gamma \in \mathcal{N}_v^d$ acts like integrating $(2a)$-times and hence it is called finitely smoothing (c.f. Natterer (1984)). Furthermore, if the eigenfunctions of $\Gamma$ are $\{\psi_j\}$, then $\Gamma \in \mathcal{N}_v^d$ holds if and only if the eigenvalues $\lambda_j$ of $\Gamma$ satisfy $\lambda_j \asymp |j|^{-2a}$, which is the case considered, for example, in Crambes et al. (2008). On the other hand in the exponential case it can easily be seen that the link condition $\Gamma \in \mathcal{N}_v^d$ implies $\mathcal{R}(\Gamma) \subset W_p$ for all $p > 0$, therefore the operator $\Gamma$ is called infinitely smoothing (c.f. Main (1994)). Moreover, if the eigenfunctions of $\Gamma$ are $\{\psi_j\}$, then $\Gamma \in \mathcal{N}_v^d$ holds if and only if the eigenvalues $\lambda_j$ of $\Gamma$ satisfy $\lambda_j \asymp \exp(-j^{2a})$. To the best of our knowledge this case has not been considered yet in the literature. Since in both cases the basic regularity assumption 2.1 is satisfied, the lower bounds presented in the next assertion follow directly from Theorem 2.3. Here and subsequently, we write $a_n \lesssim b_n$ when there exists $C > 0$ such that $a_n \leq C b_n$ for all sufficiently large $n \in \mathbb{N}$ and $a_n \sim b_n$ when $a_n \sim b_n$ and $b_n \lesssim a_n$ simultaneously.

**PROPOSITION 3.1.** Under the assumptions of Theorem 2.3 we have for any estimator $\tilde{\beta}$

(i) in the polynomial case, i.e. $v_1 = 1$ and $v_j = |j|^{-2a}$, $j \geq 2$, for some $a > 1/2$, that

$$\sup_{\beta \in W_p} \{ \mathbb{E} \| \tilde{\beta} - \beta \|_1^2 \} \gtrsim n^{-(2p+2a)/(2p+2a+1)},$$

(ii) in the exponential case, i.e. $v_1 = 1$ and $v_j = \exp(-|j|^{2a})$, $j \geq 2$, for some $a > 0$, that

$$\sup_{\beta \in W_p} \{ \mathbb{E} \| \tilde{\beta} - \beta \|_1^2 \} \gtrsim n^{-1} \log n^{1/2a}.$$
On the other hand, if the dimension parameter \( m \) and the threshold \( \gamma \) in the definition of the estimator \( \hat{\beta} \) given in (1.6) are chosen appropriately, then, by applying Theorem 2.3, the rates of the lower bound given in the last assertion also provide, up to a constant, the upper bound of the risk of the estimator \( \hat{\beta} \), which is summarized in the next proposition.

**Proposition 3.2.** Under the assumptions of Theorem 2.3 consider the estimator \( \hat{\beta} \)

(i) in the polynomial case, i.e. \( v_1 = 1 \) and \( v_j = |j|^{-2a} \), \( j \geq 2 \), for some \( a > 1/2 \), with \( m \sim n^{1/(2p+2a+1)} \) and threshold \( \gamma = n \). If in addition \( k \geq 2 + 8/(2p + 2a - 1) \), then

\[
\sup_{\beta \in W_p^q} \{ \mathbb{E} \| \hat{\beta} - \beta \|_1^2 \} \lesssim n^{-(2p+2a)/(2p+2a+1)},
\]

(ii) in the exponential case, i.e. \( v_1 = 1 \) and \( v_j = \exp(-|j|^{2a}) \), \( j \geq 2 \), for some \( a > 0 \), with \( m \sim (\log n)^{1/(2a)} \) and threshold \( \gamma = n \). Then

\[
\sup_{\beta \in W_p^q} \{ \mathbb{E} \| \hat{\beta} - \beta \|_1^2 \} \lesssim n^{-1}(\log n)^{1/2a}.
\]

We have thus proved that these rates are optimal and the proposed estimator \( \hat{\beta} \) is minimax optimal in both cases. It is worth noting that replacing the condition \( \gamma = n \) by \( \gamma = cn \) with \( c > 0 \) appropriately chosen, Proposition 3.2 remains true when \( p = 0 \), that is to say when \( \beta \) is just supposed to be square integrable.

**Remark 3.1.** It is of interest to compare our results with those of Crambes et al. (2008) who measure the performance of their estimator in terms of squared prediction error. In their notations the decay of the eigenvalues of \( \Gamma \) is assumed to be of order \((|j|^{2q-1})\), i.e., \( q = a - 1/2 \). Furthermore they suppose the slope function to be \( m \) times continuously differentiable, i.e., \( m = p \). By using this parametrization we see that our results in the polynomial case imply the same rate of convergence in probability of the prediction error as it is presented in Crambes et al. (2008). However, from our general results follows a lower and an upper bound of the MPE not only in the polynomial case but also in the exponential case.

Furthermore, we shall emphasize the interesting influence of the parameters \( p \) and \( a \) characterizing the smoothness of \( \beta \) and the smoothing properties of \( \Gamma \), respectively. As we see from Propositions 3.1 and 3.2 in the polynomial case an increasing value of \( p \) leads to a faster optimal rate. In other words, as expected, a smoother regression function can be faster estimated. The situation in the exponential case is extremely different. It seems rather surprising that, contrary to the polynomial case, in the exponential case the optimal rate of convergence does not depend on the value of \( p \), however this dependence is clearly hidden in the constant. Furthermore, the parameter \( m \) does not even depend on the value of \( p \). Thereby, the proposed estimator is automatically adaptive, i.e., it does not involve an a-priori knowledge of the degree of smoothness of the slope function \( \beta \). However, the choice of the smoothing parameter depends on the value \( a \) specifying the decay of \{\( v_j \)\}. Note further that in both cases an increasing value of \( a \) leads to a faster optimal rate of convergence, i.e., we may call \( 1/a \) as degree of ill-posedness (c.f. Natterer (1984)).

**Estimation of the derivatives.** Let us consider now the estimation of derivatives of the slope function \( \beta \). It is well-known, that for any function \( g \) belonging to a Sobolev-ellipsoid \( W^p_s \) the Sobolev norm \( \| g \|_s \) for each \( 0 \leq s \leq p \) is equivalent to the \( L^2 \)-norm of the \( s \)-th weak derivative \( g^{(s)} \), i.e., \( \| g^{(s)} \| \). Thereby, the results in the previous Section imply again a lower bound as well as an upper bound of the \( L^2 \)-risk for the estimation of the \( s \)-th weak derivative of \( \beta \). In the following we consider again the two particular cases of
polynomial and exponential decreasing rates for the sequence of weights \((v_j)\). The next assertion summarizes then lower bounds for the \(L^2\)-risk for the estimation of the \(s\)-th weak derivative of \(\beta\) in both cases.

