Econometric Theory, **32**, 2016, 612–654.
doi:10.1017/S0266466614000966

ADAPTIVE ESTIMATION OF FUNCTIONALS IN NONPARAMETRIC INSTRUMENTAL REGRESSION

CHRISTOPH BREUNIG
Humboldt-Universität zu Berlin

JAN JOHANNES
CREST-Ensai and Université catholique de Louvain

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.

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

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

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

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

We would like to thank three anonymous referees and the Co-Editor for comments and suggestions that helped to improve the paper. This work was supported by DFG-SNF research group FOR916, the IAP research network no. P7/06 of the Belgian Government (Belgian Science Policy) and the ARC contract 12/17-045 of the "Communauté française de Belgique", granted by the Académie universitaire Louvain. Address correspondence to Christoph Breunig, Humboldt-Universität zu Berlin, Department of Economics, Spandauer Str. 1, 10178 Berlin, Germany, e-mail: christoph.breunig@hu-berlin.de.

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 $\ell : \mathcal{F} \to \mathbb{R}$. Important examples discussed in this paper are weighted average derivatives or point evaluation functionals 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 toward 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, Florens, and Renault (2007), Darolles, Fan, Florens, and Renault (2011), and Florens, Johannes, and Van Bellegem (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), semiparametric modeling (Florens, Johannes, and Van Bellegem, 2012), or a quasi-Baysian estimation approach (Florens and Simoni, 2012) to name but a few. For example, Ai and Chen (2003), Blundell, Chen, and Kristensen (2007), Chen and Reiß (2011) or Newey and Powell (2003) consider sieve minimum distance estimators of φ, while Darolles et al. (2011), Hall and Horowitz (2005), Gagliardini and Scaillet (2012) 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 (cf. Efromovich and Koltchinskii, 2001 or Hoffmann and Reiß, 2008) has been proposed by Johannes and Schwarz (2010). But assuming an independent and identical distributed (iid.) sample of (Y, Z, W) the estimation of the structural function φ as a whole involves the inversion of the conditional expectation operator of Z given W. This inverse problem is ill-posed in general (cf. 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 a linear functional at the usual parametric rate of convergence. It is remarkable to note, that, we do not face an ill-posed inverse problem when estimating the structural function if the sample of (Y,Z,W) is a time series formed by integrated and long memory processes. It is shown in Wang and Phillips (2009a), Wang and Phillips (2009b), and Wang and Phillips (2015) that in this situation the limit theory (even under endogeneity and without instruments) is the same as that of kernel regression in the iid. case (with minor allowance for a long run variance).

The nonparametric estimation of linear functionals from Gaussian white noise observations is a subject of considerable literature (cf. 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. Moreover, this model involves the additional difficulty of dealing with an unknown operator. 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 (2012) 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 (2012) and, when φ is not necessarily identified, by Severini and Tripathi (2012). 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). This is in line with Chen and Pouzo (2013) who study inference of functionals when $n^{1/2}$-consistency fails.

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 $\widehat{\varphi}_m$ was 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.,

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

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, adaptive estimators have been proposed by Loubes and Marteau (2009), Johannes

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

and Schwarz (2010), and Horowitz (2014). Johannes and Schwarz (2010) consider an adaptive estimator based on a model selection approach (cf. Barron, Birgé, and Massart, 1999 and its detailed discussion in Massart, 2007) which attains the minimax optimal rate. The estimator of Loubes and Marteau (2009) attains this rate within a logarithmic term. Both papers crucially rely on the a-priori knowledge of the eigenfunctions which yields an orthogonal series estimator involving the estimated singular values of the conditional expectation operator. In econometric applications, however, the eigenfunctions of this operator are unknown. Recently, Horowitz (2014) proposed an adaptive estimation procedure which is based on minimizing the asymptotic integrated mean-square error and does not involve the knowledge of the eigenfunctions of the operator.

For estimating linear functionals of the structural function φ, adaptive estimation procedures are not yet available. We propose a new method that is different from the above, does not involve a-priori knowledge of the eigenfunctions of the operator, and allows for a polynomial or exponential decay of its singular values. The methodology combines a model selection approach and Lepski's method (cf. Lepski, 1990). It is inspired by the recent work of Goldenshluger and Lepski (2011). To be more precise, the adaptive choice $\widehat{m}$ is defined as the minimizer of a random penalized contrast criterion[1], i.e.,

$$\widehat{m} := \operatorname*{arg\,min}_{1 \leqslant m \leqslant \widehat{M}_n} \left\{ \widehat{\Psi}_m + \widehat{\mathrm{pen}}_m \right\} \tag{1.2a}$$

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

$$\widehat{\Psi}_m := \max_{m \leqslant m' \leqslant \widehat{M}_n} \left\{ |\widehat{\ell}_{m'} - \widehat{\ell}_m|^2 - \widehat{\mathrm{pen}}_{m'} \right\}. \tag{1.2b}$$

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 $\mathcal{P}$. The appearance of the logarithmic factor within the rate is a known fact in the context of local estimation. Brown and Low (1996) show that it is unavoidable in the context of nonparametric Gaussian regression and, hence it is widely considered as an acceptable price for adaptation. This factor is also present in the work of Goldenshluger and Pereverzev (2000) where Lepski's method is applied in the presence of indirect Gaussian observations.

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. In Section 4 we construct the random upper

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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. Finally, Section 5 presents the results of a Monte Carlo Simulation study to illustrate the finite sample properties of our adaptive estimation procedure. 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 $L^2_Z = \{\phi : \mathbb{E}[\phi^2(Z)] < \infty\}$ to $L^2_W = \{\psi : \mathbb{E}[\psi^2(W)] < \infty\}$ (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^2_W. In what follows we always assume that there exists a unique solution $\varphi \in L^2_Z$ of equation (2.1), i.e., g belongs to the range of T, and that the null space of T is trivial (cf. Engl, Hanke, and Neubauer, 2000 or Carrasco et al., 2007 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 (cf. Carrasco et al., 2007). Note that the proof of minimax optimality of our estimator does not rely on this assumption but it is used for the illustrations and remarks below. It is well known that T is not compact if Z and W have elements in common. If T is compact then 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^2_W. 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^2_Z with respect to a prespecified orthonormal basis $\{e_j\}_{j\geqslant 1}$ in L^2_Z, which we formalize now. Given a positive sequence of weights $w := (w_j)_{j\geqslant 1}$ we define the weighted norm $\|\phi\|^2_w := \sum_{j\geqslant 1} w_j \langle\phi, e_j\rangle^2_Z$, $\phi \in L^2_Z$, the completion $\mathcal{F}_w$ of L^2_Z with respect to $\|\cdot\|_w$ and the ellipsoid $\mathcal{F}^r_w := \{\phi \in \mathcal{F}_w : \|\phi\|^2_w \leqslant r\}$ with radius $r > 0$. We shall stress that the basis $\{e_j\}_{j\geqslant 1}$ does not necessarily correspond to the eigenfunctions of T^*T where T^* denotes the adjoint operator of T. In the following we write $a_n \lesssim b_n$ when there exists a generic constant

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

$C > 0$ such that $a_n \leqslant C\, b_n$ 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 $\gamma := (\gamma_j)_{j \geqslant 1}$, 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 $1/\gamma := (\gamma_j^{-1})_{j \geqslant 1}$ (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 $\omega := (\omega_j)_{j \geqslant 1}$ satisfying $\inf_{j \geqslant 1}\{\omega_j \gamma_j\} > 0$. Obviously, $\mathcal{F}_\omega$ is a subset of $\mathcal{F}_{1/\gamma}$.

Furthermore, as usual in the context of 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 $\upsilon := (\upsilon_j)_{j \geqslant 1}$ and $d \geqslant 1$ we define the subset $\mathcal{T}_d^\upsilon$ of $\mathcal{T}$ by

$$\mathcal{T}_d^\upsilon := \Big\{ T \in \mathcal{T} : \quad \|\phi\|_\upsilon^2 / d \leqslant \|T\phi\|_W^2 \leqslant d\, \|\phi\|_\upsilon^2, \quad \forall \phi \in L_Z^2 \Big\}. \tag{2.2}$$

Notice first that any operator $T \in \mathcal{T}_d^\upsilon$ is injective if the sequence υ is strictly positive. Furthermore, for all $T \in \mathcal{T}_d^\upsilon$ it follows that $\upsilon_j / d \leqslant \|T e_j\|_W^2 \leqslant d\, \upsilon_j$ for all $j \geqslant 1$. If $(s_j)_{j \geqslant 1}$ denotes the ordered sequence of singular values of T then it can be seen that $\upsilon_j / d \leqslant s_j^2 \leqslant d\, \upsilon_j$. In other words, the sequence υ 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 υ, 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 $\gamma := (\gamma_j)_{j \geqslant 1}$, $\omega := (\omega_j)_{j \geqslant 1}$, and $\upsilon := (\upsilon_j)_{j \geqslant 1}$ be strictly positive sequences of weights with $\gamma_1 = \omega_1 = \upsilon_1 = 1$ such that γ is nondecreasing with $|j|^3 \gamma_j^{-1} = o(1)$ as $j \to \infty$, ω satisfies $\inf_{j \geqslant 1}\{\omega_j \gamma_j\} > 0$ and υ is a nonincreasing sequence.

