

I N S T I T U T D E S T A T I S T I Q U E
B I O S T A T I S T I Q U E E T
S C I E N C E S A C T U A R I E L L E S
(I S B A)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN



D I S C U S S I O N
P A P E R

2013/58

Adaptive estimation of functionals in
nonparametric instrumental regression

BREUNIG, C. AND J. JOHANNES

Adaptive estimation of functionals in nonparametric instrumental regression.

CHRISTOPH BREUNIG*

Universität Mannheim

JAN JOHANNES*

Université catholique de Louvain

September 6, 2011

We consider the problem of estimating the value $\ell(\varphi)$ of a linear functional, where the structural function φ models a nonparametric relationship in presence of instrumental variables. We propose a plug-in estimator which is based on a dimension reduction technique and additional thresholding. It is shown that this estimator is consistent and can attain the minimax optimal rate of convergence under additional regularity conditions. This, however, requires an optimal choice of the dimension parameter m depending on certain characteristics of the structural function φ and the joint distribution of the regressor and the instrument, which are unknown in practice. We propose a fully data driven choice of m which combines model selection and Lepski's method. We show that the adaptive estimator attains the optimal rate of convergence up to a logarithmic factor. The theory in this paper is illustrated by considering classical smoothness assumptions and we discuss examples such as pointwise estimation or estimation of averages of the structural function φ .

Keywords: Nonparametric regression, Instrument, Linear functional, Optimal rates of convergence, Sobolev space, finitely and infinitely smoothing operator, Adaptive estimation, Model selection, Lepski's method.

AMS subject classifications: Primary 62G08; secondary 62G20.

This work was supported by the DFG-SNF research group FOR916 and by the IAP research network no. P6/03 of the Belgian Government (Belgian Science Policy).

1. Introduction

We consider estimation of the value of a linear functional of the structural function φ in a nonparametric instrumental regression model. The structural function characterizes the

*Lehrstuhl für Statistik, Abteilung Volkswirtschaftslehre, L7, 3-5, 68131 Mannheim, Germany, e-mail: cbreunig@staff.mail.uni-mannheim.de

*Institut de statistique, biostatistique et sciences actuarielles (ISBA), Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgique, e-mail: jan.johannes@uclouvain.be

dependency of a response Y on the variation of an explanatory random variable Z by

$$Y = \varphi(Z) + U \quad \text{with} \quad \mathbb{E}[U|Z] \neq 0 \quad (1.1a)$$

for some error term U . In other words, the structural function equals not the conditional mean function of Y given Z . In this model, however, a sample from (Y, Z, W) is available, where W is a random variable, an instrument, such that

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

Given some a-priori knowledge on the unknown structural function φ , captured by a function class $\mathcal{F}$, its estimation has been intensively discussed in the literature. In contrast, in this paper we are interested in estimating the value $\ell(\varphi)$ of a continuous linear functional $\ell : \mathcal{F} \rightarrow \mathbb{R}$. Important examples discussed in this paper are the weighted average derivative or the point evaluation functional which are both continuous under appropriate conditions on $\mathcal{F}$. We establish a lower bound of the maximal mean squared error for estimating $\ell(\varphi)$ over a wide range of classes $\mathcal{F}$ and functionals ℓ . As a step towards adaptive estimation, we propose in this paper a plug-in estimator of $\ell(\varphi)$ which is consistent and minimax optimal. This estimator is based on a linear Galerkin approach which involves the choice of a dimension parameter. We present a method for choosing this parameter in a data driven way combining model selection and Lepski's method. Moreover, it is shown that the adaptive estimator can attain the minimax optimal rate of convergence within a logarithmic factor.

Model (1.1a–1.1b) 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. [2011]. 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. For example, Ai and Chen [2003], Blundell et al. [2007], Chen and Reiß [2011] or Newey and Powell [2003] consider sieve minimum distance estimators of φ , while Darolles et al. [2002], Hall and Horowitz [2005], Gagliardini and Scaillet [2006] or Florens et al. [2011] study penalized least squares estimators. A linear Galerkin approach to construct an estimator of φ coming from the inverse problem community (c.f. Efromovich and Koltchinskii [2001] or Hoffmann and Reiß [2008]) has been proposed by Johannes and Schwarz [2010]. But estimating the structural function φ as a whole involves the inversion of the conditional expectation operator of Z given W and generally leads to an ill-posed inverse problem (c.f. Newey and Powell [2003] or Florens [2003]). This essentially implies that all proposed estimators have under reasonable assumptions very poor rates of convergence. In contrast, it might be possible to estimate certain local features of φ , such as the value of certain linear functionals at the usual parametric rate of convergence.

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]). However, nonparametric instrumental regression is in general not a Gaussian white noise model. On the other hand, in the former setting the parametric estimation of linear functionals has been addressed in recent years in the econometrics literature. To be more precise, under restrictive conditions on the linear functional ℓ and the joint distribution of (Z, W) it is shown in Ai and Chen [2007], Santos [2011], and Severini

and Tripathi [2010] that it is possible to construct $n^{1/2}$ -consistent estimators of $\ell(\varphi)$. In this situation, efficiency bounds are derived by Ai and Chen [2007] and, when φ is not necessarily identified, by Severini and Tripathi [2010]. We show below, however, that $n^{1/2}$ -consistency is not possible for a wide range of linear functionals ℓ and joint distributions of (Z, W) .

In this paper we establish a minimax theory for the nonparametric estimation of the value of a linear functional $\ell(\varphi)$ of the structural function φ . For this purpose we consider a plug-in estimator $\widehat{\ell}_m := \ell(\widehat{\varphi}_m)$ of $\ell(\varphi)$, where the estimator $\widehat{\varphi}_m$ is proposed by Johannes and Schwarz [2010] and the integer m denotes a dimension to be chosen appropriately. The accuracy of $\widehat{\ell}_m$ is measured by its maximal mean squared error uniformly over the classes $\mathcal{F}$ and $\mathcal{P}$, where $\mathcal{P}$ captures conditions on the unknown joint distribution P_{UZW} of the random vector (U, Z, W) , i.e., $P_{UZW} \in \mathcal{P}$. The class $\mathcal{F}$ reflects prior information on the structural function φ , e.g., its level of smoothness, and will be constructed flexible enough to characterize, in particular, differentiable or analytic functions. On the other hand, the condition $P_{UZW} \in \mathcal{P}$ specifies amongst others some mapping properties of the conditional expectation operator of Z given W implying a certain decay of its singular values. The construction of $\mathcal{P}$ allows us to discuss both a polynomial and an exponential decay of those singular values. Considering the maximal mean squared error over $\mathcal{F}$ and $\mathcal{P}$ we derive a lower bound for estimating $\ell(\varphi)$. Given an optimal choice m_n^* of the dimension we show that the lower bound is attained by $\widehat{\ell}_{m_n^*}$ up to a constant $C > 0$, i.e.,

$$\sup_{P_{UZW} \in \mathcal{P}} \sup_{\varphi \in \mathcal{F}} \mathbb{E} |\widehat{\ell}_{m_n^*} - \ell_h(\varphi)|^2 \leq C \inf_{\check{\ell}} \sup_{P_{UZW} \in \mathcal{P}} \sup_{\varphi \in \mathcal{F}} \mathbb{E} |\check{\ell} - \ell_h(\varphi)|^2$$

where the infimum on the right hand side runs over all possible estimators $\check{\ell}$. Thereby, the estimator $\widehat{\ell}_{m_n^*}$ is minimax optimal even though the optimal choice m_n^* depends on the classes $\mathcal{F}$ and $\mathcal{P}$, which are unknown in practice.

The main issue addressed in this paper is the construction of a data driven selection method for the dimension parameter which adapts to the unknown classes $\mathcal{F}$ and $\mathcal{P}$. When estimating the structural function φ as a whole Loubes and Marteau [2009] and Johannes and Schwarz [2010] propose adaptive estimators under the condition that the eigenfunctions of the unknown conditional expectation operator are a-priori given. In contrast our method does not involve this a-priori knowledge and moreover, allows for both a polynomial and an exponential decay of the associated singular values. The methodology combines a model selection approach (cf. Barron et al. [1999] and its detailed discussion in Massart [2007]) and Lepski's method (cf. Lepski [1990]). It is inspired by the recent work of Goldenschluger and Lepski [2010]. To be more precise, the adaptive choice $\widehat{m}$ is defined as the minimizer of a random penalized contrast criterion¹, i.e.,

$$\widehat{m} := \arg \min_{1 \leq m \leq \widehat{M}_n} \left\{ \widehat{\Psi}_m + \widehat{\text{pen}}_m \right\} \quad (1.2a)$$

with random integer $\widehat{M}_n$ and random penalty sequence $\widehat{\text{pen}} := (\widehat{\text{pen}}_m)_{m \geq 1}$ to be defined below and the sequence of contrast $\widehat{\Psi} := (\widehat{\Psi}_m)_{m \geq 1}$ given by

$$\widehat{\Psi}_m := \max_{m \leq m' \leq \widehat{M}_n} \left\{ |\widehat{\ell}_{m'} - \widehat{\ell}_m|^2 - \widehat{\text{pen}}_{m'} \right\}. \quad (1.2b)$$

With this adaptive choice $\widehat{m}$ at hand the estimator $\widehat{\ell}_{\widehat{m}}$ is shown to be minimax optimal within a logarithmic factor over a wide range of classes $\mathcal{F}$ and $\mathcal{P}$.

¹For a sequence $(a_m)_{m \geq 1}$ having a minimal value in $A \subset \mathbb{N}$ set $\arg \min_{m \in A} \{a_m\} := \min\{m : a_m \leq a_{m'} \forall m' \in A\}$.

The paper is organized as follows. In Section 2 we introduce our basic model assumptions and derive a lower bound for estimating the value of a linear functional in nonparametric instrumental regression. In Section 3 we show consistency of the proposed estimator first and second that it attains the lower bound up to a constant. We illustrate the general results by considering classical smoothness assumptions. The applicability of these results is demonstrated by various examples such as the estimation of the structural function at a point, of its average or of its weighted average derivative. Finally, in section 4 we construct the random upper bound $\widehat{M}_n$ and the random penalty sequence $\widehat{\text{pen}}$ used in (1.2a–1.2b) to define the data driven selection procedure for the dimension parameter m . The proposed adaptive estimator is shown to attain the lower bound within a logarithmic factor. All proofs can be found in the appendix.

2. Complexity of functional estimation: a lower bound.

2.1. Notations and basic model assumptions.

The nonparametric instrumental regression model (1.1a–1.1b) leads to a Fredholm equation of the first kind. To be more precise, let us introduce the conditional expectation operator $T\phi := \mathbb{E}[\phi(Z)|W]$ mapping $L_Z^2 = \{\phi : \mathbb{E}[\phi^2(Z)] < \infty\}$ to $L_W^2 = \{\psi : \mathbb{E}[\psi^2(W)] < \infty\}$ (which are endowed with the usual inner products $\langle \cdot, \cdot \rangle_Z$ and $\langle \cdot, \cdot \rangle_W$, respectively). Consequently, model (1.1a–1.1b) can be written as

$$g = T\varphi \tag{2.1}$$

where the function $g := \mathbb{E}[Y|W]$ belongs to L_W^2 . 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 of T , and that the null space of T is trivial (c.f. Engl et al. [2000] or Carrasco et al. [2006] in the special case of nonparametric instrumental regression). Estimation of the structural function φ is thus linked with the inversion of the operator T . Moreover, we suppose throughout the paper that T is compact which is under fairly mild assumptions satisfied (c.f. Carrasco et al. [2006]). Consequently, 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.

In this section we show that the obtainable accuracy of any estimator of the value $\ell(\varphi)$ of a linear functional can be essentially determined by regularity conditions imposed on the structural function φ and the conditional expectation operator T . In this paper these conditions are characterized by different weighted norms in L_Z^2 with respect to a pre-specified orthonormal basis $\{e_j\}_{j \geq 1}$ in L_Z^2 , which we formalize now. Given a positive sequence of weights $w := (w_j)_{j \geq 1}$ we define the weighted norm $\|\phi\|_w^2 := \sum_{j \geq 1} w_j |\langle \phi, e_j \rangle_Z|^2$, $\phi \in L_Z^2$, the completion $\mathcal{F}_w$ of L_Z^2 with respect to $\|\cdot\|_w$ and the ellipsoid $\mathcal{F}_w^r := \{\phi \in \mathcal{F}_w : \|\phi\|_w^2 \leq r\}$ with radius $r > 0$. We shall stress that the basis $\{e_j\}_{j \geq 1}$ does not necessarily correspond to the eigenfunctions of T . In the following we write $a_n \lesssim b_n$ when there exists a generic constant $C > 0$ such that $a_n \leq C b_n$ for 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.

Minimal regularity conditions. Given a nondecreasing sequence of weights $\gamma := (\gamma_j)_{j \geq 1}$, we suppose, here and subsequently, that the structural function φ belongs to the ellipsoid $\mathcal{F}_\gamma^\rho$ for some $\rho > 0$. The ellipsoid $\mathcal{F}_\gamma^\rho$ captures all the prior information (such as smoothness)

about the unknown structural function φ . Observe that the dual space of $\mathcal{F}_\gamma$ can be identified with $\mathcal{F}_{1/\gamma}$ where $1/\gamma := (\gamma_j^{-1})_{j \geq 1}$ (cf. Krein and Petunin [1966]). To be more precise, for all $\phi \in \mathcal{F}_\gamma$ the value $\langle h, \phi \rangle_Z$ is well defined for all $h \in \mathcal{F}_{1/\gamma}$ and by Riesz's Theorem there exists a unique $h \in \mathcal{F}_{1/\gamma}$ such that $\ell(\phi) = \langle h, \phi \rangle_Z =: \ell_h(\phi)$. In certain applications one might not only be interested in the performance of an estimation procedure of $\ell_h(\varphi)$ for a given representer h , but also for h varying over the ellipsoid $\mathcal{F}_\omega^\tau$ with radius $\tau > 0$ for a nonnegative sequence $\omega := (\omega_j)_{j \geq 1}$ satisfying $\inf_{j \geq 1} \{\omega_j \gamma_j\} > 0$. Obviously, $\mathcal{F}_\omega$ is a subset of $\mathcal{F}_{1/\gamma}$.

Furthermore, as usual in the context of ill-posed inverse problems, we specify some mapping properties of the operator under consideration. Denote by $\mathcal{T}$ the set of all compact operators mapping L_Z^2 into L_W^2 . Given a sequence of weights $v := (v_j)_{j \geq 1}$ and $d \geq 1$ we define the subset $\mathcal{T}_d^v$ of $\mathcal{T}$ by

$$\mathcal{T}_d^v := \left\{ T \in \mathcal{T} : \quad \|\phi\|_v^2/d \leq \|T\phi\|_W^2 \leq d \|\phi\|_v^2, \quad \forall \phi \in L_Z^2 \right\}. \quad (2.2)$$

Notice first that any operator $T \in \mathcal{T}_d^v$ is injective if the sequence v is strictly positive. Furthermore, for all $T \in \mathcal{T}_d^v$ it follows that $v_j/d \leq \|Te_j\|_W^2 \leq dv_j$ for all $j \geq 1$ and if $(s_j)_{j \geq 1}$ denotes the ordered sequence of singular values of T then it holds $v_j/d \leq s_j^2 \leq dv_j$. In other words, the sequence v specifies the decay of the singular values of T . In what follows, all the results are derived under regularity conditions on the structural function φ and the conditional expectation operator T described through the sequence γ and v , respectively. We provide illustrations of these conditions below by assuming a “regular decay” of these sequences. The next assumption summarizes our minimal regularity conditions on these sequences.

ASSUMPTION 1. *Let $\gamma := (\gamma_j)_{j \geq 1}$, $\omega := (\omega_j)_{j \geq 1}$ and $v := (v_j)_{j \geq 1}$ be strictly positive sequences of weights with $\gamma_1 = \omega_1 = v_1 = 1$ such that γ is nondecreasing with $|j|^3 \gamma_j^{-1} = o(1)$ as $j \rightarrow \infty$, ω satisfies $\inf_{j \geq 1} \{\omega_j \gamma_j\} > 0$ and v is a nonincreasing sequence with $\lim_{j \rightarrow \infty} v_j = 0$.*

REMARK 2.1. We illustrate Assumption 1 for typical choices of γ and v usually studied in the literature (c.f. Hall and Horowitz [2005], Chen and Reiß [2011] or Johannes et al. [2011]), that is,

- (pp) $\gamma_j \sim |j|^{2p}$ with $p > 3/2$, $v_j \sim |j|^{-2a}$, $a > 0$, and
 - (i) $[h]_j^2 \sim |j|^{-2s}$, $s > 1/2 - p$ or
 - (ii) $\omega_j \sim |j|^{2s}$, $s > -p$.
- (pe) $\gamma_j \sim |j|^{2p}$, $p > 3/2$ and $v_j \sim \exp(-|j|^{2a})$, $a > 0$, and
 - (i) $[h]_j^2 \sim |j|^{-2s}$, $s > 1/2 - p$ or
 - (ii) $\omega_j \sim |j|^{2s}$, $s > -p$.
- (ep) $\gamma_j \sim \exp(|j|^{2p})$, $p > 0$ and $v_j \sim |j|^{-2a}$, $a > 0$, and
 - (i) $[h]_j^2 \sim |j|^{-2s}$, $s \in \mathbb{R}$ or
 - (ii) $\omega_j \sim |j|^{2s}$, $s \in \mathbb{R}$.

Note that condition $|j|^3 \gamma_j^{-1} = o(1)$ as $j \rightarrow \infty$ is automatically satisfied in case of (ep). In the other two cases this condition states under classical smoothness assumptions that, roughly speaking, the structural function φ has to be differentiable. $\square$

In order to formulate below the lower as well as the upper bound let us define for $x \geq 1$

$$m_x^* := \arg \min_{m \in \mathbb{N}} \left\{ \frac{\max \left\{ \frac{v_m}{\gamma_m}, x^{-1} \right\}}{\min \left\{ \frac{v_m}{\gamma_m}, x^{-1} \right\}} \right\}, \quad \alpha_x^* := \max \left\{ \frac{v_{m_x^*}}{\gamma_{m_x^*}}, x^{-1} \right\}. \quad (2.3)$$

We shall see that the minimax optimal rate is determined by the sequence $\mathcal{R}^h := (\mathcal{R}_n^h)_{n \geq 1}$, in case of a fixed representer h , and $\mathcal{R}^\omega := (\mathcal{R}_n^\omega)_{n \geq 1}$ in case of a representer varying over the class $\mathcal{F}_\omega^\tau$. These sequences are given for all $x \geq 1$ by

$$\mathcal{R}_x^h := \max \left\{ \alpha_x^* \sum_{j=1}^{m_x^*} \frac{[h]_j^2}{v_j}, \sum_{j > m_x^*} \frac{[h]_j^2}{\gamma_j} \right\} \quad \text{and} \quad \mathcal{R}_x^\omega := \alpha_x^* \max_{1 \leq j \leq m_x^*} \left\{ \frac{1}{\omega_j v_j} \right\}. \quad (2.4)$$

In case of adaptive estimation the rate of convergence is given by $\mathcal{R}_{\text{adapt}}^h := (\mathcal{R}_{n(1+\log n)-1}^h)_{n \geq 1}$ and $\mathcal{R}_{\text{adapt}}^\omega := (\mathcal{R}_{n(1+\log n)-1}^\omega)_{n \geq 1}$, respectively. For ease of notation let $m_n^\circ := m_{n(1+\log n)-1}^*$ and $\alpha_n^\circ := \alpha_{n(1+\log n)-1}^*$. The bounds established below need the following additional assumption, which is satisfied in all cases used to illustrate the results.