**Proposition 3.3.** Under the assumptions of Theorem 2.4, we have for any estimator \(\hat{\beta}(s)\)

(i) in the polynomial case, i.e. \(v_1 = 1\) and \(v_j = |j|^{-2a}, \ j \geq 2, \) for some \(a > 1/2, \) that

\[
\sup_{\beta \in \mathbb{W}_p^s} \{ \mathbb{E}[\|\hat{\beta}(s) - \beta(s)\|^2] \} \geq n^{-2(2p-2s)/(2p+2a+1)},
\]

(ii) in the exponential case, i.e. \(v_1 = 1\) and \(v_j = \exp(-|j|^{2a}), \ j \geq 2, \) for some \(a > 0, \) that

\[
\sup_{\beta \in \mathbb{W}_p^s} \{ \mathbb{E}[\|\hat{\beta}(s) - \beta(s)\|^2] \} \geq (\log n)^{-(p-s)/a}.
\]

On the other hand considering the estimator \(\hat{\beta}\) given in (1.6), we only have to calculate the \(s\)-th weak derivative of \(\beta\). Given the exponential basis, which is linked to the trigonometric basis by the relation \(\exp(2i\pi kt) = 2^{-1/2}(\psi_{2k}(t) + 1)\psi_{2k+1}(t)), \) for \(k \in \mathbb{Z}\) and \(t \in [0, 1], \) with \(i^2 = -1, \) we recall that for \(0 \leq s < p\) the \(s\)-th derivative \(\beta^{(s)}(t)\) of \(\beta\) in a weak sense satisfies

\[
\beta^{(s)}(t) = \sum_{k \in \mathbb{Z}} (2\pi k)^s \left( \int_0^1 \beta(u) \exp(-2\pi ku) \, du \right) \exp(2\pi kt).
\]

Given a dimension \(m \geq 1, \) we denote now by \([\hat{\Gamma}]_m\) the \((2m+1) \times (2m+1)\) matrix with generic elements \((\hat{\Gamma}\psi_j, \psi_j), -m \leq j, \ell \leq m\) and by \([\hat{g}]_m\) the \(2m+1\) vector with elements \((\hat{g}, \psi_j), -m \leq \ell \leq m.\) Furthermore for integer \(s\) define the diagonal matrix \(\nabla_m^{1/2}\) with entries \(\nabla_m^{1/2} := (2\pi j)^s, \ -m \leq j \leq m.\) Then we consider the estimator of \(\beta^{(s)}\) defined by

\[
\begin{align*}
\hat{\beta}^{(s)} := [\hat{\beta}^{(s)}]_m^t [\psi]_m \cdot \\
[\hat{\beta}^{(s)}]_m = \begin{cases} \\
\nabla_m^{s/2} [\hat{\Gamma}]_m^{-1} [\hat{g}]_m, & \text{if } [\hat{\Gamma}]_m \text{ is nonsingular and } ||[\hat{\Gamma}]_m^{-1}\||^2 \leq \gamma, \\
\n0, & \text{otherwise.}
\end{cases}
\end{align*}
\]

Furthermore, if the dimension parameter \(m\) and the threshold \(\gamma\) in the definition of the estimator \(\hat{\beta}^{(s)}\) given in (3.3) are chosen appropriately, then by applying Theorem 2.4 the rates of the lower bound given in the last assertion provide up to a constant again the upper bound of the \(L^2\)-risk of the estimator \(\hat{\beta}^{(s)},\) which is summarized in the next Proposition. We have thus proved that these rates are optimal and the proposed estimator \(\hat{\beta}^{(s)}\) is minimax optimal in both cases.

**Proposition 3.4.** Under the assumptions of Theorem 2.4, consider the estimator \(\hat{\beta}^{(s)}\)

(i) in the polynomial case, i.e. \(v_1 = 1\) and \(v_j = |j|^{-2a}, \ j \geq 2, \) for some \(a > 1/2, \) with \(m \sim n^{1/(2p+2a+1)}\) and threshold \(\gamma = n.\) If in addition \(k \geq 2 + 8/(2p + 2a - 1,\) then

\[
\sup_{\beta \in \mathbb{W}_p^s} \{ \mathbb{E}[\|\hat{\beta}^{(s)} - \beta^{(s)}\|^2] \} \lesssim n^{-2(2p-2s)/(2p+2a+1)},
\]

(ii) in the exponential case, i.e. \(v_1 = 1\) and \(v_j = \exp(-|j|^{2a}), \ j \geq 2, \) for some \(a > 0, \) with \(m \sim (\log n)^{1/(2a)}\) and threshold \(\gamma = n.\) Then

\[
\sup_{\beta \in \mathbb{W}_p^s} \{ \mathbb{E}[\|\hat{\beta}^{(s)} - \beta^{(s)}\|^2] \} \lesssim (\log n)^{-(p-s)/a}.
\]
Remark 3.2. It is worth noting that the $L^2$-risk in estimating the slope function $\beta$, i.e., $s = 0$, has been considered in Hall and Horowitz (2007) only in the polynomial case. In their notations the decrease of the eigenvalues of $\Gamma$ is of order $(|j|^{-\alpha})$, i.e., $\alpha = 2a$. Furthermore the Fourier coefficients of the slope function decay at least with rate $j^{-\beta}$, i.e., $\beta = p + 1/2$. By using this new parametrization we see that we recover the result of Hall and Horowitz (2007) in the polynomial case with $s = 0$, but without the additional assumption $\beta > \alpha/2 + 1$ or $\beta > \alpha - 1/2$.

Furthermore, we shall discuss again the interesting influence of the parameters $p$ and $a$. As we see from Propositions 3.3 and 3.4 in both cases an decreasing of the value of $a$ or an increasing of the value $p$ leads to a faster optimal rate of convergence. Hence, in opposite to the MPE by considering the $L^2$-risk the parameter $a$ describes in both cases the degree of ill-posedness. Furthermore, the estimation of higher derivatives of the slope function, i.e. by considering a larger value of $s$, is as usual only possible with a slower optimal rate. Finally, as for the MPE in the exponential case the parameter $m$ does not depend on the values of $p$ or $s$, hence the proposed estimator is automatically adaptive.

Remark 3.3. There is an interesting hidden issue in the parametrization we have chosen. Consider a classical indirect regression model with known operator given by $\Gamma$, i.e., $Y = [\Gamma\beta](U) + \epsilon$ where $U$ has a uniform distribution on $[0, 1]$ and $\epsilon$ is white noise (for details see e.g. Mair and Ruymgaart (1996)). If in addition the operator $\Gamma$ is finitely smoothing, i.e., $(v_j)$ is polynomially decreasing with $v_j = j^{-2a}$, $j \geq 2$, then given an $n$-sample of $Y$ the optimal rate of convergence of the $W_s$-risk of any estimator of $\beta$ is of order $n^{-2(p-s)/(2(p+2a)+1)}$, since $R(\Gamma) = W_{2a}$ (c.f. Mair and Ruymgaart (1996) or Chen and Reiß (2008)). However, we have shown that in a functional linear model even with estimated operator the optimal rate is of order $n^{-2(p-s)/(2(p+a)+1)}$. Thus comparing both rates we see that in a functional linear model the covariance operator $\Gamma$ has the degree of ill-posedness $a$ while the same operator has, in the indirect regression model, a degree of ill-posedness $(2a)$. In other words in a functional linear model we do not face the complexity of an inversion of $\Gamma$ but only of its square root $\Gamma^{1/2}$. This, roughly speaking, may be seen as a multiplication of the normal equation $YX = \langle \beta, X \rangle + X\epsilon$ by the inverse of $\Gamma^{1/2}$. Remarking that $\Gamma$ is also the covariance operator associated to the error term $\epsilon X$, the multiplication by the inverse of $\Gamma^{1/2}$ leads, roughly speaking, to white noise.