Remark 2.1. We illustrate Assumption 1 for typical choices of γ and υ usually studied in the literature (cf. Hall and Horowitz, 2005, Chen and Reiß, 2011 or Johannes, Van Bellegem, and Vanhems, 2011). Let $[h]_j$ be the j-th generalized Fourier coefficient, i.e., $[h]_j := \mathbb{E}[h(Z) e_j(Z)]$, then we consider the cases

(pp) $\gamma_j \sim |j|^{2p}$ with $p > 3/2$, $\upsilon_j \sim |j|^{-2a}$, $a > 0$, and

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

(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 $\upsilon_j \sim \exp(-|j|^{2a})$, $a > 0$, 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 $\upsilon_j \sim |j|^{-2a}$, $a > 0$, 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 $j \to \infty$ is automatically satisfied for all $p > 0$ 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. Note that Hall and Horowitz (2005), who only consider the polynomial case, assume $2p + 1 > 2a > p$ with $p > 0$ and $a > 1/2$ which is more restrictive than Assumption 1 for $a \geqslant 2$.

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

$$\mathcal{R}_x^h := \max \left\{ x^{-1} \sum_{j=1}^{m_x^*} \frac{[h]_j^2}{\upsilon_j}, \sum_{j > m_x^*} \frac{[h]_j^2}{\gamma_j} \right\} \quad \text{and}$$

$$\mathcal{R}_x^\omega := x^{-1} \max_{1 \leqslant j \leqslant m_x^*} \left\{ \frac{1}{\omega_j \upsilon_j} \right\}, \tag{2.3}$$

where either $x = n$ in case of minimax optimal estimation or $x = n(1 + \log n)^{-1}$ in case of adaptive estimation. The rate $\mathcal{R}_n^h$ corresponds to the usual variance and bias decomposition of the mean square error. Here the dimension parameter m_x^* is chosen to trade off both, that is, we let for $x \geqslant 1$

$$m_x^* := \arg\min_{m \in \mathbb{N}} \left\{ \left| \frac{\upsilon_m}{\gamma_m} - x^{-1} \right| \right\}. \tag{2.4}$$

In case of adaptive estimation the rate of convergence is given by $\mathcal{R}_{\text{adapt}}^h := (\mathcal{R}_{n(1+\log n)^{-1}}^h)_{n \geqslant 1}$ and $\mathcal{R}_{\text{adapt}}^\omega := (\mathcal{R}_{n(1+\log n)^{-1}}^\omega)_{n \geqslant 1}$, respectively. For ease of notation let $m_n^\circ := m_{n(1+\log n)^{-1}}^*$. The bounds established below need the following additional assumption, which is satisfied in all cases considered in Remark 2.1.

Assumption 2. There exists a constant $0 < \kappa \leqslant 1$ such that for all $x \geqslant 1$

$$\kappa \leqslant \frac{x \, \upsilon_{m_x^*}}{\gamma_{m_x^*}} \leqslant \kappa^{-1}. \tag{2.5}$$

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

Assumption 2 implies that $n\upsilon_{m_n^*}\gamma_{m_n^*}^{-1}$ is uniformly bounded from above and away from zero. Thereby, we can write $n\upsilon_{m_n^*} \sim \gamma_{m_n^*}$.

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 $\mathbb{E}[U^4|W] \leqslant \sigma^4$ 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* υ *be sequences satisfying Assumptions 1 and 2. Suppose that* $\sup_{j\geqslant 1}\mathbb{E}[e_j^4(Z)|W] \leqslant \eta^4$, $\eta \geqslant 1$, *and* $\sigma^4 \geqslant \left(\sqrt{3}+4\rho\,\eta^2\sum_{j\geqslant 1}\gamma_j^{-1}\right)^2$. *Then for all* $n \geqslant 1$ *we have*

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

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

Note that in Theorem 2.1 and in the following results the marginal distributions of Z and W are kept fixed while only the dependency structure of the joint distribution of (Z, W) and of (U, Z, W) is allowed to vary.

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 $\sigma^4 \geqslant \left(\sqrt{3}+4\rho\,\eta^2\sum_{j\geqslant 1}\gamma_j^{-1}\right)^2$ 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.

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

In other words, this estimator attains the lower bound only if $\{f_l\}_{l\geqslant 1}$ is chosen appropriately.

Remark 2.4. The rate $\mathcal{R}^h$ of the lower bound is never faster than the $\sqrt{n}$-rate, that is, $\mathcal{R}^h_n \geqslant n^{-1}$. Moreover, using $\sum_{j>m^*_n}[h]^2_j\gamma_j^{-1} \leqslant n^{-1}\kappa^{-1}\sum_{j>m^*_n}[h]^2_j\upsilon_j^{-1}$ it can be seen that the lower bound rate is parametric if and only if $\sum_{j\geqslant 1}[h]^2_j\upsilon_j^{-1} < \infty$. This condition does not involve the sequence γ and hence, attaining a $\sqrt{n}$-rate is independent of the regularity conditions imposed on the structural function. Moreover, due to the link condition $T \in \mathcal{T}^{\upsilon}_d$ we have that Picard's condition $\sum_{j\geqslant 1}[h]^2_j\upsilon_j^{-1} < \infty$ is equivalent to h belonging to the range $\mathcal{R}(T^*)$, where T^* denotes the adjoint of T. Note that Severini and Tripathi (2012) showed in their Lemma 4.1 that $h \in \mathcal{R}(T^*)$ is necessary to obtain $\sqrt{n}$-estimability. Under appropriate conditions on φ and the joint distribution of (Y, Z, W) we show in the next section that $h \in \mathcal{R}(T^*)$ is also sufficient for $\sqrt{n}$-estimability.

The following assertion gives a lower bound over the ellipsoid $\mathcal{F}^{\tau}_{\omega}$ of representer. Consider the function $h^* := \tau\omega_{j^*}^{-1/2}e_{j^*}$ with $j^* := \arg\max_{1\leqslant j\leqslant m^*_n}\{(\omega_j\upsilon_j)^{-1}\}$ which obviously belongs to $\mathcal{F}^{\tau}_{\omega}$. 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 $n \geqslant 1$ we have*

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

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 $(\omega_j\gamma_j)_{j\geqslant 1}$ 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 $\varphi \in \mathcal{F}^{\rho}_{\gamma}$ and all $h \in \mathcal{F}^{\tau}_{\omega}$ is only possible under restrictions on the representer h in the sense that ω has to be a divergent sequence.

3. MINIMAX OPTIMAL ESTIMATION

3.1. Estimation by Dimension Reduction and Thresholding

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

Matrix and operator notations. Given $m \geqslant 1$, $\mathcal{E}_m$ and $\mathcal{F}_m$ denote the subspace of L^2_Z and L^2_W 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. $\mathcal{F}_m$)

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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 $1 \leqslant j, l \leqslant m$. 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 $\{v_j, e_j, f_j\}_{j \geqslant 1}$. 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 $1 \leqslant j, l \leqslant m$.

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 $f_j(W)$, $1 \leqslant j \leqslant m$, respectively, then it holds $[T]_{\underline{m}} = \mathbb{E}\left\{[f(W)]_{\underline{m}}[e(Z)]_{\underline{m}}^t\right\}$. Throughout the paper $[T]_{\underline{m}}$ is assumed to be nonsingular for all $m \geqslant 1$, 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 $(Y_1, Z_1, W_1), \ldots, (Y_n, Z_n, W_n)$ be an iid. sample of (Y, Z, W). Since $[T]_{\underline{m}} = \mathbb{E}\left\{[f(W)]_{\underline{m}}[e(Z)]_{\underline{m}}^t\right\}$ and $[g]_{\underline{m}} = \mathbb{E}\left\{Y[f(W)]_{\underline{m}}\right\}$ 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 $m \geqslant 1$ by

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

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 $1 \leqslant j \leqslant m$ 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 $1 \leqslant j \leqslant m$. Hence for $\int_{\mathcal{Z}} \varphi(z)h(z)dz$ the results below follow mutatis mutandis.

Note that the orthonormal bases $\{e_j\}_{j \geqslant 1}$ in L_Z^2 and $\{f_l\}_{l \geqslant 1}$ in L_W^2 depend on the marginal distributions of Z and W. As we illustrate in the following remark, these marginals are not needed to be completely known in advance as long as they satisfy additional regularity conditions.

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

Remark 3.2. Assume that the support of Z and W is confined to $[0,1]$ and denote $L^2_{[0,1]} := \{\phi : \int_0^1 \phi^2(z)dz < \infty\}$. If one assumes in addition that $L^2_{[0,1]} \subset L^2_Z$ and $L^2_W \subset L^2_{[0,1]}$ then it is possible to consider the restriction of T onto $L^2_{[0,1]}$. Note that this condition is satisfied if the density of Z is bounded from above and the density of W is uniformly bounded away from zero. For a detailed discussion we refer to a preliminary version of Darolles et al. (2011) or Section 2.2 of Florens et al. (2012). Further, let $\{e_j\}_{j\geqslant 1}$ and $\{f_j\}_{j\geqslant 1}$ be orthonormal bases in $L^2_{[0,1]}$. In this case, $(\mathbb{E}[e_l(Z)f_j(W)])_{j,l\geqslant 1}$ is the matrix representation of the restricted operator $(\widetilde{T}\phi)(\cdot) := \int_0^1 \phi(z)p_{ZW}(z,\cdot)dz$ on $L^2_{[0,1]}$ where p_{ZW} denotes the joint density of (Z,W). Moreover, due to Remark 3.1 the estimator of $\ell(\varphi)$ in this situation coincides with the estimator $\widehat{\ell}_m$ given in (3.1) and hence, the results below follow similarly.

In practice, assuming the supports of Z and W to be contained in $[0,1]$ can be restrictive. Our method, however, can easily be adapted, either replacing the interval $[0,1]$ by any compact subset on which the densities are bound from below and above or referring to weighted L^2 spaces as proposed by the authors mentioned above. Nevertheless, for an ease of presentation and allowing a more intuitive interpretation we consider the function space $L^2[0,1]$ as illustration.