ASSUMPTION 2. *Let γ and v be sequences such that*

$$0 < \kappa := \inf_{n \geq 1} \left\{ (\alpha_n^*)^{-1} \min \left\{ \frac{v_{m_n^*}}{\gamma_{m_n^*}}, n^{-1} \right\} \right\} \leq 1. \quad (2.5)$$

2.2. Lower bounds.

The results derived below involve assumptions on the conditional moments of the random variables U given W , captured by $\mathcal{U}_\sigma$, which contains all conditional distributions of U given W , denoted by $P_{U|W}$, satisfying $\mathbb{E}[U|W] = 0$ and $\mathbb{E}[U^4|W] \leq \sigma^4$ for some $\sigma > 0$. The next assertion gives a lower bound for the mean squared error of any estimator when estimating the value $\ell_h(\varphi)$ of a linear functional with given representer h and structural function φ in the function class $\mathcal{F}_\gamma^\rho$.

THEOREM 2.1. *Assume an iid. n -sample of (Y, Z, W) from the model (1.1a–1.1b). Let γ and v be sequences satisfying Assumptions 1 and 2. Suppose that $\sup_{j \geq 1} \mathbb{E}[e_j^4(Z)|W] \leq \eta^4$, $\eta \geq 1$, and $\sigma^4 \geq (\sqrt{3} + 4\rho\eta^2 \sum_{j \geq 1} \gamma_j^{-1})^2$. Then for all $n \geq 1$ we have*

$$\inf_{\check{\ell}} \inf_{T \in \mathcal{T}_d^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbb{E} |\check{\ell} - \ell_h(\varphi)|^2 \geq \frac{\kappa}{4} \min \left(\frac{1}{2d}, \rho \right) \mathcal{R}_n^h$$

where the first infimum runs over all possible estimators $\check{\ell}$.

REMARK 2.2. In the proof of the lower bound we consider a test problem based on two hypothetical structural functions. For each test function the condition $\sigma^4 \geq (\sqrt{3} + 4\rho\eta^2 \sum_{j \geq 1} \gamma_j^{-1})^2$ ensures a certain complexity of the hypothetical model in a sense that it allows for Gaussian residuals. This specific case is only needed to simplify the calculation of the distance between distributions corresponding to different structural functions. A similar assumption has been used by Chen and Reiß [2011] 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 .

On the other hand, below we derive an upper bound assuming that $P_{U|W}$ belongs to $\mathcal{U}_\sigma$ and that the joint distribution of (Z, W) fulfills in addition Assumption 3. Obviously in this situation Theorem 2.1 provides a lower bound for any estimator as long as σ is sufficiently large. $\square$

REMARK 2.3. The regularity conditions imposed on the structural function φ and the conditional expectation operator T involve only the basis $\{e_j\}_{j \geq 1}$ in L_Z^2 . Therefore, the lower bound derived in Theorem 2.1 does not capture the influence of the basis $\{f_l\}_{l \geq 1}$ in L_W^2 used below to construct an estimator of the value $\ell_h(\varphi)$. In other words, this estimator attains the lower bound only if $\{f_l\}_{l \geq 1}$ is chosen appropriately. $\square$

REMARK 2.4. The rate $\mathcal{R}^h$ of the lower bound is never faster than the parametric rate, that is, $\mathcal{R}_n^h \geq n^{-1}$. Moreover, it is easily seen that the lower bound rate is parametric if and only if $\sum_{j \geq 1} [h]_j^2 v_j^{-1} < \infty$. This condition does not involve the sequence γ and hence, attaining a parametric rate is independent of the regularity conditions we are willing to impose on the structural function. $\square$

The following assertion gives a lower bound over the ellipsoid $\mathcal{F}_\omega^\tau$ of representer. Consider the function $h^* := \tau \omega_{j^*}^{-1/2} e_{j^*}$ with $j^* := \arg \max_{1 \leq j \leq m_n^*} \{(\omega_j v_j)^{-1}\}$ which obviously belongs to $\mathcal{F}_\omega^\tau$. Corollary 2.2 follows then by calculating the value of the lower bound in Theorem 2.1 for the specific representer h^* and, hence we omit its proof.

COROLLARY 2.2. *Let the assumptions of Theorem 2.1 be satisfied. Then for all $n \geq 1$ we have*

$$\inf_{\check{\ell}} \inf_{T \in \mathcal{T}_d^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho, h \in \mathcal{F}_\omega^\tau} \mathbb{E} |\check{\ell} - \ell_h(\varphi)|^2 \geq \frac{\tau \kappa}{4} \min \left(\frac{1}{2d}, \rho \right) \mathcal{R}_n^\omega$$

where the first infimum runs over all possible estimators $\check{\ell}$.

REMARK 2.5. If the lower bound given in Corollary 2.2 tends to zero then $(\omega_j \gamma_j)_{j \geq 1}$ is a divergent sequence. In other words, without any additional restriction on φ , that is, $\gamma \equiv 1$, consistency of an estimator of $\ell_h(\varphi)$ uniformly over all $\varphi \in \mathcal{F}_\gamma^\rho$ and all $h \in \mathcal{F}_\omega^\tau$ is only possible under restrictions on the representer h in the sense that ω has to be a divergent sequence. This obviously reflects the ill-posedness of the underlying inverse problem. $\square$

3. Minimax optimal estimation.

3.1. Estimation by dimension reduction and thresholding.

In addition to the basis $\{e_j\}_{j \geq 1}$ in L_Z^2 used to establish the lower bound we consider now also a second basis $\{f_l\}_{l \geq 1}$ in L_W^2 .

Matrix and operator notations. Given $m \geq 1$, $\mathcal{E}_m$ 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 K from L_Z^2 to L_W^2 we denote its inverse by K^{-1} and its adjoint by K^* . If we restrict $F_m K E_m$ to an operator from $\mathcal{E}_m$ to $\mathcal{F}_m$, then it can be represented by a matrix $[K]_{\underline{m}}$ with entries $[K]_{l,j} = \langle K e_j, f_l \rangle_W$ for $1 \leq j, l \leq m$.

Its spectral norm is denoted by $\|[K]_{\underline{m}}\|$, its inverse by $[K]_{\underline{m}}^{-1}$ and its transposed by $[K]_{\underline{m}}^t$. We write I for the identity operator and ∇_v for the diagonal operator with singular value decomposition $\{v_j, e_j, f_j\}_{j \geq 1}$. Respectively, given functions $\phi \in L_Z^2$ and $\psi \in L_W^2$ we define by $[\phi]_{\underline{m}}$ and $[\psi]_{\underline{m}}$ m -dimensional vectors with entries $[\phi]_j = \langle \phi, e_j \rangle_Z$ and $[\psi]_l = \langle \psi, f_l \rangle_W$ for $1 \leq j, l \leq m$.

Consider the conditional expectation operator T associated with (Z, W) . If $[e(Z)]_{\underline{m}}$ and $[f(W)]_{\underline{m}}$ denote random vectors with entries $e_j(Z)$ and $f_j(W)$, $1 \leq j \leq m$, respectively, then it holds $[T]_{\underline{m}} = \mathbb{E}[f(W)]_{\underline{m}}[e(Z)]_{\underline{m}}^t$. Throughout the paper $[T]_{\underline{m}}$ is assumed to be nonsingular for all $m \geq 1$, so that $[T]_{\underline{m}}^{-1}$ always exists. Note that it is a nontrivial problem to determine when such an assumption holds (cf. Efromovich and Koltchinskii [2001] and references therein).

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

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

Then the estimator of the linear functional $\ell_h(\varphi)$ is defined for all $m \geq 1$ by

$$\widehat{\ell}_m := \begin{cases} [h]_{\underline{m}}^t [\widehat{T}]_{\underline{m}}^{-1} [\widehat{g}]_{\underline{m}}, & \text{if } [\widehat{T}]_{\underline{m}} \text{ is nonsingular and } \|[\widehat{T}]_{\underline{m}}^{-1}\| \leq \sqrt{n}, \\ 0, & \text{otherwise.} \end{cases} \quad (3.1)$$

In fact, the estimator $\widehat{\ell}_m$ is obtained from the linear functional $\ell_h(\varphi)$ by replacing the unknown structural function φ by an estimator proposed by Johannes and Schwarz [2010].

REMARK 3.1. If Z is continuously distributed one might be also interested in estimating the value $\int_{\mathcal{Z}} \varphi(z) h(z) dz$ where $\mathcal{Z}$ is the support of Z . Assume that this integral and also $\int_{\mathcal{Z}} h(z) e_j(z) dz$ for $1 \leq j \leq m$ are well defined. Then we can cover the problem of estimating $\int_{\mathcal{Z}} \varphi(z) h(z) dz$ by simply replacing $[h]_{\underline{m}}$ in the definition of $\widehat{\ell}_m$ by a m -dimensional vector with entries $\int_{\mathcal{Z}} h(z) e_j(z) dz$ for $1 \leq j \leq m$. Hence for $\int_{\mathcal{Z}} \varphi(z) h(z) dz$ the results below follow mutatis mutandis. $\square$

Moment assumptions. Besides the link condition (2.2) for the conditional expectation operator T we need moment conditions on the basis, more specific, on the random variables $e_j(Z)$ and $f_l(W)$ for $j, l \geq 1$, which we summarize in the next assumption.

ASSUMPTION 3. *There exists $\eta \geq 1$ such that the joint distribution of (Z, W) satisfies*

- (i) $\sup_{j \in \mathbb{N}} \mathbb{E}[e_j^2(Z)|W] \leq \eta^2$ and $\sup_{l \in \mathbb{N}} \mathbb{E}[f_l^4(W)] \leq \eta^4$;
- (ii) $\sup_{j, l \in \mathbb{N}} \mathbb{E}[e_j(Z)f_l(W) - \mathbb{E}[e_j(Z)f_l(W)]]^k \leq \eta^k k!$, $k = 3, 4, \dots$

Note that condition (ii) is also known as Cramer's condition, which is sufficient to obtain an exponential bound for large deviations of the centered random variable $e_j(Z)f_l(W) - \mathbb{E}[e_j(Z)f_l(W)]$ (c.f. Bosq [1998]). Moreover, any joint distribution of (Z, W) satisfies Assumption 3 for sufficiently large η if the basis $\{e_j\}_{j \geq 1}$ and $\{f_l\}_{l \geq 1}$ are uniformly bounded, which holds, for example, for the trigonometric basis considered in Subsection 3.4.

3.2. Consistency.

The next assertion summarizes sufficient conditions to ensure consistency of the estimator $\widehat{\ell}_m$ introduced in (3.1). Let $\varphi_m \in \mathcal{E}_m$ with $[\varphi_m]_{\underline{m}} = [T]_{\underline{m}}^{-1}[g]_{\underline{m}}$. Obviously, up to the threshold, the estimator $\widehat{\ell}_m$ is the empirical counterpart of $\ell_h(\varphi_m)$. In Proposition 3.1 consistency of the estimator $\widehat{\ell}_m$ is only obtained under the condition

$$\|\varphi - \varphi_m\|_\gamma = o(1) \text{ as } m \rightarrow \infty \quad (3.2)$$

which does not hold true in general. Obviously (3.2) implies the convergence of $\ell_h(\varphi_m)$ to $\ell_h(\varphi)$ as m tends to infinity for all $h \in \mathcal{F}_{1/\gamma}$.

PROPOSITION 3.1. *Assume an iid. n -sample of (Y, Z, W) from the model (1.1a–1.1b) with $P_{U|W} \in \mathcal{U}_\sigma$ and joint distribution of (Z, W) fulfilling Assumption 3. Let the dimension parameter m_n satisfy $m_n^{-1} = o(1)$, $m_n = o(n)$,*

$$\|[h]_{\underline{m}_n}^t [T]_{\underline{m}_n}^{-1}\|^2 = o(n), \text{ and } m_n^3 \|[T]_{\underline{m}_n}^{-1}\|^2 = O(n) \text{ as } n \rightarrow \infty. \quad (3.3)$$

If (3.2) holds true then $\mathbb{E}|\widehat{\ell}_{m_n} - \ell_h(\varphi)|^2 = o(1)$ as $n \rightarrow \infty$ for all $\varphi \in \mathcal{F}_\gamma$ and $h \in \mathcal{F}_{1/\gamma}$.

Notice that condition (3.2) also involves the basis $\{f_l\}_{l \geq 1}$ in L_W^2 . In what follows, we introduce an alternative but stronger condition to guarantee (3.2) which extends the link condition (2.2). We denote by $\mathcal{T}_{d,D}^v$ for some $D \geq d$ the subset of $\mathcal{T}_d^v$ given by

$$\mathcal{T}_{d,D}^v := \left\{ T \in \mathcal{T}_d^v : \sup_{m \in \mathbb{N}} \|[\nabla_v]_{\underline{m}}^{1/2} [T]_{\underline{m}}^{-1}\|^2 \leq D \right\}. \quad (3.4)$$

REMARK 3.2. If $T \in \mathcal{T}_d^v$ and if in addition its singular value decomposition is given by $\{s_j, e_j, f_j\}_{j \geq 1}$ then for all $m \geq 1$ the matrix $[T]_{\underline{m}}$ is diagonalized with diagonal entries $[T]_{j,j} = s_j$, $1 \leq j \leq m$. In this situation it is easily seen that $\sup_{m \in \mathbb{N}} \|[\nabla_v]_{\underline{m}}^{1/2} [T]_{\underline{m}}^{-1}\|^2 \leq d$ and, hence T satisfies the extended link condition (3.4), that is, $T \in \mathcal{T}_{d,D}^v$. Furthermore, it holds $\mathcal{T}_d^v = \mathcal{T}_{d,D}^v$ for suitable $D > 0$, if T is a small perturbation of $\nabla_v^{1/2}$ or if T is strictly positive (c.f. Efromovich and Koltchinskii [2001] or Cardot and Johannes [2010], respectively). $\square$

REMARK 3.3. Once both basis $\{e_j\}_{j \geq 1}$ and $\{f_l\}_{l \geq 1}$ are specified the extended link condition (3.4) restricts the class of joint distributions of (Z, W) such that (3.2) holds true. Moreover, under (3.4) the estimator $\widehat{\varphi}_m$ of φ proposed by Johannes and Schwarz [2010] can attain the minimax optimal rate. In this sense, given a joint distribution of (Z, W) a basis $\{f_l\}_{l \geq 1}$ satisfying condition (3.4) can be interpreted as optimal instruments (c.f. Newey [1990]). $\square$

REMARK 3.4. For each pre-specified basis $\{e_j\}_{j \geq 1}$ we can theoretically construct a basis $\{f_l\}_{l \geq 1}$ such that (3.4) is equivalent to the link condition (2.2). To be more precise, if $T \in \mathcal{T}_d^v$, which involves only the basis $\{e_j\}_{j \geq 1}$, then the fundamental inequality of Heinz [1951] implies $\|(T^*T)^{-1/2}e_j\|_Z^2 \leq dv_j^{-1}$. Thereby, the function $(T^*T)^{-1/2}e_j$ is an element of L_Z^2 and, hence $f_j := T(T^*T)^{-1/2}e_j$, $j \geq 1$, belongs to L_W^2 . Then it is easily checked that $\{f_l\}_{l \geq 1}$ is a basis of the closure of the range of T which may be completed to a basis of L_W^2 . Obviously $[T]_{\underline{m}}$ is symmetric and moreover, strictly positive since $\langle Te_j, f_l \rangle_W = \langle (T^*T)^{1/2}e_j, e_l \rangle_Z$ for all $j, l \geq 1$. Thereby, we can apply Lemma A.3 in Cardot and Johannes [2010] which gives $\mathcal{T}_d^v = \mathcal{T}_{d,D}^v$ for sufficiently large D . We are currently exploring the data driven choice of the basis $\{f_l\}_{l \geq 1}$. $\square$

Under the extended link condition (3.4) the next assertion summarizes sufficient conditions to ensure consistency.

COROLLARY 3.2. *The conclusion of Proposition 3.1 still holds true without imposing condition (3.2), if the sequence v satisfies Assumption 1, the conditional expectation operator T belongs to $\mathcal{T}_{d,D}^v$, and (3.3) is substituted by*

$$\sum_{j=1}^{m_n} [h]_j^2 v_j^{-1} = o(n) \quad \text{and} \quad m_n^3 = O(nv_{m_n}) \quad \text{as } n \rightarrow \infty. \quad (3.5)$$

3.3. An upper bound.

The last assertions show that the estimator $\widehat{\ell}_m$ defined in (3.1) is consistent for all structural functions and representers belonging to $\mathcal{F}_\gamma$ and $\mathcal{F}_{1/\gamma}$, respectively. The following theorem provides now an upper bound if φ belongs to an ellipsoid $\mathcal{F}_\gamma^\rho$. In this situation the rate $\mathcal{R}^h$ of the lower bound given in Theorem 2.1 provides up to a constant also an upper bound of the estimator $\widehat{\ell}_{m_n^*}$. Thus we have proved that the rate $\mathcal{R}^h$ is optimal and, hence $\widehat{\ell}_{m_n^*}$ is minimax optimal.

THEOREM 3.3. *Assume an iid. n -sample of (Y, Z, W) from the model (1.1a–1.1b) with joint distribution of (Z, W) fulfilling Assumption 3. Let Assumptions 1 and 2 be satisfied. Suppose that the dimension parameter m_n^* given by (2.3) satisfies*

$$(m_n^*)^3 \max \left\{ |\log \mathcal{R}_n^h|, (\log m_n^*) \right\} = o(\gamma_{m_n^*}), \quad \text{as } n \rightarrow \infty, \quad (3.6)$$

then we have for all $n \geq 1$

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbb{E} |\widehat{\ell}_{m_n^*} - \ell_h(\varphi)|^2 \leq C \mathcal{R}_n^h$$

for a constant $C > 0$ only depending on the classes $\mathcal{F}_\gamma^\rho$, $\mathcal{T}_{d,D}^v$, the constants σ , η and the representer h .

