On rate optimal local estimation in nonparametric instrumental regression.

CHRISTOPH BREUNIG* and JAN JOHANNES*

February 12, 2009

Abstract

We consider the problem of estimating the value of a linear functional in nonparametric instrumental regression, where in the presence of an instrument W a response Y is modeled in dependence of an endogenous explanatory variable Z. The proposed estimator is based on dimension reduction and additional thresholding. The minimax optimal rate of convergence of the estimator is derived assuming that the structural function and the representer of the linear functional belong to some ellipsoids which are in a certain sense linked to the conditional expectation operator of Z given W. We illustrate these results by considering classical smoothness assumptions.

Keywords: Nonparametric regression, Instrument, Linear functional,
Linear Galerkin approach, Optimal rates of convergence,
Sobolev space, finitely and infinitely smoothing operator.
JEL classifications: Primary C14; secondary C30.

1 Introduction

Nonparametric instrumental regression models have attract a growing attention recently in the econometrics and statistics literature (c.f. Florens [2003], Darolles et al. [2002], Newey and Powell [2003], Hall and Horowitz [2007] or Blundell et al. [2007] to name only a few). To be precise, these models deal with situations where the depends of a response Y on the variation of an endogenous vector Z of explanatory variables is characterized by

$$Y = \varphi(Z) + U \tag{1.1a}$$

for some error term U, and there exists an exogenous vector of instruments W such that

$$\mathbb{E}[U|W] = 0 . \tag{1.1b}$$

The nonparametric relationship is thereby modeled by the structural function φ. Typical examples leading to such situation are given by error-in-variable models, simultaneous equations or treatment models with endogeneous selection. However, it is worth noting that in the presence of instrumental variables the model equations (1.1a–1.1b) are the natural generalization of a standard parametric model (see, e.g., Amemiya [1974]) to the nonparametric

---

*Universität Heidelberg, Institut für Angewandte Mathematik, Im Neuenheimer Feld, 294, D-69120 Heidelberg, Germany. Corresponding author Email: johannes@statlab.uni-heidelberg.de.

situation. This extension has been introduced first by Florens [2003] and Newey and Powell [2003], while its identification has been studied e.g. in Carrasco et al. [2006], Darolles et al. [2002] and Florens et al. [2007]. It is interesting to note that recent applications and extensions of this approach include nonparametric tests of exogeneity (Blundell and Horowitz [2007]), quantile regression models (Horowitz and Lee [2007]), or semiparametric modeling (Florens et al. [2009]) to name but a few.

The nonparametric estimation of the structural function φ given a sample of (Y, Z, W) has been intensively studied in the literature. For example, Ai and Chen [2003], Blundell et al. [2007] or Newey and Powell [2003] consider sieve minimum distance estimator, while Darolles et al. [2002], Hall and Horowitz [2005], Gagliardini and Scaillet [2006] or Florens et al. [2007] study penalized least squares estimator. However, as it has been noticed by Newey and Powell [2003] and Florens [2003], the nonparametric estimation of the structural function φ generally leads to an ill-posed inverse problem. Precisely, consider the model equations (1.1a–1.1b), then taking the conditional expectation with respect to the instruments W on both sides in equation (1.1a) leads to the conditional moment equation:

$$\mathbb{E}\left[Y|W\right] = \mathbb{E}\left[\varphi(Z)|W\right]. \tag{1.2}$$

Therefore, the estimation of the structural function φ is linked to the inversion of equation (1.2), which is under fairly mild assumptions not stable and hence an ill-posed inverse problem (for a comprehensive review of inverse problems in econometrics we refer to Carrasco et al. [2006]).

The instability of the conditional moment equation (1.2) essentially implies that all proposed estimators of the structural function φ have under reasonable assumptions very poor rates of convergence. In other words, even relatively large sample sizes may not be of much help in accurately estimating the structural function φ. In contrast, it might be possible to estimate certain local features of φ, such as the value of a linear functional $\ell_h(\varphi) := \mathbb{E}\left[h(Z)\varphi(Z)\right]$ with respect to some given representer h, at the usual parametric rate of convergence. Take as an example the case of an endogenous regressor Z uniformly distributed on $[0, 1]$. In this situation rather than estimating the structural function φ itself one may be interested in its average value $\int_a^b \varphi(t)dt$ over a certain interval $[a, b]$ which equals the value $\ell_h(\varphi)$ of a linear functional with representer given by the characteristic function $h = \mathbb{1}_{[a,b]}$. Then it is of interest to characterize the attainable accuracy of any estimator, for example, in terms of the mean squared error (MSE), which obviously depends on the representer h and the conditions imposed on φ. It is worth noting, that the nonparametric estimation of the value of a linear functional from Gaussian white noise observations is a subject of considerable literature (c.f. Speckman [1979], Li [1982] or Ibragimov and Has'minskii [1984] in case of direct observations, while in case of indirect observations we refer to Donoho and Low [1992], Donoho [1994] or Goldenshluger and Pereverzev [2000] and references therein). However, as far as we know this question has not yet been addressed in nonparametric instrumental regression, which in general is not a Gaussian white noise model.

The objective of this paper is the nonparametric estimation of the value $\ell_h(\varphi)$ of a linear functional based on an independent and identically distributed (i.i.d.) sample of (Y, Z, W) obeying (1.1a–1.1b). In this paper we follow an often in the literature used approach to construct an estimator of the value of a linear functional. That is, we replace in $\ell_h(\varphi)$ the unknown structural function φ by an estimator. Therefore, let us first motivate the estimator of φ (for its asymptotic properties we refer to Johannes [2009]). Suppose for

a moment that the structural function can be developed by using only m pre-specified functions $e_1, \ldots, e_m$, say $\varphi = \sum_{j=1}^m [\varphi]_j e_j$, where now the coefficients $[\varphi]_1, \ldots, [\varphi]_m$ are only unknown. Thereby, the conditional moment equation (1.2) reduces to a multivariate linear conditional moment equation, that is, $\mathbb{E}[Y|W] = \sum_{j=1}^m [\varphi]_j \mathbb{E}[e_j(Z)|W]$. Notice that solving this equation is a classical textbook problem in econometrics (c.f. Pagan and Ullah [1999]). One popular approach is to replace the conditional moment equation by unconditional once. Therefore, given m functions $f_1, \ldots, f_m$ one may consider m unconditional moment equations in place of the multivariate conditional moment equation, that is, $\mathbb{E}[Y f_l(W)] = \sum_{j=1}^m [\varphi]_j \mathbb{E}[e_j(Z) f_l(W)]$, $l = 1, \ldots, m$. Notice that once the functions $\{f_l\}_{l=1}^m$ are chosen all the unknown quantities in the unconditional moment equations can be straightforward estimated by replacing the theoretical expectation by its empirical counterpart. Moreover, a least squares solution of the estimated equation leads then under very mild assumptions to a consistent and asymptotic normal estimator of the parameter vector $([\varphi]_j)_{j=1}^m$. Furthermore, the choice of the functions $\{f_l\}_{l=1}^m$ directly influences the asymptotic variance of the estimator and thus the question of optimal instruments arises (c.f. Newey and Powell [2003]). However, our objective is the estimation of the value of a linear functional. For simplicity suppose the regressor Z is uniformly distributed on $[0,1]$ and the linear functional is given by the representer $h = \mathbb{1}_{[a,b]}$, that is, $\ell_h(\varphi) = \int_a^b \varphi(t) dt$. In case $\varphi = \sum_{j=1}^m [\varphi]_j e_j$ the value of the linear functional writes $\ell_h(\varphi) = \sum_{j=1}^m [h]_j [\varphi]_j$ where the coefficients $[h]_j := \int_a^b e_j(t) dt$, $1 \leqslant j \leqslant m$, are known. A natural estimator of $\ell_h(\varphi)$ is then defined by replacing the unknown coefficients $[\varphi]_j$ by their least squares estimators. This approach is very simple and the estimator can be calculated with most statistical software. However, it has a major default, since in most situations there is an infinite number of functions $\{e_j\}_{j \geqslant 1}$ and associated coefficients $([\varphi]_j)_{j \geqslant 1}$ needed to develop the structural function φ. The choice of the functions $\{e_j\}_{j \geqslant 1}$ reflects now the a priori information (such as smoothness) about the structural function φ. However, if we consider also an infinite number of functions $\{f_l\}_{l \geqslant 1}$ then for each $m \geqslant 1$ we could still consider the least squares estimator described above. Notice, that the dimension m plays here the role of a smoothing parameter and we may hope that the estimator of the structural function φ (hence of the value $\ell_h(\varphi)$) is also consistent as m tends suitably to infinity. Unfortunately, if $\varphi_m := \sum_{j=1}^m [\varphi_m]_j e_j$ denotes a least squares solution of the reduced unconditional moment equations, that is, the vector of coefficients $([\varphi_m]_j)_{j=1}^m$ minimizes the quantity $\sum_{l=1}^m \{\mathbb{E}[Y f_l(W)] - \sum_{j=1}^m \beta_j \mathbb{E}[e_j(Z) f_l(W)]\}^2$ over all $(\beta_j)_{j=1}^m$. Then, φ_m converges to the true structural function as m tends to infinity only under an additional assumption (defined below) on the basis $\{f_j\}_{j \geqslant 1}$. In this paper we show under this additional assumption that in terms of the MSE a plug-in estimator of $\ell_h(\varphi)$ using a least squares estimator of φ based on a dimension reduction together with an additional thresholding is consistent and can attain optimal rates of convergences. It is worth to note that all the results in this paper are obtained without an additional smoothness assumption on the joint density of (Y, Z, W). In fact we do even not impose that a joint density exists.

The paper is organized in the following way. In Section 2 we introduce our basic assumptions and derive a lower bound for estimating the value of a linear functional based on an i.i.d. sample obeying the model equations (1.1a–1.1b). In Section 3 under certain moment assumptions we show in terms of the MSE first consistency of the proposed estimator and second its minimax-optimality. We illustrate the general results in Section 4 by considering classical smoothness assumptions. All proofs can be found in the Appendix.

3

2 Complexity of local estimation: a lower bound.

2.1 Basic model assumptions.

It is convenient to rewrite the moment equation (1.2) in terms of an operator between Hilbert spaces. Let us first introduce the Hilbert Spaces

$$L_Z^2 = \{\phi : \mathbb{R}^p \to \mathbb{R};\ \|\phi\|_Z^2 := \mathbb{E}\,[\phi^2(Z)] < \infty\},$$
$$L_W^2 = \{\psi : \mathbb{R}^q \to \mathbb{R};\ \|\psi\|_W^2 := \mathbb{E}\,[\psi^2(W)] < \infty\}$$

which are endowed with corresponding inner products $\langle \phi, \tilde{\phi} \rangle_Z = \mathbb{E}\,[\phi(Z)\tilde{\phi}(Z)]$, $\phi, \tilde{\phi} \in L_Z^2$, and $\langle \psi, \tilde{\psi} \rangle_W = \mathbb{E}\,[\psi(W)\tilde{\psi}(W)]$, $\psi, \tilde{\psi} \in L_W^2$, respectively. Then the conditional expectation of Z given W defines a linear operator $T\phi := \mathbb{E}\,[\phi(Z)|W]$, $\phi \in L_Z^2$, which maps L_Z^2 into L_W^2. Thereby the moment equation (1.2) can be written as

$$g := \mathbb{E}\,[Y|W] = \mathbb{E}\,[\varphi(Z)|W] =: T\varphi \tag{2.1}$$

where the function g belongs to L_W^2. Estimation of the structural function φ is thus linked with the inversion of the conditional expectation operator T and, hence called an inverse problem. Moreover, we suppose throughout the paper that the operator T is compact which is under fairly mild assumptions satisfied (c.f. Carrasco et al. [2006]). Consequently, unlike in a multivariate linear instrumental regression model, a continuous generalized inverse of T does not exist as long as the range of the operator T is an infinite dimensional subspace of L_W^2. This corresponds to the setup of ill-posed inverse problems (with the additional difficulty that T is unknown and, hence has to be estimated). In what follows we always assume that there exists a unique solution $\varphi \in L_Z^2$ of equation (2.1), i.e., g belongs to the range $\mathcal{R}(T)$ of T, and that the null space $\mathcal{N}(T)$ of T is trivial or equivalently T is injective (for a detailed discussion in the context of inverse problems see Chapter 2.1 in Engl et al. [2000], while in the special case of a nonparametric instrumental regression we refer to Carrasco et al. [2006]). Furthermore, we suppose that the representer h of the linear functional $\ell_h(\cdot) := \langle \cdot, h \rangle_Z$ of interest is an element of L_Z^2 as well. Then it is straightforward to see, that the value of the linear functional $\ell_h(\varphi)$ is identified if and only if h belongs to the orthogonal complement $\mathcal{N}(T)^\perp$ of the null space $\mathcal{N}(T)$. Hence, for all $h \in L_Z^2$ the identification is in particular guaranteed under the assumption of an injective conditional expectation operator T.

2.2 Notations and regularity assumptions.

In this section we show that the obtainable accuracy of any estimator of the value $\ell_h(\varphi)$ of a linear functional can be essentially determined by additional regularity conditions imposed on the structural function φ, the representer h and the conditional expectation operator T. In this paper these conditions are characterized through different weighted norms in L_Z^2 with respect to a pre-specified orthonormal basis $\{e_j\}_{j \geqslant 1}$ in L_Z^2, which we formalize now. Given a strictly positive sequence of weights $w := (w_j)_{j \geqslant 1}$ and a constant $c > 0$ we denote for all $r \in \mathbb{R}$ by $\mathcal{F}_{w^r}^c$ the ellipsoid defined by

$$\mathcal{F}_{w^r}^c := \Big\{ \phi \in L_Z^2 : \sum_{j=1}^{\infty} w_j^r |\langle \phi, e_j \rangle_Z|^2 =: \|\phi\|_{w^r}^2 \leqslant c \Big\}. \tag{2.2}$$

Furthermore, let $\mathcal{F}_{w^r} := \{\phi \in L_Z^2 : \|\phi\|_{w^r}^2 < \infty\}$. It is worth noting, that in case $w \equiv 1$ we have $\|\phi\|_{w^r} = \|\phi\|_Z$ for all $\phi \in L_Z^2$ and hence the set $\mathcal{F}_w^c$ denotes an ellipsoid in L_Z^2 which does not impose additional restrictions.