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 $j,l \geqslant 1$, which we summarize in the next assumption.

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

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

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

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 $e_j(Z)f_l(W) - \mathbb{E}[e_j(Z)f_l(W)]$ (cf. Bosq, 1998). Moreover, any joint distribution of (Z,W) satisfies Assumption 3 for sufficiently large η if the basis $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$ 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 us introduce the function $\varphi_m \in \mathcal{E}_m$ which is uniquely defined by the vector of coefficients $[\varphi_m]_{\underline{m}} = [T]_{\underline{m}}^{-1}[g]_{\underline{m}}$ and $[\varphi]_j = 0$ for $j \geqslant m+1$. Obviously, up to the threshold, the estimator $\widehat{\ell}_m$ is the empirical

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

counterpart of $\ell_h(\varphi_m)$. In Proposition 3.1 consistency of the estimator $\widehat{\ell}_m$ is only obtained under the condition

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

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 $m_n^{-1} = o(1)$, $m_n = o(n)$,*

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

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

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

$$\mathcal{T}^v_{d,D} := \left\{ T \in \mathcal{T}^v_d : [T]_{\underline{m}} \text{is nonsingular and } \|[\nabla_v]_{\underline{m}}^{1/2} [T]_{\underline{m}}^{-1}\|^2 \leqslant D \text{ for all } m \geqslant 1 \right\}. \tag{3.4}$$

Remark 3.3. If $T \in \mathcal{T}^v_d$ and if in addition its singular value decomposition is given by $\{s_j, e_j, f_j\}_{j \geqslant 1}$ then for all $m \geqslant 1$ the matrix $[T]_{\underline{m}}$ is diagonalized with diagonal entries $[T]_{j,j} = s_j$, $1 \leqslant j \leqslant m$. In this situation it is easily seen that $\sup_{m \in \mathbb{N}} \|[\nabla_v]_{\underline{m}}^{1/2} [T]_{\underline{m}}^{-1}\|^2 \leqslant d$ and, hence T satisfies the extended link condition (3.4), that is, $T \in \mathcal{T}^v_{d,D}$. Furthermore, it holds $\mathcal{T}^v_d = \mathcal{T}^v_{d,D}$ for suitable $D > 0$, if T is a small perturbation of $\nabla_v^{1/2}$ or if T is strictly positive (cf. Efromovich and Koltchinskii, 2001 or Cardot and Johannes, 2010, respectively).

Remark 3.4. Once both basis $\{e_j\}_{j \geqslant 1}$ and $\{f_l\}_{l \geqslant 1}$ 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 $\{f_l\}_{l \geqslant 1}$ satisfying condition (3.4) can be interpreted as optimal instruments (cf. Newey, 1990).

Remark 3.5. For each pre-specified basis $\{e_j\}_{j \geqslant 1}$ we can theoretically construct a basis $\{f_l\}_{l \geqslant 1}$ such that (3.4) is equivalent to the link condition (2.2). To be more precise, if $T \in \mathcal{T}^v_d$, which involves only the basis $\{e_j\}_{j \geqslant 1}$, then the fundamental inequality of Heinz (1951) implies $\|(T^*T)^{-1/2} e_j\|^2_Z \leqslant d\, v_j^{-1}$. Thereby, the function $(T^*T)^{-1/2} e_j$ is an element of L^2_Z and, hence $f_j := T(T^*T)^{-1/2} e_j$, $j \geqslant 1$,

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

belongs to L^2_W. Then it is easily checked that $\{f_l\}_{l\geqslant 1}$ is a basis of the closure of the range of T which may be completed to a basis of L^2_W. 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 $j,l \geqslant 1$. Thereby, we can apply Lemma A.3 in Cardot and Johannes (2010) which gives $\mathcal{T}^{\upsilon}_d = \mathcal{T}^{\upsilon}_{d,D}$ for sufficiently large D. Another approach relies on $f_j = Te_j$ which corresponds to $T^*g = T^*T\varphi$. In this case, an additional smoothing parameter is required.

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 υ satisfies Assumption 1, the conditional expectation operator T belongs to $\mathcal{T}^{\upsilon}_{d,D}$, and (3.3) is substituted by*

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

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 $\mathcal{F}^{\rho}_{\gamma}$. 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 (2.4) satisfies*

$$(m^*_n)^3 \max\left\{|\log \mathcal{R}^h_n|, (\log m^*_n)\right\} = o(\gamma_{m^*_n}), \quad \text{as } n \to \infty, \tag{3.6}$$

then we have for all $n \geqslant 1$

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

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

The next result gives sufficient conditions for $\sqrt{n}$-estimability of $\ell_h(\varphi)$. The next corollary is a direct consequence of Theorem 3.3 and Remark 2.4, hence its proof is omitted.

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

COROLLARY 3.4. *Let the assumptions of Theorem 3.3 be satisfied. If in addition $h \in \mathcal{R}(T^*)$ then we have for all $n \geqslant 1$*

$$\sup_{T \in \mathcal{T}^{\upsilon}_{d,D}} \sup_{P_{U|W} \in \mathcal{U}_{\sigma}} \sup_{\varphi \in \mathcal{F}^{\rho}_{\gamma}} \mathbb{E}\,|\widehat{\ell}_{m^*_n} - \ell_h(\varphi)|^2 \leqslant \mathcal{C}\, n^{-1},$$

where $\mathcal{C}$ is as in Theorem 3.3.

Remark 3.6. The last result together with Remark 2.4 established equivalence between condition $h \in \mathcal{R}(T^*)$ and $\sqrt{n}$-estimability of $\ell_h(\varphi)$ under appropriate conditions on φ and the joint distribution of (Y, Z, W) (as conjectured in Chapter 4, Remark (ii) of Severini and Tripathi, 2012). As illustrated in the next subsection, depending on the severeness of the ill-posedness $\sqrt{n}$-estimability could not be possible even for smooth functionals. In the polynomial case (pp), condition $h \in \mathcal{R}(T^*)$ holds true only if $s > a + 1/2$. In case of (ep), $h \in \mathcal{R}(T^*)$ only if the representer h is analytic. In contrast to our framework, the estimation procedure of Santos (2011) crucially relies on condition $h \in \mathcal{R}(T^*)$ which implies the existence of a function $\vartheta \in L^2_W$ such that $\ell_h(\varphi) = \mathbb{E}[Y\vartheta(W)]$.

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

COROLLARY 3.5. *Let the assumptions of Theorem 3.3 be satisfied where we substitute condition (3.6) by $(m^*_n)^3 \max\left\{|\log \mathcal{R}^{\omega}_n|, (\log m^*_n)\right\} = o(\gamma_{m^*_n})$ as $n \to \infty$. Then we have*

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

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

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^2_Z and L^2_W equal $L^2[0, 1]$. Moreover, as a basis $\{e_j\}_{j \geqslant 1}$ in $L^2[0, 1]$ we choose the trigonometric basis given by

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

In this subsection also the second basis $\{f_l\}_{l \geqslant 1}$ is given by the trigonometric basis. In this situation, the moment conditions formalized in Assumption 3 are automatically fulfilled.

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

Recall the typical choices of the sequences γ, ω, and υ 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 (cf. Neubauer, 1988a, 1988b). In case of (ep) it is well known that $\mathcal{F}_\gamma$ contains only analytic functions if $p > 1$(cf. 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 υ. Easy calculus shows that any operator T satisfying the link condition (2.2) acts like integrating (a)-times and hence is called *finitely smoothing* (cf. Natterer, 1984). In case of (pe) we consider an exponential decay of υ and it can easily be seen that $T \in \mathcal{T}_d^\upsilon$ implies $\mathcal{R}(T) \subset C^\infty[0,1]$, therefore the operator T is called *infinitely smoothing* (cf. 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 3.6. *Assume an iid. n-sample of* (Y, Z, W) *from the model (1.1a–1.1b) with* $T \in \mathcal{T}_{d,D}^\upsilon$ *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*

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

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

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

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

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

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

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

Remark 3.7. As we see from Proposition 3.6, if the value of a increases the obtainable optimal rate of convergence decreases. Therefore, the parameter a is often called *degree of ill-posedness* (cf. 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, $p \geqslant 3/2$ is needed) if the representer is smoother than the degree of ill-posedness of T, i.e., *(i)* $s \geqslant a - 1/2$ and *(ii)* $s \geqslant a$. 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 υ.

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

Remark 3.8. It is of interest to compare our results with those of Hall and Horowitz (2005) or Chen and Reiss (2011) who consider the estimation of the structural function as a whole. In the notations of Hall and Horowitz (2005), who consider only the case (pp), the decay of the eigenvalues of T^*T is assumed to be of order $j^{-\alpha}$, that is, $\alpha = 2a$ with $\alpha > 1$. Furthermore, they suppose a decay of the coefficients of the structural function of order $j^{-\beta}$, that is, $\beta = p + 1/2$ with $\beta > 1/2$. By using this parametrization, Hall and Horowitz (2005) obtain in the case (pp) the minimax rate of convergence $n^{-2p/(2a+2p+1)}$ (see also Chen and Reiß, 2011). Let us compare this rate when estimating φ at a point $t_0 \in [0,1]$ (cf. Example 3.1). Here, we have $s = 0$ and hence, obtain the minimax rate of convergence $n^{-(2p-1)/(2a+2p)}$. Roughly speaking, one looses $1/2$ of smoothness, which corresponds to the loss of smoothness of Sobolev embeddings in Hölder spaces. For any representer h with $2s > (2a+1)/(2a+2p+1)$, however, the rate of convergence for estimating $\ell_h(\varphi)$ in the case (pp) is faster than estimating φ as a whole.