4 Concluding remarks and perspectives

We have proposed in this work a new kind of estimation procedures for the regression function and its derivatives in the functional linear model and proved they can attain optimal rates of convergence.

These estimators depend on two parameters which play the role of smoothing parameters, the dimension $m$ of the projection space and the threshold value $\gamma$. Building data driven rules that can permit to choose automatically the values of these parameters is certainly a topic that deserves further attention and one promising direction is to adapt the selection technique proposed in Efromovich and Koltchinskii (2001), Goldenshluger and Pereverzev (2000) and Tsybakov (2000).

Another point of interest is to extend the thresholding approach in order to consider different thresholding rules for different coordinates in the considered basis. This could lead for instance with wavelet basis to estimators that would adapt to sparseness as well as varying regularity of the regression function.
A Appendix: Proofs

A.1 Proofs of Section 2

We begin by defining and recalling notations to be used in the proofs of this section. Given $m > 0$, a Galerkin solution of $g = \Gamma \beta$ is denoted by $\tilde{\beta}^m \in \Psi_m$ (see equation [1.7]). Furthermore, we use the notations

$$
\tilde{\beta}^m := [\tilde{\beta}^m]_{m}[\psi]_{m}(\cdot) \quad \text{with} \quad [\tilde{\beta}^m]_{m} := [\beta^m]_{m} \{ \| [\tilde{\Gamma}]_{m}^{-1} \| \leq \gamma \},
$$

$$
[\tilde{\Gamma}]_{m} = \frac{1}{n} \sum_{i=1}^{n} [X_i]_{m}[X_i]_{m}^{T}, \quad [\tilde{X}_i]_{m} := [\Gamma]_{m}^{-1/2}[X_i]_{m}, \quad [\tilde{\Gamma}]_{m} := \frac{1}{n} \sum_{i=1}^{n} [\tilde{X}_i]_{m}[\tilde{X}_i]_{m}^{T},
$$

$$
[\Xi_n]_{m} := [\tilde{\Gamma}]_{m} - I_{m}, \quad [T_n]_{m} := \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \beta - \beta^m \rangle [X_i]_{m}, \quad [W_n]_{m} := \frac{\sigma}{n} \sum_{i=1}^{n} \epsilon_i [X_i]_{m},
$$

(A.1)

where $[\tilde{\beta}]_{m} - [\tilde{\Gamma}]_{m} [\beta^m]_{m} = [T_n]_{m} + [W_n]_{m}$ with $E[T_n]_{m} = [\Gamma(\beta - \beta^m)]_{m} = 0$ and $E[W_n]_{m} = 0$, $E[\tilde{\Gamma}]_{m} = [\Gamma]_{m}$, $[\tilde{\Gamma}]_{m} = [\Gamma]_{m}^{-1/2} [\Gamma]_{m}^{-1/2} = [\Gamma]_{m}$ and hence $E[\Xi_n]_{m} = 0$. Moreover, let us introduce the events

$$
\Omega := \{ \| [\tilde{\Gamma}]_{m}^{-1} \| \leq \gamma \}, \quad \Omega_{1/2} := \{ \| [\Xi_n]_{m} \| \leq 1/2 \}
$$

$$
\Omega^c := \{ \| [\tilde{\Gamma}]_{m}^{-1} \| > \gamma \} \quad \text{and} \quad \Omega^c_{1/2} = \{ \| [\Xi_n]_{m} \| > 1/2 \}. \quad (A.2)
$$

Observe that $\Omega_{1/2} \subset \Omega$ in case $\gamma \geq 2 \| [\Gamma]_{m}^{-1} \|$. Indeed, if $\| [\Xi_n]_{m} \| \leq 1/2$ then the identity $[\tilde{\Gamma}]_{m} = [\Gamma]_{m}^{-1/2} \{ I + [\Xi_n]_{m} \} [\Gamma]_{m}^{-1/2}$ implies by the usual Neumann series argument that $\| [\tilde{\Gamma}]_{m}^{-1} \| \leq 2 \| [\Gamma]_{m}^{-1} \|$. Thereby, if $\gamma \geq 2 \| [\Gamma]_{m}^{-1} \|$, then we have $\Omega_{1/2} \subset \Omega$. These results will be used below without further reference.

We shall prove in the end of this section the two technical Lemma [A.1] and [A.2] which are used in the following proofs.

Proof of the consistency.

**Proof of Proposition 2.1.** The proof is based on the decomposition

$$
E\| \tilde{\beta} - \beta \|_{\omega}^2 \leq 2 \{ E\| \tilde{\beta} - \tilde{\beta}^m \|_{\omega}^2 + E\| \tilde{\beta}^m - \beta \|_{\omega}^2 \}. \quad (A.3)
$$

Since $\gamma \geq 2 \| [\Gamma]_{m}^{-1} \|$ it follows that $\Omega^c \subset \Omega^c_{1/2}$ and hence

$$
E\| \tilde{\beta}^m - \beta \|_{\omega}^2 \leq 2 \{ \| \beta^m - \beta \|_{\omega}^2 + \| \beta^m \|_{\omega}^2 \} P(\Omega^c_{1/2}), \quad (A.4)
$$

On the other hand we show below for some constant $C > 0$ the following bound

$$
E\| \tilde{\beta} - \tilde{\beta}^m \|_{\omega}^2 \leq C \cdot \| [\text{Diag}(\omega)]_{m}^{1/2} [\Gamma]_{m}^{-1/2} \| (m/n) \eta \left\{ \sigma^2 + \| \beta - \beta^m \|_{\omega}^2 \right\} \left\{ 1 + \gamma^2 m^2/n \eta^{-1/2} (P(\Omega^c_{1/2}))^{1/2} \right\}, \quad (A.5)
$$