Minimal regularity conditions. Let $\gamma := (\gamma_j)_{j\geqslant 1}$ and $\omega := (\omega_j)_{j\geqslant 1}$ denote two sequences of weights. Then we suppose, here and subsequently, that the structural function φ belongs to the ellipsoid $\mathcal{F}_\gamma^\rho$ for some $\rho > 0$ and that the representer h of the linear functional ℓ_h is an element of the ellipsoid $\mathcal{F}_\omega^\tau$ for some $\tau > 0$. The ellipsoids $\mathcal{F}_\gamma^\rho$ and $\mathcal{F}_\omega^\tau$ capture all the prior information (such as smoothness) about the unknown structural function φ and the given representer h respectively. Furthermore, as usual in the context of ill-posed inverse problems, we specify the mapping properties of the conditional expectation operator T. Therefore, consider the sequence $(\|Te_j\|_W)_{j\geqslant 1}$, which converges to zero since T is compact. In what follows we impose restrictions on the decay of this sequence. Denote by $\mathcal{T}$ the set of all injective compact operator mapping L_Z^2 into L_W^2. Given a sequence of weights $\upsilon := (\upsilon_j)_{j\geqslant 1}$ and $d \geqslant 1$ we define the subset $\mathcal{T}_\upsilon^d$ of $\mathcal{T}$ by

$$\mathcal{T}_\upsilon^d := \left\{ T \in \mathcal{T} : \quad \|\phi\|_\upsilon^2/d \leqslant \|T\phi\|_W^2 \leqslant d\,\|\phi\|_\upsilon^2, \quad \forall \phi \in L_Z^2 \right\}. \tag{2.3}$$

Notice that for all $T \in \mathcal{T}_\upsilon^d$ it follows that[1] $\|Te_j\|_W^2 \asymp_d \upsilon_j$. Hence, the sequence $(\upsilon_j)_{j\geqslant 1}$ has to be strictly positive since T is injective. Furthermore, let us denote by $T^* : L_W^2 \to L_Z^2$ the adjoint of T which satisfies $T^*\psi = \mathbb{E}\left[\psi(W)|Z\right]$. If now $T \in \mathcal{T}$ and $\{\lambda_j, e_j\}_{j\geqslant 1}$ is an eigenvalue decomposition of T^*T. Then the condition $T \in \mathcal{T}_\upsilon^d$ is satisfied if and only if $\lambda_j \asymp_d \upsilon_j$. In other words, in this situation the sequence υ specifies the decay of the eigenvalues of T^*T. In what follows all the results are derived under regularity conditions on the structural function φ, the representer h and the conditional expectation operator T described through the sequence γ, ω and υ respectively. However, we provide below illustrations of these conditions by assuming a "regular decay" of these sequences. The next assumption summarizes our minimal regularity conditions on these sequences.

ASSUMPTION 2.1. *Let* $\gamma := (\gamma_j)_{j\geqslant 1}$, $\omega := (\omega_j)_{j\geqslant 1}$ *and* $\upsilon := (\upsilon_j)_{j\geqslant 1}$ *be strictly positive sequences of weights with* $\gamma_1 = 1$, $\omega_1 = 1$ *and* $\upsilon_1 = 1$ *such that* γ *and* ω *are non decreasing and* υ *is non increasing. Furthermore, there exists a constant* $\Lambda \geqslant 1$ *such that* $\upsilon_m \sup_{1\leqslant j\leqslant m}\{\upsilon_j^{-1}\omega_j^{-1}\} \leqslant \Lambda \max(\omega_m^{-1}, \upsilon_m)$ *for all* $m \in \mathbb{N}$.

We shall stress that $\mathcal{F}_\gamma^\rho$ is just an ellipsoid in L_Z^2 in case $\gamma \equiv 1$, hence in this situation there is not an additional regularity condition on the structural function φ imposed. Furthermore, the last condition in Assumption 2.1 is obviously satisfied with $\Lambda = 1$ if the sequence $(\upsilon_j\omega_j)_{j\geqslant 1}$ is either monotonically decreasing or increasing.

2.3 The lower bound.

In the proof of the next theorem we show that an one-dimensional subproblem captures the full difficulty in estimating a linear functional in nonparametric instrumental regression. In other words, there exist two sequences of structural functions $\varphi_{1,n}, \varphi_{2,n} \in \mathcal{F}_\gamma^\rho$, which are statistically not consistently distinguishable, and a sequence of representer $h_n \in \mathcal{F}_\omega^\tau$ such that $|\ell_{h_n}(\varphi_{1,n}) - \ell_{h_n}(\varphi_{2,n})|^2 \geqslant C\delta_n$, where δ_n is the optimal rate of convergence. Moreover, we obtain the following lower bound under the additional assumption that there exist error terms $U_{i,n}$, $i = 1, 2$, such that the conditional distribution of $\varphi_{i,n} - T\varphi_{i,n} + U_{i,n}$ given

---

[1]We write $a \asymp_d b$ if $d^{-1} \leqslant b/a \leqslant d$.

the instrument W is Gaussian with mean zero and variance one. A similar assumption has recently been used by Chen and Reiß [2008] in order to derive a lower bound for the estimation of the structural function φ itself. In particular the authors show that in opposite to the present work an one-dimensional subproblem is not sufficient to describe the full difficulty in estimating φ.

THEOREM 2.1. *Assume an n-sample of (Y, Z, W) from the model (1.1a–1.1b) with error term U belonging to $\mathcal{U}_\sigma := \{U : \mathbb{E}\, U|W = 0 \text{ and } \mathbb{E}\, U^4|W \leqslant \sigma^4\}$, $\sigma > 0$. Let γ, ω and υ be sequences satisfying Assumption 2.1. Suppose that the conditional expectation operator T associated (Z, W) belongs to $\mathcal{T}_\upsilon^d$, $d \geqslant 1$, and that $\sup_{j \geqslant} \mathbb{E}\,[e_j^4(Z)|W] \leqslant \eta$, $\eta \geqslant 1$. Let $m_* := m_*(n) \in \mathbb{N}$ and $\delta_n^* := \delta_n^*(n) \in \mathbb{R}^+$ be chosen such that for some $\triangle \geqslant 1$*

$$1/\triangle \leqslant \frac{\gamma_{m_*}}{n\, \upsilon_{m_*}} \leqslant \triangle \quad and \quad \delta_n^* := \gamma_{m_*}^{-1}\omega_{m_*}^{-1}. \tag{2.4}$$

If in addition σ is sufficiently large then for any estimator $\breve{\ell}$ we have

$$\sup_{U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \sup_{h \in \mathcal{F}_\omega^\tau} \left\{ \mathbb{E}\, |\breve{\ell} - \ell_h(\varphi)|^2 \right\} \geqslant \max\left(\delta_n^*, \frac{1}{n}\right) \frac{\tau}{4\,\triangle} \min\left(\frac{\sigma_0^2}{2\,d}, \frac{\rho}{\triangle}\right).$$

REMARK 2.1. In the last theorem the additional moment condition $\sup_{j \geqslant 1} \mathbb{E}\,[e_j^4(Z)|W] \leqslant \eta$ is obviously satisfied if the basis functions $\{e_j\}$ are uniformly bounded (e.g. the trigonometric basis considered in Section 4). However, if V denotes Gaussian random variable with mean zero and variance one, which is independent of (Z, W), then the additional moment condition ensures that for all structural functions of the form $\varphi = a \cdot e_j \in \mathcal{F}_\gamma^\rho$ with $j \geqslant 1$ and $a \in \mathbb{R}$, the error term $U := V - \varphi(Z) + [T\varphi](W)$ belongs to $\mathcal{U}_\sigma$ for all sufficiently large σ. This specific case is only needed to simplify the calculation of the distance between distributions corresponding to different structural functions. On the other hand, below we derive an upper bound assuming that the error term U belongs to $\mathcal{U}_\sigma$ and that the joint distribution of (Z, W) fulfills additional moment conditions. Obviously in this situation Theorem 2.1 provides a lower bound for any estimator as long as σ is sufficiently large. Furthermore, it is worth noting that this lower bound tends only to zero if $(\omega_j \gamma_j)_{j \geqslant 1}$ is a divergent sequence. In other words, in case $\gamma \equiv 1$, i.e., without any additional restriction on φ, consistency of an estimator of $\ell_h(\varphi)$ uniformly over all $\varphi \in \mathcal{F}_\gamma^\rho$ is only possible under restrictions on the representer $h \in \mathcal{F}_\omega^\tau$, that is, ω is a divergent sequence. This obviously reflects the ill-posedness of the underlying inverse problem. Finally, it is important to note that the regularity conditions imposed on the structural function φ, the representer h and the conditional expectation operator T involve only the basis $\{e_j\}_{j \geqslant 1}$ in L_Z^2. Therefore, the lower bound derived in Theorem 2.1 does not capture the influence of the basis $\{f_l\}_{l \geqslant 1}$ in L_W^2 used to construct the estimator. In other words, an estimator of the value $\ell_h(\varphi)$ can only attain this lower bound if $\{f_l\}_{l \geqslant 1}$ is appropriate chosen. $\square$

3 Minimax-optimal local estimation: the general case.

3.1 Estimation by dimension reduction and thresholding.

In addition to the basis $\{e_j\}_{j \geqslant 1}$ in L_Z^2 considered in the last section we introduce now also a second basis $\{f_l\}_{l \geqslant 1}$ in L_W^2. We derive in this section the asymptotic properties of the estimator under minimal assumptions on those basis. Precisely, we show first consistency

of the proposed estimator under fairly mild additional moment assumptions. In particular, we do not impose any regularity assumption on both the structural function φ and the representer h. In a second step we suppose that the structural function φ and the representer h belong to some ellipsoid $\mathcal{F}_\gamma^\rho$ and $\mathcal{F}_\omega^\tau$ respectively, and that the conditional expectation satisfies a link condition, i.e., $T \in \mathcal{T}_d^v$. Furthermore, we introduce an additional condition linked to the basis $\{f_l\}_{l \geqslant 1}$. Then under stronger moment conditions we show that the proposed estimator attains the lower bound derived in the last section. However, all these results are illustrated in the next section by considering classical smoothness assumptions.

Matrix and operator notations. Given $m \geqslant 1$, $\mathcal{E}_m$ and $\mathcal{F}_m$ denote the subspace of L_Z^2 and L_W^2 spanned by the functions $\{e_j\}_{j=1}^m$ and $\{f_l\}_{l=1}^m$, respectively. E_m and $E_m^\perp$ (resp. F_m and $F_m^\perp$) denote the orthogonal projections on $\mathcal{E}_m$ (resp. $\mathcal{F}_m$) and its orthogonal complement $\mathcal{E}_m^\perp$ (resp. $\mathcal{F}_m^\perp$), respectively. Given an operator (matrix) K, $\|K\|$ denotes its operator norm . The inverse operator (matrix) of K is denoted by K^{-1}, the adjoint (transposed) operator (matrix) of K by K^*. The identity operator (matrix) is denoted by I. $[\phi]$, $[\psi]$ and $[K]$ denote the (infinite) vector and matrix of the function $\phi \in L_Z^2$, $\psi \in L_W^2$ and the operator $K : L_Z^2 \to L_W^2$ with the entries $[\phi]_j = \langle \phi, e_j \rangle$, $[\psi]_l = \langle \psi, f_l \rangle$ and $[K]_{l,j} = \langle K e_j, f_l \rangle$, respectively. The upper m subvector and $m \times m$ submatrix of $[\phi]$, $[\psi]$ and $[K]$ is denoted by $[\phi]_{\underline{m}}$, $[\psi]_{\underline{m}}$ and $[K]_{\underline{m}}$, respectively. Note, that $[K^*]_{\underline{m}} = [K]_{\underline{m}}^*$. The diagonal matrix with entries v is denoted by $\mathrm{Diag}(v)$. Clearly, $[E_m \phi]_{\underline{m}} = [\phi]_{\underline{m}}$ and if we restrict $F_m K E_m$ to an operator from $\mathcal{E}_m$ into $\mathcal{F}_m$, then it has the matrix $[K]_{\underline{m}}$.

Consider the conditional expectation operator T associated to the regressor Z and the instrument W. If $[e(Z)]$ and $[f(W)]$ denote the infinite random vector with entries $e_j(Z)$ and $f_j(W)$ respectively, then $[T]_{\underline{m}} = \mathbb{E}\,[f(W)]_{\underline{m}}[e(Z)]_{\underline{m}}^*$ which is throughout the paper assumed to be non singular for all $m \geqslant 1$ (or, at least for large enough m), so that $[T]_{\underline{m}}^{-1}$ always exists. Note that it is a nontrivial problem to determine when such an assumption holds (see e.g. Efromovich and Koltchinskii [2001] and references therein). Under this assumption the notation T_m^{-1} is used for the operator from L_W^2 into L_Z^2, whose matrix in the basis $\{e_j\}_{j \geqslant 1}$ and $\{f_l\}_{l \geqslant 1}$ has the entries $([T]_{\underline{m}}^{-1})_{j,l}$ for $1 \leqslant j, l \leqslant m$ and zeros otherwise.