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 $\sum_{j \geqslant 1} \gamma_j^{-1} < \infty$ (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 $\widehat{\ell}_m$ introduced in (3.1) writes for all $m \geqslant 1$ as

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

where $\widehat{\varphi}_m$ is an estimator proposed by Johannes and Schwarz (2010). Let $p \geqslant 3/2$ and $a > 0$. Then the estimator $\widehat{\varphi}_{m_n^*}(t_0)$ attains within a constant the minimax optimal rate of convergence $\mathcal{R}^{h_{t_0}}$. Applying Proposition 3.6 gives

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

In case of (ep) we obtain a rate of convergence that attains the parametric rate within a logarithmic term. This is in particular remarkable since the representer of the point evaluation functional does not even belong to L_Z^2.

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

$j \geqslant 1$ and, hence $[h]_j^2 \sim j^{-2}$. Again we assume that $p \geqslant 3/2$ and $a > 0$. Then the mean squared error of the estimator $\widehat{\ell}_{m_n^*} = \int_0^b \widehat{\varphi}_{m_n^*}(t)dt$ 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

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

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

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

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 $a < 1/2$. On the other hand, only logarithmic rates of convergence can be achieved for averages in case of (pe). This illustrates the difficulty of recovering only partial information of the structural function in the severely ill-posed case.

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 $\widehat{\ell}_{m_n^*} = -\int_0^1 \widehat{\varphi}_{m_n^*}(t)h(t)dt$ 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 $0 \leqslant t \leqslant 1$. 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 $j \geqslant 1$ 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.

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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 $\mathrm{pen} := (\mathrm{pen}_m)_{m \geqslant 1}$, 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,

$$\widetilde{m} := \operatorname*{arg\,min}_{1 \leqslant m \leqslant M_n} \left\{ \Psi_m + \mathrm{pen}_m \right\}, \tag{4.1a}$$

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

$$\Psi_m := \max_{m \leqslant m' \leqslant M_n} \left\{ |\widehat{\ell}_{m'} - \widehat{\ell}_m|^2 - \mathrm{pen}_{m'} \right\}. \tag{4.1b}$$

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 $1 \leqslant m \leqslant M_n$

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

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

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

where $(\cdot)_+$ denotes the positive part of a function. Combining the last estimate with the previous reduction scheme yields for all $1 \leqslant m \leqslant M_n$

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

where $\mathrm{bias}_m := \sup_{m' \geqslant m} |\ell_h(\varphi_{m'} - \varphi)|^2$. We will prove below that $\mathrm{pen}_m + \mathrm{bias}_m$ is of the order $\mathcal{R}^h_{n(1+\log n)^{-1}}$. 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 $[h]_1 \neq 0$.

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

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

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* $\varsigma_m^2 = 74\big(\mathbb{E}[Y^2] + \max_{1\leqslant m'\leqslant m} \|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2\big)$, *then we define*

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

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 $\mathbb{E}[U|W] = 0$, $\mathbb{E}[U^2|W] \leqslant \sigma^2$, and Cramer's condition hold, i.e.,

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

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

Assumption 4. There exists $\eta \geqslant 1$ such that the distribution of W satisfies

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

We now present an upper bound for $\widehat{\ell}_{\widetilde{m}}$. As Goldenshluger and Pereverzev (2000) 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* $(m_n^\circ)^3 \max_{1\leqslant j\leqslant m_n^\circ} [h]_j^2 = o(a_n v_{m_n^\circ})$ *as* $n \to \infty$. *Then we have for all* $n \geqslant 1$

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

where $\mathcal{C}$ *is as in Theorem 3.3.*

Remark 4.1. In all examples studied below the condition $(m_n^\circ)^3 \max_{1\leqslant j\leqslant m_n^\circ} [h]_j^2 = o(a_n v_{m_n^\circ})$ 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.

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

selection method. For this purpose, we construct an estimator for ς_m^2 by replacing the unknown quantities by their empirical analog, that is,

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

With the nondecreasing sequence $(\widehat{\varsigma}_m^2)_{m\geqslant 1}$ 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 $n \geqslant 1$ define*

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

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 $m \geqslant 1$ an empirical analog of pen_m *by*

$$\widehat{\mathrm{pen}}_m := 204\,\widehat{\varsigma}_m^2(1+\log n)n^{-1} \max_{1\leqslant m'\leqslant m} \|[h]_{\underline{m}'}^t[\widehat{T}]_{\underline{m}'}^{-1}\|^2. \tag{4.4}$$

Before we establish the next upper bound we introduce

$$M_n^+ := \min\Big\{2 \leqslant m \leqslant M_n^h : \upsilon_m^{-1}m^3 \max_{1\leqslant j\leqslant m}[h]_j^2 > 4Da_n\Big\} - 1 \tag{4.5}$$

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 attain the minimax rate of convergence $\mathcal{R}^h$ within a logarithmic term.

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

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

where $\mathcal{C}$ is as in Theorem 3.3.

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

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 $M_n^\omega := \max\{1 \leqslant m \leqslant \lfloor n^{1/4}\rfloor : \max_{1\leqslant j\leqslant m}(\omega_j^{-1}) \leqslant n\}$. In the definition of the bounds

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

$\widehat{M}_n$, M_n^+, and M_n^- (cf. Appendix 4) we replace M_n^h and $\max_{1\leqslant j\leqslant m}[h]_j^2$ by M_n^ω and $\max_{1\leqslant j\leqslant m}\omega_j^{-1}$, respectively. Consequently, by employing $\|h\|_{1/\gamma}^2 \leqslant \tau$ and $\mathcal{R}_n^h \leqslant \tau\,\mathcal{R}_n^\omega$ 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* $n \geqslant 1$

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

where $\mathcal{C}$ *is as in Corollary 3.5.*

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 $\mathcal{R}^h_{\mathrm{adapt}}$ and $\mathcal{R}^\omega_{\mathrm{adapt}}$. 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 3.6.

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

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

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

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

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

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

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

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

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

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

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 $\mathcal{R}^{h_{t_0}}_{\text{adapt}}$. Applying Proposition 4.4 gives

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

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 $\mathcal{R}^h_{\text{adapt}}$. Applying Proposition 4.4 gives

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

5. MONTE CARLO EXPERIMENTS

In this section, we examine the finite sample properties of our estimation procedure. We study first the point evaluation functional and thereafter, an average of the structural function. As in Subsection 3.4, we consider the case where Z and W are both scalar and $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$ coincide with the trigonometric basis. Moreover, we generate the joint density of (Z,W) by the multivariate function $p_{ZW}(z,w) = C_\upsilon [e(z)]^t_{\underline{k}}\left([I]_{\underline{k}} + A_k\right)[\nabla_\upsilon]^{1/2}_{\underline{k}}[f(w)]_{\underline{k}}$ where C_υ is a normalization constant, $(\upsilon_j)_{j\geqslant 1}$ is a nondecreasing sequence which is specified below, and $k = 100$. Here, A_k is a randomly generated $k \times k$ matrix with spectral

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

norm $1/2$. Due to the construction of the joint density p_{ZW} the link condition $T \in \mathcal{T}_d^v$ is satisfied for all $\phi \in \mathcal{E}_k$. Note that if A_k equals the zero matrix then this would correspond to the situation where the eigenfunctions of T coincide with the bases $\{e_j\}_{j\geqslant 1}$ and $\{f_l\}_{l\geqslant 1}$. We generate samples of size $n = 1000$ using the relationship $Y = \mathbb{E}[\varphi(Z)|W] + V$ where $V \sim N(0, 0.01)$. The number of Monte Carlo replications is 1000.

In particular, we want to study the performance of our estimators in finite samples when the dimension parameter m is chosen by our adaptive procedure given in (1.2a–1.2b). The constants in the definition of the adaptive procedure, though suitable for the theory, may be chosen much smaller in practice. Here, we replace in definition of $\widehat{\text{pen}}$ (given in (4.4)) and $\widehat{\varsigma}_m^2$ the constants 204 and 74 by 5 and 1, respectively. Without this adjustment, we found that the penalty dominates the criterion for all reasonable sample sizes. Our proposed adjustment also works well for smaller sample sizes as we illustrate below. In addition, we adjust the upper bound $\widehat{M}$ in the following way. We replace a_n (given in Definition 4.1) by $40n(1+\log n)^{-1}$. The modification of a_n ensures that the upper bound $\widehat{M}_n$ is sufficiently large. Our results are not sensitive to a larger choice of a_n.

Point wise estimation. Let us consider the problem of pointwise estimation of $\varphi(z) = 10z^2 \sin(\pi z)$ for $z \in [0, 1]$ over an equidistantly spaced grid of length 50. We truncate its infinite dimensional vector of coefficients at a sufficiently large integer, say 100. In Figure 1, we compare the performance of the estimators with optimal parameter m_n^* (in the first column) and data driven parameter $\widehat{m}$ (in the second column). More precisely, at each point t_0 of the grid we choose m_n^* as the minimizer of the empirical mean of $|\widehat{\ell}_m - \ell_{h_{t_0}}(\varphi)|^2$. The first row represents (pp) with $v_j = j^{-1}$ while the second depicts (pe) with $v_j = \exp(-j)$. In case of (pp), the pointwise 95%–confidence bands are sufficiently tight to make significant statements about the curvature of φ. Not surprisingly, in case of (pe) the pointwise confidence bands are much wider. But also in this case the pointwise median of the adaptive estimators is very close to φ.