The following assertion states an upper bound uniformly over the class $\mathcal{F}_\omega^\tau$ of representer. Observe that $\|h\|_{1/\gamma}^2 \leq \tau$ and $\mathcal{R}_n^h \leq \tau \alpha_n^* \max_{1 \leq j \leq m_n^*} \{(\omega_j v_j)^{-1}\} = \tau \mathcal{R}_n^\omega$ for all $h \in \mathcal{F}_\omega^\tau$. Employing these estimates the proof of the next result follows line by line the proof of Theorem 3.3 and is thus omitted.

COROLLARY 3.4. *Let the assumptions of Theorem 3.3 be satisfied where we substitute condition (3.6) by $(m_n^*)^3 \max \{ |\log \mathcal{R}_n^\omega|, (\log m_n^*) \} = o(\gamma_{m_n^*})$ as $n \rightarrow \infty$. Then we have*

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho, h \in \mathcal{F}_\omega^\tau} \mathbb{E} |\widehat{\ell}_{m_n^*} - \ell_h(\varphi)|^2 \leq C \mathcal{R}_n^\omega$$

for a constant $C > 0$ only depending on the classes $\mathcal{F}_\gamma^\rho$, $\mathcal{F}_\omega^\tau$, $\mathcal{T}_{d,D}^v$ and the constants σ , η .

3.4. Illustration by classical smoothness assumptions.

Let us illustrate our general results by considering classical smoothness assumptions. To simplify the presentation we follow Hall and Horowitz [2005], and suppose that the marginal distribution of the scalar regressor Z and the scalar instrument W are uniformly distributed on the interval $[0, 1]$. All the results below can be easily extended to the multivariate case.

In the univariate case, however, both Hilbert spaces L_Z^2 and L_W^2 equal $L^2[0, 1]$. Moreover, as a basis $\{e_j\}_{j \geq 1}$ in $L^2[0, 1]$ we choose the trigonometric basis given by

$$e_1 := 1, \quad e_{2j}(t) := \sqrt{2} \cos(2\pi jt), \quad e_{2j+1}(t) := \sqrt{2} \sin(2\pi jt), \quad t \in [0, 1], \quad j \in \mathbb{N}.$$

In this subsection also the second basis $\{f_l\}_{l \geq 1}$ is given by the trigonometric basis. In this situation, the moment conditions formalized in Assumption 3 are automatically fulfilled. Recall the typical choices of the sequences γ , ω , and v introduced in Remark 2.1. If $\gamma_j \sim |j|^{2p}$, $p > 0$, as in case (pp) and (pe), then $\mathcal{F}_\gamma$ coincides with the Sobolev space of p -times differential periodic functions (c.f. Neubauer [1988a,b]). In case of (ep) it is well known that $\mathcal{F}_\gamma$ contains only analytic functions if $p > 1$ (c.f. Kawata [1972]). Furthermore, we consider two special cases describing a “regular decay” of the unknown singular values of T . In case of (pp) and (ep) we consider a polynomial decay of the sequence v . Easy calculus shows that any operator T satisfying the link condition (2.2) acts like integrating (a) -times and hence is called *finitely smoothing* (c.f. Natterer [1984]). In case of (pe) we consider an exponential decay of v and it can easily be seen that $T \in \mathcal{T}_d^v$ implies $\mathcal{R}(T) \subset C^\infty[0, 1]$, therefore the operator T is called *infinitely smoothing* (c.f. Mair [1994]). In the next assertion we present the order of sequences $\mathcal{R}^h$ and $\mathcal{R}^\omega$ which were shown to be minimax-optimal.

PROPOSITION 3.5. *Assume an iid. n -sample of (Y, Z, W) from the model (1.1a–1.1b) with $T \in \mathcal{T}_{d,D}^v$ and $P_{U|W} \in \mathcal{U}_\sigma$. Then for the example configurations of Remark 2.1 we obtain*

(pp) $m_n^* \sim n^{1/(2p+2a)}$ and

$$(i) \quad \mathcal{R}_n^h \sim \begin{cases} n^{-(2p+2s-1)/(2p+2a)}, & \text{if } s - a < 1/2, \\ n^{-1} \log n, & \text{if } s - a = 1/2, \\ n^{-1}, & \text{otherwise,} \end{cases}$$

$$(ii) \quad \mathcal{R}_n^\omega \sim \max(n^{-(p+s)/(p+a)}, n^{-1}).$$

(pe) $m_n^* \sim \log(n(\log n)^{-p/a})^{1/(2a)}$ and

$$(i) \quad \mathcal{R}_n^h \sim (\log n)^{-(2p+2s-1)/(2a)},$$

$$(ii) \quad \mathcal{R}_n^\omega \sim (\log n)^{-(p+s)/a}.$$

(ep) $m_n^* \sim \log(n(\log n)^{-a/p})^{1/(2p)}$ and

$$(i) \quad \mathcal{R}_n^h \sim \begin{cases} n^{-1} (\log n)^{(2a-2s+1)/(2p)}, & \text{if } s - a < 1/2, \\ n^{-1} \log(\log n), & \text{if } s - a = 1/2, \\ n^{-1}, & \text{otherwise,} \end{cases}$$

$$(ii) \quad \mathcal{R}_n^\omega \sim \max(n^{-1} (\log n)^{(a-s)/p}, n^{-1}).$$

REMARK 3.5. As we see from Proposition 3.5, 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 or s leads to a faster optimal rate. Moreover, in the cases (pp) and (ep) the parametric rate n^{-1} is obtained independent of the smoothness assumption imposed on the structural function φ (however, $p \geq 3/2$ is needed) if the representer is smoother than the degree of ill-posedness of T , i.e., (i) $s \geq a - 1/2$ and (ii) $s \geq a$. Moreover, it is easily seen that if $[h]_j \sim \exp(-|j|^s)$ or $\omega_j \sim \exp(-|j|^{2s})$, $s > 0$, then the minimax convergence rates are always parametric for any polynomial sequences γ and v . $\square$

EXAMPLE 3.1. Suppose we are interested in estimating the value $\varphi(t_0)$ of the structural function φ evaluated at a point $t_0 \in [0, 1]$. Consider the representer given by $h_{t_0} = \sum_{j=1}^{\infty} e_j(t_0)e_j$. Let $\varphi \in \mathcal{F}_\gamma$. Since $\sum_{j \geq 1} \gamma_j^{-1} < \infty$ (cf. Assumption 1) it holds $h \in \mathcal{F}_{1/\gamma}$ and hence the point evaluation functional in $t_0 \in [0, 1]$, i.e., $\ell_{h_{t_0}}(\varphi) = \varphi(t_0)$, is well defined. In this case, the estimator $\hat{\ell}_m$ introduced in (3.1) writes for all $m \geq 1$ as

$$\hat{\varphi}_m(t_0) := \begin{cases} [e(t_0)]_{\underline{m}}^t [\hat{T}]_{\underline{m}}^{-1} [\hat{g}]_{\underline{m}}, & \text{if } [\hat{T}]_{\underline{m}} \text{ is nonsingular and } \|[\hat{T}]_{\underline{m}}^{-1}\| \leq \sqrt{n}, \\ 0, & \text{otherwise} \end{cases}$$

where $\hat{\varphi}_m$ is an estimator proposed by Johannes and Schwarz [2010]. Let $p \geq 3/2$ and $a > 0$. Then the estimator $\hat{\varphi}_{m_n}^*(t_0)$ attains within a constant the minimax optimal rate of convergence $\mathcal{R}^{h_{t_0}}$. Applying Proposition 3.5 gives

$$\begin{aligned} (\text{pp}) \quad \mathcal{R}_n^{h_{t_0}} &\sim n^{-(2p-1)/(2p+2a)}, \\ (\text{ep}) \quad \mathcal{R}_n^{h_{t_0}} &\sim (\log n)^{-(2p-1)/(2a)}, \\ (\text{ep}) \quad \mathcal{R}_n^{h_{t_0}} &\sim n^{-1}(\log n)^{(2a+1)/(2p)}. \end{aligned} \quad \square$$

EXAMPLE 3.2. We want to estimate the average value of the structural function φ over a certain interval $[0, b]$ with $0 < b < 1$. The linear functional of interest is given by $\ell_h(\varphi) = \int_0^b \varphi(t)dt$ with representer $h := \mathbb{1}_{[0,b]}$. Its Fourier coefficients are given by $[h]_1 = b$, $[h]_{2j} = (\sqrt{2}\pi j)^{-1} \sin(2\pi j b)$, $[h]_{2j+1} = -(\sqrt{2}\pi j)^{-1} \cos(2\pi j b)$ for $j \geq 1$ and, hence $[h]_j^2 \sim j^{-2}$. Again we assume that $p \geq 3/2$ and $a > 0$. Then the mean squared error of the estimator $\hat{\ell}_{m_n} = \int_0^b \hat{\varphi}_{m_n}^*(t)dt$ is bounded up to a constant by the minimax rate of convergence $\mathcal{R}^h$. In the three cases the order of $\mathcal{R}_n^h$ is given by

$$\begin{aligned} (\text{pp}) \quad \mathcal{R}_n^h &\sim \begin{cases} n^{-(2p+1)/(2p+2a)}, & \text{if } a > 1/2, \\ n^{-1} \log n, & \text{if } a = 1/2, \\ n^{-1}, & \text{otherwise,} \end{cases} \\ (\text{ep}) \quad \mathcal{R}_n^h &\sim (\log n)^{-(2p+1)/(2a)}, \\ (\text{ep}) \quad \mathcal{R}_n^h &\sim \begin{cases} n^{-1}(\log n)^{(2a-1)/(2p)}, & \text{if } a > 1/2, \\ n^{-1} \log(\log n), & \text{if } a = 1/2, \\ n^{-1}, & \text{otherwise.} \end{cases} \end{aligned}$$

As in the direct regression model where the average value of the regression function can be estimated with rate n^{-1} we obtain the parametric rate in the case of (pp) and (ep) if $a < 1/2$. $\square$

EXAMPLE 3.3. Consider estimation of the weighted average derivative of the structural function φ with weight function H , i.e., $\int_0^1 \varphi'(t)H(t)dt$. This functional is useful not only for estimating scaled coefficients of an index model, but also to quantify the average slope of structural functions. Assume that the weight function H is continuously differentiable and vanishes at the boundary of the support of Z , i.e., $H(0) = H(1) = 0$. Integration by parts gives $\int_0^1 \varphi'(t)H(t)dt = -\int_0^1 \varphi(t)h(t)dt = -\ell_h(\varphi)$ with representer h given by the derivative of H . The weighted average derivative estimator $\hat{\ell}_{m_n} = -\int_0^1 \hat{\varphi}_{m_n}^*(t)h(t)dt$ is minimax optimal. As an illustration consider the specific weight function $H(t) = 1 - (2t - 1)^2$ with derivative $h(t) = 4(1 - 2t)$ for $0 \leq t \leq 1$. It is easily seen that the Fourier coefficients of the

representer h are $[h]_1 = 0$, $[h]_{2j} = 0$, $[h]_{2j+1} = 4\sqrt{2}(\pi j)^{-1}$ for $j \geq 1$ and, thus $[h]_{2j+1}^2 \sim j^{-2}$. Thus, for the particular choice of the weight function H the estimator $\widehat{\ell}_{m_n^*}$ attains up to a constant the optimal rate $\mathcal{R}^h$, which was already specified in Example 3.2. $\square$

4. Adaptive estimation

In this section we derive an adaptive estimation procedure for the value of the linear function $\ell_h(\varphi)$. This procedure is based on the estimator $\widehat{\ell}_{\widehat{m}}$ given in (3.1) with dimension parameter $\widehat{m}$ selected as a minimizer of the data driven penalized contrast criterion (1.2a–1.2b). The selection criterion (1.2a–1.2b) involves the random upper bound $\widehat{M}_n$ and the random penalty sequence $\widehat{\text{pen}}$ which we introduce below. We show that the estimator $\widehat{\ell}_{\widehat{m}}$ attains the minimax rate of convergence within a logarithmic term. Moreover, we illustrate the cost due to adaption by considering classical smoothness assumptions.

In an intermediate step we do not consider the estimation of unknown quantities in the penalty function. Let us therefore consider a deterministic upper bound M_n and a deterministic penalty sequence $\text{pen} := (\text{pen}_m)_{m \geq 1}$, which is nondecreasing. These quantities are constructed such that they can be easily estimated in a second step. As an adaptive choice $\widetilde{m}$ of the dimension parameter m we propose the minimizer of a penalized contrast criterion, that is,

$$\widetilde{m} := \arg \min_{1 \leq m \leq M_n} \{\Psi_m + \text{pen}_m\} \quad (4.1a)$$

where the random sequence of contrast $\Psi := (\Psi_m)_{m \geq 1}$ is defined by

$$\Psi_m := \max_{m \leq m' \leq M_n} \left\{ |\widehat{\ell}_{m'} - \widehat{\ell}_m|^2 - \text{pen}_{m'} \right\}. \quad (4.1b)$$

The fundamental idea to establish an appropriate upper bound for the risk of $\widehat{\ell}_{\widetilde{m}}$ is given by the following reduction scheme. Let us denote $m \wedge m' := \min(m, m')$. Due to the definition of Ψ and $\widetilde{m}$ we deduce for all $1 \leq m \leq M_n$

$$\begin{aligned} |\widehat{\ell}_{\widetilde{m}} - \ell_h(\varphi)|^2 &\leq 3 \left\{ |\widehat{\ell}_{\widetilde{m}} - \widehat{\ell}_{\widetilde{m} \wedge m}|^2 + |\widehat{\ell}_{\widetilde{m} \wedge m} - \widehat{\ell}_m|^2 + |\widehat{\ell}_m - \ell_h(\varphi)|^2 \right\} \\ &\leq 3 \left\{ \Psi_m + \text{pen}_{\widetilde{m}} + \Psi_{\widetilde{m}} + \text{pen}_m + |\widehat{\ell}_m - \ell_h(\varphi)|^2 \right\} \\ &\leq 6 \{ \Psi_m + \text{pen}_m \} + 3 |\widehat{\ell}_m - \ell_h(\varphi)|^2 \end{aligned}$$

where the right hand side does not depend on the adaptive choice $\widetilde{m}$. Since the penalty sequence pen is nondecreasing we obtain

$$\Psi_m \leq 6 \max_{m \leq m' \leq M} \left(|\widehat{\ell}_{m'} - \ell_h(\varphi_{m'})|^2 - \frac{1}{6} \text{pen}_{m'} \right)_+ + 3 \max_{m \leq m' \leq M_n} |\ell_h(\varphi_m - \varphi_{m'})|^2.$$

Combing the last estimate with the previous reduction scheme yields for all $1 \leq m \leq M_n$

$$|\widehat{\ell}_{\widetilde{m}} - \ell_h(\varphi)|^2 \leq 7 \text{pen}_m + 78 \text{bias}_m + 42 \max_{m \leq m' \leq M} \left(|\widehat{\ell}_{m'} - \ell_h(\varphi_{m'})|^2 - \frac{1}{6} \text{pen}_{m'} \right)_+ \quad (4.2)$$

where $\text{bias}_m := \sup_{m' \geq m} |\ell_h(\varphi_{m'} - \varphi)|^2$. We will prove below that $\text{pen}_m + \text{bias}_m$ is of the order $\mathcal{R}_{n(1+\log n)}^h$. Moreover, we will bound the right hand side term appropriately with the help of Bernstein's inequality.

Let us now introduce the upper bound M_n and sequence of penalty pen_m used in the penalized contrast criterion (4.1a–4.1b). In the following, assume without loss of generality that $[h]_1 \neq 0$.

DEFINITION 4.1. For all $n \geq 1$ let $a_n := n^{1-1/\log(2+\log n)}(1 + \log n)^{-1}$ and $M_n^h := \max\{1 \leq m \leq \lfloor n^{1/4} \rfloor : \max_{1 \leq j \leq m} [h]_j^2 \leq n[h]_1^2\}$ then we define

$$M_n := \min \left\{ 2 \leq m \leq M_n^h : m^3 \|[T]_{\underline{m}}^{-1}\|^2 \max_{1 \leq j \leq m} [h]_j^2 > a_n \right\} - 1$$

where we set $M_n := M_n^h$ if the min runs over an empty set. Thus, M_n takes values between 1 and M_n^h . Let $\varsigma_m^2 = 74(\mathbb{E}[Y^2] + \max_{1 \leq m' \leq m} \|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2)$, then we define

$$\text{pen}_m := 24 \varsigma_m^2 (1 + \log n) n^{-1} \max_{1 \leq m' \leq m} \|[h]_{\underline{m}'}^t [T]_{\underline{m}'}^{-1}\|^2. \quad (4.3)$$

To apply Bernstein's inequality we need another assumption regarding the error term U . This is captured by the set $\mathcal{U}_\sigma^\infty$ for some $\sigma > 0$, which contains all conditional distributions $P_{U|W}$ such that $\mathbb{E}[U|W] = 0$, $\mathbb{E}[U^2|W] \leq \sigma^2$, and Cramer's condition hold, i.e.,

$$\mathbb{E}[|U|^k|W] \leq \sigma^k k!, \quad k = 3, 4, \dots$$

Moreover, besides Assumption 3 we need the following Cramer condition which is in particular satisfied if the basis $\{f_l\}_{l \geq 1}$ are uniformly bounded.

ASSUMPTION 4. There exists $\eta \geq 1$ such that the distribution of W satisfies

$$\sup_{j,l \in \mathbb{N}} \mathbb{E} |f_j(W) f_l(W) - \mathbb{E}[f_j(W) f_l(W)]|^k \leq \eta^k k!, \quad k = 3, 4, \dots$$

We now present an upper bound for $\widehat{\ell}_{\widetilde{m}}$. As usual in the context of adaptive estimation of functionals we face a logarithmic loss due to the adaptation.

THEOREM 4.1. Assume an iid. n -sample of (Y, Z, W) from the model (1.1a–1.1b) with $\mathbb{E}[Y^2] > 0$. Let Assumptions 1–4 be satisfied. Suppose that $(m_n^\circ)^3 \max_{1 \leq j \leq m_n^\circ} [h]_j^2 = o(a_n v_{m_n^\circ})$ as $n \rightarrow \infty$. Then we have for all $n \geq 1$

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma^\infty} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbb{E} |\widehat{\ell}_{\widetilde{m}} - \ell_h(\varphi)|^2 \leq \mathcal{C} \mathcal{R}_{n(1+\log n)}^h$$

for a constant $\mathcal{C} > 0$ only depending on the classes $\mathcal{F}_\gamma^\rho$, $\mathcal{T}_{d,D}^v$, the constants σ , η and the representer h .