Definition of the estimator. Let $(Y_1, Z_1, W_1), \ldots, (Y_n, Z_n, W_n)$ be an i.i.d. sample of (Y, Z, W). Since $[T]_{\underline{m}} = \mathbb{E}\,[f(W)]_{\underline{m}}[e(Z)]_{\underline{m}}^*$ and $[g]_{\underline{m}} = \mathbb{E}\,Y[f(W)]_{\underline{m}}$ we constuct estimators by using their empirical counterparts, that is,

$$[\widehat{T}]_{\underline{m}} := (1/n) \sum_{i=1}^n [f(W_i)]_{\underline{m}}[e(Z_i)]_{\underline{m}}^* \quad \text{and} \quad [\widehat{g}]_{\underline{m}} := (1/n) \sum_{i=1}^n Y_i [f(W_i)]_{\underline{m}}. \tag{3.1}$$

Then the estimator of the linear functional $\ell_h(\varphi)$ is defined by

$$\widehat{\ell}_h := \begin{cases} [h]_{\underline{m}}^t [\widehat{T}]_{\underline{m}}^{-1} [\widehat{g}]_{\underline{m}}, & \text{if } [\widehat{T}]_{\underline{m}} \text{ is nonsingular and } \|[\widehat{T}]_{\underline{m}}^{-1}\| \leqslant \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{3.2}$$

where the dimension parameter $m = m(n)$ and the threshold $\alpha = \alpha(n)$ have to tend to infinity as the sample size n increases. In fact, the estimator $\widehat{\ell}_h$ is obtained from the linear functional $\ell_h(\varphi)$ by replacing the unknown structural function φ by an estimator proposed by Johannes [2009], which takes its inspiration in the linear Galerkin approach coming from the inverse problem community (c.f. Efromovich and Koltchinskii [2001] or Hoffmann and Reiß [2008]).

7

3.2 Consistency.

We start by providing minimal conditions used to proof consistency of the estimator. More specific, we formalize first additional moment assumptions on the basis under consideration.

ASSUMPTION A1. *The joint distribution of* (Z, W) *satisfies* $\sup_{j \in \mathbb{N}} \mathbb{E}\,[f_j^2(W)|Z] \leqslant \eta^2$ *and* $\sup_{j,l \in \mathbb{N}} \mathbb{V}\mathrm{ar}(e_j(Z)f_l(W)) \leqslant \eta^2$ *for some* $\eta \geqslant 1$.

It is worth noting that the Assumption A1 is always fulfilled in case both basis are uniformly bounded (e.g. in case of the trigonometric basis considered in Section 4). The next assertion summarizes our minimal conditions to ensure consistency of the estimator $\widehat{\ell}_h$ introduced in (3.2).

PROPOSITION 3.1. *Assume an n-sample of* (Y, Z, W) *from the model (1.1a–1.1b). Suppose that the error term* U *satisfies* $\mathbb{E}\,U^2|W \leqslant \sigma^2$ *with* $\sigma > 0$ *and that the joint distribution of* (Z, W) *fulfills Assumption A1. Let* $\widehat{\ell}_h$ *be defined with dimension* $m := m(n)$ *and threshold* $\alpha := \alpha(n)$ *satisfying* $\alpha \geqslant 2\|[T]_{\underline{m}}^{-1}\|$ *and as* $n \to \infty$ *that* $1/m = o(1)$ *and* $m^2\alpha^2 = o(n)$. *If in addition* $\sup_{m \in \mathbb{N}} \|T_m^{-1} F_m T E_m^\perp\| < \infty$, *then we have* $\mathbb{E}\,|\widehat{\ell}_h - \ell_h(\varphi)|^2 = o(1)$ *as* $n \to \infty$.

The last result shows consistency of the estimator without an a priori regularity assumption on both the structural function φ and the representer h. However, consistency is only obtained under the condition $\sup_{m \in \mathbb{N}} \|T_m^{-1} F_m T E_m^\perp\| < \infty$, which is known to be necessary to ensure L^2-convergence of the least squares solution $\varphi_m = \sum_{j=1}^m [\varphi_m]_j e_j$ with $[\varphi_m]_{\underline{m}} = [T]_{\underline{m}}^{-1}[g]_{\underline{m}}$ to the structural function φ as $m \to \infty$. Notice that this condition involves now also the basis $\{f_l\}_{l \geqslant 1}$ in L_W^2. In what follows we introduce an alternative but stronger condition to guarantee the L^2-consistency which extends the link condition (2.3), that is, $T \in \mathcal{T}_d^v$. We denote by $\mathcal{T}_{d,D}^v$ for some $D \geqslant d$ the subset of $\mathcal{T}_d^v$ given by

$$\mathcal{T}_{d,D}^v := \Big\{ T \in \mathcal{T}_d^v : \quad \sup_{m \in \mathbb{N}} \|[\mathrm{Diag}(v)]_{\underline{m}}^{1/2}[T]_{\underline{m}}^{-1}\|^2 \leqslant D \Big\}. \tag{3.3}$$

REMARK 3.1. If $\{\sqrt{\lambda_j}, e_j, f_j\}_{j \geqslant 1}$ is a singular value decomposition of $T \in \mathcal{T}$ then for all $m \geqslant 1$ the matrix $[T]_{\underline{m}}$ is diagonalized with diagonal entries $[T]_{j,j} = \sqrt{\lambda_j}$, $1 \leqslant j \leqslant m$. Therefore, the link condition (2.3) holds true, that is, $T \in \mathcal{T}_d^v$, if and only if $\lambda_j \asymp_d v_j$ for all $j \in \mathbb{N}$. Moreover, it is easily seen that $\sup_{m \in \mathbb{N}} \|[\mathrm{Diag}(v^{1/2})]_{\underline{m}}[T]_{\underline{m}}^{-1}\|^2 \leqslant d$ and hence the extended link condition (3.3) is fulfilled, that is, $T \in \mathcal{T}_{d,D}^v$ for all $D \geqslant d$. Furthermore, the extended link condition equals the link condition ($\mathcal{T}_d^v = \mathcal{T}_{d,D}^v$ for suitable $D > 0$), if $[T]$ is only a small perturbation of $\mathrm{Diag}(v^{1/2})$ or if T is strictly positive (for a detailed discussion we refer to Efromovich and Koltchinskii [2001]) and Cardot and Johannes [2008] respectively).

We shall stress that once both basis $\{e_j\}_{j \geqslant 1}$ and $\{f_l\}_{l \geqslant 1}$ are specified the extended link condition (3.3) restricts the class of joint distributions of (Z, W) to those for which the least squares solution φ_m is L^2-consistent. Moreover, it is shown in Johannes [2009], that under the extended link condition a least squares estimator of φ based on a dimension reduction together with an additional thresholding can attain minimax-optimal rates of convergence. In this sense, given a joint distribution of (Z, W) a basis $\{f_l\}_{l \geqslant 1}$ satisfying the extended link condition can be interpreted as optimal instruments. However, for each pre-specified basis $\{e_j\}_{j \geqslant 1}$ we can theoretically construct a basis $\{f_l\}_{l \geqslant 1}$ such that the extended link condition is not a stronger restriction than the link condition (2.3). To be more precise, if $T \in \mathcal{T}_d^v$, which involves only the basis $\{e_j\}_{j \geqslant 1}$, then it is not hard to see that the fundamental inequality of Heinz [1951] implies $\|(T^*T)^{-1/2} e_j\|^2 \asymp_d v_j^{-1}$ for all $j \geqslant 1$. Thereby, the

8

function $(T^*T)^{-1/2}e_j$ is an element of L_Z^2 and hence there exists $f_j := T(T^*T)^{-1/2}e_j \in L_W^2$, $j \geqslant 1$. Then it is easily checked that $\{f_l\}_{l \geqslant 1}$ is an orthonormal system and moreover a basis of the closure of the range $\mathcal{R}(T)$ of T. Hence by taking any basis of the orthogonal complement $\mathcal{R}(T)^\perp$ of $\mathcal{R}(T)$ we may complete the orthonormal set $\{f_l\}_{l \geqslant 1}$ to become a basis of L_W^2. Then it is straightforward to see that $[T]_{\underline{m}}$ is symmetric and moreover strictly positive, since $\langle Te_j, f_l \rangle_W = \langle Te_j, T(T^*T)^{-1/2}e_l \rangle_W = \langle (T^*T)^{1/2}e_j, e_l \rangle_Z$ for all $j, l \geqslant 1$. Thereby, we can apply Lemma A.3 in Cardot and Johannes [2008] which gives $\mathcal{T}_d^v = \mathcal{T}_{d,D}^v$ for all sufficiently large D. $\qquad\square$

Under the extended link condition (3.3), that is, $T \in \mathcal{T}_{d,D}^v$, the next assertion summarizes minimal conditions to ensure consistency.

COROLLARY 3.2. *Let the assumptions of Proposition 3.1 be satisfied and assume in addition that $T \in \mathcal{T}_{d,D}^v$. If $\widehat{\ell}_h$ is defined with threshold $\alpha = 2\sqrt{D/v_m}$ and dimension $m := m(n)$ such that $m^2/(nv_m) = o(1)$ and $1/m = o(1)$. Then we have $\mathbb{E}\,|\widehat{\ell}_h - \ell_h(\varphi)|^2 = o(1)$, as $n \to \infty$.*

3.3 The upper bound.

The last assertions show that the estimator $\widehat{\ell}_h$ defined in (3.2) is consistent without any additional regularity conditions both on structural function and representer. The following theorem provides now an upper bound if these conditions are given through ellipsoids $\mathcal{F}_\gamma^\rho$ and $\mathcal{F}_\omega^\tau$ for the structural function and the representer respectively together with an extended link condition (3.3) for the conditional expectation operator T. Furthermore, the result is derived under stronger moment conditions on the basis, more specific, on the random vector $[e(Z)]$ and $[f(W)]$, which we formalize first.

ASSUMPTION A2. *There exists $\eta \geqslant 1$ such that the joint distribution of (Z, W) satisfies*

(i) $\sup_{j \in N} \mathbb{E}\,[e_j^2(Z)|W] \leqslant \eta^2$ *and* $\sup_{l \in N} \mathbb{E}\,[f_l^4(W)] \leqslant \eta^4$;

(ii) $\sup_{j,l \in \mathbb{N}} \mathbb{V}\mathrm{ar}(e_j(Z)f_l(W)) \leqslant \eta^2$ *and*
$\sup_{j,l \in \mathbb{N}} \mathbb{E}\,|e_j(Z)f_l(W) - \mathbb{E}\,[e_j(Z)f_l(W)]|^8 \leqslant 8!\,\eta^6\, \mathbb{V}\mathrm{ar}(e_j(Z)f_l(W))$.

It is worth noting that again any joint distribution of (Z, W) satisfies Assumption A2 for sufficiently large η if the basis $\{e_j\}_{j \geqslant 1}$ and $\{f_l\}_{l \geqslant 1}$ are uniformly bounded. Here and subsequently, we write $a_n \lesssim b_n$ when there exists $C > 0$ such that $a_n \leqslant C\,b_n$ for all sufficiently large $n \in \mathbb{N}$ and $a_n \sim b_n$ when $a_n \lesssim b_n$ and $b_n \lesssim a_n$ simultaneously.

THEOREM 3.3. *Assume an n-sample of (Y, Z, W) from the model (1.1a–1.1b) with error term $U \in \mathcal{U}_\sigma$, $\sigma > 0$. Suppose that the joint distribution of (Z, W) fulfills Assumption A2 for some $\eta \geqslant 1$ and that the associated conditional expectation operator $T \in \mathcal{T}_{d,D}^v$, $d, D \geqslant 1$, where the sequences γ, ω and v satisfy Assumption 2.1. Let $m_* := m_*(n)$ and $\delta_n^* := \delta_n^*(n)$ be such that (2.4) holds for some $\triangle \geqslant 1$. Consider the estimator $\widehat{\ell}_h$ with dimension $m := m_*$ and threshold $\alpha^2 := n\max(1, 4\,D\triangle/\gamma_m)$. If in addition $\Gamma := \sum_{j \in \mathbb{N}} \gamma_j^{-1} < \infty$, then we have*

$$\sup_{\varphi \in \mathcal{F}_\gamma^\rho} \sup_{h \in \mathcal{F}_\omega^\tau} \mathbb{E}\,|\widehat{\ell}_h - \ell_h(\varphi)|^2 \lesssim d\,D^2\,\Lambda\,\triangle\,\eta^4\,\{\sigma^2 + dD\Gamma\}\,\rho\,\tau$$

$$\left\{1 + m_*^3/\gamma_{m_*} + m_*^3 \Big| P\Big(\|[\widehat{T}]_{\underline{m_*}} - [T]_{\underline{m_*}}\|^2 > v_{m_*}/(4D)\Big)\Big|^{1/4}\right\}$$

$$\left\{\max\Big(\delta_n^*, 1/n\Big) + P\Big(\|[\widehat{T}]_{\underline{m_*}} - [T]_{\underline{m_*}}\|^2 > v_{m_*}/(4D)\Big)\right\}.$$

We shall stress that the bound in the last theorem is non asymptotic. However, it does not establish the optimality of the estimator compared with the lower bound in Theorem 2.1. But, the bound in Theorem 3.3 can be improved by imposing a moment condition stronger than Assumption A2. To be more precise, consider the centered random variable $e_j(Z)f_l(W) - \mathbb{E}[e_j(Z)f_l(W)]$. Then Assumption A2 (ii) states that its 8th moment is uniformly bound over $j, l \in \mathbb{N}$. In the next Assumption we suppose that these random variables satisfy uniformly Cramer's condition, which is known to be sufficient to obtain an exponential bound for their large deviations (c.f. Bosq [1998]).