From Figure 1 we see that the confidence bands are tighter in case of (pp) than in case (pe). Further, the pointwise confidence bands are wider in case of adaptive estimation but, in case of $v_j = j^{-1}$, are close to the bands obtained by minimax optimal estimation. Not surprisingly, the confidence intervals are wider at the boundary. It may seem odd that the confidence intervals can be very narrow in the interval $[0.3, 0.6]$ and sometimes do not even contain the true functional value. To explain this, note that the choice of the dimension parameter m_n^* is driven by the nonlinearity of φ and the degree of ill posedness. Between 0.3 and 0.6 we expect that the curve can be well approximated by just one relatively linear basis function. The second and third basis functions of the trigonometric basis are already relatively nonlinear and thus, due to the ill posed inverse problem generate a large variance. This is why in Figure 1(c) the empirical risk was minimized for taking just the first (constant) basis function. The accuracy of estimation can be

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

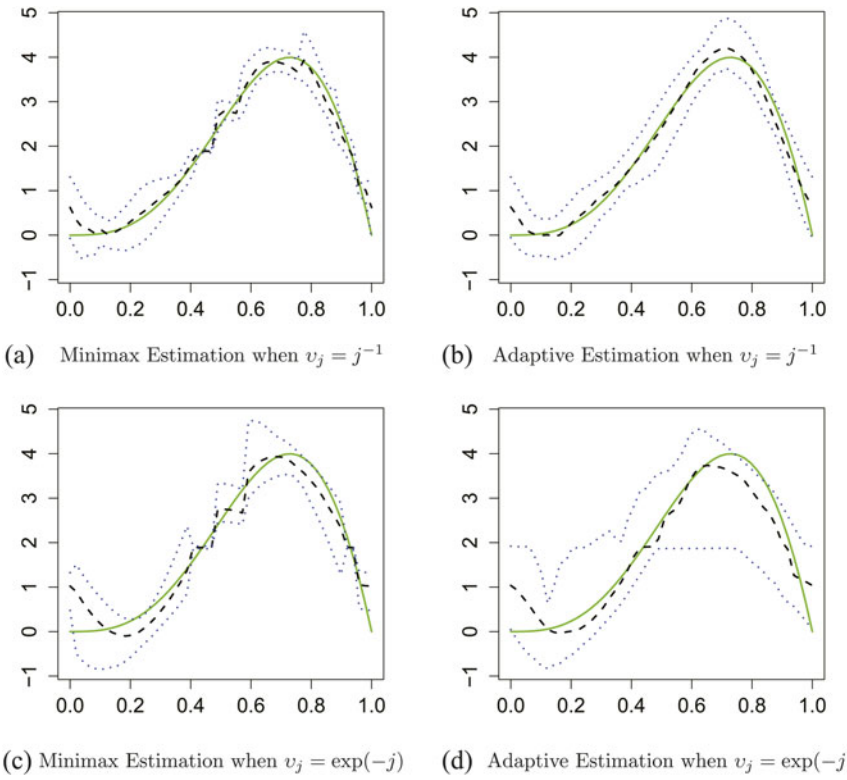

FIGURE 1. The green solid, black dashed, and blue dotted lines show φ, pointwise median of the estimators, and their pointwise 95% confidence band.

improved by choosing different basis functions such as Legendre polynomials or Wavelets.

Estimation of averages. We now consider the estimation of averages of the structural function. In the following, we consider the structural function $\varphi(z) = \sum_{j=1}^{100}(-1)^{j+1} j^{-2} e_j(z)$. We consider the problem of estimating the value of the linear functional $\int_0^{0.75} \varphi(z)dz \approx 0.898$. The empirical means from a Monte Carlo simulation are displayed in Table 1. Here, we choose m_n^* as the minimizer of the empirical mean of $|\widehat{\ell}_m - \int_0^{0.75} \varphi(z)dz|^2$. From Table 1 we see that the difference of the empirical means of $|\widehat{\ell}_{m_n^*} - \ell_h(\varphi)|^2$ and $|\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2$ are small.

6. CONCLUSION

In this paper, we propose a minimax optimal estimation procedure for linear functionals in nonparametric instrumental regression. Taking into account that this procedure still relies on an optimal choice of a smoothing parameter, we complete

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

TABLE 1. Results of Monte Carlo simulations

Model v_j	*Sample Size*	*Empirical mean of* $\lvert\widehat{\ell}_{m_n^*} - \ell_h(\varphi)\rvert^2$	$\lvert\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)\rvert^2$
j^{-1}	200	0.0218	0.0202
	1000	0.0058	0.0070
j^{-2}	200	0.0784	0.0770
	1000	0.0317	0.0300
j^{-3}	200	0.1295	0.1404
	1000	0.0931	0.1058
j^{-4}	200	0.1462	0.1533
	1000	0.1288	0.1462
$\exp(-j)$	200	0.0627	0.0619
	1000	0.0214	0.0313
$\exp(-2j)$	200	0.1275	0.1479
	1000	0.1080	0.1362
$\exp(-3j)$	200	0.1521	0.1555
	1000	0.1341	0.1538

the procedure by a data driven selection of this parameter. The main result established in the paper states that the fully data driven estimator can attain minimax-optimal rates of convergence up to a logarithmic factor, which is a widely accepted price to pay for local estimation. Obviously, the derivation of adaptive confidence intervals for the linear functional $\ell(\varphi)$ is only one amongst the many interesting questions for further research and we are currently exploring this topic.

NOTE

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

REFERENCES

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

Ai, C. & X. Chen (2007) Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics* 141, 5–43.

Ai, C. & X. Chen (2012) The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics* 170(2), 442–457.

Barron, A., L. Birgé, & P. Massart (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields* 113(3), 301–413.

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

Blundell, R., X. Chen, & D. Kristensen (2007) Semi-nonparametric IV estimation of shape-invariant engel curves. *Econometrica* 75(6), 1613–1669.

Blundell, R. & J.L. Horowitz (2007) A nonparametric test of exogeneity. *Review of Economic Studies* 74, 1035–1058.

Bosq, D. (1998) *Nonparametric Statistics for Stochastic Processes*. Springer.

Brown, L.D. & M.G. Low (1996) A constrained risk inequality with applications to nonparametric functional estimation. *Annals of Statistics* 24(6), 2524–2535.

Cardot, H. & J. Johannes (2010) Thresholding projection estimators in functional linear models. *Journal of Multivariate Analysis* 101, 395–408.

Carrasco, M., J.-P. Florens, & E. Renault (2007) Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. *Handbook of Econometrics*, vol. 6. North Holland.

Chen, X. & D. Pouzo (2013) Sieve Quasi Likelihood Ratio Inference on Semi/Nonparametric Conditional Moment Models1. Technical report, Cowles Foundation for Research in Economics, Yale University.

Chen, X. & M. Reiß (2011) On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory* 27, 497–521.

Darolles, S., Y. Fan, J.P. Florens, & E. Renault (2011) Nonparametric instrumental regression. *Econometrica* 79(5), 1541–1565. ISSN 1468-0262.

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

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

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

Engl, H.W., M. Hanke, & A. Neubauer (2000) *Regularization of Inverse problems.* Kluwer Academic.

Florens, J.-P. (2003) Inverse problems and structural econometrics: The example of instrumental variables. *Advances in Economics and Econometrics: Theory and Applications – Eight World Congress*, Volume 36 of Econometric Society Monographs, vol. 2, pp. 284. Cambridge University Press.

Florens, J.P., J. Johannes, & S. Van Bellegem (2011) Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory* 27, 522–545.

Florens, J.-P., J. Johannes, & S. Van Bellegem (2012) Instrumental regression in partially linear models. *The Econometrics Journal*, 15(2), 304–324.

Florens, J.-P. & A. Simoni (2012) Nonparametric estimation of an instrumental regression: A quasi-bayesian approach based on regularized posterior. *Journal of Econometrics* 170(2), 458–475.

Gagliardini, P. & O. Scaillet (2012) Tikhonov regularization for nonparametric instrumental variable estimators. *Journal of Econometrics* 167(1), 61–75.

Goldenshluger, A. & O. Lepski (2011) Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Annals of Statistics* 39(3), 1608–1632.

Goldenshluger, A. & S.V. Pereverzev (2000) Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations. *Probability Theory and Related Fields* 118, 169–186.

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

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

Hoffmann, M. & M. Reiß (2008) Nonlinear estimation for linear inverse problems with error in the operator. *Annals of Statistics* 36(1), 310–336.

Horowitz, J.L. (2014) Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics* 180, 158–173.

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

Ibragimov, I. & R. Has'minskii (1984) On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory of Probability and its Applications* 29, 18–32.