where by applying Markov’s inequality [A.12] in Lemma [A.1] implies $P(\Omega^c_{1/2}) \leq C \eta m^2/n$ for some $C > 0$. Moreover, $\| [\Gamma]_{m} \|^2 \leq \| \Gamma \|^2$ and $\| [\text{Diag}(\omega)]_{m}^{1/2} [\Gamma]_{m}^{-1/2} \|^2 \leq \gamma \sup_{1 \leq j \leq m} \{ \omega_j \}$
since $\gamma \geq 2\|\Gamma\|_m^{-1/2}^2$, which by combination of (A.4) and (A.5) leads to the estimate
\[
E\|\hat{\beta} - \beta\|_\omega^2 \leq C \left\{ \|\beta - \beta\|_\omega^2 + \|\beta - \beta\|_\omega^2 (m^2/n) \eta 
+ \gamma \sup_{1 \leq j \leq m} \{ \omega_j \} (m/n) \eta \{ \sigma^2 + \|\beta - \beta\|_\omega^2 \|X\|^2 \} \{1 + \gamma^2 (m^3/n^1+1/2) \|\Gamma\|_\omega^2 \} \right\} (A.6)
\]for some $C > 0$. Furthermore, for each $\beta \in \mathcal{W}_\omega$, we have $\|\beta - \beta\|_\omega = o(1)$ as $m \to \infty$, which can be realized as follows. Since $\|\Pi^\perp_m \beta\|_\omega = o(1)$ and $\|\Pi^\perp_m \beta\|_\omega = o(1)$ as $m \to \infty$ by using Lebesgue’s dominated convergence theorem, the assertion follows from the identity $[\Pi_m \beta - \beta]^m = -[\Gamma_m]^{-1} [\Pi_m \beta]^m$ by using that $\|\Pi_m \beta - \beta\|_\omega \leq \|\Pi_m \beta\|_\omega \sup_m \|\Gamma_m^{-1} \Pi_m \Gamma_m^m\|_\omega = O(\|\Pi_m \beta\|_\omega)$. Consequently, the conditions on $m$ and $\gamma$ ensure the convergence to zero as $n \to \infty$ of the bound given in (A.6), which proves the result.

Proof of (A.6). From the identity $[\beta]_m - [\Gamma_m]^{-1} [\beta]^m = [T_n]_m + [W_n]_m$ it follows that
\[
E\|\beta - \beta|^m_\omega^2 = E\|\text{Diag} (\omega)^{1/2} \{ [\Gamma]^{-1}_m + [\Pi]^{-1}_m ([\Gamma]_m - [\Gamma]_m) [\Gamma]^{-1}_m \} \{ [T_n]_m + [W_n]_m \|^2 \Omega.
\]Since $2\|\Gamma]^{-1}_m \|| \leq \gamma$ we have $\Omega_{1/2} \subset \Omega$, and hence by using $\|\text{Diag} (\omega)^{1/2} \Omega \| \leq \gamma^2$ we obtain
\[
E\|\beta - \beta|^m_\omega^2 \leq 3 \|\text{Diag} (\omega)^{1/2} \{ [\Gamma]^{-1}_m + [\Pi]^{-1}_m ([\Gamma]_m - [\Gamma]_m) [\Gamma]^{-1}_m \} \| \leq \gamma^2 \|\{ [T_n]_m + [W_n]_m \|^2 \Omega_{1/2}.
\]

From (A.10)-(A.12) in Lemma A.2 together with $\|\{ I + [\Sigma]_m \}^{-1} \||\Sigma]_m\|_\Omega_{1/2} \leq 1$ follows then (A.5), which completes the proof.

Proof of Corollary 2.2. The link condition $\Gamma \in \mathcal{N}^d_\omega$ implies $2\|\Gamma]^{-1}_m \|| \leq 8 \delta^3/\nu_m = \gamma$, $\|\text{Diag} (\omega)^{1/2} \{ [\Gamma]^{-1}_m \| \leq 4 \delta^3 \sup_{1 \leq j \leq m} \omega_j/\nu_j \}$ and $\|\Gamma\|_m^2 \leq d^2$ by using the estimates (A.16), (A.17) and (A.18) in Lemma A.3 respectively. Therefore, by combination of (A.4) and (A.5) in the proof of Proposition 2.1 we obtain
\[
E\|\hat{\beta} - \beta\|_\omega^2 \leq C \left\{ \|\beta - \beta\|_\omega^2 + \|\beta - \beta\|_\omega^2 (m^2/n) \eta \sup_{1 \leq j \leq m} \{ \omega_j/\nu_j \} (m/n) \right.
\eta \{ \sigma^2 + \|\beta - \beta\|_\omega^2 \|X\|^2 \} \left. \{1 + \gamma^2 (m^3/n^1+1/2) \|\Gamma\|_\omega^2 \} \right\} (A.7)
\]for some $C > 0$. By using the identity $[\Pi_m \beta - \beta]^m = -[\Gamma_m]^{-1} [\Pi_m \beta]^m$ and the estimate (A.23) in the proof of Lemma A.3 with $b \equiv \omega$ the link condition $\Gamma \in \mathcal{N}^d_\omega$ implies further that $\|\Gamma_m^{-1} \Pi_m \Gamma_m^m\|_\omega^2 = \sup_{\|\beta\|_\omega = 1} \|\Pi_m \beta - \beta\|_\omega^2 \leq 2(1 + d^2)$ for all $m \in \mathbb{N}$. Therefore we have $\|\beta - \beta\|_\omega = o(1)$ as $m \to \infty$ for each $\beta \in \mathcal{W}_\omega$. Consequently, the conditions on $m$ and $\gamma$ ensure the convergence to zero as $n \to \infty$ of the bound given in (A.7), which proves the result.

Proof of the lower bound.

Proof of Theorem 2.3. Let $X_i$, $i \in \mathbb{N}$, be i.i.d. copies of $X$ with associated covariance operator $\Gamma$ belonging to $\mathcal{N}^d_\omega$. Then for each $j$, $[X]_j$ is centered and has variance $E[X_j]^2 = \langle \Gamma \psi_j, \psi_j \rangle \leq \nu_j \delta$. This result will be used below without further reference. Consider independent error terms $\epsilon_i \sim \mathcal{N}(0,1)$, $i \in \mathbb{N}$, which are independent of the random
functions \( \{X_i\} \). Let \( \theta \in \{-1,1\}^{m_*} \), where \( m_* := m_*(n) \in \mathbb{N} \) satisfies (2.4) for some \( \Delta \geq 1 \). Define a \( m_* \)-vector \( u \) of coefficients \( u_j \) satisfying \( \text{[A.14]} \) in Lemma \( \text{[A.2]} \). For each \( \theta \) we consider a slope function \( \beta^\theta := \sum_{j=1}^{m_*} \theta_j u_j \) by using \( \text{[A.13]} \) in Lemma \( \text{[A.2]} \). Consequently, for each \( \theta \) the random variables \( (Y_i, X_i) \) with \( Y_i := \int_0^1 \beta^\theta(s) X_i(s) ds + \sigma\epsilon_i, \) \( i = 1, \ldots, n \), form a sample of the model (1.1) and we denote its joint distribution by \( P_{\theta} \).