ASSUMPTION A3. *There exists $\eta \geqslant 1$ such that the joint distribution of (Z, W) satisfies Assumption A2 and in addition*

(iii) $\sup_{j,l \in \mathbb{N}} \mathbb{E}\,|e_j(Z)f_l(W) - \mathbb{E}[e_j(Z)f_l(W)]|^k \leqslant \eta^{k-2}k!\,\mathbb{V}\mathrm{ar}(e_j(Z)f_l(W)),\ k = 3, 4, \ldots.$

It is well-known that Cramer's condition is in particular fulfilled if the random variable $e_j(Z)f_l(W) - \mathbb{E}[e_j(Z)f_l(W)]$ is bounded. Hence in case the basis $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$ are uniformly bounded it follows again that any joint distribution of (Z, W) satisfies Assumption A3 for sufficiently large η. On the other hand, in Lemma A.5 in the Appendix we show that Assumption A3 implies an exponential bound on the large deviation probability $P(\|[\widehat{T}]_{\underline{m}} - [T]_{\underline{m}}\|^2 > \upsilon_m/(4D))$. Thereby, if the sequences γ, ω and υ have the following additional properties

$$m_*^2(\log \gamma_{m_*})\gamma_{m_*}^{-1} = o(1), \quad m_*^2(\log \min(\upsilon_{m_*}^{-1}, \omega_{m_*}))\gamma_{m_*}^{-1} = o(1), \quad m_*^3\gamma_{m_*}^{-1} = O(1) \quad \text{as } n \to \infty, \tag{3.4}$$

where $m_* := m_*(n)$ and $\delta_n^* := \delta_n^*(m_*)$ are given by (2.4), then the large deviation probability tends to zero more quickly than $\max(\delta_n^*, 1/n)$. In this situation it is not hard to see that $\max(\delta_n^*, 1/n)$ is the order of the upper bound given in Theorem 3.3. Hence, the rate $\max(\delta_n^*, 1/n)$ is optimal and $\widehat{\ell}_h$ is minimax-optimal, which is summarized in the next assertion.

THEOREM 3.4. *Suppose that the assumptions of Theorem 3.3 are satisfied. In addition assume that the joint distribution of (Z, W) fulfills Assumption A3 and that the sequences γ, ω and υ have the properties (3.4). Then, we have*

$$\sup_{\varphi \in \mathcal{F}_\gamma^\rho} \sup_{h \in \mathcal{F}_\omega^\tau} \mathbb{E}\,|\widehat{\ell}_h - \ell_h(\varphi)|^2 \lesssim d\,D^2\,\Lambda \triangle \eta^4 \{\sigma^2 + dD\Gamma\}\,\rho\,\tau\,\max(\delta_n^*, n^{-1}).$$

REMARK 3.2. It is worth noting that the bound in the last result is again non asymptotic. Furthermore, from Theorem 2.1 and 3.4 follows that the estimator $\widehat{\ell}_h$ attains the optimal rate $\max(\delta_n^*, n^{-1})$ (hence is minimax-optimal) for all sequences γ, ω and υ satisfying both the minimal regularity conditions summarized in Assumption 2.1 and the additional properties (3.4). We shall emphasize the interesting influence of the sequences γ, ω and υ. As we see from Theorem 2.1 and 3.4, if the sequence υ decreases more quickly to zero then the obtainable optimal rate of convergence decreases. On the other hand, a faster increasing sequence γ or ω leads to a faster optimal rate. In other words, as expected, values of a linear functional given by a structural function or representer satisfying a stronger regularity condition can be estimated faster.

Note furthermore, if the eigenfunctions of the operator T are given by $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$, then $T \in \mathcal{T}_{d,D}^\upsilon$ holds if and only if the corresponding singular values $[T]_{jj} =$

$\langle Te_j, f_j \rangle$, $j \geqslant 1$, satisfy $[T]_{jj}^2 \asymp_d v_j$. Hence, in this situation the optimal rate obtained in the last assertion is linked to the decay of the singular values of T. However, the set $\mathcal{T}_{d,D}^v$ contains also operators with eigenfunctions not given by $\{e_j\}_{j \geqslant 1}$ and $\{f_l\}_{l \geqslant 1}$. Then their corresponding eigenvalues may decay far slower than the sequence of weights v. Moreover, it is straightforward to show, that by using a projection onto the basis $\{e_j\}_{j \geqslant 1}$ and $\{f_l\}_{l \geqslant 1}$ instead of their eigenfunctions, the obtainable rate of convergence given in Theorem 3.4 may be far slower than the rate obtained by using the eigenfunctions (see e.g. Johannes and Schenk [2009] in the context of functional linear model). However, the rate in Theorem 3.4 is optimal since the eigenfunctions are generally unknown.

Finally, since the sequence γ increases it follows that in Theorem 3.3 and hence also in Theorem 3.4 for all large enough n the threshold $\alpha = n$ is used to construct the estimator $\widehat{\ell}_h$. On the other hand, the choice of the dimension m depends on the sequences γ and v characterizing the regularity conditions imposed on the structural function and the conditional expectation operator which are in practice not known. Building data driven rules that can permit to choose automatically the value of m 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] or Tsybakov [2000]. □

4 Minimax-optimal estimation under classical smoothness assumptions.

In this section we shall describe the prior information about the unknown structural function φ and the given representer h by their level of smoothness. In order to simplify the presentation we follow Hall and Horowitz [2005] (where also a more detailed discussion of this assumption can be found), and suppose that the marginal distribution of the scalar regressor Z and the scalar instrument W are uniformly distributed on the interval $[0, 1]$. It is worth noting that all the results below can be straightforward extended to the multivariate case. However, in the univariate case it follows that both Hilbert spaces L_Z^2 and L_W^2 equal $L^2[0, 1]$, which is endowed with the usual norm $\|\cdot\|$ and inner product $\langle \cdot, \cdot \rangle$.

In the last sections we have seen that the choice of the basis $\{e_j\}_{j \geqslant 1}$ is directly linked to the a priori assumptions we are willing to impose on the structural function and the representer. In case of classical smoothness assumptions it is natural to consider the trigonometric basis

$$e_1 :\equiv 1, \ e_{2j}(s) := \sqrt{2}\cos(2\pi j s), \ e_{2j+1}(s) := \sqrt{2}\sin(2\pi j s), s \in [0, 1], \ j \in \mathbb{N}, \qquad (4.1)$$

which can be realized as follows. Let us introduce the Sobolev space of periodic functions $\mathcal{W}_r$, $r \geqslant 0$, which for integer r is given by

$$\mathcal{W}_r = \left\{ f \in H_p : f^{(j)}(0) = f^{(j)}(1), \quad j = 0, 1, \ldots, r - 1 \right\},$$

where $H_r := \{ f \in L^2[0, 1] : f^{(r-1)}$ absolutely continuous $, f^{(r)} \in L^2[0, 1] \}$ is a Sobolev space. If we consider now $\mathcal{F}_{w^r}$ given in (2.2) with weight sequence $w_1 = 1$, $w_j = |j|^2$, $j \geqslant 2$, and trigonometric basis $\{e_j\}$, then it is well-known that the subset $\mathcal{F}_{w^r}$ coincides with the Sobolev space of periodic functions $\mathcal{W}_r$ (c.f. Neubauer [1988a,b], Mair and Ruymgaart [1996] or Tsybakov [2004]). Therefore, let us denote by $\mathcal{W}_r^c := \mathcal{F}_{w^r}^c$, $c > 0$ an ellipsoid in the Sobolev space $\mathcal{W}_r$. We use in case $r = 0$ again the convention that $\mathcal{W}_r^c$ denotes an ellipsoid

in $L^2[0,1]$. In the rest of this section we suppose that the unknown structural function φ and the given representer h are $p \geqslant 0$ and $s \geqslant 0$ times differentiable, respectively. More precisely, the prior information about φ and h are characterized by the Sobolev ellipsoid $\mathcal{W}_p^\rho$, $\rho > 0$, and $\mathcal{W}_s^\tau$, $\tau > 0$, respectively.

Furthermore, to illustrate the general results in Section 3 we consider two special cases describing a "regular decay" of the sequence υ, which characterizes the mapping properties of the associated conditional expectation operator. Precisely, we assume in the following the sequence υ to be either polynomially decreasing, i.e., $\upsilon_1 = 1$ and $\upsilon_j = |j|^{-2a}$, $j \geqslant 2$, or exponentially decreasing, i.e., $\upsilon_1 = 1$ and $\upsilon_j = \exp(-|j|^{2a})$, $j \geqslant 2$, for some $a > 0$. In the polynomial case easy calculus shows that any operator T satisfying the link condition (2.3), that is $T \in \mathcal{T}_d^\upsilon$, acts like integrating (a)-times and hence it is called *finitely smoothing* (c.f. Natterer [1984]). On the other hand in the exponential case it can easily be seen that $T \in \mathcal{T}_\upsilon^d$ implies $\mathcal{R}(T) \subset \mathcal{W}_r$ for all $r > 0$, therefore the operator T is called *infinitely smoothing* (c.f. Mair [1994]). It is worth noting that these are the usually studied cases in the literature (c.f. Hall and Horowitz [2005], Chen and Reiß [2008] or Johannes et al. [2007] in the context of nonparametric estimation of the structural function itself). However, the general results in the last section can be also applied considering more sophisticated sequences. Nevertheless, since in both cases the minimal regularity conditions given in Assumption 2.1 are satisfied, the lower bounds presented in the next assertion follow directly from Theorem 2.1.

THEOREM 4.1. *Under the assumptions of Theorem 2.1 we have for any estimator $\breve{\ell}$*

(i) *in the polynomial case, i.e. $\upsilon_1 = 1$ and $\upsilon_j = |j|^{-2a}$, $j \geqslant 2$, for some $a > 0$, that*

$$\sup_{U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{W}_p^\rho} \sup_{h \in \mathcal{W}_s^\tau} \{\mathbb{E}\, |\breve{\ell} - \ell_h(\varphi)|^2\} \gtrsim \max(n^{-(p+s)/(p+a)}, n^{-1}),$$

(ii) *in the exponential case, i.e. $\upsilon_1 = 1$ and $\upsilon_j = \exp(-|j|^{2a})$, $j \geqslant 2$, for some $a > 0$, that*

$$\sup_{U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{W}_p^\rho} \sup_{h \in \mathcal{W}_s^\tau} \{\mathbb{E}\, |\breve{\ell} - \ell_h(\varphi)|^2\} \gtrsim (\log n)^{-(p+s)/a}.$$

Let us introduce now the second basis $\{f_l\}_{l \geqslant 1}$, which is in this section also given by the trigonometric basis. In this situation the additional moment conditions formalized in Assumption A1-A3 are automatically fulfilled since both basis $\{e_j\}_{j \geqslant 1}$ and $\{f_l\}_{l \geqslant 1}$ are uniformly bounded. However, we suppose that the associated conditional expectation operator T satisfies the extended link condition (3.3), that is, $T \in \mathcal{T}_{d,D}^\upsilon$. Thereby, we restrict the set of possible joint distributions of (Z, W) to those having the trigonometric basis as optimal instruments. On the other hand, if the dimension m and the threshold α in the definition of the estimator $\widehat{\ell}_h$ given in (3.2) are chosen appropriate, then by applying Theorem 3.4 the rates of the lower bound given in the last assertion provide up to a constant also the upper bound of the risk of $\widehat{\ell}_h$, which is summarized in the next theorem. We have thus proved that these rates are optimal and the proposed estimator $\widehat{\ell}_h$ is minimax-optimal in both cases.

THEOREM 4.2. *Assume an n-sample of (Y, Z, W) from the model (1.1a–1.1b) with error term $U \in \mathcal{U}_\sigma$, $\sigma > 0$, and associated conditional expectation operator $T \in \mathcal{T}_{d,D}^\upsilon$, $d, D \geqslant 1$. Consider the estimator $\widehat{\ell}_h$ given in (3.2)*

(i) *in the polynomial case, i.e. $\upsilon_1 = 1$ and $\upsilon_j = |j|^{-2a}$, $j \geqslant 2$, for some $a > 0$, with $m \sim n^{1/(2p+2a)}$ and threshold $\alpha \sim n$. If in addition $p \geqslant 3/2$ then*

$$\sup_{\varphi \in \mathcal{W}_p^\rho, h \in \mathcal{W}_s^\tau} \{\mathbb{E}\, |\widehat{\ell}_h - \ell_h(\varphi)|^2\} \lesssim \max(n^{-(p+s)/(p+a)}, n^{-1}),$$

(ii) *in the exponential case, i.e. $\upsilon_1 = 1$ and $\upsilon_j = \exp(-|j|^{2a})$, $j \geqslant 2$, for some $a > 0$, with $m \sim (\log n)^{1/(2a)}$ and threshold $\alpha \sim n$. If in addition $p \geqslant 3/2$, then*

$$\sup_{\varphi \in \mathcal{W}_p^\rho, h \in \mathcal{W}_s^\tau} \{\mathbb{E}\,|\widehat{\ell}_h - \ell_h(\varphi)|^2\} \lesssim (\log n)^{-(p+s)/a}.$$

REMARK 4.1. We shall emphasize the interesting influence of the parameters p, s and a characterizing the smoothness of φ, h and the smoothing properties of T respectively. As we see from Theorem 4.1 and 4.2, if the value of a increases the obtainable optimal rate of convergence decreases. Therefore, the parameter a is often called *degree of ill-posedness* (c.f. Natterer [1984]). On the other hand, an increasing of the value $p + s$ leads to a faster optimal rate. In other words, as expected, values of a linear functional given by a smoother structural function or representer can be estimated faster. Moreover, in the polynomial case independent of the imposed smoothness assumption on the slope parameter (only $p \geqslant 3/2$ is needed) the parametric rate n^{-1} is obtained if and only if the representer is smoother than the degree of ill-posedness of T, i.e., $s \geqslant a$. The situation is different in the exponential case. As long as the representer h is only finitely times differentiable, then due to Theorem 4.1 and 4.2 the optimal rate of convergence is logarithmic. However, if we restrict the class of representers even more, e.g. by considering $\mathcal{F}_\omega^\tau$ with weights $\omega_1 := 1$, $\omega_j = \exp(|j|^{2q}), j \geqslant 2$, which contains only analytic functions given $q > 1$ (c.f. Kawata [1972]). Then faster rates are possible. Again independent of the imposed smoothness assumption on the structural parameter (again $p \geqslant 3/2$ is needed) the parametric rate n^{-1} is obtained if and only if the representer h is smoother than the degree of ill-posedness of T, e.g., $q \geqslant a$. Finally, in opposite to the polynomial case in the exponential case the smoothing parameter m does not depend on the value of p. It follows that the proposed estimator is automatically adaptive, i.e., it does not depend on an a-priori knowledge of the degree of smoothness of the structural function φ. However, the choice of the smoothing parameter depends on the smoothing properties of T, i.e., the value of a. $\qquad\square$

A Appendix

A.1 Proofs of Section 2.