REMARK 4.1. In all examples studied below the condition $(m_n^\circ)^3 \max_{1 \leq j \leq m_n^\circ} [h]_j^2 = o(a_n v_{m_n^\circ})$ as n tends to infinity is satisfied if the structural function φ is sufficiently smooth. More precisely, in case of (pp) it suffices to assume $3 < 2p + 2 \min(0, s)$. On the other hand, in case of (pe) or (ep) this condition is automatically fulfilled. $\square$

In the following definition we introduce empirical versions of the integer M_n and the penalty sequence pen . Thereby, we complete the data driven penalized contrast criterion (1.2a–1.2b). This allows for a completely data driven selection method. For this purpose, we construct an estimator for ς_m^2 by replacing the unknown quantities by their empirical analogon, that is,

$$\widehat{\varsigma}_m^2 := 74 \left(n^{-1} \sum_{i=1}^n Y_i^2 + \max_{1 \leq m' \leq m} \|\widehat{[T]}_{\underline{m}}^{-1} [\widehat{g}]_{\underline{m}}\|^2 \right).$$

With the nondecreasing sequence $(\xi_m^2)_{m \geq 1}$ at hand we only need to replace the matrix $[T]_{\underline{m}}$ by its empirical counterpart (cf. Subsection 3.1).

DEFINITION 4.2. Let a_n and M_n^h be as in Definition 4.1 then for all $n \geq 1$ define

$$\widehat{M}_n := \min \left\{ 2 \leq m \leq M_n^h : m^3 \|\widehat{T}_{\underline{m}}^{-1}\|^2 \max_{1 \leq j \leq m} [h]_j^2 > a_n \right\} - 1$$

where we set $\widehat{M}_n := M_n^h$ if the min runs over an empty set. Thus, $\widehat{M}_n$ takes values between 1 and M_n^h . Then we introduce for all $m \geq 1$ an empirical analogon of pen_m by

$$\widehat{\text{pen}}_m := 204 \xi_m^2 (1 + \log n) n^{-1} \max_{1 \leq m' \leq m} \|[h]_{\underline{m}'}^t [\widehat{T}_{\underline{m}'}^{-1}]\|^2. \quad (4.4)$$

Before we establish the next upper bound we introduce

$$M_n^+ := \min \left\{ 2 \leq m \leq M_n^h : v_m^{-1} m^3 \max_{1 \leq j \leq m} [h]_j^2 > 4Da_n \right\} - 1 \quad (4.5)$$

where $M_n^+ := M_n^h$ if the min runs over an empty set. Thus, M_n^+ takes values between 1 and M_n^h . As in the partial adaptive case we do not attain up to a constant the minimax rate of convergence $\mathcal{R}^h$. A logarithmic term must be paid for adaption as we see in the next assertion.

THEOREM 4.2. Let the assumptions of Theorem 4.1 be satisfied. Additionally suppose that $(M_n^+ + 1)^2 \log n = o(nv_{M_n^++1})$ as $n \rightarrow \infty$ and $\sup_{j \geq 1} \mathbb{E} |e_j(Z)|^{20} \leq \eta^{20}$. Then for all $n \geq 1$ we have

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma^\infty} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbb{E} |\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2 \leq \mathcal{C} \mathcal{R}_{n(1+\log n)}^h$$

for a constant $\mathcal{C} > 0$ only depending on the classes $\mathcal{F}_\gamma^\rho$, $\mathcal{T}_{d,D}^v$, the constants σ , η and the representer h .

REMARK 4.2. Note that below in all examples illustrating Theorem 4.2 the condition $(M_n^+ + 1)^2 \log n = o(nv_{M_n^++1})$ as n tends to infinity is automatically satisfied. $\square$

As in the case of minimax optimal estimation we now present an upper bound uniformly over the class $\mathcal{F}_\omega^\tau$ of representer. For this purpose define $M_n^\omega := \max\{1 \leq m \leq \lfloor n^{1/4} \rfloor : \max_{1 \leq j \leq m} (\omega_j^{-1}) \leq n\}$. In the definition of the bounds $\widehat{M}_n$, M_n^+ , and M_n^- (cf. Appendix 4) we replace M_n^h and $\max_{1 \leq j \leq m} [h]_j^2$ by M_n^ω and $\max_{1 \leq j \leq m} \omega_j^{-1}$, respectively. Consequently, by employing $\|h\|_{1/\gamma}^2 \leq \tau$ and $\mathcal{R}_n^h \leq \tau \mathcal{R}_n^\omega$ for all $h \in \mathcal{F}_\omega^\tau$ the next result follows line by line the proof of Theorem 4.2 and hence its proof is omitted.

COROLLARY 4.3. Under the conditions of Theorem 4.2 we have for all $n \geq 1$

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma^\infty} \sup_{\varphi \in \mathcal{F}_\gamma^\rho, h \in \mathcal{F}_\omega^\tau} \mathbb{E} |\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2 \leq \mathcal{C} \mathcal{R}_{n(1+\log n)}^\omega$$

where the constant $\mathcal{C} > 0$ depends on the parameter spaces $\mathcal{F}_\gamma^\rho$, $\mathcal{F}_\omega^\tau$, $\mathcal{T}_{d,D}^v$, and the constants σ , η .

Illustration by classical smoothness assumptions. Let us illustrate the cost due to adaption by considering classical smoothness assumptions as discussed in Subsection 3.4. In Theorem 4.2 and Corollary 4.3, respectively, we have seen that the adaptive estimator $\widehat{\ell}_{\widehat{m}}$ attains within a constant the rates $\mathcal{R}_{\text{adapt}}^h$ and $\mathcal{R}_{\text{adapt}}^\omega$. Let us now present the orders of these rates by considering the example configurations of Remark 2.1. The proof of the following result is omitted because of the analogy with the proof of Proposition 3.5.

PROPOSITION 4.4. *Assume an iid. n -sample of (Y, Z, W) from the model (1.1a–1.1b) with conditional expectation operator $T \in \mathcal{T}_{d,D}^v$, error term U such that $P_{U|W} \in \mathcal{U}_\sigma^\infty$, and $\mathbb{E}[Y^2] > 0$. Then for the example configurations of Remark 2.1 we obtain*

(pp) *if in addition $3 < 2p + 2\min(s, 0)$ that $m_n^\circ \sim (n(1 + \log n)^{-1})^{1/(2p+2a)}$ and*

$$(i) \mathcal{R}_{n(1+\log n)^{-1}}^h \sim \begin{cases} (n^{-1}(1 + \log n))^{(2p+2s-1)/(2p+2a)}, & \text{if } s - a < 1/2 \\ n^{-1}(1 + \log n)^2, & \text{if } s - a = 1/2 \\ n^{-1}(1 + \log n), & \text{if } s - a > 1/2, \end{cases}$$

$$(ii) \mathcal{R}_{n(1+\log n)^{-1}}^\omega \sim \max((n^{-1}(1 + \log n))^{(p+s)/(p+a)}, n^{-1}(1 + \log n)).$$

(pe) $m_n^\circ \sim \log(n(1 + \log n)^{-(a+p)/a})^{1/2a}$ and

$$(i) \mathcal{R}_{n(1+\log n)^{-1}}^h \sim (1 + \log n)^{-(2p+2s-1)/(2a)},$$

$$(ii) \mathcal{R}_{n(1+\log n)^{-1}}^\omega \sim (1 + \log n)^{-(p+s)/a}.$$

(ep) $m_n^\circ \sim \log(n(1 + \log n)^{-(a+p)/p})^{1/2p}$ and

$$(i) \mathcal{R}_{n(1+\log n)^{-1}}^h \sim \begin{cases} n^{-1}(1 + \log n)^{(2a+2p-2s+1)/(2p)}, & \text{if } s - a < 1/2 \\ n^{-1}(1 + \log n)(\log \log n), & \text{if } s - a = 1/2 \\ n^{-1}(1 + \log n), & \text{if } s - a > 1/2, \end{cases}$$

$$(ii) \mathcal{R}_{n(1+\log n)^{-1}}^\omega \sim \max(n^{-1}(\log n)^{(a+p-s)/p}, n^{-1}(1 + \log n)).$$

Let us revisit Examples 3.1 and 3.2. In the following, we apply the general theory to adaptive pointwise estimation and adaptive estimation of averages of the structural function φ .

EXAMPLE 4.1. Consider the point evaluation functional $\ell_{h_{t_0}}(\varphi) = \varphi(t_0)$, $t_0 \in [0, 1]$, introduced in Example 3.1. In this case, the estimator $\widehat{\ell}_{\widehat{m}}$ with dimension parameter $\widehat{m}$ selected as a minimizer of criterion (1.2a–1.2b) writes as

$$\widehat{\varphi}_{\widehat{m}}(t_0) := \begin{cases} [e(t_0)]_{\widehat{m}}^t [\widehat{T}]_{\widehat{m}}^{-1} [\widehat{g}]_{\widehat{m}}, & \text{if } [\widehat{T}]_{\widehat{m}} \text{ is nonsingular and } \|[\widehat{T}]_{\widehat{m}}^{-1}\| \leq \sqrt{n}, \\ 0, & \text{otherwise} \end{cases}$$

where $\widehat{\varphi}_m$ is an estimator proposed by Johannes and Schwarz [2010]. Then $\widehat{\varphi}_{\widehat{m}}(t_0)$ attains within a constant the rate of convergence $\mathcal{R}_{\text{adapt}}^{h_{t_0}}$. Applying Proposition 4.4 gives

$$(pp) \mathcal{R}_{n(1+\log n)^{-1}}^{h_{t_0}} \sim (n^{-1}(1 + \log n))^{(2p-1)/(2p+2a)},$$

$$(ep) \mathcal{R}_{n(1+\log n)^{-1}}^{h_{t_0}} \sim (1 + \log n)^{-(2p-1)/(2a)},$$

$$(ep) \mathcal{R}_{n(1+\log n)^{-1}}^{h_{t_0}} \sim n^{-1}(1 + \log n)^{(2a+2p+1)/(2p)}.$$

□

EXAMPLE 4.2. Consider the linear functional $\ell_h(\varphi) = \int_0^b \varphi(t)dt$ with representer $h := \mathbb{1}_{[0,b]}$ introduced in Example 3.2. The mean squared error of the estimator $\widehat{\ell}_{\widehat{m}} = \int_0^b \widehat{\varphi}_{\widehat{m}}(t)dt$ is bounded up to a constant by $\mathcal{R}_{\text{adapt}}^h$. Applying Proposition 4.4 gives

$$\begin{aligned} \text{(pp)} \quad \mathcal{R}_{n(1+\log n)^{-1}}^h &\sim \begin{cases} (n^{-1}(1+\log n))^{(2p+1)/(2p+2a)}, & \text{if } a > 1/2, \\ n^{-1}(1+\log n)^2, & \text{if } a = 1/2, \\ n^{-1}(1+\log n), & \text{otherwise,} \end{cases} \\ \text{(ep)} \quad \mathcal{R}_{n(1+\log n)^{-1}}^h &\sim (1+\log n)^{-(2p+1)/(2a)}, \\ \text{(ep)} \quad \mathcal{R}_{n(1+\log n)^{-1}}^h &\sim \begin{cases} n^{-1}(1+\log n)^{(2a+2p-1)/(2p)}, & \text{if } a > 1/2, \\ n^{-1}(1+\log n)(\log \log n), & \text{if } a = 1/2, \\ n^{-1}(1+\log n), & \text{otherwise.} \end{cases} \end{aligned}$$

□

A. Appendix

A.1. Proof of the lower bound given in Section 2.

PROOF OF THEOREM 2.1. Define the function $\varphi_* := \left(\frac{\zeta \kappa \alpha_n^*}{\sum_{l=1}^{m_n^*} [h]_l^2 v_l^{-1}} \right)^{1/2} \sum_{j=1}^{m_n^*} [h]_j v_j^{-1} e_j$ with $\zeta := \min(1/(2d), \rho)$. Since $(\gamma_j^{-1} v_j)_{j \geq 1}$ is nonincreasing and by using the definition of κ given in (2.5) it follows that φ_* and in particular $\varphi_\theta := \theta \varphi_*$ for $\theta \in \{-1, 1\}$ belong to $\mathcal{F}_\gamma^\rho$. 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) . Consider $U_\theta := [T\varphi_\theta](W) - \varphi_\theta(Z) + V$, then $P_{U_\theta|W}$ belongs to $\mathcal{U}_\sigma$ for all $\sigma^4 \geq (\sqrt{3} + 4\rho \sum_{j \geq 1} \gamma_j^{-1} \eta^2)^2$, which can be realized as follows. Obviously, we have $\mathbb{E}[U_\theta|W] = 0$. Moreover, we have $\sup_j \mathbb{E}[e_j^4(Z)|W] \leq \eta^4$ implies $\mathbb{E}[\varphi_\theta^4(Z)|W] \leq \rho^2 (\sum_{j \geq 1} \gamma_j^{-1})^2 \mathbb{E}[e_j^4(Z)|W] \leq \rho^2 \eta^4 (\sum_{j \geq 1} \gamma_j^{-1})^2$ and thus, $|[T\varphi_\theta](W)|^4 \leq \mathbb{E}[\varphi_\theta^4(Z)|W] \leq \rho^2 \eta^4 (\sum_{j \geq 1} \gamma_j^{-1})^2$. From the last two bounds we deduce $\mathbb{E}[U_\theta^4|W] \leq 16 \mathbb{E}[\varphi_\theta^4(Z)|W] + 6 \text{Var}(\varphi_\theta(Z)|W) + 3 \leq (\sqrt{3} + 4\rho \eta^2 \sum_{j \geq 1} \gamma_j^{-1})^2$. Consequently, for each θ iid. copies (Y_i, Z_i, W_i) , $1 \leq i \leq n$, 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 and by $\mathbb{E}_\theta$ the expectation with respect to P_θ . In case of P_θ the conditional distribution of Y given W is Gaussian with mean $[T\varphi_\theta](W)$ and variance 1. The log-likelihood of P_1 with respect to P_{-1} is given by

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

Since $T \in \mathcal{T}_d^v$ the Kullback-Leibler divergence satisfies $KL(P_1, P_{-1}) \leq \mathbb{E}_1[\log(dP_1/dP_{-1})] = 2n\|T\varphi_*\|_W^2 \leq 2nd\|\varphi_*\|_v^2$. It is well known that the Hellinger distance $H(P_1, P_{-1})$ satisfies $H^2(P_1, P_{-1}) \leq KL(P_1, P_{-1})$ and thus, employing again the definition of κ we have

$$H^2(P_1, P_{-1}) \leq 2nd \sum_{j=1}^{m_n^*} [\varphi_*]_j^2 v_j = 2nd \frac{\zeta \kappa \alpha_n^*}{\sum_{l=1}^{m_n^*} [h]_l^2 v_l^{-1}} \sum_{j=1}^{m_n^*} \frac{[h]_j^2}{v_j} = 2d\zeta \frac{\kappa \alpha_n^*}{n^{-1}} \leq 2d\zeta \leq 1. \quad (\text{A.1})$$

Consider the Hellinger affinity $\rho(P_1, P_{-1}) = \int \sqrt{dP_1 dP_{-1}}$ then for any estimator $\check{\ell}$ it holds

$$\begin{aligned} \rho(P_1, P_{-1}) &\leq \int \frac{|\check{\ell} - \ell_h(\varphi_1)|}{2|\ell_h(\varphi_*)|} \sqrt{dP_1 dP_{-1}} + \int \frac{|\check{\ell} - \ell_h(\varphi_{-1})|}{2|\ell_h(\varphi_*)|} \sqrt{dP_1 dP_{-1}} \\ &\leq \left(\int \frac{|\check{\ell} - \ell_h(\varphi_1)|^2}{4|\ell_h(\varphi_*)|^2} dP_1 \right)^{1/2} + \left(\int \frac{|\check{\ell} - \ell_h(\varphi_{-1})|^2}{4|\ell_h(\varphi_*)|^2} dP_{-1} \right)^{1/2}. \end{aligned} \quad (\text{A.2})$$

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

$$\mathbb{E}_1 |\check{\ell} - \ell_h(\varphi_1)|^2 + \mathbb{E}_{-1} |\check{\ell} - \ell_h(\varphi_{-1})|^2 \geq \frac{1}{2} |\ell_h(\varphi_*)|^2. \quad (\text{A.3})$$

Obviously, $|\ell_h(\varphi_*)|^2 = \zeta \kappa \alpha_n^* \sum_{j=1}^{m_n^*} [h]_j^2 v_j^{-1}$. From (A.3) together with the last identity we conclude for any possible estimator $\check{\ell}$

$$\begin{aligned} \sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_U | W \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbb{E} |\check{\ell} - \ell_h(\varphi)|^2 &\geq \sup_{\theta \in \{-1, 1\}} \mathbb{E}_\theta |\check{\ell} - \ell_h(\varphi_*^{(\theta)})|^2 \\ &\geq \frac{1}{2} \left\{ \mathbb{E}_1 |\check{\ell} - \ell_h(\varphi_1)|^2 + \mathbb{E}_{-1} |\check{\ell} - \ell_h(\varphi_{-1})|^2 \right\} \\ &\geq \frac{\kappa}{4} \min \left(\frac{1}{2d}, \rho \right) \alpha_n^* \sum_{j=1}^{m_n^*} [h]_j^2 v_j^{-1}. \end{aligned} \quad (\text{A.4})$$

Consider now $\tilde{\varphi}_* := \left(\frac{\zeta \kappa}{\sum_{l > m_n^*} [h]_l^2 \gamma_l^{-1}} \right)^{1/2} \sum_{j > m_n^*} [h]_j \gamma_j^{-1} e_j$, which belongs to $\mathcal{F}_\gamma^\rho$ since $\kappa \leq 1$ and $\zeta \leq \rho$. Moreover, since $(\gamma_j^{-1} v_j)_{j \geq 1}$ is nonincreasing and by using the definition of κ given in (2.5) we have

$$2nd \sum_{j > m_n^*} [\tilde{\varphi}_*]_j^2 v_j = 2nd \frac{\zeta \kappa}{\sum_{l > m_n^*} [h]_l^2 \gamma_l^{-1}} \sum_{j > m_n^*} \frac{[h]_j^2 v_j}{\gamma_j^2} \leq 2d \zeta \frac{\kappa}{\gamma_{m_n^*}^* v_{m_n^*}^{-1}} \leq 2d \zeta \leq 1.$$

Thereby, following line by line the proof of (A.4) we obtain for any possible estimator $\check{\ell}$

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_U | W \in \mathcal{U}_\sigma} \sup_{\varphi \in \mathcal{F}_\gamma^\rho} \mathbb{E} |\check{\ell} - \ell_h(\varphi)|^2 \geq \frac{1}{4} |\ell_h(\tilde{\varphi}_*)|^2 = \frac{\kappa}{4} \min \left(\frac{1}{2d}, \rho \right) \sum_{j > m_n^*} [h]_j^2 \gamma_j^{-1}.$$

Combining, the last estimate and (A.4) implies the result of the theorem, which completes the proof. $\square$

A.2. Proofs of Section 3.