Furthermore, for \( j = 1, \ldots, m_* \) and each \( \theta \) we introduce \( \theta^{(j)} \) by \( \theta^{(j)} = \theta \) for \( j \neq l \) and \( \theta^{(j)} = -\theta_j \). As in case of \( P_{\theta} \) the conditional distribution of \( Y_i \) given \( X_i \) is Gaussian with mean \( \sum_{j=1}^{m_*} \theta_j u_j |X_i|_j \) and variance \( \sigma^2 \) it is easily seen that the log-likelihood of \( P_{\theta^{(j)}} \) w.r.t. \( P_{\theta} \) is given by

\[
\log \left( \frac{dP_{\theta^{(j)}}}{dP_{\theta}} \right) = -\frac{1}{\sigma^2} \sum_{j=1}^{n} \left\{ Y_j - \sum_{l=1}^{m_*} \theta_l u_l |X_i|_l \right\} \theta_j u_j |X_i|_j - \frac{2}{\sigma^2} \sum_{i=1}^{n} u_j^2 |X_i|_j^2
\]

and its expectation w.r.t. \( P_{\theta} \) satisfies \( \mathbb{E}_{P_{\theta}} \left[ \log \left( \frac{dP_{\theta^{(j)}}}{dP_{\theta}} \right) \right] \geq -2nd u_j^2 v_j / \sigma^2 \). In terms of Kullback-Leibler divergence this means \( KL(P_{\theta^{(j)}}, P_{\theta}) \leq 2nd u_j^2 v_j / \sigma^2 \). Since the Hellinger distance \( H(P_{\theta^{(j)}}, P_{\theta}) \) satisfies \( H^2(P_{\theta^{(j)}}, P_{\theta}) \leq KL(P_{\theta^{(j)}}, P_{\theta}) \) it follows from (A.15) in Lemma \( \text{[A.2]} \) that

\[
H^2(P_{\theta^{(j)}}, P_{\theta}) \leq \frac{2nd}{\sigma^2} \cdot u_j^2 \cdot v_j \leq 1, \quad j = 1, \ldots, m_*.
\] (A.8)

Consider the Hellinger affinity \( \rho(P_{\theta^{(j)}}, P_{\theta}) = \int \sqrt{dP_{\theta^{(j)}} dP_{\theta}} \), then we obtain for any estimator \( \tilde{\beta} \) of \( \beta \) that

\[
\rho(P_{\theta^{(j)}}, P_{\theta}) \leq \int \frac{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2}{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2} \sqrt{dP_{\theta^{(j)}}} dP_{\theta} + \int \frac{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2}{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2} \sqrt{dP_{\theta^{(j)}}} dP_{\theta}
\]

\[
\leq \left( \int \frac{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2}{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2} dP_{\theta^{(j)}} \right)^{1/2} + \left( \int \frac{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2}{\left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2} dP_{\theta} \right)^{1/2}.
\] (A.9)

Due to the identity \( \rho(P_{\theta^{(j)}}, P_{\theta}) = 1 - \frac{1}{2} H^2(P_{\theta^{(j)}}, P_{\theta}) \) combining (A.8) with (A.9) yields

\[
\left\{ \mathbb{E}_{\theta^{(j)}} \left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2 + \mathbb{E}_{\theta} \left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2 \right\} \geq \frac{1}{2} u_j^2, \quad j = 1, \ldots, m_*.
\]

From this we conclude for each estimator \( \tilde{\beta} \) that

\[
\sup_{\beta \in \mathbb{W}_\theta^\omega} \mathbb{E} \left| \tilde{\beta} - \beta \right|^2 \geq \sup_{\theta \in \{-1,1\}^{m_*}} \mathbb{E}_{\theta} \left| \tilde{\beta} - \beta^\theta \right|^2
\]

\[
\geq \frac{1}{2m_*} \sum_{\theta \in \{-1,1\}^{m_*}} \sum_{j=1}^{m_*} \omega_j \mathbb{E}_{\theta} \left| \langle \beta - \beta^\theta, \psi_j \rangle \right|^2
\]

\[
= \frac{1}{2m_*} \sum_{\theta \in \{-1,1\}^{m_*}} \sum_{j=1}^{m_*} \omega_j \left( \frac{1}{2} \left( \mathbb{E}_{\theta^{(j)}} \left| \langle \beta - \beta^{\theta}, \psi_j \rangle \right|^2 + \mathbb{E}_{\theta^{(j)}} \left| \langle \beta - \beta^{\theta^{(j)}}, \psi_j \rangle \right|^2 \right) \right)
\]

\[
\geq \frac{1}{4} \sum_{j=1}^{m_*} u_j^2 \cdot \omega_j \geq \frac{1}{4} \cdot \min \left\{ \frac{\sigma^2}{2d}, \frac{1}{\Delta} \right\} \cdot \delta_n^\omega,
\]

where the last inequality follows from (A.15) in Lemma \( \text{[A.2]} \) which completes the proof. \( \square \)
Proof of Theorem 2.4. Our proof starts with the observation that the link condition $\Gamma \in \mathcal{N}_v^d$ implies $2\|\Gamma^{-1}\| \leq 8d^3/v_m$, $\|\text{Diag}(\omega)\|^{1/2} \|\Gamma^{-1/2}\|^{1/2} \leq 4d^3\sup_{1 \leq j \leq m}\{\omega_j/v_j\}$ and $\|\Gamma\|^{2} \leq d^2$ by using the estimates (A.16), (A.17) and (A.18) in Lemma A.3 respectively. Moreover, for all $X \in \mathcal{X}_{4k}^\kappa$ by applying Markov’s inequality (A.12) in Lemma A.1 we have $P(\Omega_{1/2}^{k}) \leq C\eta m^{2k}/n^{k}$ for some $C > 0$. Furthermore, by using the definition of $m_*$ the condition $m = m_*$ implies $1/v_m \leq n\Delta/b_m$, and hence $\gamma = n\max(1,8d^3\Delta/b_m) \geq 2\|\Gamma^{-1}\|$. Therefore, from (A.4) and (A.5) in the proof of Proposition 2.1 follows

$$
E\|\tilde{\beta} - \beta\|_\omega^2 \leq C \left\{ \|\beta^{m*} - \beta\|_\omega^2 + \|\beta^{m*}\|_\omega^2 (m_*^{2k}/n^k) \eta + d^3 \sup_{1 \leq j \leq m_*}\{\omega_j/v_j\} (m_*/n) \right. \\
\left. \eta \{\sigma^2 + \|\beta - \beta^{m*}\|^2 \mathbb{E}\|X\|^2 \} \{1 + m_*^{2+k}/(n^k/2-1)d^2 \Delta^2\}\right\}
$$

for some $C > 0$. Consequently, the definition of $\delta^*_n$ by using (A.19) in Lemma A.3 i.e., $\|\beta - \beta^{m*}\|_\omega^2 \leq 10d^3 \rho \delta^*_n$, and $\mathbb{E}\|X\|^2 \leq d\Lambda$, implies

$$
E\|\tilde{\beta} - \beta\|_\omega^2 \leq C \delta^*_n \eta d^{16} \Delta^2 \{\sigma^2 + \rho \Lambda\} \\
\left\{ 1 + m_*^{2k}/(\delta^*_n n^k) + m_*/(\delta^*_n n) \sup_{1 \leq j \leq m_*}\{\omega_j/v_j\} \right\} \left\{ 1 + m_*^{2+k}/(n^k/2-1)\right\}
$$

Thereby, the result follows from the condition (2.6) which ensures that the factors in braces are bounded as $n \to \infty$, which completes the proof.

Technical assertions.
The following two lemma gather technical results used in the proof of Proposition 2.1 Theorem 2.3 and Theorem 2.4.