Consider the conditional expectation operator T associated to the regressor Z and the instrument W, then $\mathbb{E}\,|[Te_j](W)|^2 = \|Te_j\|_W^2$, $j \in \mathbb{N}$. Therefore, if the link condition (2.3), that is $T \in \mathcal{T}_d^\upsilon$, is satisfied, then it follows that $\mathbb{E}\,|[Te_j](W)|^2 \asymp_d \upsilon_j$, for all $j \in \mathbb{N}$. This result will be used below without further reference. We shall prove at the end of this section the technical Lemma A.1 used in the next proof.

Proof of the lower bound.

PROOF OF THEOREM 2.1. We show below for any estimator $\breve{\ell}$ only based on an n-sample of (Y, Z, W) from the model (1.1a–1.1b) the following two lower bounds:

$$\sup_{U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \sup_{h \in \mathcal{F}_\omega^\tau} \mathbb{E}\,|\breve{\ell} - \ell_h(\varphi)|^2 \geqslant \delta_n^* \frac{\tau}{4\triangle} \min\left(\frac{1}{2d}, \frac{\rho}{\triangle}\right), \qquad (A.1)$$

$$\sup_{U \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \sup_{h \in \mathcal{F}_\omega^\tau} \mathbb{E}\,|\breve{\ell} - \ell_h(\varphi)|^2 \geqslant \frac{1}{n} \frac{\tau}{4} \min\left(\frac{1}{2d}, \rho\right). \qquad (A.2)$$

Consequently, the result follows by combination of these two lower bounds.

Proof of (A.1). Consider (Z, W) with associated $T \in \mathcal{T}_d^\upsilon$. Define the structural function $\varphi_* := [\varphi_*]_{m_*} e_{m_*}$, where m_* satisfies (2.4) for some $\triangle \geqslant 1$ and $[\varphi_*]_{m_*}$ is given in (A.7)

13

(Lemma A.1). Then from (A.9) in Lemma A.1 follows $\varphi_* \in \mathcal{F}_\gamma^\rho$ and thus $\varphi_*^{(\theta)} := \theta\varphi_* \in \mathcal{F}_\gamma^\rho$ with $\theta \in \{-1, 1\}$. Let V be a Gaussian random variable with mean zero and variance one ($V \sim \mathcal{N}(0, 1)$) which is independent of (Z, W). Then $U_\theta := [T\varphi_*^{(\theta)}](W) - \varphi_*^{(\theta)}(Z) + V$ belongs to $\mathcal{U}_\sigma$ for all sufficiently large σ, since $\mathbb{E}\, U_\theta|W = 0$ and $\mathbb{E}\,[U_\theta^4|W] \leqslant 8\{16\rho^2\eta + 3\}$. Consequently, for each θ i.i.d. copies (Y_i, Z_i, W_i), $1 \leqslant i \leqslant n$, of (Y, Z, W) with $Y := \varphi_*^\theta(Z) + U_\theta$ form an n-sample of the model (1.1a–1.1b) and we denote their joint distribution by P_θ. In case of P_θ the conditional distribution of Y_i given W_i is then Gaussian with mean $\theta[T\varphi_*](W_i)$ and variance 1. Then, it is easily seen that the log-likelihood of P_1 with respect to P_{-1} is given by

$$\log\Big(\frac{dP_1}{dP_{-1}}\Big) = \sum_{i=1}^n 2(Y_i - [T\varphi_*](W_i))[T\varphi_*](W_i) + \sum_{i=1}^n 2|[T\varphi_*](W_i)|^2.$$

Its expectation with respect to P_1 satisfies $\mathbb{E}_{P_1}[\log(dP_1/dP_{-1})] = 2n\|T\varphi_*\|^2 \leqslant 2nd[\varphi_*]_{m_*}^2 v_{m_*}$ by using $T \in \mathcal{T}_d^v$. In terms of Kullback-Leibler divergence this means $KL(P_1, P_{-1}) \leqslant 2\, d\, n\, [\varphi_*]_{m_*}^2 v_{m_*}$. Since the Hellinger distance $H(P_1, P_{-1})$ satisfies $H^2(P_1, P_{-1}) \leqslant KL(P_1, P_{-1})$ it follows from (A.9) in Lemma A.1 that

$$H^2(P_1, P_{-1}) \leqslant 2\, d\, n\, [\varphi_*]_{m_*}^2 v_{m_*} \leqslant 1. \tag{A.3}$$

Consider the Hellinger affinity $\rho(P_1, P_{-1}) = \int \sqrt{dP_1 dP_{-1}}$ then we obtain for any estimator $\breve{\ell}$ and for all $h \in \mathcal{F}_\omega^\tau$ that

$$\rho(P_1, P_{-1}) \leqslant \int \frac{|\breve{\ell} - \ell_h(\varphi_*^{(1)})|}{2|\ell_h(\varphi_*)|}\sqrt{dP_1 dP_{-1}} + \int \frac{|\breve{\ell} - \ell_h(\varphi_*^{(-1)})|}{2|\ell_h(\varphi_*)|}\sqrt{dP_1 dP_{-1}}$$

$$\leqslant \Big(\int \frac{|\breve{\ell} - \ell_h(\varphi_*^{(1)})|^2}{4|\ell_h(\varphi_*)|^2}dP_1\Big)^{1/2} + \Big(\int \frac{|\breve{\ell} - \ell_h(\varphi_*^{(-1)})|^2}{4|\ell_h(\varphi_*)|^2}dP_{-1}\Big)^{1/2}. \tag{A.4}$$

Due to the identity $\rho(P_1, P_{-1}) = 1 - \frac{1}{2}H^2(P_1, P_{-1})$ combining (A.3) with (A.4) yields

$$\Big\{\mathbb{E}_{P_1}|\breve{\ell} - \ell_h(\varphi_*^{(1)})|^2 + \mathbb{E}_{P_{-1}}|\breve{\ell} - \ell_h(\varphi_*^{(-1)})|^2\Big\} \geqslant \frac{1}{2}|\ell_h(\varphi_*)|^2. \tag{A.5}$$

Consider now the representer $h_* := [h_*]_{m_*}e_{m_*}$, where $[h_*]_{m_*}^2 := \tau/\omega_{m_*}$. Then by construction $h_* \in \mathcal{F}_\omega^\tau$ and $|\ell_{h_*}(\varphi_*)|^2 = [h_*]_{m_*}^2[\varphi_*]_{m_*}^2 \geqslant (\tau/\triangle)\min(1/(2d), \rho/\triangle)\delta_n^*$ by using (A.9) in Lemma A.1. From (A.5) together with the last estimate we conclude that

$$\sup_{U \in \mathcal{U}_\sigma}\sup_{\varphi \in \mathcal{F}_\gamma^\rho}\sup_{h \in \mathcal{F}_\omega^\tau}\mathbb{E}\,|\breve{\ell} - \ell_h(\varphi)|^2 \geqslant \sup_{\theta \in \{-1, 1\}}\mathbb{E}_{P_\theta}|\breve{\ell} - \ell_{h_*}(\varphi_*^{(\theta)})|^2$$

$$\geqslant \frac{1}{2}\Big\{\mathbb{E}_{P_1}|\breve{\ell} - \ell_{h_*}(\varphi_*^{(1)})|^2 + \mathbb{E}_{P_{-1}}|\breve{\ell} - \ell_{h_*}(\varphi_*^{(-1)})|^2\Big\}$$

$$\geqslant (1/4)\,[h_*]_{m_*}^2[\varphi_*]_{m_*}^2 \geqslant (\delta_n^*/4)(\tau/\triangle)\min(1/(2d), \rho/\triangle),$$

which proves (A.1). The proof of (A.2) is similar to the proof of (A.1), but uses (A.8) in Lemma A.1 rather than (A.9). To be more precise, we define the structural function $\varphi_* := [\varphi_*]_1 e_1$, and the representer $[h_*] := [h_*]_1 e_1$, where $[\varphi_*]_1$ and $[h_*]_1$ are given in (A.6) (Lemma A.1). Then by following along the same lines as in the proof of (A.1) we obtain (A.2), which completes the proof. $\qquad\square$

14

LEMMA A.1. *Consider sequences υ, γ and ω satisfying Assumption 2.1. Let m_* and δ_n^* be such that (2.4) holds true for some $\triangle \geqslant 1$. If we define*

$$[h_*]_1^2 := \tau, \qquad [\varphi_*]_1^2 := \frac{\xi_1}{n}, \quad with \quad \xi_1 := \min\left\{\frac{1}{2d}, \rho\right\}, \tag{A.6}$$

$$[h_*]_{m_*}^2 := \frac{\tau}{\omega_{m_*}} \quad and \quad [\varphi_*]_{m_*}^2 := \frac{\xi}{n \cdot \upsilon_{m_*}}, \quad where \quad \xi := \min\left\{\frac{1}{2d}, \frac{\rho}{\triangle}\right\}. \tag{A.7}$$

Then we have

$$2dn\upsilon_1[\varphi_*]_1^2 \leqslant 1; \ \gamma_1[\varphi_*]_1^2 \leqslant \rho; \ [h_*]_1^2\,[\varphi_*]_1^2 \geqslant (1/n)\,\tau\,\min(1/(2d), \rho); \tag{A.8}$$

$$2dn\upsilon_{m_*}[\varphi_*]_{m_*}^2 \leqslant 1; \ \gamma_{m_*}[\varphi_*]_{m_*}^2 \leqslant \rho; \ [h_*]_{m_*}^2[\varphi_*]_{m_*}^2 \geqslant \delta_n^*\,(\tau/\triangle)\min(1/(2d), \rho/\triangle). \tag{A.9}$$

PROOF. We only prove (A.9). The proof of (A.8) follows analogously and we omit the details. The first inequality in (A.9) is obtained trivially by using the definition of ξ. The second and third inequality in (A.9) follows from the definition of m_* and δ_n^* given in (2.4), i.e., $\gamma_{m_*}[\varphi_*]_{m_*}^2 \leqslant \xi \triangle$ and $[h_*]_{m_*}^2[\varphi_*]_{m_*}^2 = \xi\,\tau\,(\gamma_{m_*}/(n\upsilon_{m_*}))\,\delta_n^* \geqslant \xi(\tau/\triangle)\,\delta_n^*$, together with the definition of ξ, which completes the proof. $\qquad\square$

A.2 Proofs of Section 3.

We begin by defining and recalling notations to be used in the proofs of this section. Given $m > 0$, denote $\varphi_m := \sum_{j=1}^m [\varphi_m]_j e_j$ with $[\varphi_m]_{\underline{m}} = [T]_{\underline{m}}^{-1}[g]_{\underline{m}}$ which is well-defined since $[T]_{\underline{m}}$ is non singular. Then, the identities $[T(\varphi - \varphi_m)]_{\underline{m}} = 0$ and $[\varphi_m - E_m\varphi]_{\underline{m}} = [T]_{\underline{m}}^{-1}[TE_m^\perp\varphi]_{\underline{m}}$ hold true. Furthermore, let $[\Xi]_{\underline{m}} := [\widehat{T}]_{\underline{m}} - [T]_{\underline{m}}$ and define vector $[B]_{\underline{m}}$ and $[S]_{\underline{m}}$ by

$$[B]_j := \frac{1}{n}\sum_{i=1}^n U_i f_j(W_i), \ [S]_j := \frac{1}{n}\sum_{i=1}^n f_j(W_i)\{\varphi(Z_i) - [\varphi_m]_{\underline{m}}^t[e]_{\underline{m}}(Z_i)\}, \ 1 \leqslant j \leqslant m, \tag{A.10}$$

where $[\widehat{g}]_{\underline{m}} - [\widehat{T}]_{\underline{m}}[\varphi_m]_{\underline{m}} = [B]_{\underline{m}} + [S]_{\underline{m}}$. Note that $\mathbb{E}[B]_{\underline{m}} = 0$ due to the mean independence, i.e., $\mathbb{E}(U|W) = 0$, and that $\mathbb{E}[S]_{\underline{m}} = [T\varphi]_{\underline{m}} - [T\varphi_m]_{\underline{m}} = 0$. Moreover, let us introduce the events

$$\Omega := \{\|[\widehat{T}]_{\underline{m}}^{-1}\| \leqslant \alpha\}, \quad \Omega_{1/2} := \{\|[\Xi]_{\underline{m}}\|\|[T]_{\underline{m}}^{-1}\| \leqslant 1/2\}$$
$$\Omega^c := \{\|[\widehat{T}]_{\underline{m}}^{-1}\| > \alpha\} \quad and \quad \Omega_{1/2}^c = \{\|[\Xi]_{\underline{m}}\|\|[T]_{\underline{m}}^{-1}\| > 1/2\}. \tag{A.11}$$

Observe that $\Omega_{1/2} \subset \Omega$ in case $\alpha \geqslant 2\|[T]_{\underline{m}}^{-1}\|$. Indeed, if $\|[\Xi]_{\underline{m}}\|\|[T]_{\underline{m}}^{-1}\| \leqslant 1/2$ then the identity $[\widehat{T}]_{\underline{m}} = [T]_{\underline{m}}\{I + [T]_{\underline{m}}^{-1}[\Xi_n]_{\underline{m}}\}$ implies by the usual Neumann series argument that $\|[\widehat{T}]_{\underline{m}}^{-1}\| \leqslant 2\|[T]_{\underline{m}}^{-1}\|$. Thereby, if $\alpha \geqslant 2\|[T]_{\underline{m}}^{-1}\|$, 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 four technical Lemma (A.2 – A.5) which are used in the following proofs.