We begin by defining and recalling notations to be used in the proofs of this section. For $m \geq 1$ recall $\varphi_m = \sum_{j=1}^m [\varphi_m]_j e_j$ with $[\varphi_m]_{\underline{m}} = [T]_{\underline{m}}^{-1} [g]_{\underline{m}}$ keeping in mind that $[T]_{\underline{m}}$ is nonsingular. Then the identities $[T(\varphi - \varphi_m)]_{\underline{m}} = 0$ and $[\varphi_m - E_m \varphi]_{\underline{m}} = [T]_{\underline{m}}^{-1} [T E_m^\perp \varphi]_{\underline{m}}$ hold true. We denote $Q_m := [\hat{T}]_{\underline{m}} - [T]_{\underline{m}}$ and $V_m := [\hat{g}]_{\underline{m}} - [\hat{T}]_{\underline{m}} [\varphi_m]_{\underline{m}} = n^{-1} \sum_{i=1}^n (U_i + \varphi(Z_i) - \varphi_m(Z_i)) [f(W_i)]_{\underline{m}}$, where obviously $\mathbb{E} V_m = 0$. Moreover, let us introduce the events

$$\begin{aligned} \Omega_m &:= \{ \|\hat{T}_{\underline{m}}^{-1}\| \leq \sqrt{n} \}, \quad \mathcal{U}_m := \{ \sqrt{m} \|Q_m\| \|\hat{T}_{\underline{m}}^{-1}\| \leq 1/2 \} \\ \Omega_m^c &:= \{ \|\hat{T}_{\underline{m}}^{-1}\| > \sqrt{n} \} \quad \text{and} \quad \mathcal{U}_m^c = \{ \sqrt{m} \|Q_m\| \|\hat{T}_{\underline{m}}^{-1}\| > 1/2 \}. \end{aligned}$$

Observe that if $\sqrt{m}\|Q_m\|\|[T]_{\underline{m}}^{-1}\| \leq 1/2$ then the identity $[\widehat{T}]_{\underline{m}} = [T]_{\underline{m}}\{I + [T]_{\underline{m}}^{-1}Q_m\}$ implies by the usual Neumann series argument that $\|[\widehat{T}]_{\underline{m}}^{-1}\| \leq 2\|[T]_{\underline{m}}^{-1}\|$. Thereby, if $\sqrt{n} \geq 2\|[T]_{\underline{m}}^{-1}\|$ we have $\bar{\Omega}_m \subset \Omega_m$. These results will be used below without further reference. We shall prove at the end of this section four technical Lemmata (A.1 – A.4) which are used in the following proofs. Furthermore, we will denote by C universal numerical constants and by $C(\cdot)$ constants depending only on the arguments. In both cases, the values of the constants may change from line to line.

Proof of the consistency.

PROOF OF PROPOSITION 3.1. Consider for all $m \geq 1$ the decomposition

$$\begin{aligned} \mathbb{E}|\widehat{\ell}_m - \ell_h(\varphi)|^2 &= \mathbb{E}|\widehat{\ell}_m - \ell_h(\varphi)|^2 \mathbb{1}_{\Omega_m} + |\ell_h(\varphi)|^2 P(\Omega_m^c) \\ &\leq 2\mathbb{E}|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m} + 2|\ell_h(\varphi_m - \varphi)|^2 + |\ell_h(\varphi)|^2 P(\Omega_m^c) \end{aligned} \quad (\text{A.5})$$

where we bound each term separately. Let $\bar{\Omega}_m := \{\|Q_m\|\|[T]_{\underline{m}}^{-1}\| \leq 1/2\}$ and let $\bar{\Omega}_m^c$ denote its complement. By employing $\|[\widehat{T}]_{\underline{m}}^{-1}\| \mathbb{1}_{\bar{\Omega}_m} \leq 2\|[T]_{\underline{m}}^{-1}\|$ and $\|[\widehat{T}]_{\underline{m}}^{-1}\|^2 \mathbb{1}_{\Omega_m} \leq n$ it follows that

$$\begin{aligned} |\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m} &\leq 2|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1} V_m|^2 + 2|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1} Q_m [\widehat{T}]_{\underline{m}}^{-1} V_m|^2 \mathbb{1}_{\Omega_m} (\mathbb{1}_{\bar{\Omega}_m} + \mathbb{1}_{\bar{\Omega}_m^c}) \\ &\leq 2|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1} V_m|^2 + 2\|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\|^2 \left\{ 4\|[T]_{\underline{m}}^{-1}\|^2 \|Q_m\|^2 \|V_m\|^2 + n\|Q_m\|^2 \|V_m\|^2 \mathbb{1}_{\bar{\Omega}_m^c} \right\}. \end{aligned}$$

Thus, from estimate (A.9), (A.10), and (A.11) in Lemma A.1 we infer

$$\begin{aligned} \mathbb{E}|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m} &\leq C(\gamma)n^{-1}\|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\|^2 \eta^4 (\sigma^2 + \|\varphi - \varphi_m\|_\gamma^2) \\ &\quad \times \left\{ 1 + \frac{m^3}{n} \|[T]_{\underline{m}}^{-1}\|^2 + m^3 P^{1/4}(\bar{\Omega}_m^c) \right\}. \end{aligned} \quad (\text{A.6})$$

Let $m = m_n$ satisfying $m_n^{-1} = o(1)$, $m_n = o(n)$, and condition (3.3). We have $\sqrt{n} \geq 2\|[T]_{\underline{m}_n}^{-1}\|$ and thus, $\Omega_{m_n}^c \subset \bar{\Omega}_{m_n}^c$ for n sufficiently large. From Lemma A.3 it follows that $m_n^{12} P(\bar{\Omega}_{m_n}^c) \leq 2 \exp \{ -m_n (32\eta^2 n^{-1} m_n^3 \|[T]_{\underline{m}_n}^{-1}\|^2)^{-1} + 14 \log m_n \} = O(1)$ as $n \rightarrow \infty$ since $m_n (4n^{-1} m_n^3 \|[T]_{\underline{m}_n}^{-1}\|^2)^{-1} \leq 4\eta^2 n$ for n sufficiently large. Thus, in particular $P(\Omega_{m_n}^c) = o(1)$. Consequently, as $n \rightarrow \infty$ we obtain $\mathbb{E}|\widehat{\ell}_{m_n} - \ell_h(\varphi_{m_n})|^2 \mathbb{1}_{\Omega_{m_n}} = o(1)$ since $\|[h]_{\underline{m}_n}^t [T]_{\underline{m}_n}^{-1}\|^2 = o(n)$. Moreover, as $n \rightarrow \infty$ it holds $|\ell_h(\varphi_{m_n}) - \ell_h(\varphi)|^2 \leq \|h\|_{1/\gamma} \|\varphi - \varphi_{m_n}\|_\gamma = o(1)$ due to condition (3.2), and $|\ell_h(\varphi)|^2 P(\Omega_{m_n}^c) \leq \|h\|_{1/\gamma} \|\varphi\|_\gamma P(\Omega_{m_n}^c) = o(1)$. This together with decomposition (A.5) proves the result. $\square$

PROOF OF COROLLARY 3.2. The assertion follows directly from Proposition 3.1, it only remains to check conditions (3.2) and (3.3). We make use of decomposition $\|\varphi - \varphi_m\|_\gamma \leq \|E_m^\perp \varphi\|_\gamma + \|E_m \varphi - \varphi_m\|_\gamma$. As in the proof of Lemma A.2 we conclude $\|E_m \varphi - \varphi_m\|_\gamma^2 \leq \|E_m^\perp \varphi\|_\gamma \sup_m \sup_{\|\phi\|_\gamma=1} \|T_m^{-1} F_m T E_m^\perp \phi\|_\gamma \leq Dd \|E_m^\perp \varphi\|_\gamma$. By using Lebesgue's dominated convergence theorem we observe $\|E_m^\perp \varphi\|_\gamma = o(1)$ as $m \rightarrow \infty$ and hence (3.2) holds. Condition $T \in \mathcal{T}_{d,D}^v$ implies $\|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\|^2 \leq D \sum_{j=1}^m [h]_j^2 v_j^{-1}$ and $\|[T]_{\underline{m}}^{-1}\|^2 \leq Dv_m^{-1}$ for all $m \geq 1$ since v is nonincreasing. Thereby, condition (3.5) implies condition (3.3), which completes the proof. $\square$

Proof of the upper bound.

PROOF OF THEOREM 3.3. The proof is based on inequality (A.5). Applying estimate (A.14) in Lemma A.2 gives $|\ell_h(\varphi_m - \varphi)|^2 \leq 2\rho \{ \sum_{j>m} [h]_j^2 \gamma_j^{-1} + Dd v_m \gamma_m^{-1} \sum_{j=1}^m [h]_j^2 v_j^{-1} \}$ for all $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_{1/\gamma}$. Since $|\ell_h(\varphi)|^2 \leq \|\varphi\|_\gamma^2 \|h\|_{1/\gamma}^2$ and $\|\varphi\|_\gamma^2 \leq \rho$ we conclude

$$\begin{aligned} \mathbb{E} |\widehat{\ell}_m - \ell_h(\varphi)|^2 &\leq 2 \mathbb{E} |\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m} \\ &\quad + 4\rho \left\{ \sum_{j>m} [h]_j^2 \gamma_j^{-1} + dD \frac{v_m}{\gamma_m} \sum_{j=1}^m [h]_j^2 v_j^{-1} \right\} + \rho \|h\|_{1/\gamma}^2 P(\Omega_m^c). \end{aligned} \quad (\text{A.7})$$

By employing $\|Q_m[\widehat{T}_m]^{-1}\|^2 \mathbb{1}_{\Omega_m} \leq m^{-1}$ and $\|[\widehat{T}_m]^{-1}\|^2 \mathbb{1}_{\Omega_m} \leq n$ it follows that

$$\begin{aligned} |\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m} &\leq 2|[h]_m^t [T]_m^{-1} V_m|^2 + 2m^{-1} \|[h]_m^t [T]_m^{-1}\|^2 \|V_m\|^2 \\ &\quad + 2n \|[h]_m^t [T]_m^{-1}\|^2 \|Q_m\|^2 \|V_m\|^2 \mathbb{1}_{\Omega_m^c}. \end{aligned}$$

Due to $T \in \mathcal{T}_{d,D}^v$ and $\varphi \in \mathcal{F}_\gamma^\rho$ we have $\|[h]_m^t [T]_m^{-1}\|^2 \leq D \sum_{j=1}^m [h]_j^2 / v_j$ and $\|\varphi - \varphi_m\|_\gamma^2 \leq 2\rho(1 + Dd)$ (cf. (A.13) in Lemma A.2), respectively. Thereby, similarly to the proof of Proposition 3.1 we get

$$\mathbb{E} |\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m} \leq C(\gamma) D(\sigma^2 + \eta^2 d D \rho) n^{-1} \sum_{j=1}^m [h]_j^2 v_j^{-1} \left\{ 1 + m^3 P(\mathcal{U}_m^c)^{1/4} \right\}.$$

Combining the last estimate with (A.7) yields

$$\begin{aligned} \mathbb{E} |\widehat{\ell}_m - \ell_h(\varphi)|^2 &\leq C(\gamma) D(\sigma^2 + \eta^2 d D \rho) \max \left\{ \sum_{j>m} [h]_j^2 \gamma_j^{-1}, \max \left(\frac{v_m}{\gamma_m}, n^{-1} \right) \sum_{j=1}^m [h]_j^2 v_j^{-1} \right\} \\ &\quad \times \left\{ 1 + m^3 P(\mathcal{U}_m^c)^{1/4} \right\} + \rho \|h\|_{1/\gamma}^2 P(\Omega_m^c). \end{aligned} \quad (\text{A.8})$$

Consider now the optimal choice $m = m_n^*$ defined in (2.3), then we have

$$\begin{aligned} \mathbb{E} |\widehat{\ell}_{m_n^*} - \ell_h(\varphi)|^2 &\leq C(\gamma) D \left\{ \sigma^2 + \rho(\eta^2 d D + \|h\|_{1/\gamma}^2) \right\} \mathcal{R}_n^h \\ &\quad \times \left\{ 1 + (m_n^*)^3 P(\mathcal{U}_{m_n^*}^c)^{1/4} + (\mathcal{R}_n^h)^{-1} P(\Omega_{m_n^*}^c) \right\} \end{aligned}$$

and hence, the assertion follows by making use of Lemma A.4. $\square$

Technical assertions.

The following paragraph gathers technical results used in the proofs of Section 3. Below we consider the set $\mathbb{S}^m := \{s \in \mathbb{R}^m : \|s\| = 1\}$.

LEMMA A.1. Suppose that $P_{U|W} \in \mathcal{U}_\sigma$ and that the joint distribution of (Z, W) satisfies Assumption 3. If in addition $\varphi \in \mathcal{F}_\gamma^\rho$ with γ satisfying Assumption 1, then for all $m \geq 1$ we have

$$\sup_{s \in \mathbb{S}^m} \mathbb{E} |s^t V_m|^2 \leq 2n^{-1} (\sigma^2 + C(\gamma) \eta^2 \|\varphi - \varphi_m\|_\gamma^2), \quad (\text{A.9})$$

$$\mathbb{E} \|V_m\|^4 \leq C(\gamma) (n^{-1} m \eta^2 (\sigma^2 + \|\varphi - \varphi_m\|_\gamma^2))^2, \quad (\text{A.10})$$

$$\mathbb{E} \|Q_m\|^8 \leq C (n^{-1} m^2 \eta^2)^4. \quad (\text{A.11})$$

PROOF. Proof of (A.9). Since $(\{U_i + \varphi(Z_i) - \varphi_m(Z_i)\} \sum_{j=1}^m s_j f_j(W_i)), 1 \leq i \leq n$, are iid. with mean zero we have $\mathbb{E}|s^t V_m|^2 = n^{-1} \mathbb{E}|\{U + \varphi(Z) - \varphi_m(Z)\} \sum_{j=1}^m s_j f_j(W)|^2$. Then (A.9) follows from $\mathbb{E}[U^2|W] \leq (\mathbb{E}[U^4|W])^{1/2} \leq \sigma^2$ and from Assumption 3 (i), i.e., $\sup_{j \in \mathbb{N}} \mathbb{E}[e_j^2(Z)|W] \leq \eta^2$. Indeed, applying condition $|j|^3 \gamma_j^{-1} = o(1)$ (cf. Assumption 1) gives $\sum_{j \geq 1} \gamma_j^{-1} \leq C(\gamma)$ and thus,

$$\begin{aligned} \mathbb{E}|\{\varphi(Z) - \varphi_m(Z)\} \sum_{j=1}^m s_j f_j(W)|^2 &\leq \|\varphi - \varphi_m\|_\gamma^2 \sum_{l=1}^\infty \gamma_l^{-1} \mathbb{E}|e_l(Z) \sum_{j=1}^m s_j f_j(W)|^2 \\ &\leq C(\gamma) \eta^2 \|\varphi - \varphi_m\|_\gamma^2 \sum_{j=1}^m s_j^2 = C(\gamma) \eta^2 \|\varphi - \varphi_m\|_\gamma^2. \end{aligned}$$

Proof of (A.10). Observe that for each $1 \leq j \leq m$, $(\{U_i + \varphi(Z_i) - \varphi_m(Z_i)\} f_j(W_i)), 1 \leq i \leq n$, are iid. with mean zero. It follows from Theorem 2.10 in Petrov [1995] that $\mathbb{E}\|V_m\|^4 \leq Cn^{-2}m^2 \sup_{j \in \mathbb{N}} \mathbb{E}|\{U + \varphi(Z) - \varphi_m(Z)\} f_j(W)|^4$. Thereby, (A.10) follows from $\mathbb{E}[U^4|W] \leq \sigma^4$ and $\sup_{j \in \mathbb{N}} \mathbb{E}[f_j^4(W)] \leq \eta^4$ together with $\mathbb{E}|\{\varphi(Z) - \varphi_m(Z)\} f_j(W)|^4 \leq C(\gamma) \eta^4 \|\varphi - \varphi_m\|_\gamma^4$, which can be realized as follows. Since $[T(\varphi - \varphi_m)]_j = 0$ we have $\{\varphi(Z) - \varphi_m(Z)\} f_j(W) = \sum_{l \geq 1} [\varphi - \varphi_m]_l \{e_l(Z) f_j(W) - [T]_{j,l}\}$. Furthermore, Assumption 3 (ii), i.e., $\sup_{j,l \in \mathbb{N}} \mathbb{E}|e_l(Z) f_j(W) - [T]_{j,l}|^4 \leq 4!\eta^4$, implies

$$\begin{aligned} \mathbb{E}|\{\varphi(Z) - \varphi_m(Z)\} f_j(W)|^4 &\leq \|\varphi - \varphi_m\|_\gamma^4 \mathbb{E} \left| \sum_{l \geq 1} \gamma_l^{-1} |e_l(Z) f_j(W) - [T]_{j,l}|^2 \right|^2 \\ &\leq C(\gamma) \eta^4 \|\varphi - \varphi_m\|_\gamma^4. \end{aligned}$$

Proof of (A.11). The random variables $(e_l(Z_i) f_j(W_i) - [T]_{j,l}), 1 \leq i \leq n$, are iid. with mean zero for each $1 \leq j, l \leq m$. Hence, Theorem 2.10 in Petrov [1995] implies $\mathbb{E}\|Q_m\|^8 \leq Cn^{-4}m^8 \sup_{j,l \in \mathbb{N}} \mathbb{E}|e_l(Z) f_j(W) - [T]_{j,l}|^8$ and thus, the assertion follows from Assumption 3 (ii), which completes the proof. $\square$

LEMMA A.2. If $T \in \mathcal{T}_{d,D}^v$ and $\varphi \in \mathcal{F}_\gamma^\rho$, then for all $m \geq 1$ we have

$$\|E_m \varphi - \varphi_m\|_\gamma^2 \leq D d \rho, \quad (\text{A.12})$$

$$\|\varphi - \varphi_m\|_\gamma^2 \leq 2(1 + D d) \rho, \quad (\text{A.13})$$

$$|\langle h, \varphi - \varphi_m \rangle_Z|^2 \leq 2\rho \sum_{j>m} \frac{[h]_j^2}{\gamma_j} + 2D d \rho \frac{v_m}{\gamma_m} \sum_{j=1}^m \frac{[h]_j^2}{v_j}. \quad (\text{A.14})$$