**Lemma A.1.** Suppose $X \in \mathcal{X}_{4k}^\kappa$ and $\epsilon \in \mathcal{E}_{4k}^\kappa$, $k \in \mathbb{N}$. Then for some constant $C > 0$ only depending on $k$ we have

$$
\mathbb{E}\|\Gamma^{-1/2}W_{n,m}\|^{2k} \leq C \cdot \frac{m^k}{n^k} \cdot \sigma^{2k} \cdot \eta, \quad (A.10)
$$

$$
\mathbb{E}\|\Gamma^{-1/2}T_{n,m}\|^{2k} \leq C \cdot \frac{m^k}{n^k} \cdot \|\beta - \beta^{m}\|^{2k} \cdot (\mathbb{E}\|X\|^2)^k \cdot \eta, \quad (A.11)
$$

$$
\mathbb{E}\|\Xi_{n,m}\|^{2k} \leq C \cdot \eta \cdot \frac{m^{2k}}{n^k}, \quad (A.12)
$$

$$
\mathbb{E}\|\{\Gamma - \hat{\Gamma}\|^{1/2}\}^{1/2}\|^{2k} \leq C \cdot \eta \cdot \frac{m^{2k}}{n^k} \cdot (\mathbb{E}\|X\|^2)^k \quad (A.13)
$$

**Proof.** Let $\tilde{W} := [\Gamma^{-1/2}W_{n,m}]$, then $\mathbb{E}\|\Gamma^{-1/2}W_{n,m}\|^{2k} \leq m^{k-1} \sum_{j=1}^{m} \mathbb{E}\tilde{W}_j^{2k}$, where $\tilde{W}_j = (1/n) \sum_{i=1}^{n} \sigma_i [\tilde{X}_i]_j$. The random variables $(\epsilon_i [\tilde{X}_i]_j)_j$, $1 \leq i \leq n$, are independent and identically distributed (i.i.d.) with mean zero. Since $X \in \mathcal{X}_{4k}^\kappa$ and $\epsilon \in \mathcal{E}_{4k}^\kappa$, (A.10) follows from Theorem 2.10 in Petrov (1995), that is, $\mathbb{E}\tilde{W}_j^{2k} \leq Cn^{-k}\sigma^{2k}\mathbb{E}[\epsilon [\tilde{X}_j]_j]^{2k} \leq Cn^{-k}\sigma^{2k}\eta$ for some constant $C > 0$ only depending on $k$.

**Proof of (A.11).** Due to $\mathbb{E}(\beta - \beta m, X)[X]_m = [\Gamma(\beta - \beta m)]_m = 0$, i.e., the random variables $(\beta - \beta^m, X_i)[X_i]_m$, $1 \leq i \leq n$, are i.i.d. with mean zero. Furthermore, we claim
that $X \in X_\eta^{4k}$ implies $\mathbb{E} \| \langle \beta - \beta^m, X \rangle \|_{2k}^2 \leq C \cdot \eta \cdot \| \beta - \beta^m \|_{2k}^2 (\mathbb{E} \| X \|)^2$, for each $j \in \mathbb{N}$. Then the estimate (A.11) follows in analogy to (A.10). Indeed, we have $\mathbb{E} \| \langle \beta, j \rangle \|_{4k}^2 \leq \| \beta - \beta^m \|_{4k} \sum_{j_1} \cdots \sum_{j_{2k}} \mathbb{E} \prod_{l=1}^{2k} \| \langle X, j \rangle \|_{1/2}^2$ which imply together the assertion by using the Cauchy-Schwarz inequality.

Proof of (A.12). From the identity $(\Xi,n,m)_{j,l} = (1/n) \sum_{i=1}^{n} \{ \langle X, j \rangle - \delta_{jl} \}$ with $\delta_{jl} = 1$ if $j = l$ and zero otherwise, we conclude $\mathbb{E}(\Xi,n,m)_{j,l}^2 \leq C' n^{-k} \mathbb{E} \| X \| X \| X \| - \delta_{jl} \|^2$. Thus $X \in X_\eta^{4k}$ implies $\mathbb{E}(\Xi,n,m)_{j,l}^2 \leq m^{2(k-1)} \sum_{j,l} \mathbb{E}(\Xi,n,m)_{j,l}^2 \leq C m^{2k} n^{-k} \eta$. The estimate (A.13) follows by using the identity $\{ \langle \Gamma \rangle_m - \langle \Gamma \rangle_n \} \langle \Gamma \rangle_m^{-1/2} = \langle \Gamma \rangle_m^{-1/2} \Xi_n,m$ from (A.12), which completes the proof. \hfill \Box

**Lemma A.2.** Let $m_* \in \mathbb{N}$ and $\delta_n^* \in \Xi$ be chosen such that (2.4) is satisfied for some $\Delta \geq 1$. Consider a (infinite) vector $u$ with components $u_j$ satisfying

$$u_j^2 = \frac{\zeta}{n \cdot u_j}, \quad j \in \mathbb{N}, \quad \text{with} \quad \zeta := \min \left\{ \frac{\sigma^2}{(2d)} \right\} ,$$

then under Assumption 2.1 we have for all $j \in \mathbb{N}$

$$\frac{2nd}{\sigma^2} u_j^2 \leq 1, \quad \sum_{j=1}^{m_*} u_j^2 \beta_j \leq \rho, \quad \text{and} \quad \sum_{j=1}^{m_*} u_j^2 \omega_j \geq \min \left\{ \frac{\sigma^2}{(2d)} \right\} \frac{\rho}{\Delta} \frac{\delta_n^*}{\Delta} .$$

**Proof.** The first inequality in (A.15) follows trivially by using the definition of $\zeta$, while the definition of $m_*$ given in (2.4) together with Assumption 2.1, i.e., $(b_j/\omega_j)$ is non-decreasing, implies the second, i.e., $\sum_{j=1}^{m_*} u_j^2 \beta_j \leq \zeta b_{m_*} / \omega_{m_*} \sum_{j=1}^{m_*} \omega_j / (n v_j) \leq \zeta \Delta \leq \rho$. To deduce the third estimate from the definition of $m_*$ and $\delta_n^*$ observe that $\sum_{j=1}^{m_*} u_j^2 \omega_j = \delta_n^* \zeta b_{m_*} / \omega_{m_*} \sum_{j=1}^{m_*} \omega_j / (n v_j) \geq \delta_n^* \zeta / \Delta$, which proves the lemma. \hfill \Box

**Lemma A.3.** Suppose the sequences $b$, $\omega$ and $v$ satisfy Assumption 2.1. Let $\Gamma \in \mathcal{N}_\nu^d$. Then

$$\sup_{m \in \mathbb{N}} \left\{ u_m \| \langle \Gamma \rangle^m_\nu^{-1/2} \|_2^2 \right\} \leq \left\{ (2d^2 (2d^4 + 3))^{1/2} \right\} 4d^3 ,$$

$$\sup_{m \in \mathbb{N}} \left\{ \| \text{Diag}(v) \|^{1/2}_m \| \langle \Gamma \rangle^m_\nu^{-1/2} \|_2^2 \right\} \leq \left\{ (2d^2 (2d^4 + 3))^{1/2} \right\} 4d^3 ,$$

$$\sup_{m \in \mathbb{N}} \left\{ \| \text{Diag}(v) \|^{1/2}_m \| \langle \Gamma \rangle^m_\nu^{-1/2} \|_2^2 \right\} \leq d .$$