Johannes, J. & M. Schwarz (2010) Adaptive Nonparametric Instrumental Regression by Model Selection. Technical report, Université catholique de Louvain.
Johannes, J., S. Van Bellegem, & A. Vanhems (2011) Convergence rates for ill-posed inverse problems with an unknown operator. *Econometric Theory* 27, 472–496.
Kawata, T. (1972) *Fourier Analysis in Probability Theory.* Academic Press.
Krein, S. & Y.I. Petunin (1966) Scales of banach spaces. *Russian Mathematical Surveys* 21, 85–169.
Lepski, O.V. (1990) On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and its Applications* 35, 454–466.
Li, K. (1982) Minimaxity of the method of regularization of stochastic processes. *Annals of Statistics* 10, 937–942.
Loubes, J.-M. & C. Marteau (2009) Oracle Inequalities for Instrumental Variable Regression. Technical report, Toulouse.
Mair, B.A. (1994) Tikhonov regularization for finitely and infinitely smoothing operators. *SIAM Journal on Mathematical Analysis* 25, 135–147.
Massart, P. (2007) *Concentration Inequalities and Model Selection. Ecole d'Eté de Probabilités de Saint-Flour XXXIII – 2003.* Lecture Notes in Mathematics, vol. 1896. Springer, xiv, pp. 337.
Natterer, F. (1984) Error bounds for Tikhonov regularization in Hilbert scales. *Applicable Analysis* 18, 29–37.
Neubauer, A. (1988a) 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.
Neubauer, A. (1988b) When do Sobolev spaces form a Hilbert scale? *Proceedings of the American Mathematical Society* 103(2), 557–562.
Newey, W.K. (1990) Efficient instrumental variables estimation of nonlinear models. *Econometrica* 58, 809–837.
Newey, W.K. & J.L. Powell (2003) Instrumental variable estimation of nonparametric models. *Econometrica* 71, 1565–1578.
Petrov, V.V. (1995) *Limit Theorems of Probability Theory. Sequences of Independent Random Variables.* Oxford Studies in Probability, 4th ed. Clarendon Press.
Santos, A. (2011) Instrumental variable methods for recovering continuous linear functionals. *Journal of Econometrics* 161, 129–146.
Severini, T.A. & G. Tripathi (2012) Efficiency bounds for estimating linear functionals of nonparametric regression models with endogenous regressors. *Journal of Econometrics* 170(2), 491–498, ISSN 0304-4076.
Speckman, P. (1979) Minimax estimation of linear functionals in a Hilbert space. Unpublished manuscript.
Wang, Q. & P.C. Phillips (2009a) Asymptotic theory for local time density estimation and nonparametric cointegrating regression. *Econometric Theory* 25, 710–738.
Wang, Q. & P.C. Phillips (2009b) Structural nonparametric cointegrating regression. *Econometrica* 77(6), 1901–1948.
Wang, Q. & P.C. Phillips (2015) Nonparametric cointegrating regression with endogeneity and long memory. *Econometric Theory* 33, 359–401.

A. APPENDIX

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

Proof of Theorem 2.1. Define the function $\varphi_* := \zeta^{1/2}\left(n\sum_{l=1}^{m_n^*}[h]_l^2 \upsilon_l^{-1}\right)^{-1/2}\sum_{j=1}^{m_n^*}[h]_j \upsilon_j^{-1}e_j$ with $\zeta := \min(1/(2d), \rho)$. Since $\left(\gamma_j^{-1}\upsilon_j\right)_{j\geqslant 1}$ is nonincreasing and by Assumption 2 it follows that φ_* and in particular $\varphi_\theta := \theta\varphi_*$ for $\theta \in \{-1, 1\}$ belong to

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

$\mathcal{F}_\gamma^\rho$. 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 $\sigma^4 \geqslant \big(\sqrt{3} + 4\rho \sum_{j\geqslant 1} \gamma_j^{-1}\eta^2\big)^2$, which can be realized as follows. Obviously, we have $\mathbb{E}[U_\theta|W] = 0$. Moreover, we have $\sup_j \mathbb{E}[e_j^4(Z)|W] \leqslant \eta^4$ implies $\mathbb{E}[\varphi_\theta^4(Z)|W] \leqslant \rho^2\big(\sum_{j\geqslant 1}\gamma_j^{-1}\big)^2 \mathbb{E}[e_j^4(Z)|W] \leqslant \rho^2\eta^4\big(\sum_{j\geqslant 1}\gamma_j^{-1}\big)^2$ and thus, $|[T\varphi_\theta](W)|^4 \leqslant \mathbb{E}[\varphi_\theta^4(Z)|W] \leqslant \rho^2\eta^4(\sum_{j\geqslant 1}\gamma_j^{-1})^2$. From the last two bounds we deduce $\mathbb{E}[U_\theta^4|W] \leqslant 16\,\mathbb{E}[\varphi_\theta^4(Z)|W] + 6\,\mathbb{V}\mathrm{ar}(\varphi_\theta(Z)|W) + 3 \leqslant \big(\sqrt{3} + 4\rho\,\eta^2 \sum_{j\geqslant 1}\gamma_j^{-1}\big)^2$. Consequently, for each θ iid. copies (Y_i, Z_i, W_i), $1 \leqslant i \leqslant n$, of (Y,Z,W) with $Y := \varphi_\theta(Z) + U_\theta$ form an n-sample of the model (1.1a–1.1b) and we denote their joint distribution by P_θ 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 $KL(P_1, P_{-1}) \leqslant \mathbb{E}_1[\log(dP_1/dP_{-1})] = 2n\|T\varphi_*\|_W^2 \leqslant 2nd\|\varphi_*\|_v^2$. It is well known that the Hellinger distance $H(P_1, P_{-1})$ satisfies $H^2(P_1,P_{-1}) \leqslant KL(P_1,P_{-1})$ and thus

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

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

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

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

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

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

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

$$2nd\sum_{j>m_n^*}[\widetilde{\varphi}_*]_j^2\upsilon_j = 2nd\frac{\zeta\kappa}{\sum_{l>m_n^*}[h]_l^2\gamma_l^{-1}}\sum_{j>m_n^*}\frac{[h]_j^2\upsilon_j}{\gamma_j^2} \leqslant 2nd\zeta\frac{\kappa}{\gamma_{m_n^*}\upsilon_{m_n^*}^{-1}} \leqslant 2d\zeta \leqslant 1.$$

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

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

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

A.2. Proofs of Section 3

We begin by defining and recalling notations to be used in the proofs of this section. For $m \geqslant 1$ 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 $[\varphi_m - E_m\varphi]_{\underline{m}} = [T]_{\underline{m}}^{-1}[TE_m^\perp\varphi]_{\underline{m}}$ hold true. We denote $Q_m := [\widehat{T}]_{\underline{m}} - [T]_{\underline{m}}$ and $V_m := [\widehat{g}]_{\underline{m}} - [\widehat{T}]_{\underline{m}}[\varphi_m]_{\underline{m}} = n^{-1}\sum_{i=1}^n(U_i+\varphi(Z_i)-\varphi_m(Z_i))[f(W_i)]_{\underline{m}}$, where obviously $\mathbb{E}V_m = 0$. Moreover, let us introduce the events

$$\Omega_m := \{\|[\widehat{T}]_{\underline{m}}^{-1}\| \leqslant \sqrt{n}\}, \quad \mho_m := \{\sqrt{m}\|Q_m\|\|[T]_{\underline{m}}^{-1}\| \leqslant 1/2\}$$
$$\Omega_m^c := \{\|[\widehat{T}]_{\underline{m}}^{-1}\| > \sqrt{n}\} \quad \text{and} \quad \mho_m^c = \{\sqrt{m}\|Q_m\|\|[T]_{\underline{m}}^{-1}\| > 1/2\}.$$

Observe that if $\sqrt{m}\|Q_m\|\|[T]_{\underline{m}}^{-1}\| \leqslant 1/2$ 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 $\|[\widehat{T}]_{\underline{m}}^{-1}\| \leqslant 2\|[T]_{\underline{m}}^{-1}\|$. Thereby, if $\sqrt{n} \geqslant 2\|[T]_{\underline{m}}^{-1}\|$ we have $\mho_m \subset \Omega_m$. 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 $m \geqslant 1$ the decomposition

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

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

it follows that

$$
\begin{aligned}
&|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \, \mathbb{1}_{\Omega_m} \\
&\quad \leqslant 2\left|[h]^t_{\underline{m}}[T]^{-1}_{\underline{m}} V_m\right|^2 + 2\left|[h]^t_{\underline{m}}[T]^{-1}_{\underline{m}} Q_m [\widehat{T}]^{-1}_{\underline{m}} V_m\right|^2 \mathbb{1}_{\Omega_m}\left(\mathbb{1}_{\overline{\mho}_m} + \mathbb{1}_{\overline{\mho}^c_m}\right) \\
&\quad \leqslant 2|[h]^t_{\underline{m}}[T]^{-1}_{\underline{m}} V_m|^2 + 2\|[h]^t_{\underline{m}}[T]^{-1}_{\underline{m}}\|^2 \Big\{4\|[T]^{-1}_{\underline{m}}\|^2 \|Q_m\|^2 \|V_m\|^2 + n\|Q_m\|^2\|V_m\|^2 \, \mathbb{1}_{\overline{\mho}^c_m}\Big\}.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}\,|\widehat{\ell}_m - \ell_h(\varphi_m)|^2 \, \mathbb{1}_{\Omega_m} &\leqslant C(\gamma) n^{-1} \|[h]^t_{\underline{m}}[T]^{-1}_{\underline{m}}\|^2 \eta^4 \big(\sigma^2 + \|\varphi - \varphi_m\|^2_\gamma\big) \\
&\quad \times \left\{1 + \frac{m^3}{n}\|[T]^{-1}_{\underline{m}}\|^2 + m^3 P^{1/4}(\overline{\mho}^c_m)\right\}. \qquad \textbf{(A.6)}
\end{aligned}
$$

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

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 $\|\varphi - \varphi_m\|_\gamma \leqslant \|E^\perp_m \varphi\|_\gamma + \|E_m\varphi - \varphi_m\|_\gamma$. As in the proof of Lemma A.2 we conclude $\|E_m\varphi - \varphi_m\|^2_\gamma \leqslant \|E^\perp_m\varphi\|_\gamma \sup_m \sup_{\|\phi\|_\gamma = 1} \|T_m^{-1} F_m T E^\perp_m \phi\|_\gamma \leqslant Dd\,\|E^\perp_m\varphi\|_\gamma$. By using Lebesgue's dominated convergence theorem we observe $\|E^\perp_m\varphi\|_\gamma = o(1)$ as $m \to \infty$ and hence (3.2) holds. Condition $T \in \mathcal{T}^v_{d,D}$ implies $\|[h]^t_{\underline{m}}[T]^{-1}_{\underline{m}}\|^2 \leqslant D\sum_{j=1}^m [h]_j^2 {v_j}^{-1}$ and $\|[T]^{-1}_{\underline{m}}\|^2 \leqslant D v_m^{-1}$ for all $m \geqslant 1$ since v is nonincreasing. Thereby, condition (3.5) implies condition (3.3), which completes the proof. ∎