PROOF. Consider (A.12). Since $T \in \mathcal{T}_{d,D}^v$ the identity $[E_m \varphi - \varphi_m]_{\underline{m}} = -[T]_{\underline{m}}^{-1} [T E_m^\perp \varphi]_{\underline{m}}$ implies $\|E_m \varphi - \varphi_m\|_v^2 \leq D \|T E_m^\perp \varphi\|_W^2 \leq D d \|E_m^\perp \varphi\|_v^2$. Consequently,

$$\|E_m \varphi - \varphi_m\|_v^2 \leq D d \gamma_m^{-1} v_m \|\varphi\|_\gamma^2 \quad (\text{A.15})$$

because $(\gamma_j^{-1} v_j)_{j \geq 1}$ is nonincreasing and thus, $\|E_m \varphi - \varphi_m\|_\gamma^2 \leq \gamma_m v_m^{-1} \|E_m \varphi - \varphi_m\|_v^2$. By combination of the last estimate and (A.15) we obtain the assertion (A.12). By employing the decomposition $\|\varphi - \varphi_m\|_\gamma^2 \leq 2\|\varphi - E_m \varphi\|_\gamma^2 + 2\|E_m \varphi - \varphi_m\|_\gamma^2$ the bound (A.13) follows from (A.12) and $\|\varphi - E_m \varphi\|_\gamma^2 \leq \|\varphi\|_\gamma^2$. It remains to show (A.14). Applying the Cauchy-Schwarz inequality gives $|\langle h, \varphi - E_m \varphi \rangle_Z|^2 \leq \|\varphi\|_\gamma^2 \sum_{j>m} [h]_j^2 \gamma_j^{-1}$ and $|\langle h, E_m \varphi - \varphi_m \rangle_Z|^2 \leq D d \|\varphi\|_\gamma^2 v_m \gamma_m^{-1} \sum_{j=1}^m [h]_j^2 v_j^{-1}$ by (A.15). Thereby (A.14) follows from the inequality $|\langle h, \varphi - \varphi_m \rangle_Z|^2 \leq 2|\langle h, \varphi - E_m \varphi \rangle_Z|^2 + 2|\langle h, E_m \varphi - \varphi_m \rangle_Z|^2$, which completes the proof. $\square$

LEMMA A.3. Suppose that the joint distribution of (Z, W) satisfies Assumption 3. Then for all $n \geq 1$ and $m \geq 1$ we have

$$P(m^{-2}n\|Q_m\|^2 \geq t) \leq 2 \exp\left(-\frac{t}{8\eta^2} + 2 \log m\right) \quad \text{for all } 0 < t \leq 4\eta^2 n. \quad (\text{A.16})$$

Proof. Our proof starts with the observation that for all $j, l \in \mathbb{N}$ the condition (ii) in Assumption 3 implies for all $t > 0$

$$P\left(\left|\sum_{i=1}^n \{e_j(Z_i)f_l(W_i) - \mathbb{E}[e_j(Z)f_l(W)]\}\right| \geq t\right) \leq 2 \exp\left(\frac{-t^2}{4n\eta^2 + 2\eta t}\right),$$

which is just Bernstein's inequality (cf. Bosq [1998]). This implies for all $0 < t \leq 2\eta n$

$$\sup_{j, l \in \mathbb{N}} P\left(\left|\sum_{i=1}^n \{e_j(Z_i)f_l(W_i) - \mathbb{E}[e_j(Z)f_l(W)]\}\right| \geq t\right) \leq 2 \exp\left(-\frac{t^2}{8\eta^2 n}\right). \quad (\text{A.17})$$

It is well-known that $m^{-1}\|[A]_{\underline{m}}\| \leq \max_{1 \leq j, l \leq m} |[A]_{j, l}|$ for any $m \times m$ matrix $[A]_{\underline{m}}$. Combining the last estimate and (A.17) we obtain for all $0 < t \leq 2\eta n^{1/2}$

$$\begin{aligned} P(m^{-1}n^{1/2}\|Q_m\| \geq t) &\leq \sum_{j, l=1}^m P\left(\left|\sum_{i=1}^n (e_j(Z_i)f_l(W_i) - \mathbb{E}[e_j(Z)f_l(W)])\right| \geq n^{1/2}t\right) \\ &\leq 2 \exp\left(-\frac{t^2}{8\eta^2} + 2 \log m\right). \end{aligned}$$

□

LEMMA A.4. Under the conditions of Theorem 3.3 we have for all $n \geq 1$

$$(m_n^*)^{12} P(\mathcal{U}_{m_n^*}^c) \leq C(\gamma, v, \eta, D) \quad (\text{A.18})$$

$$(\mathcal{R}_n^h)^{-1} P(\Omega_{m_n^*}^c) \leq C(\gamma, v, \eta, h, D). \quad (\text{A.19})$$

PROOF. Proof of (A.18). Since $\|[T]_{\underline{m}}^{-1}\|^2 \leq Dv_m^{-1}$ due to $T \in \mathcal{T}_{d, D}^v$ it follows from Lemma A.3 for all $m, n \geq 1$ that

$$P(\mathcal{U}_m^c) \leq P\left(m^{-2}n\|Q_m\|^2 > \frac{nv_m}{4Dm^3}\right) \leq 2 \exp\left(-\frac{nv_m}{32D\eta^2 m^3} + 2 \log m\right)$$

since $(4Dm^3v_m^{-1})^{-1} \leq 1 \leq 4\eta^2$ for all $m \geq 1$. Due to condition (3.6) there exists $n_0 \geq 1$ such that $nv_{m_n^*} \geq 448D\eta^2(m_n^*)^3 \log m_n^*$ for all $n \geq n_0$. Consequently, $(m_n^*)^{12} P(\mathcal{U}_{m_n^*}^c) \leq 2$ for all $n \geq n_0$, while trivially $(m_n^*)^{12} P(\mathcal{U}_{m_n^*}^c) \leq (m_{n_0}^*)^{12}$ for all $n \leq n_0$, which gives (A.18) since n_0 and $m_{n_0}^*$ depend on γ, v, η and D only.

Consider (A.19). Let $n_0 \in \mathbb{N}$ such that $\max\{|\log \mathcal{R}_n^h|, (\log m_n^*)\}(m_n^*)^3 \leq nv_{m_n^*}(96D\eta^2)^{-1}$ for all $n \geq n_0$. Observe that $\mathcal{U}_m \subset \Omega_m$ if $n \geq 4Dv_m^{-1}$. Since $(m_n^*)^{-3}nv_{m_n^*} \geq 96D\eta^2$ for all $n \geq n_0$ it follows $nv_{m_n^*} \geq 4D$ for all $n \geq n_0$ and hence $(\mathcal{R}_n^h)^{-1} P(\Omega_{m_n^*}^c) \leq (\mathcal{R}_n^h)^{-1} P(\mathcal{U}_{m_n^*}^c) \leq 2$ for all $n \geq n_0$ as in the proof of (A.18). Combining the last estimate and the elementary inequality $(\mathcal{R}_n^h)^{-1} P(\Omega_{m_n^*}^c) \leq (\mathcal{R}_{n_0}^h)^{-1}$ for all $n \leq n_0$ shows (A.19) since n_0 depends on γ, v, η, h and D only, which completes the proof. □

A.3. Proofs of Section 3.4

PROOF OF PROPOSITION 3.5. Proof of (pp). From the definition of m_n^* in (2.3) it follows $m_n^* \sim n^{1/(2p+2a)}$. Consider case (i). The condition $s - a < 1/2$ implies $n^{-1} \sum_{j=1}^{m_n^*} |j|^{2a-2s} \sim n^{-1} (m_n^*)^{2a-2s+1} \sim n^{-(2p+2s-1)/(2p+2a)}$ and moreover, $\sum_{j>m_n^*} |j|^{-2p-2s} \sim n^{-(2p+2s-1)/(2p+2a)}$ since $p + s > 1/2$. If $s - a = 1/2$ then $n^{-1} \sum_{j=1}^{m_n^*} |j|^{2a-2s} \sim n^{-1} \log(n^{1/(2p+2a)})$ and $\sum_{j>m_n^*} |j|^{-2p-2s} \sim n^{-1}$. In the case of $s - a > 1/2$ it follows that $\sum_{j=1}^{m_n^*} |j|^{2a-2s}$ is bounded whereas $\sum_{j>m_n^*} |j|^{-2p-2s} \lesssim n^{-1}$ and hence, $\mathcal{R}_n^h \sim n^{-1}$. To prove (ii) we make use of Corollary 2.2. We observe that if $s - a \geq 0$ the sequence ωv is bounded from below, and hence $\mathcal{R}_n^\omega \sim n^{-1}$. Otherwise, the condition $s - a < 0$ implies $\mathcal{R}_n^\omega \sim n^{-(p+s)/(p+a)}$. Proof of (pe). Note that m_n^* satisfies $m_n^* \sim \log(n(\log n)^{-p/a})^{1/(2a)}$. In order to prove (i), we calculate that $\sum_{j>m_n^*} |j|^{-2p-2s} \sim (\log n)^{(-2p-2s+1)/(2a)}$ and $n^{-1} \sum_{j=1}^{m_n^*} \exp(|j|^{2a}) |j|^{-2s} \lesssim (\log n)^{(-2p-2s+1)/(2a)}$. In case (ii) we immediately obtain $\mathcal{R}_n^\omega \sim (\log n)^{-(p+s)/a}$. Proof of (ep). It holds true $m_n^* \sim \log(n(\log n)^{-a/p})^{1/(2p)}$. Consider case (i). If $s - a < 1/2$ then $n^{-1} \sum_{j=1}^{m_n^*} |j|^{2a-2s} \sim n^{-1} (\log n)^{(2a-2s+1)/(2p)}$. If $s - a = 1/2$ we conclude $n^{-1} \sum_{j=1}^{m_n^*} |j|^{2a-2s} \sim n^{-1} \log(\log(n))$. On the other hand, the condition $s - a > 1/2$ implies that $\sum_{j=1}^{m_n^*} |j|^{2a-2s}$ is bounded and thus, we obtain the parametric rate n^{-1} . Moreover, it is easily seen that $\sum_{j>m_n^*} |j|^{-2s} \exp(-|j|^{2p}) \lesssim n^{-1} \sum_{j=1}^{m_n^*} |j|^{2a-2s}$. In case (ii) if $s - a \geq 0$ then the sequence ωv is bounded from below as mentioned above and thus, $\mathcal{R}_n^\omega \sim n^{-1}$. If $s - a < 0$ then $\mathcal{R}_n^\omega \sim n^{-1} (\log n)^{(a-s)/p}$, which completes the proof. $\square$

A.4. Proofs of Section 4

At the end of this section we shall prove six technical Lemmata (A.7 – A.12) which are used in the following proofs. Let us introduce a nondecreasing sequence $\Delta := (\Delta_m)_{m \geq 1}$ and its empirical analogon $\hat{\Delta} := (\hat{\Delta}_m)_{m \geq 1}$ by $\Delta_m := \max_{1 \leq m' \leq m} \|[h]_{\underline{m}'}^t, [T]_{\underline{m}'}^{-1}\|^2$ and $\hat{\Delta}_m := \max_{1 \leq m' \leq m} \|[h]_{\underline{m}'}^t, [\hat{T}]_{\underline{m}'}^{-1}\|^2$, respectively. Similarly to M_n^+ introduced in (4.5) we define

$$M_n^- := \min \left\{ 2 \leq m \leq M_n^h : 4Dv_m^{-1}m^3 \max_{1 \leq j \leq m} [h]_j^2 > a_n \right\} - 1 \quad (\text{A.20})$$

where we set $M_n^- := M_n^h$ if the set is empty. Thus, M_n^- takes values between 1 and M_n^h . In the following $\mathcal{C} > 0$ denotes a constant only depending on the classes $\mathcal{F}_\gamma^\rho$, $\mathcal{T}_{d,D}^v$, the constants σ , η and the representer h . For ease of notation, the value of $\mathcal{C} > 0$ may change from line to line.

PROOF OF THEOREM 4.1. The proof of the theorem is based on inequality (4.2). Observe that by Lemma A.10 we have $M_n^- \leq M_n \leq M_n^+$. Due to condition $(m_n^\circ)^3 \max_{1 \leq j \leq m_n^\circ} [h]_j^2 = o(a_n v_{m_n^\circ})$ as $n \rightarrow \infty$ there exists $n_0 \geq 1$ only depending on h , γ , and v such that for all $n \geq n_0$ it holds $m_n^\circ \leq M_n^-$. We distinguish in the following the cases $n \geq n_0$ and $n < n_0$. First, consider $n \geq n_0$. Applying Corollary A.6 together with estimate (4.2) implies

$$\mathbb{E}|\hat{\ell}_{\tilde{m}} - \ell_h(\varphi)|^2 \leq \mathcal{C} \left\{ \text{pen}_{m_n^\circ} + \text{bias}_{m_n^\circ} + n^{-1} \right\}.$$

From the definition of pen_m we infer $\text{pen}_m \leq 24(3\rho + 2\sigma^2)(1 + \log n)n^{-1}D \sum_{j=1}^m [h]_j^2 v_j^{-1}$ since $T \in \mathcal{T}_{d,D}^v$, $U \in \mathcal{U}_\sigma^\infty$, and $\varphi \in \mathcal{F}_\gamma^\rho$. Moreover, since $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_{1/\gamma}$ estimate (A.14) in

Lemma A.2 implies for all $1 \leq m \leq M_n^-$ that $\text{bias}_m \leq \min_{1 \leq m' \leq M_n^-} 2\rho \left\{ \sum_{j>m'} [h]_j^2 \gamma_j^{-1} + dDv_{m'} \gamma_{m'}^{-1} \sum_{j=1}^{m'} [h]_j^2 v_j^{-1} \right\}$.
Consequently,

$$\mathbb{E} |\widehat{\ell}_{\widetilde{m}} - \ell_h(\varphi)|^2 \leq \mathcal{C} \left\{ \max \left(\sum_{j>m_n^\circ} [h]_j^2 \gamma_j^{-1}, \alpha_n^\circ \sum_{j=1}^{m_n^\circ} [h]_j^2 v_j^{-1} \right) + n^{-1} \right\}.$$

Consider now $n < n_0$. Observe that for all $1 \leq m \leq M_n^h$ it holds

$$\begin{aligned} |\widehat{\ell}_m - \ell_h(\varphi)|^2 &\leq 2|[h]_{\underline{m}}^t [\widehat{T}]_{\underline{m}}^{-1} V_m|^2 \mathbb{1}_{\Omega_m} + 2(|\ell_h(\varphi_m - \varphi)|^2 + |\ell_h(\varphi)|^2 \mathbb{1}_{\Omega_m^c}) \\ &\leq 2n \|[h]_{\underline{M}_n^h}\|^2 \|V_{M_n^h}\|^2 + 2(|\ell_h(\varphi_m - \varphi)|^2 + |\ell_h(\varphi)|^2 \mathbb{1}_{\Omega_m^c}). \end{aligned} \quad (\text{A.21})$$

From the definition of M_n^h we infer $\|[h]_{\underline{M}_n^h}\|^2 \leq [h]_1^2 n^{5/4}$. Hence inequality (A.10) in Lemma A.1, inequality (A.13) in Lemma A.2 and Lemma A.12 yield for all $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_{1/\gamma}$

$$n \mathbb{E} |\widehat{\ell}_{\widetilde{m}} - \ell_h(\varphi)|^2 \leq 2[h]_1^2 n^{9/5} \|V_{M_n^h}\|^2 + 6\rho \|h\|_{1/\gamma}^2 (1 + Dd)n \leq \mathcal{C},$$

which proves the result. $\square$

LEMMA A.5. Consider $(\widetilde{\text{pen}}_m)_{m \geq 1}$ with $\widetilde{\text{pen}}_m := 24(24\mathbb{E}[U^2] + 96\eta^2 \rho m^3 \gamma_m^{-1})(1 + \log n)n^{-1}$. Then under the conditions of Theorem 4.1 we have for all $n \geq 1$

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_{U|W} \in \mathcal{U}_\sigma^\infty} \mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} \left(|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 - \frac{1}{6} \widetilde{\text{pen}}_m \right)_+ \leq \mathcal{C} n^{-1}.$$

PROOF. Similarly to the proof of Theorem 3.3 we obtain the decomposition

$$\begin{aligned} |\widehat{\ell}_m - \ell_h(\varphi_m)|^2 &\leq 2|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1} V_m|^2 + 2m^{-1} \|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\|^2 \|V_m\|^2 + \\ &\quad 2n \|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1} Q_m\|^2 \|V_m\|^2 \mathbb{1}_{\Omega_m^c} + |\ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m^c}. \end{aligned}$$

Observe that $\|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\|^2 \leq \Delta_m$ for all $m \geq 1$ and hence, we have for all $m_n^\circ \leq m \leq M_n^+$

$$\begin{aligned} \left(|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 - \frac{1}{6} \widetilde{\text{pen}}_m \right)_+ &\leq 2\Delta_m \left(\frac{|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1} V_m|^2}{\|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\|^2} - \frac{\widetilde{\text{pen}}_m}{24\Delta_m} \right)_+ \\ &\quad + 2\Delta_m \left(\frac{\|V_m\|^2}{m} - \frac{\widetilde{\text{pen}}_m}{24\Delta_m} \right)_+ + 2n\Delta_m \|Q_m\|^2 \|V_m\|^2 \mathbb{1}_{\Omega_m^c} + |\ell_h(\varphi_m)|^2 \mathbb{1}_{\Omega_m^c} \\ &=: I_m + II_m + III_m + IV_m. \end{aligned}$$

Consider the first two right hand side terms. We calculate

$$\mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} (I_m + II_m) \leq 4 \max_{m_n^\circ \leq m \leq M_n^+} \sup_{s \in \mathbb{S}^m} \mathbb{E} \left(|s^t V_m|^2 - \frac{\widetilde{\text{pen}}_m}{24\Delta_m} \right)_+ + \sum_{m=1}^{M_n^+} \Delta_m.$$

From the definition of $\widetilde{\text{pen}}$ we infer for all $s \in \mathbb{S}^m$ and $m_n^\circ \leq m \leq M_n^+$

$$\begin{aligned} n \mathbb{E} \left(|s^t V_m|^2 - \frac{\widetilde{\text{pen}}_m}{24\Delta_m} \right)_+ &\leq 2 \mathbb{E} \left((n^{-1/2} \sum_{i=1}^n U_i s^t [f(W_i)]_{\underline{m}})^2 - 12 \mathbb{E}[U^2] (1 + \log n) \right)_+ \\ &\quad + 2 \mathbb{E} \left((n^{-1/2} \sum_{i=1}^n (\varphi(Z_i) - \varphi_m(Z_i)) s^t [f(W_i)]_{\underline{m}})^2 - 48\eta^2 \rho m^3 \gamma_m^{-1} (1 + \log n) \right)_+ \\ &\leq C(\sigma, \eta, \gamma, \rho, D) n^{-1} \end{aligned}$$

where the last inequality follows from Lemma A.7 and A.8. Due to the definition of M_n^+ and since Δ is nondecreasing we have $n^{-1} \sum_{m=1}^{M_n^+} \Delta_m \leq D(n\nu_{M_n^+})^{-1} (M_n^+)^2 \max_{1 \leq j \leq M_n^+} [h]_j^2 \leq 4D^2$. Consequently, $\mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} (I_m + II_m) \leq \mathcal{C}n^{-1}$. Further, we obtain for $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_{1/\gamma}$

$$\begin{aligned} \mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} (III_m) &\leq n\Delta_{M_n^+} (\mathbb{E} \|Q_{M_n^+}\|^8)^{1/4} (\mathbb{E} \|V_{M_n^+}\|^4)^{1/2} P^{1/4} \left(\bigcup_{m=1}^{M_n^+} \mathcal{U}_m^c \right) \\ &\leq C(\gamma) \eta^4 (\sigma^2 + (1 + Dd)\rho) n^{-1} \Delta_{M_n^+} (M_n^+)^3 P^{1/4} \left(\bigcup_{m=1}^{M_n^+} \mathcal{U}_m^c \right) \end{aligned}$$

where the last inequality is due to Lemma A.1 and

$$\mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} (IV_m) \leq \rho \|h\|_{1/\gamma}^2 P \left(\bigcup_{m=1}^{M_n^+} \Omega_m^c \right).$$