Proof of the consistency.

PROOF OF PROPOSITION 3.1. Let $\widehat{\ell}_h^\alpha := \ell_h(\varphi_m)\mathbb{1}\{\|[\widehat{T}]_{\underline{m}}^{-1}\| \leqslant \alpha\}$. Then the proof is based on the decomposition

$$\mathbb{E}\,|\widehat{\ell}_h - \ell_h(\varphi)|^2 \leqslant 2\{\mathbb{E}\,|\widehat{\ell}_h - \widehat{\ell}_h^\alpha|^2 + \mathbb{E}\,|\widehat{\ell}_h^\alpha - \ell_h(\varphi)|^2\}. \tag{A.12}$$

Under the assumption $\alpha \geqslant 2\|[T]_{\underline{m}}^{-1}\|$ we show below that for all $n \geqslant 1$

$$\mathbb{E}\,|\widehat{\ell}_h - \widehat{\ell}_h^{\alpha}|^2 \leqslant 2\|h\|^2 \cdot \frac{m\,\alpha^2}{n} \cdot (\eta \cdot \|\varphi - \varphi_m\|^2 + \sigma^2), \tag{A.13}$$

$$\mathbb{E}\,|\widehat{\ell}_h^{\alpha} - \ell_h(\varphi)|^2 \leqslant 2\|h\|^2 \left\{ \|\varphi - \varphi_m\|^2 + \|\varphi_m\|^2 \cdot \eta \cdot \frac{m^2\,\alpha^2}{n} \right\}. \tag{A.14}$$

Moreover, we have $\|\varphi - \varphi_m\| = o(1)$ as $m \to \infty$, which can be realized as follows. Consider the decomposition $\|\varphi - \varphi_m\| \leqslant \|E_m^{\perp}\varphi\| + \|E_m\varphi - \varphi_m\|$, where $\|E_m^{\perp}\varphi\| = o(1)$ by using Lebesgue's dominated convergence theorem. The consistency of φ_m follows then from $\|E_m\varphi - \varphi_m\| \leqslant \|E_m^{\perp}\varphi\| \sup_m\|T_m^{-1}F_mTE_m^{\perp}\| = O(\|E_m^{\perp}\varphi\|)$. Consequently, the conditions on m and α ensure the convergence to zero as $n \to \infty$ of the bound given in (A.13) and (A.14), respectively, which proves the result.

Proof of (A.13). By making use of the identity $[\widehat{g}]_{\underline{m}} - [\widehat{T}]_{\underline{m}}[\varphi_m]_{\underline{m}} = [B]_{\underline{m}} + [S]_{\underline{m}}$ the Cauchy-Schwarz inequality and $\|[\widehat{T}]_{\underline{m}}^{-1}\|\mathbb{1}_{\Omega} \leqslant \alpha$ imply together

$$\mathbb{E}\,|\widehat{\ell}_h - \widehat{\ell}_h^{\alpha}|^2 \leqslant \|h\|^2 \cdot \alpha^2 \cdot \mathbb{E}\,\|[B]_{\underline{m}} + [S]_{\underline{m}}\|^2.$$

and hence (A.13) follows from (A.21) and (A.22) in Lemma A.2.

The estimate (A.14) follows from the decomposition

$$\mathbb{E}\,|\widehat{\ell}_h^{\alpha} - \ell_h(\varphi)|^2 \leqslant 2\|h\|^2\{\|\varphi - \varphi_m\|^2 + \|\varphi_m\|^2 P(\Omega^c)\},$$

where we claim that $P(\Omega^c) \leqslant 4\eta\,m^2\|[T]_{\underline{m}}^{-1}\|^2/n \leqslant \eta\,m^2\,\alpha^2/n$. Indeed, since $\alpha \geqslant 2\|[T]_{\underline{m}}^{-1}\|$ it follows that $\Omega^c \subset \Omega_{1/2}^c$ and thus by applying Markov's inequality we obtain from (A.23) in Lemma A.2 the estimate, which completes the proof. $\qquad\square$

PROOF OF COROLLARY 3.2. By combination of the identity $[\varphi_m - E_m\varphi]_{\underline{m}} = [T]_{\underline{m}}^{-1}[TE_m^{\perp}\varphi]_{\underline{m}}$ and the estimate (A.31) in the proof of Lemma A.4 with $\gamma \equiv 1$ the extended link condition (3.3), that is $T \in \mathcal{T}_{d,D}^{v}$, implies $\|T_m^{-1}F_mTE_m^{\perp}\|^2 = \sup_{\|\varphi\|=1}\|E_m\varphi - \varphi_m\|^2 \leqslant D\,d$. Moreover, $2\|[T]_{\underline{m}}^{-1}\| \leqslant 2\|[\mathrm{Diag}(v)]_{\underline{m}}^{-1/2}\|\|[\mathrm{Diag}(v)]_{\underline{m}}^{1/2}[T]_{\underline{m}}^{-1}\| \leqslant 2\sqrt{D/v_m} = \alpha$ since v is non increasing. By using these estimates the result follows directly from Proposition 3.1. $\qquad\square$

Proof of the upper bound.

PROOF OF THEOREM 3.3. Our proof starts with the observation that by using the definition (2.4) of m_*, that is, $1/v_{m_*} \leqslant n\triangle/\gamma_{m_*}$, the condition on the dimension $m = m_*$ implies that $m^3/(nv_m) \leqslant \triangle m^3/\gamma_m$ and that the threshold satisfies both $\alpha^2 = n\max(1, 4\,D\,\triangle/\gamma_m) \geqslant 4\|[T]_m^{-1}\|^2$ and $\alpha^2/n \leqslant 4\,D\,\triangle$. On the other hand, we show below under the condition $\alpha \geqslant 2\|[T]_{\underline{m}}^{-1}\|$ the following two bounds:

$$\mathbb{E}\,|\widehat{\ell}_h - \widehat{\ell}_h^{\alpha}|^2 \leqslant (C/n)\,\|[\mathrm{Diag}(\omega v)]_{\underline{m}}^{-1}\|\,\|h\|_{\omega}^2\,D\,\eta^4\,(\sigma^2 + \Gamma\,\|\varphi - \varphi_m\|_{\gamma}^2)$$
$$\left\{ 1 + 4D\frac{m^3}{v_mn} + \frac{\alpha^2m^3}{n}|P(\Omega_{1/2}^c)|^{1/4} \right\}, \tag{A.15}$$

$$\mathbb{E}\,|\widehat{\ell}_h^{\alpha} - \ell_h(\varphi)|^2 \leqslant 2\Big\{ \|h\|^2\|\varphi_m\|^2 P(\Omega_{1/2}^c) + \gamma_m^{-1}\max(\omega_m^{-1}, v_m)\,2D\,d\,\Lambda\|\varphi\|_{\gamma}^2\,\|h\|_{\omega}^2 \Big\}, \tag{A.16}$$

for some generic constant $C > 0$ uniformly for all $n \in \mathbb{N}$, where $\|\varphi_m\|^2 \leqslant 2\{\|\varphi - \varphi_m\|^2 + \|\varphi\|^2\} \leqslant 2\{2Dd + 1\}\|\varphi\|_\gamma^2 \leqslant 6Dd\|\varphi\|_\gamma^2$ and $\|\varphi - \varphi_m\|_\gamma^2 \leqslant 2Dd\|\varphi\|_\gamma^2$ due to (A.29) in Lemma A.4. Thus, by using $\Omega_{1/2}^c \subset \{\|[\Xi]_{\underline{m}}\|^2 > \upsilon_m/(4D)\}$ and $\|[\mathrm{Diag}(\omega \upsilon)]_{\underline{m}}^{-1/2}\|^2 \leqslant \Lambda \upsilon_m^{-1} \max(\omega_m^{-1}, \upsilon_m)$ (Assumption 2.1) it follows again from the decomposition (A.12) by combination of (A.15) and (A.16) that uniformly for all $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_\omega^\tau$

$$\mathbb{E}\,|\widehat{\ell}_h - \ell_h(\varphi)|^2 \leqslant C\,d\,D^2\,\Lambda\,\eta^4\,\{\sigma^2 + d\,D\,\Gamma\}\left\{1 + \triangle\,\frac{m^3}{\gamma_m} + \triangle\,m^3|P(\Omega_{1/2}^c)|^{1/4}\right\}\tau\,\rho$$
$$\cdot\left\{(n\upsilon_m)^{-1}\max(\omega_m^{-1}, \upsilon_m) + \gamma_m^{-1}\max(\omega_m^{-1}, \upsilon_m) + P(\|[\Xi]_{\underline{m}}\|^2 > \upsilon_m/(4D))\right\}.$$

The result follows now from $\{(n\upsilon_{m_*})^{-1} + \gamma_{m_*}^{-1}\}\max(\omega_{m_*}^{-1}, \upsilon_{m_*}) \leqslant 2\triangle\max(\delta_n^*, 1/n)$ by using the definition of δ_n^* given in (2.4).

Proof of (A.15). By making use of the identity $[\widehat{g}]_{\underline{m}} - [\widehat{T}]_{\underline{m}}[\varphi_m]_{\underline{m}} = [B]_{\underline{m}} + [S]_{\underline{m}}$ it follows

$$\widehat{\ell}_h - \widehat{\ell}_h^\alpha = [h]_{\underline{m}}^t\left\{[T]_{\underline{m}}^{-1} + [T]_{\underline{m}}^{-1}([T]_{\underline{m}} - [\widehat{T}]_{\underline{m}})[\widehat{T}]_{\underline{m}}^{-1}\right\}\{[B]_{\underline{m}} + [S]_{\underline{m}}\}\mathbb{1}_\Omega$$
$$= [h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\{[B]_{\underline{m}} + [S]_{\underline{m}}\}\mathbb{1}_\Omega - [h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}[\Xi]_{\underline{m}}[\widehat{T}]_{\underline{m}}^{-1}\{[B]_{\underline{m}} + [S]_{\underline{m}}\}\mathbb{1}_\Omega$$

where (A.24) and (A.25) in Lemma A.3 with $z^t := [h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}/\|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\|$ imply together

$$\mathbb{E}\,|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\{[B]_{\underline{m}} + [S]_{\underline{m}}\}|^2 \leqslant (2/n)\cdot\|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\|^2\,\eta^2\,(\sigma^2 + \Gamma\,\|\varphi - \varphi_m\|_\gamma^2). \quad (\text{A.17})$$

On the other hand we show below that there exists a generic constant $C > 0$ such that

$$\mathbb{E}\,|[h]_{\underline{m}}^t[\widehat{T}]_{\underline{m}}^{-1}[\Xi]_{\underline{m}}[T]_{\underline{m}}^{-1}\{[B]_{\underline{m}} + [S]_{\underline{m}}\}\mathbb{1}_\Omega|^2 \leqslant (C/n)\,\|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\|^2\,\eta^4\,(\sigma^2 + \Gamma\,\|\varphi - \varphi_m\|_\gamma^2)$$
$$\left\{4D\frac{m^3}{\upsilon_m n} + \frac{\alpha^2 m^3}{n}|P(\Omega_{1/2}^c)|^{1/4}\right\} \quad (\text{A.18})$$

Consequently, the inequality (A.15) follows by combination of (A.17) and (A.18) together with $\|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\|^2 \leqslant \|[h]_{\underline{m}}^t[\mathrm{Diag}(\upsilon)]_{\underline{m}}^{-1/2}\|^2 D \leqslant \|h\|_\omega^2\,\|[\mathrm{Diag}(\omega\upsilon)]_{\underline{m}}^{-1/2}\|^2 D$ since $T \in \mathcal{T}_{d,D}^\upsilon$.

The proof of (A.18) starts with the observations that $T \in \mathcal{T}_{d,D}^\upsilon$ implies $\|[\widehat{T}]_m^{-1}\|\mathbb{1}_{\Omega_{1/2}} \leqslant 2\|[T]_m^{-1}\| \leqslant 2\sqrt{D/\upsilon_m}$ and that $\|[\widehat{T}]_m^{-1}\|\mathbb{1}_\Omega \leqslant \alpha$. By using these estimates we obtain

$$\mathbb{E}\,|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}[\Xi]_{\underline{m}}[\widehat{T}]_{\underline{m}}^{-1}\{[B]_{\underline{m}} + [S]_{\underline{m}}\}\mathbb{1}_\Omega|^2$$
$$\leqslant \|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\|^2\left\{4D\upsilon_m^{-1}\mathbb{E}\,\|[\Xi]_{\underline{m}}\|^2\|[B]_{\underline{m}} + [S]_{\underline{m}}\|^2\mathbb{1}_{\Omega_{1/2}} + \alpha^2\mathbb{E}\,\|[\Xi]_{\underline{m}}\|^2\|[B]_{\underline{m}} + [S]_{\underline{m}}\|^2\mathbb{1}_{\Omega_{1/2}^c}\right\}$$
$$\leqslant \|[h]_{\underline{m}}^t[T]_{\underline{m}}^{-1}\|^2\left\{4D\upsilon_m^{-1}\big(\mathbb{E}\,\|[\Xi]_{\underline{m}}\|^4\big)^{1/2} + \alpha^2\big(\mathbb{E}\,\|[\Xi]_{\underline{m}}\|^8\big)^{1/4}P(\Omega_{1/2}^c)^{1/4}\right\}\big(\mathbb{E}\,\|[B]_{\underline{m}} + [S]_{\underline{m}}\|^4\big)^{1/2}.$$

Consequently, (A.26), (A.27) and (A.28) in Lemma A.28 imply together (A.18).