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 $|\ell_h(\varphi_m - \varphi)|^2 \leqslant 2\rho\left\{\sum_{j>m}[h]_j^2 {\gamma_j}^{-1} + Dd\, v_m \gamma_m^{-1} \sum_{j=1}^m [h]_j^2 {v_j}^{-1}\right\}$ for all $\varphi \in \mathcal{F}^\rho_\gamma$ and $h \in \mathcal{F}_{1/\gamma}$. Since $|\ell_h(\varphi)|^2 \leqslant \|\varphi\|^2_\gamma\,\|h\|^2_{1/\gamma}$ and $\|\varphi\|^2_\gamma \leqslant \rho$ we conclude

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

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

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

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

Combining the last estimate with (A.7) yields

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

Consider now the optimal choice $m=m_n^*$ defined in (2.4), then we have

$$\begin{aligned}\mathbb{E}\,|\widehat{\ell}_{m_n^*} - \ell_h(\varphi)|^2 &\leqslant C(\gamma,\kappa)D\Big\{\sigma^2+\rho\big(\eta^2 d\,D+\|h\|^2_{1/\gamma}\big)\Big\}\mathcal{R}_n^h\\ &\quad\times\Big\{1+(m_n^*)^3\,P(\mho^c_{m_n^*})^{1/4}+(\mathcal{R}_n^h)^{-1}\,P(\Omega^c_{m_n^*})\Big\}\end{aligned}$$

and hence, the assertion follows by making use of Lemma A.4. ■

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 $\varphi\in\mathcal{F}^{\rho}_{\gamma}$ with γ satisfying Assumption 1, then for all $m\geqslant 1$ we have*

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

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

$$\mathbb{E}\|Q_m\|^8 \leqslant C\,\big(n^{-1}m^2\,\eta^2\big)^4. \qquad \textbf{(A.11)}$$

Proof. Proof of (A.9). Since $(\{U_i+\varphi(Z_i)-\varphi_m(Z_i)\}\sum_{j=1}^{m}s_j f_j(W_i))$, $1\leqslant i\leqslant n$, 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 $\mathbb{E}[U^2|W]\leqslant(\mathbb{E}[U^4|W])^{1/2}\leqslant\sigma^2$ and from Assumption 3 (i), i.e., $\sup_{j\in\mathbb{N}}\mathbb{E}[e_j^2(Z)|W]\leqslant\eta^2$. Indeed, applying condition $|j|^3\gamma_j^{-1}=o(1)$ (cf. Assumption 1) gives $\sum_{j\geqslant 1}\gamma_j^{-1}\leqslant C(\gamma)$ and thus,

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

$$\mathbb{E}|\{\varphi(Z) - \varphi_m(Z)\} f_j(W)|^4 \leqslant \|\varphi - \varphi_m\|_\gamma^4 \, \mathbb{E}\Big|\sum_{l\geqslant 1} \gamma_l^{-1} |e_l(Z) f_j(W) - [T]_{j,l}|^2\Big|^2$$
$$\leqslant C(\gamma)\eta^4\|\varphi - \varphi_m\|_\gamma^4.$$

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

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

$$\|E_m\varphi - \varphi_m\|_\gamma^2 \leqslant D\, d\, \rho, \tag{A.12}$$

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

$$|\langle h, \varphi - \varphi_m\rangle_Z|^2 \leqslant 2\rho \sum_{j>m} \frac{[h]_j^2}{\gamma_j} + 2\, D\, d\, \rho\, \frac{\upsilon_m}{\gamma_m} \sum_{j=1}^{m} \frac{[h]_j^2}{\upsilon_j}. \tag{A.14}$$

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

$$\|E_m\varphi - \varphi_m\|_\upsilon^2 \leqslant D\, d\, \gamma_m^{-1}\upsilon_m\|\varphi\|_\gamma^2 \tag{A.15}$$

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

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

$$P\big(m^{-2}n\|Q_m\|^2 \geqslant t\big) \leqslant 2\exp\Big(-\frac{t}{8\eta^2} + 2\log m\Big) \quad \text{for all } 0 < t \leqslant 4\eta^2 n. \tag{A.16}$$

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

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

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

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

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

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

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

$$(m_n^*)^{12}P(\mho_{m_n^*}^c) \leqslant C(\gamma, \upsilon, \eta, D) \tag{A.18}$$

$$(\mathcal{R}_n^h)^{-1}P(\Omega_{m_n^*}^c) \leqslant C(\gamma, \upsilon, \eta, h, D). \tag{A.19}$$

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

$$P(\mho_m^c) \leqslant P\left(m^{-2}n\|Q_m\|^2 > \frac{n\upsilon_m}{4Dm^3}\right) \leqslant 2\exp\left(-\frac{n\upsilon_m}{32D\eta^2m^3} + 2\log m\right)$$

since $(4Dm^3\upsilon_m^{-1})^{-1} \leqslant 1 \leqslant 4\eta^2$ for all $m \geqslant 1$. Due to condition (3.6) there exists $n_0 \geqslant 1$ such that $n\upsilon_{m_n^*} \geqslant 448D\eta^2(m_n^*)^3\log m_n^*$ for all $n \geqslant n_0$. Consequently, $(m_n^*)^{12}P(\mho_{m_n^*}^c) \leqslant 2$ for all $n \geqslant n_0$, while trivially $(m_n^*)^{12}P(\mho_{m_n^*}^c) \leqslant (m_{n_0}^*)^{12}$ for all $n \leqslant n_0$, which gives (A.18) since n_0 and $m_{n_0}^*$ depend on γ, υ, η, and D only.

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

A.3. Proofs of Section 3.4

Proof of Proposition 3.6. Proof of (pp). From the definition of m_n^* in (2.4) 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,

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

$\sum_{j>m_n^*}|j|^{-2p-2s} \sim n^{-(2p+2s-1)/(2p+2a)}$ 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 $s-a \geqslant 0$ the sequence $\omega\upsilon$ 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 $s-a \geqslant 0$ then the sequence $\omega\upsilon$ 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. ■

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 $\Delta := (\Delta_m)_{m\geqslant 1}$ and its empirical analog $\widehat{\Delta} := (\widehat{\Delta}_m)_{m\geqslant 1}$ by $\Delta_m := \max_{1\leqslant m'\leqslant m}\|[h]_{\underline{m}'}^t[T]_{\underline{m}'}^{-1}\|^2$ and $\widehat{\Delta}_m := \max_{1\leqslant m'\leqslant m}\|[h]_{\underline{m}'}^t[\widehat{T}]_{\underline{m}'}^{-1}\|^2$, respectively. Similarly to M_n^+ introduced in (4.5) we define

$$M_n^- := \min\left\{2\leqslant m\leqslant M_n^h : 4D\upsilon_m^{-1}m^3 \max_{1\leqslant j\leqslant m}[h]_j^2 > a_n\right\} - 1 \qquad \textbf{(A.20)}$$

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 $\mathcal{F}_\gamma^\rho$, $\mathcal{T}_{d,D}^\upsilon$, 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 $M_n^- \leqslant M_n \leqslant M_n^+$. Due to condition $(m_n^\circ)^3\max_{1\leqslant j\leqslant m_n^\circ}[h]_j^2 = o(a_n\upsilon_{m_n^\circ})$ as $n\to\infty$ there exists $n_0\geqslant 1$ only depending on h, γ, and υ such that for all $n\geqslant n_0$ it holds $m_n^\circ \leqslant M_n^-$. We distinguish in the following the cases $n\geqslant n_0$ and $n<n_0$. First, consider $n\geqslant n_0$. Applying Corollary A.6 together with estimate (4.2) implies

$$\mathbb{E}|\widehat{\ell}_{\widetilde{m}} - \ell_h(\varphi)|^2 \leqslant \mathcal{C}\left\{\mathrm{pen}_{m_n^\circ} + \mathrm{bias}_{m_n^\circ} + n^{-1}\right\}.$$

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

estimate (A.14) in Lemma A.2 implies for all $1 \leqslant m \leqslant M_n^-$ that $\mathrm{bias}_m \leqslant \min_{1\leqslant m'\leqslant M_n^-} 2\rho\{\sum_{j>m'}[h]_j^2\gamma_j^{-1} + dD\upsilon_{m'}\gamma_{m'}^{-1}\sum_{j=1}^{m'}[h]_j^2\upsilon_j^{-1}\}$.