Now applying $n^{-1} \Delta_{M_n^+} (M_n^+)^3 \leq 4D^2$ and Lemma A.9 gives $\mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} (III_m + IV_m) \leq \mathcal{C}n^{-1}$, which completes the proof. $\square$

COROLLARY A.6. *Under the conditions of Theorem 4.1 we have for all $n \geq 1$*

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{P_U|W \in \mathcal{U}_\sigma^\infty} \mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} \left(|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 - \frac{1}{6} \text{pen}_m \right)_+ \leq \mathcal{C}n^{-1}.$$

Proof. Observe that $m^3 \gamma_m^{-1} = o(1)$ and $\|\varphi - \varphi_m\|_Z^2 = o(1)$ as $m \rightarrow \infty$ due to Assumption 1 and $T \in \mathcal{T}_{d,D}^v$ (cf. proof of Corollary 3.2), respectively. Thereby, there exists a constant n_0 only depending on γ , ρ , and η such that for all $n \geq n_0$ and $m \geq m_n^\circ$ we have

$$24 \mathbb{E}[U^2] + 96\eta^2 \rho m^3 \gamma_m^{-1} \leq 72 (\mathbb{E}[Y^2] + \|\varphi_m\|_Z^2 + \|\varphi - \varphi_m\|_Z^2) + 96\eta^2 \rho m^3 \gamma_m^{-1} \leq \zeta_m^2. \quad (\text{A.22})$$

We distinguish in the following the cases $n < n_0$ and $n \geq n_0$. First, consider $n < n_0$. Due to $n^{-1} \sum_{m=1}^{M_n^+} \Delta_m \leq 4D^2$ and inequality (A.9) in Lemma A.1 we calculate for all $s \in \mathbb{S}^m$

$$\sum_{m=1}^{M_n^+} \Delta_m \mathbb{E} \left(|s^t V_m|^2 - \frac{\text{pen}_m}{24\Delta_m} \right)_+ \leq \sum_{m=1}^{M_n^+} \Delta_m \mathbb{E} |s^t V_m|^2 \leq 8n_0 D^2 (\sigma^2 + C(\gamma) \eta^2 \|\varphi - \varphi_m\|_\gamma^2) n^{-1}.$$

Therefore, following line by line the proof of Lemma A.5 it is easily seen that it holds $n \mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} (|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 - \frac{1}{6} \text{pen}_m)_+ \leq \mathcal{C}$. Consider now $n \geq n_0$. Inequality (A.22) implies $\widehat{\text{pen}}_m \leq \text{pen}_m$ and thus, $(|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 - \frac{1}{6} \text{pen}_m)_+ \leq (|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 - \frac{1}{6} \widehat{\text{pen}}_m)_+$ for all $m_n^\circ \leq m \leq M_n^+$. Thus, from Lemma A.5 we infer $n \mathbb{E} \max_{m_n^\circ \leq m \leq M_n^+} (|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 - \frac{1}{6} \widehat{\text{pen}}_m)_+ \leq \mathcal{C}$, which completes the proof of the corollary. $\square$

PROOF OF THEOREM 4.2. Similarly to the proof of Theorem 4.1 and since $\widehat{\text{pen}}$ is a non-decreasing sequence we have for all $1 \leq m \leq \widehat{M}_n$

$$|\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2 \lesssim \widehat{\text{pen}}_m + \text{bias}_m + \max_{m \leq m' \leq \widehat{M}_n} \left(|\widehat{\ell}_{m'} - \ell_h(\varphi_{m'})|^2 - \frac{1}{6} \widehat{\text{pen}}_{m'} \right)_+.$$

Let us introduce the set

$$\mathcal{A} := \{\text{pen}_m \leq \widehat{\text{pen}}_m \leq 8 \text{pen}_m, \quad 1 \leq m \leq M_n^+\} \cap \{M_n^- \leq \widehat{M}_n \leq M_n^+\},$$

then we conclude for all $1 \leq m \leq M_n^-$

$$|\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2 \mathbb{1}_{\mathcal{A}} \lesssim \text{pen}_m + \text{bias}_m + \max_{m \leq m' \leq M_n^+} \left(|\widehat{\ell}_{m'} - \ell_h(\varphi_{m'})|^2 - \frac{1}{6} \text{pen}_{m'} \right)_+.$$

Thereby, similarly as in the proof of Theorem 4.1 we obtain for all $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_{1/\gamma}$ the upper bound for all $n \geq 1$

$$\mathbb{E} |\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2 \mathbb{1}_{\mathcal{A}} \leq \mathcal{C} \mathcal{R}_{n(1+\log n)}^h. \quad (\text{A.23})$$

Let us now evaluate the risk of the adaptive estimator $\widehat{\ell}_{\widehat{m}}$ on $\mathcal{A}^c$. From the definition of M_n^h we infer $\|[h]_{M_n^h}\|^2 \leq [h]_1^2 n M_n^h$. Consequently, inequality (A.21) together with (A.10) in Lemma A.1, (A.13) in Lemma A.2 and Lemma A.12 yields for all $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_{1/\gamma}$

$$\mathbb{E} |\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2 \mathbb{1}_{\mathcal{A}^c} \leq 2 [h]_1^2 n^2 M_n^h (\mathbb{E} \|V_{M_n^h}\|^4)^{1/2} P(\mathcal{A}^c)^{1/2} + 6\rho \|h\|_{1/\gamma}^2 (1 + Dd) P(\mathcal{A}^c) \leq \mathcal{C} n^{-1}.$$

The result follows by combining the last inequality with (A.23). $\square$

Technical assertions.

The following paragraph gathers technical results used in the proofs of Section 4. In the following we denote $\xi_s(w) := \sum_{j=1}^m s_j f_j(w)$ where $s \in \mathbb{S}^m = \{s \in \mathbb{R}^m : \|s\| = 1\}$.

LEMMA A.7. *Let Assumptions 3 and 4 hold. Then for all $n \geq 1$ and $1 \leq m \leq \lfloor n^{1/4} \rfloor$ we have*

$$\sup_{P_{U|W} \in \mathcal{U}_\sigma^\infty} \sup_{s \in \mathbb{S}^m} \mathbb{E} \left[\left(\frac{1}{n} \left| \sum_{i=1}^n U_i \xi_s(W_i) \right|^2 - 12 \mathbb{E}[U^2](1 + \log n) \right)_+ \right] \leq C(\sigma, \eta) n^{-1}.$$

Proof. Let us denote $\delta = 12 \mathbb{E}[U^2](1 + \log n)$. Since the error term U satisfies Cramer's condition we may apply Bernstein's inequality and since $\mathbb{E}[U^2|W] \leq \sigma^2$ we have

$$\begin{aligned} & \mathbb{E} \left[\left(\frac{1}{n} \left| \sum_{i=1}^n U_i \xi_s(W_i) \right|^2 - \delta \right)_+ | W_1, \dots, W_n \right] \\ &= \int_0^\infty P \left(\sum_{i=1}^n U_i \xi_s(W_i) \geq \sqrt{n(t + \delta)} | W_1, \dots, W_n \right) dt \\ &\leq \int_0^\infty \exp \left(\frac{-n(t + \delta)}{8\sigma^2 \sum_{i=1}^n |\xi_s(W_i)|^2} \right) dt + \int_0^\infty \exp \left(\frac{-\sqrt{n(t + \delta)}}{4\sigma \max_{1 \leq i \leq n} |\xi_s(W_i)|} \right) dt. \quad (\text{A.24}) \end{aligned}$$

Consider the first summand of (A.24). Let us introduce the set

$$\mathcal{B} := \left\{ \forall 1 \leq j, l \leq m : \left| n^{-1} \sum_{i=1}^n f_j(W_i) f_l(W_i) - \delta_{jl} \right| \leq \frac{\log n}{3\sqrt{n}} \right\}$$

where $\delta_{jl} = 1$ if $j = l$ and zero otherwise. Applying Cauchy-Schwarz's inequality twice we observe on $\mathcal{B}$ for all $n \geq 1$ and $1 \leq m \leq M_n^+$

$$|n^{-1} \sum_{i=1}^n |\xi_s(W_i)|^2 - 1| \mathbb{1}_{\mathcal{B}} \leq \sum_{j,l=1}^m |z_j| |z_l| n^{-1} \sum_{i=1}^n f_j(W_i) f_l(W_i) - \delta_{jl} \mathbb{1}_{\mathcal{B}} \leq \frac{1}{2}$$

since $n^{-1/4} \log n \leq 3/2$ for all $n \geq 1$. Thereby, it holds $n^{-1} \sum_{i=1}^n |\xi_s(W_i)|^2 \mathbb{1}_{\mathcal{B}} \leq 3/2$ and thus,

$$n \mathbb{E} \left[\int_0^\infty \exp \left(\frac{-n(t+\delta)}{8\sigma^2 \sum_{i=1}^n |\xi_s(W_i)|^2} \right) dt \mathbb{1}_{\mathcal{B}} \right] \leq 12\sigma^2 \exp \left(\log n - \frac{\delta}{12\sigma^2} \right) \leq 6\sigma^2. \quad (\text{A.25})$$

On the complement of $\mathcal{B}$ observe that $\sup_{j,l} \text{Var}(f_j(W) f_l(W)) < \eta^2$ due that Assumption 3 (i) and thus, Assumption 4 together with Bernstein's inequality yields

$$\begin{aligned} P(\mathcal{B}^c) &\leq \sum_{j,l=1}^m P \left(3 \left| \sum_{i=1}^n f_j(W_i) f_l(W_i) - \delta_{jl} \right| > \sqrt{n} \log n \right) \\ &\leq 2m^2 \exp \left(- \frac{n(\log n)^2}{36n\eta^4 + 6\eta\sqrt{n} \log n} \right) \leq 2 \exp \left(2 \log m - \frac{(\log n)^2}{42\eta^4} \right). \end{aligned}$$

By Assumption 3 (i) it holds $\mathbb{E} |\xi_s(W)|^4 \leq \mathbb{E} |\sum_{j=1}^m f_j^2(W)|^2 \leq m^2 \eta^4$. Thereby

$$n \mathbb{E} \left[\int_0^\infty \exp \left(\frac{-n(t+\delta)}{8\sigma^2 \sum_{i=1}^n |\xi_s(W_i)|^2} \right) dt \mathbb{1}_{\mathcal{B}^c} \right] \leq 8\sigma^2 n (\mathbb{E} |\xi_s(W_1)|^4 P(\mathcal{B}^c))^{1/2} \leq 12\sigma^2 \eta^2 \quad (\text{A.26})$$

for all $n \geq \exp(126\eta^4)$ and $1 \leq m \leq \lfloor n^{1/4} \rfloor$. For $n < \exp(126\eta^4)$ it holds $n \mathbb{E} [|\xi_s(W_1)|^2 \mathbb{1}_{\mathcal{B}^c}] < \exp(126\eta^4)$. Consider the second summand of (A.24). Since $\exp(-1/x)$, $x > 0$, is a concave function and $\mathbb{E} |\xi_s(W)|^4 \leq m^2 \eta^4$ we deduce for all $1 \leq m \leq \lfloor n^{1/4} \rfloor$

$$\begin{aligned} \mathbb{E} \left[\int_0^\infty \exp \left(\frac{-\sqrt{n(t+\delta)}}{4\sigma \max_{1 \leq i \leq n} |\xi_s(W_i)|} \right) dt \right] &\leq \int_0^\infty \exp \left(\frac{-\sqrt{n(t+\delta)}}{4\sigma \mathbb{E} \max_{1 \leq i \leq n} |\xi_s(W_i)|} \right) dt \\ &\leq \int_0^\infty \exp \left(\frac{-\sqrt{n(t+\delta)}}{4\sigma (n \mathbb{E} |\xi_s(W)|^4)^{1/4}} \right) dt \leq \int_0^\infty \exp \left(\frac{-n^{1/4} \sqrt{(t+\delta)}}{4\sigma \eta \sqrt{m}} \right) dt \\ &\leq 8\sigma \eta \sqrt{m/n} \exp \left(\frac{-n^{1/4} \sqrt{\delta}}{4\sigma \eta \sqrt{m}} \right) \left(n^{1/4} \sqrt{\delta} + 4\sigma \eta \sqrt{m} \right) \leq C(\sigma, \eta) n^{-1}. \quad (\text{A.27}) \end{aligned}$$

The assertion follows now by combining inequality (A.24) with (A.25), (A.26), and (A.27). $\square$

LEMMA A.8. *Let Assumptions 1 and 3 hold. Then for all $n \geq 1$ and $m \geq 1$ we have*

$$\sup_{T \in \mathcal{T}_{d,D}^v} \sup_{s \in \mathbb{S}^m} \mathbb{E} \left[\left(\frac{1}{n} \left| \sum_{i=1}^n (\varphi(Z_i) - \varphi_m(Z_i)) \xi_s(W_i) \right|^2 - 48\eta^2 \rho \frac{m^3}{\gamma_m} (1 + \log n) \right)_+ \right] \leq C(\eta, \gamma, \rho, D) n^{-1}.$$

Proof. Let us consider a sequence $w := (w_j)_{j \geq 1}$ with $w_j := j^2$. Since $[T(\varphi - \varphi_m)]_{\underline{m}} = 0$ we conclude for $m \geq 1$, $s \in \mathbb{S}^m$, and $k = 2, 3, \dots$ that

$$\begin{aligned} \mathbb{E} |(\varphi(Z) - \varphi_m(Z))\xi_s(W)|^k &= \mathbb{E} \left| \sum_{l=1}^{\infty} [\varphi - \varphi_m]_l \sum_{j=1}^m s_j (e_l(Z) f_j(W) - [T]_{jl}) \right|^k \\ &\leq \|\varphi - \varphi_m\|_w^k \mathbb{E} \left| \sum_{l=1}^{\infty} w_l^{-1} \sum_{j=1}^m (e_l(Z) f_j(W) - [T]_{jl}) \right|^{k/2} \\ &\leq \|\varphi - \varphi_m\|_w^k m^{k/2} (\pi/\sqrt{6})^k \sup_{j,l \in \mathbb{N}} \mathbb{E} |e_l(Z) f_j(W) - [T]_{jl}|^k \end{aligned}$$

where due to Assumption 3 (i) $\sup_{j,l \in \mathbb{N}} \text{Var}(e_l(Z) f_j(W)) \leq \eta^2$ and due to Assumption 3 (ii) it holds $\sup_{j,l \in \mathbb{N}} \mathbb{E} |e_l(Z) f_j(W) - [T]_{jl}|^k \leq k! \eta^k$ for $k \geq 3$. Moreover, similarly to the proof of (A.13) in Lemma A.2 we conclude $m^{k/2} \|\varphi - \varphi_m\|_w^k \leq (m^3 \gamma_m^{-1})^{k/2} (2 + 2Dd)^{k/2} \rho^{k/2}$. Let us denote $\mu_m := \eta(1 + Dd) \sqrt{6\rho m^3 \gamma_m^{-1}}$. Consequently, for all $m \geq 1$ we have $\mathbb{E} |(\varphi(Z) - \varphi_m(Z))\xi_s(W)|^2 \leq \mu_m^2$ and

$$\sup_{s \in \mathbb{S}^m} \mathbb{E} |(\varphi(Z) - \varphi_m(Z))\xi_s(W)|^k \leq \mu_m^k k! \text{ for } k = 3, 4, \dots \quad (\text{A.28})$$

Now Bernstein's inequality gives for all $m \geq 1$

$$\begin{aligned} \sup_{s \in \mathbb{S}^m} \mathbb{E} \left[\left(\frac{1}{n} \left| \sum_{i=1}^n (\varphi(Z_i) - \varphi_m(Z_i)) \xi_s(W_i) \right|^2 - 8\mu_m^2 (1 + \log n) \right)_+ \right] \\ \leq 2 \int_0^\infty \exp \left(-\frac{(t + \delta)}{8\mu_m^2} \right) dt + 2 \int_0^\infty \exp \left(-\frac{\sqrt{n(t + \delta)}}{4\mu_m} \right) dt \\ \leq 16\mu_m^2 \exp(-\log n) + 16\mu_m n^{-1/2} \exp \left(-\frac{\sqrt{n(1 + \log n)}}{2} \right) (4\mu_m + \sqrt{8n\mu_m^2 (1 + \log n)}) \\ \leq C(\eta, \gamma, \rho, D) n^{-1} \end{aligned}$$

and thus, the assertion follows. $\square$

LEMMA A.9. *Let $T \in \mathcal{T}_{d,D}^v$. Then for all $n \geq 1$ it holds*

$$P \left(\bigcup_{m=1}^{M_n^+} \mathcal{U}_m^c \right) \leq C(h, v, \eta, D) n^{-4}, \quad (\text{A.29})$$

$$P \left(\bigcup_{m=1}^{M_n^+} \Omega_m^c \right) \leq C(h, v, \eta, D) n^{-1}. \quad (\text{A.30})$$

Proof. Proof of (A.29). Since $T \in \mathcal{T}_{d,D}^v$ we have $\|[T]_{\underline{m}}^{-1}\|^2 \leq Dv_m^{-1}$ and thus, exploiting Lemma A.3 together with the definition of M_n^+ gives

$$n^4 P \left(\bigcup_{m=1}^{M_n^+} \mathcal{U}_m^c \right) \leq 2 \exp \left(-\frac{1}{48\eta D} \frac{nv_{M_n^+}}{(M_n^+)^3} + 3 \log M_n^+ + 4 \log n \right) \leq C(h, v, \eta, D).$$