Proof of (A.16). Following along the lines of the proof of (A.14) we obtain

$$\mathbb{E}\,|\widehat{\ell}_h^\alpha - \ell_h(\varphi)|^2 \leqslant 2\{|\langle h, \varphi - \varphi_m\rangle|^2 + \|h\|^2\|\varphi_m\|^2 P(\Omega_{1/2}^c)\}.$$

Then, due to $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_\omega^\tau$ the estimate (A.30) in Lemma A.4 implies (A.16), which completes the proof. $\qquad\square$

PROOF OF THEOREM 3.4. The result follows from Theorem 3.3 since $m_*^3 \gamma_{m_*}^{-1} = O(1)$ by using the additional properties (3.4) and

$$m_*^3 \Big| P\Big(\|[\widehat{T}]_{\underline{m_*}} - [T]_{\underline{m_*}}\|^2 > \upsilon_{m_*}/(4D) \Big) \Big|^{1/4} = O(1), \tag{A.19}$$

$$P\Big(\|[\widehat{T}]_{\underline{m_*}} - [T]_{\underline{m_*}}\|^2 > \frac{\upsilon_{m_*}}{4D} \Big) = O(\max(\delta_n^*, 1/n)), \tag{A.20}$$

which can be realized as follows. Consider first (A.20). From the definition of m_* follows $n\upsilon_{m_*}\gamma_{m_*}^{-1} \geqslant \triangle^{-1}$. By using this estimate together with (A.35) in Lemma A.5 we conclude

$$m_*^3 \Big| P\Big(\|[\widehat{T}]_{\underline{m_*}} - [T]_{\underline{m_*}}\|^2 > \upsilon_{m_*}/(4D) \Big) \Big|^{1/4}$$
$$\leqslant 2^{1/4} \exp\{-(n\upsilon_{m_*}m_*^{-2})/(80D\eta^2) + (7/2)\log m_*\}$$
$$\leqslant 2^{1/4} \exp\Big\{-\frac{\gamma_{m_*}}{m_*^2}\Big(\frac{1}{80D\eta^2\triangle} - \frac{m_*^3}{\gamma_{m_*}}\frac{(7/2)\log m_*}{m_*}\Big)\Big\}.$$

Consequently, (A.19) follows also from the conditions (3.4), that is, $m_*^3 \gamma_{m_*}^{-1} = O(1)$. Consider (A.20). From the definition of m_* moreover follows $\min(\delta_n^{*-1}, n) \leqslant \triangle\gamma_{m_*} \min(\upsilon_{m_*}^{-1}, \omega_{m_*})$. This estimate and $n\upsilon_{m_*}/\gamma_{m_*} \geqslant 1/\triangle$ together with (A.35) in Lemma A.5 implies now

$$\min(\delta_n^{*-1}, n)P(\|[\widehat{T}]_{\underline{m_*}} - [T]_{\underline{m_*}}\|^2 > \upsilon_{m_*}/(4D))$$
$$\leqslant 2\exp\{-(n\upsilon_{m_*}m_*^{-2})/(20D\eta^2) + 2\log m_* + \log\min(\delta_n^{*-1}, n)\}$$
$$\leqslant 2\exp\Big\{-\frac{\gamma_{m_*}}{m_*^2}\Big(\frac{1}{20D\eta^2\triangle} - \frac{m_*^3}{\gamma_{m_*}}\frac{2\log m_* + \log\triangle}{m_*}$$
$$- \frac{m_*^2 \log\gamma_{m_*}}{\gamma_{m_*}} - \frac{m_*^2 \log\min(\upsilon_{m_*}^{-1}, \omega_{m_*})}{\gamma_{m_*}}\Big)\Big\}.$$

Thus, the estimate (A.20) follows again by using the additional properties (3.4), which completes the proof. $\square$

Technical assertions.

The following paragraph gathers technical results used in the proof of Section 3.

LEMMA A.2. *Suppose that the error term U satisfies $\mathbb{E}[U^2|W] \leqslant \sigma^2$, $\sigma > 0$ and that the joint distribution of (Z, W) fulfills Assumption A1. Then for all $m \in \mathbb{N}$ we have*

$$\mathbb{E}\|[B]_{\underline{m}}\|^2 \leqslant (m/n) \cdot \sigma^2, \tag{A.21}$$

$$\mathbb{E}\|[S]_{\underline{m}}\|^2 \leqslant (m/n) \cdot \eta \cdot \|\varphi - \varphi_m\|^2, \tag{A.22}$$

$$\mathbb{E}\|[\Xi]_{\underline{m}}\|^2 \leqslant (m^2/n) \cdot \eta. \tag{A.23}$$

PROOF. Proof of (A.21) and (A.22). Consider $\mathbb{E}\|[B]_{\underline{m}}\|^2 = \sum_{j=1}^m \mathbb{E}|(1/n)\sum_{i=1}^n U_i f_j(W_i)|^2$. By using the mean independence (Assumption A1), i.e., $\mathbb{E}[U|W] = 0$, it follows that the random variables $(U_i f_j(W_i))$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero, thus $\mathbb{E}\|[B]_{\underline{m}}\|^2 = (m/n)\mathbb{E}|Uf_j(W)|^2$. Thus, $\mathbb{E}(U^2|W) \leqslant \sigma^2$ and $\mathbb{E}f_j^2(W) = 1$ imply (A.21). Consider (A.22), where for each $1 \leqslant j \leqslant m$ the random variables $(\{\varphi(Z_i) - [\varphi_m]_{\underline{m}}^t[e]_{\underline{m}}(Z_i)\}f_j(W_i))$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero. Thus $\mathbb{E}\|[S]_{\underline{m}}\|^2 \leqslant (m/n)\sup_j \mathbb{E}|\{\varphi(Z) - [\varphi_m]_{\underline{m}}^t[e]_{\underline{m}}(Z)\}f_j(W)|^2$ and, hence (A.22) follows from Assumption A1, i.e., $\sup_{j\in\mathbb{N}}\mathbb{E}[f_j^2(W)|Z] \leqslant \eta$, together with $\mathbb{E}\{\varphi(Z) - [\varphi_m]_{\underline{m}}^t[e]_{\underline{m}}(Z)\}^2 = \|\varphi - \varphi_m\|^2$.

Proof of (A.23). Let $1 \leqslant j, l \leqslant m$. Then $(e_j(Z_i)f_l(W_i) - [T]_{j,l})$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero and $\mathbb{E}\,[\Xi]^2_{j,l} = n^{-1}\mathbb{E}\,\{e_j(Z)f_l(W) - [T]_{j,l}\}^2 \leqslant n^{-1}\eta$ by Assumption A1. Consequently, (A.23) follows from the estimate $\mathbb{E}\,\|[\Xi]_{\underline{m}}\|^2 \leqslant \sum_{j,l=1}^m \mathbb{E}\,[\Xi]^2_{j,l}$, which completes the proof. $\qquad\square$

LEMMA A.3. *Let* $\mathbb{S}^m := \{z \in \mathbb{R}^m : z^t z = 1\}$. *Suppose that* $U \in \mathcal{U}_\sigma$, $\sigma > 0$ *and that the joint distribution of* (Z, W) *satisfies Assumption A2. If in addition* $\varphi \in \mathcal{F}^\rho_\gamma$ *with* $\Gamma = \sum_{j=1}^\infty \gamma_j^{-1} < \infty$, *then there exists a constant* $C > 0$ *such that for all* $m \in \mathbb{N}$

$$\sup_{z \in \mathbb{S}^m} \{\mathbb{E}\,|z^t\,[B]_{\underline{m}}|^2\} \leqslant (1/n)\,\sigma^2, \tag{A.24}$$

$$\sup_{z \in \mathbb{S}^m} \{\mathbb{E}\,|z^t\,[S]_{\underline{m}}|^2\} \leqslant (1/n)\,\eta^2\,\Gamma\,\|\varphi - \varphi_m\|_\gamma^2 \tag{A.25}$$

$$\mathbb{E}\,\|[B]_{\underline{m}}\|^4 \leqslant C \cdot \left((m/n) \cdot \sigma^2 \cdot \eta^2\right)^2, \tag{A.26}$$

$$\mathbb{E}\,\|[S]_{\underline{m}}\|^4 \leqslant C \cdot \left((m/n) \cdot \eta^2 \cdot \Gamma \cdot \|\varphi - \varphi_m\|_\gamma^2\right)^2, \tag{A.27}$$

$$\mathbb{E}\,\|[\Xi]_{\underline{m}}\|^8 \leqslant C \cdot \left((m^2/n) \cdot \eta^2\right)^4. \tag{A.28}$$

PROOF. Consider (A.24). Let $z \in \mathbb{S}^m$. By using the mean independence, i.e., $\mathbb{E}\,[U|W] = 0$, it follows that the random variables $(U_i \sum_{j=1}^m z_j f_j(W_i))$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero. Therefore, we have $\mathbb{E}\,|z^t\,[B]_{\underline{m}}|^2 = (1/n)\mathbb{E}\,|U \sum_{j=1}^m z_j f_j(W)|^2$. Then (A.24) follows from $\mathbb{E}\,(U^2|W) \leqslant (\mathbb{E}\,(U^4|W))^{1/2} \leqslant \sigma^2$ and $\mathbb{E}\,[f_j(W)f_l(W)] = \delta_{jl}$ with $\delta_{jl} = 1$ if $j = l$ and zero otherwise. Consider (A.25). Since $(f_j(W)\{\varphi(Z) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z)\})$ has mean zero, it follows that $(\{\varphi(Z_i) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z_i)\} \sum_{j=1}^m z_j f_j(W_i))$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero. Thus, $\mathbb{E}\,|z^t\,[S]_{\underline{m}}|^2 = (1/n)\mathbb{E}\,|\{\varphi(Z) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z)\} \sum_{j=1}^m z_j f_j(W)|^2$. Then (A.25) follows from Assumption A2 (i), i.e., $\sup_{l \in \mathbb{N}} \mathbb{E}\,[|e_l(Z)|^2|W] \leqslant \eta^2$, and $\mathbb{E}\,[f_j(W)f_l(W)] = \delta_{jl}$. Indeed, by using the Cauchy-Schwarz inequality and that $\sum_{j \in \mathbb{N}} \gamma_j^{-1} = \Gamma < \infty$ we have

$$\mathbb{E}\,|\{\varphi(Z) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z)\} \sum_{j=1}^m z_j f_j(W)|^2 \leqslant \|\varphi - \varphi_m\|_\gamma^2 \sum_{l \in \mathbb{N}} \gamma_l^{-1} \mathbb{E}\,|e_l(Z) \sum_{j=1}^m z_j f_j(W)|^2$$

$$\leqslant \|\varphi - \varphi_m\|_\gamma^2\,\eta^2 \sum_{l \in \mathbb{N}} \gamma_l^{-1} \sum_{j=1}^m z_j^2 = \|\varphi - \varphi_m\|_\gamma^2\,\eta^2\,\Gamma.$$

Proof of (A.26). Since $\mathbb{E}\,\|[B]_{\underline{m}}\|^4 \leqslant m \sum_{j=1}^m \mathbb{E}\,|(1/n)\sum_{i=1}^n U_i f_j(W_i)|^4$, where for each $1 \leqslant j \leqslant m$ the random variables $(U_i f_j(W_i))$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero. It follow from Theorem 2.10 in Petrov [1995] that $\mathbb{E}\,|(1/n)\sum_{i=1}^n U_i f_j(W_i)|^4 \leqslant C n^{-2}\mathbb{E}\,|U f_j(W)|^4$ for some generic constant $C > 0$. Thus, by using $\mathbb{E}\,(U^4|W) \leqslant \sigma^4$ and $\sup_{j \in \mathbb{N}} \mathbb{E}\,|f_j(W)|^4 \leqslant \eta^4$ (Assumption A2 (i)), we obtain (A.26). The proof of (A.27) follows in analogy to the proof of (A.26). Observe that for each $1 \leqslant j \leqslant m$, $(\{\varphi(Z_i) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z_i)\}f_j(W_i))$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero and $\mathbb{E}\,|\{\varphi(Z_i) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z_i)\}f_j(W_i)|^4 \leqslant \eta^4\,\Gamma^2\,\|\varphi - \varphi_m\|_\gamma^4$, which can be realized as follows. Since $[T(\varphi - \varphi_m)]_j = 0$ it follows that $\{\varphi(Z) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z)\}f_j(W) = \sum_{l \in \mathbb{N}}[\varphi - \varphi_m]_l\{e_l(Z)f_j(W) - [T]_{j,l}\}$. Furthermore, by using Assumption A2 (ii), i.e., $\sup_{j,l \in \mathbb{N}} \mathbb{E}\,|e_l(Z)f_j(W) - [T]_{j,l}|^4 \leqslant 2\eta^4$, the Cauchy-Schwarz inequality implies

$$\mathbb{E}\,|\{\varphi(Z) - [\varphi_m]^t_{\underline{m}}[e]_{\underline{m}}(Z)\}f_j(W)|^4 \leqslant \|\varphi - \varphi_m\|_\gamma^4\,\mathbb{E}\,\left|\sum_{l \in \mathbb{N}} \gamma_l^{-1}|e_l(Z)f_j(W) - [T]_{j,l}|^2\right|^2$$

$$\leqslant \|\varphi - \varphi_m\|_\gamma^4\,\Gamma^2\,2\eta^4.$$

Proof of (A.28). The random variables $(e_l(Z_i)f_j(W_i) - [T]_{j,l})$, $1 \leqslant i \leqslant n$, are i.i.d. with mean zero for each $1 \leqslant j,l \leqslant m$. Hence, Theorem 2.10 in Petrov [1995] together with Assumption A2 (ii), i.e., $\sup_{j,l\in\mathbb{N}} \mathbb{E}\, |e_l(Z)f_j(W) - [T]_{j,l}|^8 \leqslant 8!\eta^8$, implies $\sum_{j,l=1}^{m} \mathbb{E}\,[\Xi]_{j,l}^8 \leqslant Cm^2 n^{-4}\eta^8$. Consequently, (A.28) follows from the estimate $\mathbb{E}\,\|[\Xi]_{\underline{m}}\|^8 \leqslant m^6 \sum_{j,l=1}^{m} \mathbb{E}\,[\Xi]_{j,l}^8$, which completes the proof. $\qquad\square$