If in addition $\beta^m$ denotes a Galerkin solution of $g = \Gamma \beta$ with $\beta \in \mathcal{W}_\nu^{0}$, then

$$\sup_{m \in \mathbb{N}} \left\{ b_{m} / \omega_{m} \| \beta - \beta^m \|_\omega^2 \right\} \leq 2(2d^4 + 3) \rho \leq 10d^4 .$$

**Proof.** We start our proof with the observation that the link condition $\Gamma \in \mathcal{N}_\nu^d$ implies that $\Gamma$ is strictly positive and that for all $|s| \leq 1$ by using the inequality of Heinz (1951)

$$d^{-2|s|} \| f \|_{\nu^{2s}}^2 \leq \| \Gamma^s f \|_2^2 \leq d^{2|s|} \| f \|_{\nu^{2s}}^2 .$$

(20)
Consider $g \in \Psi_m$. Then (A.20) implies $\beta := \Gamma^{-1} g \in L^2[0, 1]$ by using that $\|g\|_{v^2} = \|[\text{Diag}(v)]^{-1}_{m}[g]_m\| < \infty$. Furthermore, $\beta^m = [\Gamma]^{-1}_m [g]_m$ is the unique Galerkin solution of (A.24). By using successively the first inequality of (A.20), the Galerkin condition (A.23) and the second inequality of (A.20), we obtain

$$
\|\beta - \beta^m\|_{v^2}^2 \leq d^2 \|\Gamma(\beta - \beta^m)\|^2 \leq d^2 \|\Gamma(\beta - \Pi_m \beta)\|_{v^2}^2 \leq d^4 \|\beta - \Pi_m \beta\|_{v^2}^2 \tag{A.21}
$$

Since $(v_j)$ is monotonically decreasing it follows $\|\beta - \Pi_m \beta\|_{v^2} \leq v_m^2 \|\beta\|^2$ and, hence by using (A.20) with $s = -1$ we have $\|\beta - \Pi_m \beta\|_{v^2} \leq d^2 v_m^2 \|g\|_{v^2}^2$. Combining the last estimate with (A.21) we obtain

$$
\|\beta^m - \Pi_m \beta\|_{v^2} \leq 2\{\|\beta - \beta^m\|_{v^2}^2 + \|\beta - \Pi_m \beta\|_{v^2}^2\} \leq 2d^2 (d^4 + 1) v_m^2 \|g\|_{v^2}^2
$$

which together with $\|f\|_{v^2} \leq v_m^{-2} \|f\|_{v^2}^2$ for all $f \in \Psi_m$ leads to

$$
\|\beta^m - \Pi_m \beta\|_{v^2} \leq v_m^{-2} \|\beta^m - \Pi_m \beta\|_{v^2} \leq 2d^4 (d^4 + 1) \|g\|_{v^2}^2.
$$

By using the last estimate together with $\|g\|_{v^2} = \|[\text{Diag}(v)]^{-1}_{m}[g]_m\|$ we conclude that

$$
\|\Gamma]^{-1}_{m}[g]_m\|^2 = \|\beta^m\|^2 \leq 2\{\|\beta^m - \Pi_m \beta\|^2 + \|\Pi_m \beta\|^2\} \leq 2d^4 (2d^4 + 3) \|[\text{Diag}(v)]^{-1}_{m}[g]_m\|^2, \ \forall g \in \Psi_m. \tag{A.22}
$$

Then, from (A.22) follows by using the inequality of Heinz (1951) for all $g \in \Psi_m$

$$
\|\Gamma]^{-1/2}_{m}[g]_m\|^2 \leq \{2d^2(2d^4 + 3)\}^{1/2} \|[\text{Diag}(v)]^{-1/2}_{m}[g]_m\|^2,
$$

which implies together with $\|[\text{Diag}(v)]^{-1}_{m}\| = v_m^{-1}$ the estimate (A.16), and furthermore by replacing $[g]_m$ by $[\text{Diag}(v)]^{-1/2}_{m}[g]_m$ the estimate (A.17), that is,

$$
\|\Gamma]^{-1/2}_{m}[\text{Diag}(v)]^{-1/2}_{m}[g]_m\|^2 \leq \{2d^2(2d^4 + 3)\}^{1/2} \|[g]_m\|^2, \ \forall g \in \Psi_m.
$$

Proof of (A.18). By using the second inequality of (A.20) together with $\|\Pi_m\| = 1$ we obtain

$$
\|\Gamma]^{-1}_{m}[g]_m\|^2 = \|\Pi_m \Gamma g\|^2 \leq \|\Gamma g\|^2 \leq d^2 \|g\|_{v^2}^2 = d^2 \|[\text{Diag}(v)]^{-1}_{m}[g]_m\|^2, \ \forall g \in \Psi_m
$$

and hence the inequality of Heinz (1951) implies

$$
\|\Gamma]^{-1/2}_{m}[g]_m\|^2 \leq d \|[\text{Diag}(v)]^{-1/2}_{m}[g]_m\|^2, \ \forall g \in \Psi_m.
$$

Thereby, (A.18) follows by replacing $[g]_m$ by $[\text{Diag}(v)]^{-1/2}_{m}[g]_m$, that is,

$$
\|\Gamma]^{-1/2}_{m}[\text{Diag}(v)]^{-1/2}_{m}[g]_m\|^2 \leq d \|[g]_m\|^2, \ \forall g \in \Psi_m.
$$

Proof of (A.19). Let $\beta \in \mathcal{W}^b_v$. Consider the decomposition

$$
\|\beta - \beta^m\|_{v^2} \leq 2\{\|\beta - \Pi_m \beta\|_{v^2}^2 + \|\Pi_m \beta - \beta^m\|_{v^2}^2\}.
$$

Since $(\omega_j/b_j)$ is non-increasing it follows $\|\beta - \Pi_m \beta\|_{v^2} \leq \omega_m/b_m \|\beta\|_{b}^2$, while we show below

$$
\|\Pi_m \beta - \beta^m\|_{v^2} \leq 2(1 + d^2) \omega_m/b_m \|\beta\|_{b}^2. \tag{A.23}
$$

Consequently, by combination of these two bounds the condition $\beta \in \mathcal{W}^b_v$, i.e., $\|\beta\|_{b}^2 \leq \rho$, implies (A.19). From (A.21) follows $\|\beta - \beta^m\|_{v^2} \leq d^4 \|\beta - \Pi_m \beta\|_{v^2}^2 \leq d^4 v_m^2/b_m \|\beta\|_{b}^2$ because $(v_j/b_j)$ is non-increasing, and hence,

$$
\|\Pi_m \beta - \beta^m\|_{v^2} \leq 2\{\|\beta - \beta^m\|_{v^2}^2 + \|\beta - \Pi_m \beta\|_{v^2}^2\} \leq 2(1 + d^4) v_m^2/b_m \|\beta\|_{b}^2. \tag{A.24}
$$

Furthermore, $\|\Pi_m \beta - \beta^m\|_{v^2} \leq \omega_m v_m^2 \|\Pi_m \beta - \beta^m\|_{v^2}^2$ since $(\omega_j/v_j^2)$ is non-decreasing. The last estimate and (A.24) imply now together (A.23), which completes the proof.
A.2 Proofs of Section 3

The mean prediction error.