Proof of (A.30). Due to the definition of M_n^+ there exists some $n_0 \geq 1$ such that $n \geq 4Dv_{M_n^+}^{-1}$ for all $n \geq n_0$. Thereby, condition $T \in \mathcal{T}_{d,D}^v$ implies $\max_{1 \leq m \leq M_n^+} \|[T]_{\underline{m}}^{-1}\|^2 \leq Dv_{M_n^+}^{-1} \leq n/4$ for all $n \geq n_0$. This gives $\bigcup_{m=1}^{M_n^+} \Omega_m^c \subset \bigcup_{m=1}^{M_n^+} \mathcal{V}_m^c$ and inequality (A.30) follows by making use of (A.29). If $n < n_0$ then $nP(\bigcup_{m=1}^{M_n^+} \Omega_m^c) \leq n_0$ and the assertion follows since n_0 only depends on h, v and D . $\square$

LEMMA A.10. *Let $T \in \mathcal{T}_{d,D}^v$. Then it holds $M_n^- \leq M_n \leq M_n^+$ for all $n \geq 1$.*

PROOF. Consider $M_n^- \leq M_n$. If $M_n^- = 1$ or $M_n = M_n^h$ the result is trivial. If $M_n = 1$, then clearly $M_n^- = 1$. It remains to consider $M_n^- > 1$ and $M_n^h > M_n > 1$. Due to $T \in \mathcal{T}_{d,D}^v$ it holds $\|[T]_{\underline{M_n+1}}^{-1}\|^{-2} \geq D^{-1}v_{M_n+1}$ and thus, by the definition of M_n and M_n^- it is easily seen that

$$\frac{v_{M_n^-}}{\max_{1 \leq j \leq M_n^-} [h]_j^2 (M_n^-)^3} > \frac{4v_{M_n+1}}{\max_{1 \leq j \leq M_n+1} [h]_j^2 (M_n+1)^3},$$

and thus, $M_n+1 > M_n^-$, i.e. $M_n \geq M_n^-$. Consider $M_n \leq M_n^+$. If $M_n = 1$ or $M_n^+ = M_n^h$ the result is trivial, while otherwise since $v_m^{-1} \leq \|[T]_{\underline{m}}^{-1}\|^2 \sup_{\|E_m \phi\|_v=1} \|F_m T E_m \phi\|^2 \leq D\|[T]_{\underline{m}}^{-1}\|^2$ due to condition $T \in \mathcal{T}_d^v$ with $d \leq D$ and by the definition of M_n and M_n^+ it follows

$$\frac{v_{M_n}}{\max_{1 \leq j \leq M_n} [h]_j^2 M_n^3} > \frac{4v_{M_n^++1}}{\max_{1 \leq j \leq M_n^++1} [h]_j^2 (M_n^++1)^3}.$$

Thus, $M_n^++1 > M_n$, i.e. $M_n^+ \geq M_n$, which completes the proof. $\square$

In the following, we make use of the notation $\sigma_Y^2 := \mathbb{E}[Y^2]$ and $\hat{\sigma}_Y^2 := n^{-1} \sum_{i=1}^n Y_i^2$. Further, let us introduce the events

$$\mathcal{H} := \{ \|Q_m\| \|[T]_{\underline{m}}^{-1}\| \leq 1/4 \quad \forall 1 \leq m \leq (M_n^++1) \}, \quad (\text{A.31})$$

$$\mathcal{G} := \{ \sigma_Y^2 \leq 2\hat{\sigma}_Y^2 \leq 3\sigma_Y^2 \}, \quad (\text{A.32})$$

$$\mathcal{J} := \left\{ \|[T]_{\underline{m}}^{-1} V_m\|^2 \leq \frac{1}{8} (\|[T]_{\underline{m}}^{-1} [g]_{\underline{m}}\|^2 + \sigma_Y^2) \quad \forall 1 \leq m \leq M_n^+ \right\}. \quad (\text{A.33})$$

LEMMA A.11. *Let $T \in \mathcal{T}_{d,D}^v$. Then it holds $\mathcal{H} \cap \mathcal{G} \cap \mathcal{J} \subset \mathcal{A}$.*

Proof. For all $1 \leq m \leq M_n^+$ observe that condition $\|Q_m\| \|[T]_{\underline{m}}^{-1}\| \leq 1/4$ yields by the usual Neumann series argument that $\|([I]_{\underline{m}} + Q_m [T]_{\underline{m}}^{-1})^{-1} - [I]_{\underline{m}}\| \leq 1/3$. Thus, using the identity $[\hat{T}]_{\underline{m}}^{-1} = [T]_{\underline{m}}^{-1} - [T]_{\underline{m}}^{-1} (([I]_{\underline{m}} + Q_m [T]_{\underline{m}}^{-1})^{-1} - [I]_{\underline{m}})$ we conclude

$$2\|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\| \leq 3\|[h]_{\underline{m}}^t [\hat{T}]_{\underline{m}}^{-1}\| \leq 4\|[h]_{\underline{m}}^t [T]_{\underline{m}}^{-1}\|.$$

Similarly, we have $2\|[T]_{\underline{m}}^{-1} v_m\| \leq 3\|[\hat{T}]_{\underline{m}}^{-1} v_m\| \leq 4\|[T]_{\underline{m}}^{-1} v_m\|$ for all $v_m \in \mathbb{R}^m$. Thereby, since $[\hat{T}]_{\underline{m}}^{-1} V_m = [\hat{T}]_{\underline{m}}^{-1} [\hat{g}]_{\underline{m}} - [T]_{\underline{m}}^{-1} [g]_{\underline{m}}$ we conclude

$$\begin{aligned} \|[T]_{\underline{m}}^{-1} [g]_{\underline{m}}\|^2 &\leq (32/9) \|[T]_{\underline{m}}^{-1} V_m\|^2 + 2\|[\hat{T}]_{\underline{m}}^{-1} [\hat{g}]_{\underline{m}}\|^2, \\ \|[\hat{T}]_{\underline{m}}^{-1} [\hat{g}]_{\underline{m}}\|^2 &\leq (32/9) \|[T]_{\underline{m}}^{-1} V_m\|^2 + 2\|[T]_{\underline{m}}^{-1} [g]_{\underline{m}}\|^2. \end{aligned}$$

On $\mathcal{J}$ it holds $\|[T]_{\underline{m}}^{-1}V_m\|^2 \leq \frac{1}{8}(\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + \sigma_Y^2)$. Thereby, the last two inequalities imply

$$\begin{aligned} (5/9)(\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + \sigma_Y^2) &\leq \sigma_Y^2 + 2\|\widehat{[T]}_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2, \\ \|\widehat{[T]}_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2 &\leq (22/9)\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + (4/9)\sigma_Y^2. \end{aligned}$$

On $\mathcal{G}$ it holds $\sigma_Y^2 \leq 2\widehat{\sigma}_Y^2 \leq 3\sigma_Y^2$ which gives

$$\begin{aligned} (5/9)(\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + \sigma_Y^2) &\leq (3/2)\widehat{\sigma}_Y^2 + 2\|\widehat{[T]}_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2, \\ \|\widehat{[T]}_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2 + \widehat{\sigma}_Y^2 &\leq (22/9)\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + (10/9)\sigma_Y^2. \end{aligned}$$

Combing the last two inequalities we conclude for all $1 \leq m \leq M_n^+$

$$(5/18)(\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + \sigma_Y^2) \leq \|\widehat{[T]}_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2 + \widehat{\sigma}_Y^2 \leq (22/9)(\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + \sigma_Y^2).$$

Consequently, we have

$$\mathcal{H} \cap \mathcal{G} \cap \mathcal{J} \subset \left\{ 4\Delta_m \leq 9\widehat{\Delta}_m \leq 16\Delta_m \text{ and } 5\zeta_m^2 \leq 18\widehat{\zeta}_m^2 \leq 44\zeta_m^2 \quad \forall 1 \leq m \leq M_n^+ \right\}$$

and thus, $\mathcal{H} \cap \mathcal{G} \cap \mathcal{J} \subset \{\text{pen}_m \leq \widehat{\text{pen}}_m \leq 18\text{pen}_m \mid \forall 1 \leq m \leq M_n^+\}$. Moreover, it holds $\mathcal{H} \subset \{M_n^- \leq \widehat{M}_n \leq M_n^+\}$, which can be seen as follows. Consider $\{\widehat{M}_n < M_n^-\}$. In case of $\widehat{M}_n = M_n^h$ or $M_n^- = 1$ clearly $\{\widehat{M}_n < M_n^-\} = \emptyset$. Otherwise by the definition of $\widehat{M}_n$ it holds

$$\{\widehat{M}_n < M_n^-\} = \bigcup_{m=1}^{M_n^- - 1} \{\widehat{M}_n = m\} \subset \left\{ \exists 2 \leq m \leq M_n^- : m^3 \|\widehat{[T]}_{\underline{m}}^{-1}\|^2 \max_{1 \leq j \leq m} [h]_j^2 > a_n \right\}.$$

By the definition of M_n^- and the property $\|[T]_{\underline{m}}^{-1}\|^2 \leq Dv_m^{-1}$ there exists $2 \leq m \leq M_n^-$ such that on $\{\widehat{M}_n < M_n^-\}$ it holds $\|\widehat{[T]}_{\underline{m}}^{-1}\|^2 > 4Dv_m^{-1} \geq 4\|[T]_{\underline{m}}^{-1}\|^2$ and thereby,

$$\{\widehat{M}_n < M_n^-\} \subset \left\{ \exists 2 \leq m \leq M_n^- : \|\widehat{[T]}_{\underline{m}}^{-1}\|^2 \geq 4\|[T]_{\underline{m}}^{-1}\|^2 \right\}. \quad (\text{A.34})$$

Consider $\{\widehat{M}_n > M_n^+\}$. In case of and $\widehat{M}_n = M_n^h$ or $M_n^- = 1$ clearly $\{\widehat{M}_n < M_n^-\} = \emptyset$. Otherwise, condition $T \in \mathcal{T}_d^\nu$ with $d \leq D$ implies $v_m^{-1} \leq D\|[T]_{\underline{m}}^{-1}\|^2$ as seen in the proof of Lemma A.9. Thereby, we conclude similarly as above

$$\{\widehat{M}_n > M_n^+\} \subset \left\{ \|[T]_{\underline{M_n^+ + 1}}^{-1}\|^2 \geq 4\|\widehat{[T]}_{\underline{M_n^+ + 1}}^{-1}\|^2 \right\}. \quad (\text{A.35})$$

Again applying the Neumann series argument we observe

$$\mathcal{H} \subset \left\{ \forall 1 \leq m \leq (M_n^+ + 1) : 2\|[T]_{\underline{m}}^{-1}\| \leq 3\|\widehat{[T]}_{\underline{m}}^{-1}\| \leq 4\|[T]_{\underline{m}}^{-1}\| \right\},$$

which combined with (A.34) and (A.35) yields $\{M_n^- \leq \widehat{M}_n \leq M_n^+\}^c \subset \mathcal{H}^c$ and thus, completes the proof. $\square$

LEMMA A.12. *Under the conditions of Theorem 4.2 we have for all $n \geq 1$*

$$n^4(M_n^h)^4 P(\mathcal{A}^c) \leq \mathcal{C}.$$

Proof. Due to Lemma A.11 it holds $n^4(M_n^h)^4 P(\mathcal{A}^c) \leq n^4(M_n^h)^4 \{P(\mathcal{H}^c) + P(\mathcal{J}^c) + P(\mathcal{G}^c)\}$. Therefore, the assertion follows if the right hand side is bounded by a constant $\mathcal{C}$, which we prove in the following. Consider $\mathcal{H}$. From condition $T \in \mathcal{T}_{d,D}^v$ and Lemma A.3 we infer

$$n^4(M_n^h)^4 P(\mathcal{H}^c) \leq 2 \exp \left(-\frac{1}{128D\eta} \frac{nv_{M_n^+ + 1}}{(M_n^+ + 1)^2} + 3 \log(M_n^+ + 1) + 5 \log n \right) \leq C(h, v, \eta, D) \quad (\text{A.36})$$

where the last inequality is due to condition $(M_n^+ + 1)^2 \log n = o(nv_{M_n^+ + 1})$. Consider $\mathcal{G}$. Due to condition $m^3 \gamma_m^{-1} = o(1)$ as $m \rightarrow \infty$ and $U \in \mathcal{U}_\sigma^\infty$ we observe $\mathbb{E}[Y^k] \leq 2^k (\mathbb{E}[\varphi^k(Z)] + \mathbb{E}[U^k]) \leq C(\gamma, \rho, \sigma) \sup_{j \geq 1} \mathbb{E}[e_j^k(Z)]$. Thereby, assumption $\sup_{j \geq 1} \mathbb{E}[e_j^{20}(Z)] \leq \eta^{20}$ together with Theorem 2.10 in Petrov [1995] imply

$$\begin{aligned} n^4(M_n^h)^4 P(\mathcal{G}^c) &\leq n^5 P(|\hat{\sigma}_Y^2 - \sigma_Y^2| > \sigma_Y^2/2) \leq 1024 \sigma_Y^{-20} n^5 \mathbb{E} \left| n^{-1} \sum_{i=1}^n Y_i^2 - \sigma_Y^2 \right|^{10} \\ &\leq 1024 \sigma_Y^{-20} \mathbb{E} |Y^2 - \sigma_Y^2|^{10} \leq C(\gamma, \rho, \sigma, \eta). \end{aligned} \quad (\text{A.37})$$

Consider $\mathcal{J}$. For all $m \geq 1$ observe that the centered random variables $(Y_i - \varphi(Z_i))f_j(W_i)$, $1 \leq i \leq n$, satisfy Cramer's condition (A.28) with $\mu_m = \eta(1+Dd)\sqrt{6\rho m^3 \gamma_m^{-1}} \leq C(\eta, \gamma, \rho, D)$. From (A.13) in Lemma A.2, $\varphi \in \mathcal{F}_\gamma^\rho$, and $P_{U|W} \in \mathcal{U}_\sigma^\infty$ we infer $\|\varphi_m\|_Z^2 + \sigma_Y^2 \leq 4(2+Dd)\rho + 2\sigma^2$. Moreover, it holds $\|[T]_{\underline{m}}^{-1} V_m\|^2 \leq Dv_m^{-1} \|V_m\|^2$ by employing condition $T \in \mathcal{T}_{d,D}^v$. Now Bernstein's inequality yields for all $1 \leq m \leq M_n^+$

$$\begin{aligned} n^6 P \left(\|[T]_{\underline{m}}^{-1} V_m\|^2 > (\|[T]_{\underline{m}}^{-1} [g]_{\underline{m}}\|^2 + \sigma_Y^2)/8 \right) \\ \leq n^6 \sum_{j=1}^m P \left(\left| \sum_{i=1}^n (Y_i - \varphi(Z_i))f_j(W_i) \right|^2 > \frac{n^2 v_m}{8Dm} (\|\varphi_m\|_Z^2 + \sigma_Y^2) \right) \\ \leq 2n^6 m \exp \left(-\frac{n^2 v_m m^{-1} (\|\varphi_m\|_Z^2 + \sigma_Y^2)}{32Dn\mu_m^2 + 16\mu_m n v_m^{1/2} m^{-1/2} (\|\varphi_m\|_Z^2 + \sigma_Y^2)^{1/2}} \right) \\ \leq 2 \exp \left(7 \log n - \frac{n v_{M_n^+} \sigma_Y^2}{M_n^+ C(\sigma, \eta, \gamma, \rho, D)} \right). \end{aligned}$$

Due to the definition of M_n^+ the last estimate implies $n^4(M_n^h)^4 P(\mathcal{J}^c) \leq \mathcal{C}$, which completes the proof. $\square$

References

- C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71:1795–1843, 2003.
- C. Ai and X. Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141: 5–43, 2007.
- A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields*, 113(3):301–413, 1999.

- R. Blundell and J. L. Horowitz. A nonparametric test of exogeneity. *Review of Economic Studies*, 74:1035–1058, 2007.
- R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.
- D. Bosq. *Nonparametric statistics for stochastic processes*. Springer, New York, 1998.
- H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis*, (101):395–408, 2010.
- M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, volume 6. North Holland, 2006.
- X. Chen and M. Reiß. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27:497–521, 2011.
- 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. *The Annals of Statistics*, 22: 238–270, 1994.
- D. Donoho and M. Low. Renormalization exponents and optimal pointwise rates of convergence. *The 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 *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 Belleghem. Instrumental regression in partially linear models. Discussion Paper # 0537, Institut de statistique, Université catholique de Louvain (first version 2005, revised), 2009.
- J. P. Florens, J. Johannes, and S. Van Belleghem. Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, 27:522–545, 2011.
- P. Gagliardini and O. Scaillet. Tikhonov regularization for functional minimum distance estimators. Swiss Finance Institute Research Paper No. 06-30, 2006.
- A. Goldenschluger and O. V. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Preprint ArXiv:1009.1016, 00:120*, 2010.

- A. Goldenshluger and S. V. Pereverzev. Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probability Theory and Related Fields*, 118: 169–186, 2000.
- P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33:2904–2929, 2005.
- 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.
- 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 of Probability and its Applications*, 29:18–32, 1984.
- J. Johannes and M. Schwarz. Adaptive nonparametric instrumental regression by model selection. Technical report, Université catholique de Louvain, 2010.
- J. Johannes, S. Van Bellegem, and A. Vanhems. Convergence rates for ill-posed inverse problems with an unknown operator. *Econometric Theory*, 27:472–496, 2011.
- T. Kawata. *Fourier analysis in probability theory*. Academic Press, New York, 1972.
- S. Krein and Y. I. Petunin. Scales of banach spaces. In *Russian Math. Surveys*, volume 21, pages 85–169, 1966.
- O. V. Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35:454466, 1990.
- K. Li. Minimacity of the method of regularization of stochastic processes. *The Annals of Statistics*, 10:937–942, 1982.
- J.-M. Loubes and C. Marteau. Oracle inequalities for instrumental variable regression. Technical report, Toulouse, 2009.
- B. A. Mair. Tikhonov regularization for finitely and infinitely smoothing operators. *SIAM Journal on Mathematical Analysis*, 25:135–147, 1994.
- P. Massart. *Concentration inequalities and model selection*. *Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003*. Lecture Notes in Mathematics 1896. Berlin: Springer. xiv, 337 p. , 2007.
- 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? *Proceedings of the American Mathematical Society*, 103(2):557–562, 1988a.

- A. Neubauer. An a posteriori parameter choice for Tikhonov regularization in Hilbert scales leading to optimal convergence rates. *SIAM Journal of Numerical Analysis*, 25(6): 1313–1326, 1988b.
- W. K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58:809–837, 1990.
- W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71:1565–1578, 2003.
- V. V. Petrov. *Limit theorems of probability theory. Sequences of independent random variables*. Oxford Studies in Probability. Clarendon Press., Oxford, 4. edition, 1995.
- A. Santos. Instrumental variable methods for recovering continuous linear functionals. *Journal of Econometrics*, 161:129146, 2011.
- T. A. Severini and G. Tripathi. Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. Forthcoming in *Journal of Econometrics*, 2010.
- P. Speckman. Minimax estimation of linear functionals in a Hilbert space. Unpublished manuscript., 1979.