LEMMA A.4. *Let $g = T\varphi$ and denote $\varphi_m := [T]_{\underline{m}}^{-1}[g]_{\underline{m}}$, $m \in \mathbb{N}$. If $T \in \mathcal{T}_{d,D}^\upsilon$ and $\varphi \in \mathcal{F}_\gamma^\rho$, then for all $0 \leqslant s \leqslant 1$ we obtain*

$$\sup_{m\in\mathbb{N}} \{\gamma_m^{1-s}\, \|\varphi - \varphi_m\|_{\gamma^s}^2\} \leqslant 2\, D\, d\, \rho. \tag{A.29}$$

If in addition $h \in \mathcal{F}_\omega^\tau$, then under Assumption 2.1 we have

$$\sup_{m\in\mathbb{N}} \{\gamma_m \min(\omega_m, \upsilon_m^{-1})\, |\langle h, \varphi - \varphi_m\rangle|^2\} \leqslant 2\, \Lambda\, D\, d\, \rho\, \tau. \tag{A.30}$$

PROOF. Consider the decomposition

$$\|\varphi - \varphi_m\|_{\gamma^s}^2 \leqslant 2\{\|\varphi - E_m\varphi\|_{\gamma^s}^2 + \|E_m\varphi - \varphi_m\|_{\gamma^s}^2\}.$$

Since (γ_j^{s-1}) is monotonically decreasing it follows that $\|\varphi - E_m\varphi\|_{\gamma^s}^2 \leqslant \gamma_m^{s-1} \|\varphi\|_\gamma^2$, while we show below that

$$\|E_m\varphi - \varphi_m\|_{\gamma^s}^2 \leqslant D\, d\, \gamma_m^{s-1} \|\varphi\|_\gamma^2. \tag{A.31}$$

Consequently, by combination of these two bounds the condition $\varphi \in \mathcal{F}_\gamma^\rho$, i.e., $\|\varphi\|_\gamma^2 \leqslant \rho$, implies (A.29). Consider (A.31). Since $T \in \mathcal{T}_{d,D}^\upsilon$, i.e., $\sup_{m\in\mathbb{N}}\|[\mathrm{Diag}(\upsilon)]_{\underline{m}}^{1/2}[T]_{\underline{m}}^{-1}\|^2 \leqslant D$ and $\|Tf\|^2 \leqslant d\|f\|_\upsilon^2$ for all $f \in L_Z^2$, the identity $[E_m\varphi - \varphi_m]_{\underline{m}} = -[T]_{\underline{m}}^{-1}[TE_m^\perp\varphi]_{\underline{m}}$ implies $\|E_m\varphi - \varphi_m\|_\upsilon^2 \leqslant D\|TE_m^\perp\varphi\|^2$ and hence

$$\|E_m\varphi - \varphi_m\|_\upsilon^2 \leqslant D\, d\, \gamma_m^{-1}\upsilon_m\|\varphi\|_\gamma^2 \tag{A.32}$$

because $(\gamma_j^{-1}\upsilon_j)$ is monotonically decreasing. Furthermore, since $(\gamma_j^s\upsilon_j^{-1})$ is monotonically increasing we have $\|E_m\varphi - \varphi_m\|_{\gamma^s}^2 \leqslant \gamma_m^s\upsilon_m^{-1}\|E_m\varphi - \varphi_m\|_\upsilon^2$. The inequality (A.31) follows now by combination of the last estimate and (A.32).

Proof of (A.30). By applying the Cauchy-Schwarz inequality we have

$$|\langle h, \varphi - E_m\varphi\rangle|^2 \leqslant \omega_m^{-1}\gamma_m^{-1}\, \|h\|_\omega^2\|\varphi\|_\gamma^2 \tag{A.33}$$

and by using (A.32) it follows

$$|\langle h, E_m\varphi - \varphi_m\rangle|^2 \leqslant \|h\|_\omega^2\|[\mathrm{Diag}(\omega)]_{\underline{m}}^{-1/2}[\mathrm{Diag}(\upsilon)]_{\underline{m}}^{-1/2}\|^2\|(E_m\varphi - \varphi_m)\|_\upsilon^2$$
$$\leqslant \|h\|_\omega^2 \{\sup_{1\leqslant j\leqslant m} 1/(\omega_j\upsilon_j)\}\, D\, d\, \gamma_m^{-1}\upsilon_m\, \|\varphi\|_\gamma^2. \tag{A.34}$$

Since under Assumption 2.1 there exist a constant Λ such that for all $m \in \mathbb{N}$ holds $\upsilon_m \sup_{1\leqslant j\leqslant m}\{1/(\omega_j\upsilon_j)\} \leqslant \Lambda \max(\omega_m^{-1}, \upsilon_m)$ the assertion (A.30) follows from (A.33) and (A.33), which completes the proof. $\qquad\square$

LEMMA A.5. *Suppose that the joint distribution of (Z,W) satisfies Assumption A3. If in addition the sequence υ fulfills Assumption 2.1, then for all $m \in \mathbb{N}$ we have*

$$P(\|[\Xi]_{\underline{m}}\|^2 > \upsilon_m/(4D)) \leqslant 2\exp\{-(n\upsilon_m/m^2)/(20D\eta^2) + 2\log m\}. \tag{A.35}$$

PROOF. Our proof starts with the observation that for all $j, l \in \mathbb{N}$ the condition (iii) in Assumption A3 implies for all $t > 0$

$$P(|e_j(Z)f_l(W) - \mathbb{E}\left[e_j(Z)f_l(W)\right]| \geqslant t) \leqslant 2\exp\{-t^2/(4n\operatorname{Var}(e_j(Z)f_l(W)) + 2\eta t)\},$$

which is just Bernstein's inequality (a detailed discussion can be found, for example, in Bosq [1998]). Therefore, the condition $\sup_{j,l\in\mathbb{N}} \operatorname{Var}(e_j(Z)f_l(W)) \leqslant \eta^2$ (Assumption A2 (ii)) implies now for all $t > 0$

$$\sup_{j,l\in\mathbb{N}} P(|e_j(Z)f_l(W) - \mathbb{E}\left[e_j(Z)f_l(W)\right]| \geqslant t) \leqslant 2\exp\{-t^2/(4n\eta^2 + 2\eta t)\}. \qquad \text{(A.36)}$$

On the other hand, it is well-known that $m^{-1}\|[A]_{\underline{m}}\| \leqslant \max_{1\leqslant j,l\leqslant m} |[A]_{j,l}|$ for any $m \times m$ matrix $[A]_{\underline{m}}$. Combining the last estimate and (A.36) we obtain for all $t > 0$

$$P(m^{-1}\|[\Xi]_{\underline{m}}\| \geqslant t) \leqslant \sum_{j,l=1}^{m} P(|e_j(Z)f_l(W) - \mathbb{E}\left[e_j(Z)f_l(W)\right]| \geqslant nt)$$

$$\leqslant 2\exp\{-(nt^2)/(4\eta^2 + 2\eta t) + 2\log m\}.$$

From the last estimate it follows now

$$P(\|[\Xi]_{\underline{m}}\|^2 > \upsilon_m/(4D)) \leqslant 2\exp\{-(n\upsilon_m/m^2)/(4D(4\eta^2 + (\eta/D^{1/2})(\upsilon_m^{1/2}/m)) + 2\log m\},$$

which together with Assumption 2.1, that is, $\upsilon_m^{1/2}/m \leqslant \eta D^{1/2}$, implies the result. $\qquad \square$

A.3 Proofs of Section 4

The lower bounds.

PROOF OF THEOREM 4.1. Observe that $\mathcal{W}_p^\rho = \mathcal{F}_\gamma^\rho$ and $\mathcal{W}_s^\tau = \mathcal{F}_\omega^\tau$ with weights $\gamma = (\gamma_j)_{j\geqslant 1}$ and $\omega = (\omega_j)_{j\geqslant 1}$ given by $\gamma_1 := 1, \gamma_j := |j|^{2p}$ and $\omega_1 := 1, \omega_j := |j|^{2s}, j \geqslant 2$, respectively. Obviously, the sequences γ, ω and υ given in (i) by $\upsilon = 1, \upsilon_j = |j|^{-2a}$ and (ii) by $\upsilon = 1, \upsilon_j = \exp(-|j|^{2a}), j \geqslant 2$, satisfy Assumption 2.1. Furthermore, in case (i) we have $1/(\gamma_{m_*}\upsilon_{m_*}) = m_*^{2a+2p}$. It follows that m_* and δ_n^* given in (2.4) of Theorem 2.1 satisfies $m_* \sim n^{1/(2p+2a)}$ and $\delta_n^* \sim n^{-(p+s)/(p+a)}$ respectively. On the other hand, $1/(\gamma_{m_*}\upsilon_{m_*}) = m_*^{2p}\exp(m_*^{2a})$ implies in case (ii) that $m_* \sim (\log n)^{1/(2a)}$ and $\delta_n^* \sim (\log n)^{(p-s)/a}$. Consequently, the lower bounds in Theorem 4.1 follow by applying Theorem 2.1. $\qquad \square$

The upper bounds.

PROOF OF THEOREM 4.2. Observe that in both cases the condition (3.4) is satisfied if $p \geqslant 3/2$. Since the condition on m and α ensures in both cases that $m \sim m_*$ and $\alpha \sim n$ (see proof of Theorem 4.1) the result follows from Theorem 3.4. $\qquad \square$

References

C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71:1795–1843, 2003.

T. Amemiya. The nonlinear two-stage least square estimator. *J. Econometrics*, 2:105–110, 1974.

R. Blundell and J. Horowitz. A nonparametric test of exogeneity. *Rev. Econ. Stud.*, 2007. forthcoming.

R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 2007. forthcoming.

D. Bosq. *Nonparametric statistics for stochastic processes*. Springer, New York, 1998.

H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. Technical report, University Heidelberg, (in revision), 2008.

M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In J. Heckman and E. Leamer, editors, *Handbook of Econometrics*, volume 6. North Holland, 2006.

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

S. Darolles, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. Working Paper # 228, IDEI, Université de Toulouse I, 2002.

D. Donoho. Statistical estimation and optimal recovery. *Annals of Statistics*, 22:238–270, 1994.

D. Donoho and M. Low. Renormalization exponents and optimal pointwise rates of convergence. *Annals of Statistics*, 20:944–970, 1992.

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.

J.-P. Florens. Inverse problems and structural econometrics: The example of instrumental variables. In M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, editors, *Advances in Economics and Econometrics: Theory and Applications – Eight World Congress*, volume 36 of *Econometric Society Monographs*. Cambridge University Press, 2003.

J. P. Florens, J. Johannes, and S. Van Bellegem. Identification and estimation by penalization in nonparametric instrumental regression. Submitted., 2007.

J.-P. Florens, J. Johannes, and S. Van Bellegem. Instrumental regression in partially linear models. Discussion Paper # 0537, Institut de statistique, Université catholique de Louvain (first version 2005, revised), 2009.

P. Gagliardini and O. Scaillet. Tikhonov regularization for functional minimum distance estimators. Swiss Finance Institute Research Paper No. 06-30, 2006.

A. Goldenshluger and S. V. Pereverzev. Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Prob. Theory and Rel. Fields*, 118:169–186, 2000.

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33:2904–2929, 2005.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *Annals of Statistics*, 35(1):70–91, 2007.

E. Heinz. Beiträge zur störungstheorie der spektralzerlegung. *Mathematische Annalen*, 123: 415–438, 1951.

M. Hoffmann and M. Reiß. Nonlinear estimation for linear inverse problems with error in the operator. *The Annals of Statistics*, 36(1):310–336, 2008. Discussion Paper # 990, WIAS, Berlin.

J. L. Horowitz and S. Lee. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75:1191–1208, 2007.

I. Ibragimov and R. Has'minskii. On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.*, 29:18–32, 1984.

J. Johannes. Nonparametric instrumental regression based dimension reduction and additional thresholding. Technical report, University Heidelberg, 2009. in preparation.

J. Johannes and R. Schenk. On rate optimal local estimation in functional linear model. Technical report, University Heidelberg (submitted.), 2009. http://arxiv.org/abs/0902.0645.

J. Johannes, S. Van Bellegem, and A. Vanhems. A unified approach to solve ill-posed inverse problems in econometrics. Submitted, 2007.

T. Kawata. *Fourier analysis in probability theory.* Academic Press, New York, 1972.

K. Li. Minimaxity of the method of regularization of stochastic processes. *Annals of Statistics*, 10:937–942, 1982.

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.

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.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 2003.

A. Pagan and A. Ullah. *Nonparametric Econometrics.* Cambridge University Press, 1999.

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

P. Speckman. Minimax estimation of linear functionals in a Hilbert space. Unpublished manuscript., 1979.

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.