**Proof of Proposition 3.1.** Since $\Gamma \in \mathcal{N}^d_v$, $d \geq 1$, it follows by using the inequality of Heinz (1951) that $E\|\beta - \beta^\dagger\|^2_1 \asymp_d E\|\beta - \beta^\dagger\|^2_\beta$. Therefore, we can apply the general results by considering the $W_\omega$-risk with $\omega = \nu$ as a measure of the performance of an estimator of $\beta$. Furthermore, in case (i) the definition of $b_j^\nu$ and $v_j$ imply together $(b_m^n/\omega_m) \sum_{j=1}^{m_n} \omega_j/v_j = m^{2a+2p+1}$. It follows that the condition on $m_n$ and $\delta_n^s$ given in (2.4) of Theorem 2.3 can be rewritten as $m_n \sim n^{1/(2p + 2a + 1)}$ and $\delta_n^s \sim n^{-(2p + 2a)/(2p + 2a + 1)}$. On the other hand, in case (ii) $(b_m^n/\omega_m) \sum_{j=1}^{m_n} \omega_j/v_j = m^{2p+1} \exp(m_2^a)$ implies that the condition on $m_n$ and $\delta_n^s$ writes $m_n \sim (\log n)^{1/(2a)}$ and $\delta_n^s \sim n^{-1}(\log n)^{1/(2a)}$. Consequently, the lower bounds in Proposition 3.1 follow by applying Theorem 2.3.

**Proof of Proposition 3.2.** Note, that for sufficiently large $n$ the condition on $\gamma$ in Theorem 2.4 writes $\gamma = n$ because $(b^p_j)$ is increasing. Furthermore, it is easily seen that the additional condition (2.6) is satisfied in the exponential case and for all $k \geq 2 + 8/(2p + 2a - 1)$ also in the polynomial case. Finally, since in both cases the condition on $m$ ensures that $m \sim m_n$ (see the proof of Proposition 3.1) the result follows from Theorem 2.3.

The estimation of derivatives.

**Proof of Proposition 3.3.** Since for each $0 \leq s \leq p$ we have $E\|\beta^{(s)} - \beta\|^2 \sim E\|\beta - \beta\|^2_\beta$, we can apply again the general results by considering the $W_\omega$-risk with $\omega = b^s$. In case (i) the well-known approximation $\sum_{j=1}^{m_n} \omega_j/v_j \sim m^{r+1}$ for $r > 0$ together with the definition of $b_j^p$ and $v_j$ imply $(b_m^p/\omega_m) \sum_{j=1}^{m_n} \omega_j/v_j \sim m^{2a+2p+1}$. It follows that the condition on $m_n$ and $\delta_n^s$ given in (2.4) of Theorem 2.3 writes $m_n \sim n^{1/(2p + 2a + 1)}$ and $\delta_n^s \sim n^{-(2p + 2a)/(2p + 2a + 1)}$. On the other hand, in case (ii) by applying Laplace’s Method (c.f. chapter 3.7 in Olver (1974)) the definition of $b_j$ and $v_j$ imply $(b_m^p/\omega_m) \sum_{j=1}^{m_n} \omega_j/v_j \sim m^{2p} \exp(m_2^a)$ implies that the condition on $m_n$ and $\delta_n^s$ can be rewritten as $m_n \sim (\log n)^{1/(2a)}$ and $\delta_n^s \sim n^{-1}(\log n)^{1/(2a)}$. Consequently, the lower bounds in Proposition 3.1 follow by applying Theorem 2.3.

**Proof of Proposition 3.4.** The proof follows in analogy to the proof of Proposition 3.2 and we omit the details.

References

F. Bauer, S. V. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23:52–72, 2007.

P. Besse, H. Cardot, and D. Stephenson. Autoregressive forecasting of some functional climatic variations. *Scand. J. of Statist.*, 27:673–687, 2000.

D. Bosq. *Linear Processes in Function Spaces.*, volume 149 of Lecture Notes in Statistics. Springer-Verlag, 2000.

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics & Probability Letters*, 45:11–22, 1999.
H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591, 2003.

H. Cardot, A. Mas, and P. Sarda. CLT in functional linear regression models. *Prob. Theory and Rel. Fields*, 18:325–361, 2007.

X. Chen and M. Reiß. On rate optimality for ill-posed inverse problems in econometrics. Technical report, Yale University, 2008.

C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. *Annals of Statistics*, 2008. To appear.

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for principal components analysis of a random vector function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12:136–154, 1982.

S. Efromovich and V. Koltchinskii. On inverse problems with unknown operators. *IEEE Transactions on Information Theory*, 47(7):2876–2894, 2001.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Kluwer Academic, Dordrecht, 2000.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Methods, Theory, Applications and Implementations*. Springer-Verlag, London, 2006.

J. Johannes. Nonparametric estimation in functional linear model with second order stationary regressors. Submitted, 2008.

J. Johannes and R. Schenk. Rate optimal estimation of linear functionals in functional linear model. Technical report, University Heidelberg, 2008.

A. P. Korostolev and A. B. Tsybakov. *Minimax Theory for Image Reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, 1993.

B. A. Mair. Tikhonov regularization for finitely and infinitely smoothing operators. *SIAM Journal on Mathematical Analysis*, 25:135–147, 1994.
B. A. Mair and F. H. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM Journal on Applied Mathematics*, 56(5):1424–1444, 1996.

B. D. Marx and P. H. Eilers. Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41:1–13, 1999.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *Ann. Stat.*, 33:774–805, 2005.

M. Nair, S. V. Pereverzev, and U. Tautenhahn. Regularization in Hilbert scales under general smoothing conditions. *Inverse Problems*, 21:1851–1869, 2005.

F. Natterer. Regulärisierung schlecht gestellter probleme durch projektionsverfahren. *Numer. Math.*, 28:329–341, 1997.

F. Natterer. Error bounds for Tikhonov regularization in Hilbert scales. *Applicable Analysis*, 18:29–37, 1984.

A. Neubauer. When do Sobolev spaces form a Hilbert scale? *Proc. Amer. Math. Soc.*, 103(2):557–562, 1988a.

A. Neubauer. An a posteriori parameter choice for Tikhonov regularization in Hilbert scales leading to optimal convergence rates. *SIAM J. Numer. Anal.*, 25(6):1313–1326, 1988b.

F. Olver. *Asymptotics and special functions*. Academic Press, New York, 1974.

V. V. Petrov. *Limit theorems of probability theory. Sequences of independent random variables*. Oxford Studies in Probability. Clarendon Press., Oxford, 4. edition, 1995.

C. Preda and G. Saporta. PLS regression on a stochastic process. *Computational Statistics & Data Analysis*, 48:149–158, 2005.

J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, New York, second ed. edition, 2005.

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B*, 53:539–572, 1991.

S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26:153 – 172, 2007.

A. B. Tsybakov. *Introduction à l’estimation non-paramétrique (Introduction to nonparametric estimation)*. Mathématiques & Applications (Paris). 41. Springer: Paris, 2004.

A. B. Tsybakov. On the best rate of adaptive estimation in some inverse problems. *Comptes Rendus de l’Académie des Sciences, Série I, Mathématiques*, 330:835–840, 2000.