Consequently,

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

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

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

From the definition of M_n^h we infer $\|[h]_{\underline{M_n^h}}\|^2 \leqslant [h]_1^2 n^{5/4}$. 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 $h \in \mathcal{F}_{1/\gamma}$

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

which proves the result. ∎

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

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

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

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

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

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

Consider the first two right hand side terms. We calculate

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

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

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 $n^{-1}\sum_{m=1}^{M_n^+}\Delta_m \leqslant D(n\upsilon_{M_n^+})^{-1}(M_n^+)^2$ $\max_{1\leqslant j\leqslant M_n^+}[h]_j^2 \leqslant 4D^2$. Consequently, $\mathbb{E}\max_{m_n^\circ\leqslant m\leqslant M_n^+}(I_m + II_m) \leqslant \mathcal{C}n^{-1}$. Further, we obtain for $\varphi \in \mathcal{F}_\gamma^\rho$ and $h \in \mathcal{F}_{1/\gamma}$

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

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

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

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

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

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

Proof. Observe that $m^3\gamma_m^{-1} = o(1)$ and $\|\varphi-\varphi_m\|_Z^2 = o(1)$ as $m\to\infty$ due to Assumption 1 and $T \in \mathcal{T}_{d,D}^\upsilon$ (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 $n \geqslant n_0$ and $m \geqslant m_n^\circ$ we have

$$
\begin{aligned}
24\,\mathbb{E}[U^2] + 96\eta^2\rho\, m^3\gamma_m^{-1} &\leqslant 72\left(\mathbb{E}[Y^2] + \|\varphi_m\|_Z^2 + \|\varphi-\varphi_m\|_Z^2\right) \\
&\quad + 96\eta^2\rho\, m^3\gamma_m^{-1} \leqslant \varsigma_m^2. \qquad \textbf{(A.22)}
\end{aligned}
$$

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

$$\sum_{m=1}^{M_n^+}\Delta_m\,\mathbb{E}\left(|s^t V_m|^2 - \frac{\mathrm{pen}_m}{24\Delta_m}\right)_+ \leqslant \sum_{m=1}^{M_n^+}\Delta_m\,\mathbb{E}\,|s^t V_m|^2 \leqslant 8n_0D^2\left(\sigma^2 + C(\gamma)\,\eta^2\,\|\varphi-\varphi_m\|_\gamma^2\right)n^{-1}.$$

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

Proof of Theorem 4.2. Similarly to the proof of Theorem 4.1 and since $\widehat{\mathrm{pen}}$ is a nondecreasing sequence we have for all $1 \leqslant m \leqslant \widehat{M}_n$

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

Let us introduce the set

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

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

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

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 $n \geqslant 1$

$$\mathbb{E}|\widehat{\ell}_{\widehat{m}} - \ell_h(\varphi)|^2\,\mathbb{1}_{\mathcal{A}} \leqslant \mathcal{C}\,\mathcal{R}^h_{n(1+\log n)^{-1}}. \tag{A.23}$$

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 $\|[h]_{\underline{M_n^h}}\|^2 \leqslant [h]_1^2\,n\,M_n^h$. 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 $h \in \mathcal{F}_{1/\gamma}$

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

The result follows by combining the last inequality with (A.23). ∎

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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 $n \geqslant 1$ and $1 \leqslant m \leqslant \lfloor n^{1/4} \rfloor$ we have*

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

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 $\mathbb{E}[U^2|W] \leqslant \sigma^2$ we have

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

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

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

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

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

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

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

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

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

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

for all $n \geqslant \exp(126\eta^4)$ and $1 \leqslant m \leqslant \lfloor n^{1/4}\rfloor$. 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 $x \mapsto \exp(-1/x)$, $x > 0$, is a concave function and $\mathbb{E}\,|\xi_s(W)|^4 \leqslant m^2\eta^4$ we deduce for all $1 \leqslant m \leqslant \lfloor n^{1/4}\rfloor$

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

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

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

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

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

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

where due to Assumption 3 (i) $\sup_{j,l\in\mathbb{N}}\mathbb{V}\mathrm{ar}(e_l(Z)f_j(W)) \leqslant \eta^2$ and due to Assumption 3 (ii) it holds $\sup_{j,l\in\mathbb{N}}\mathbb{E}\,|e_l(Z)f_j(W)-[T]_{jl}|^k \leqslant k!\eta^k$ for $k \geqslant 3$.

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

Moreover, similarly to the proof of (A.13) in Lemma A.2 we conclude $m^{k/2}\|\varphi-\varphi_m\|_w^k \leqslant (m^3\gamma_m^{-1})^{k/2}(2+2Dd)^{k/2}\rho^{k/2}$. Let us denote $\mu_m := \eta(1+Dd)\sqrt{6\rho\, m^3\gamma_m^{-1}}$. Consequently, for all $m \geqslant 1$ we have $\mathbb{E}|(\varphi(Z)-\varphi_m(Z))\xi_s(W)|^2 \leqslant \mu_m^2$ and

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

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

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

and thus, the assertion follows. ■

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

$$P\left(\bigcup_{m=1}^{M_n^+} \mho_m^c\right) \leqslant C(h,\upsilon,\eta,D)n^{-4}, \tag{A.29}$$

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

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

$$n^4 P\left(\bigcup_{m=1}^{M_n^+} \mho_m^c\right) \leqslant 2\exp\left(-\frac{1}{48\eta D}\frac{n\upsilon_{M_n^+}}{(M_n^+)^3} + 3\log M_n^+ + 4\log n\right) \leqslant C(h,\upsilon,\eta,D).$$

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

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

Proof. Consider $M_n^- \leqslant M_n$. 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 $T \in \mathcal{T}_{d,D}^{\upsilon}$ it holds $\|[T]_{\underline{M_n+1}}^{-1}\|^{-2} \geqslant D^{-1}\upsilon_{M_n+1}$ and thus, by the definition of M_n and M_n^- it is easily seen that

$$\frac{\upsilon_{M_n^-}}{\max\limits_{1\leqslant j\leqslant M_n^-}[h]_j^2(M_n^-)^3} > \frac{4\upsilon_{M_n+1}}{\max\limits_{1\leqslant j\leqslant M_n+1}[h]_j^2(M_n+1)^3},$$

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

$$\frac{\upsilon_{M_n}}{\max\limits_{1\leqslant j\leqslant M_n}[h]_j^2 M_n^3} > \frac{4\upsilon_{M_n^+ +1}}{\max\limits_{1\leqslant j\leqslant M_n^+ +1}[h]_j^2(M_n^+ +1)^3}.$$

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

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

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

$$\mathcal{G} := \Big\{\sigma_Y^2 \leqslant 2\widehat{\sigma}_Y^2 \leqslant 3\sigma_Y^2\Big\}, \tag{A.32}$$

$$\mathcal{J} := \Big\{\|[T]_{\underline{m}}^{-1} V_m\|^2 \leqslant \frac{1}{8}\big(\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + \sigma_Y^2\big) \quad \forall\, 1 \leqslant m \leqslant M_n^+\Big\}. \tag{A.33}$$

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

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

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

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

$$\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 \leqslant (32/9)\|[T]_{\underline{m}}^{-1}V_m\|^2 + 2\|[\widehat{T}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2,$$

$$\|[\widehat{T}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2 \leqslant (32/9)\|[T]_{\underline{m}}^{-1}V_m\|^2 + 2\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2.$$

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

$$(5/9)(\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + \sigma_Y^2) \leqslant \sigma_Y^2 + 2\|[\widehat{T}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2,$$

$$\|[\widehat{T}]_{\underline{m}}^{-1}[\widehat{g}]_{\underline{m}}\|^2 \leqslant (22/9)\|[T]_{\underline{m}}^{-1}[g]_{\underline{m}}\|^2 + (4/9)\sigma_Y^2.$$

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

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

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

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

Consequently, we have

$$\mathcal{H} \cap \mathcal{G} \cap \mathcal{J} \subset \Big\{4\,\Delta_m \leqslant 9\,\widehat{\Delta}_m \leqslant 16\,\Delta_m \text{ and } 5\,\varsigma_m^2 \leqslant 18\,\widehat{\varsigma}_m^2 \leqslant 44\,\varsigma_m^2 \quad \forall 1 \leqslant m \leqslant M_n^+\Big\}$$

and thus, $\mathcal{H} \cap \mathcal{G} \cap \mathcal{J} \subset \Big\{\mathrm{pen}_m \leqslant \widehat{\mathrm{pen}}_m \leqslant 18\,\mathrm{pen}_m \ \forall 1 \leqslant m \leqslant M_n^+\Big\}$. Moreover, it holds $\mathcal{H} \subset \big\{M_n^- \leqslant \widehat{M}_n \leqslant M_n^+\big\}$, 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

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

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

$$\big\{\widehat{M}_n < M_n^-\big\} \subset \Big\{\exists 2 \leqslant m \leqslant M_n^- : \|[\widehat{T}]_{\underline{m}}^{-1}\|^2 \geqslant 4\|[T]_{\underline{m}}^{-1}\|^2\Big\}. \tag{A.34}$$

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, condition $T \in \mathcal{T}_d^{\upsilon}$ with $d \leqslant D$ implies $\upsilon_m^{-1} \leqslant D\|[T]_{\underline{m}}^{-1}\|^2$ as seen in the proof of Lemma A.9. Thereby, we conclude similarly as above

$$\big\{\widehat{M}_n > M_n^+\big\} \subset \Big\{\|[T]_{\underline{M_n^+ + 1}}^{-1}\|^2 \geqslant 4\|[\widehat{T}]_{\underline{M_n^+ + 1}}^{-1}\|^2\Big\}. \tag{A.35}$$

Again applying the Neumann series argument we observe

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

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

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

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

Proof. Due to Lemma A.11 it holds $n^4(M_n^h)^4 P(\mathcal{A}^c) \leqslant n^4(M_n^h)^4 \big\{P(\mathcal{H}^c) + P(\mathcal{J}^c) + P(\mathcal{G}^c)\big\}$. 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 $T \in \mathcal{T}_{d,D}^{\upsilon}$ and Lemma A.3 we infer

$$\begin{aligned} n^4 (M_n^h)^4 P(\mathcal{H}^c) &\leqslant 2\exp\left(-\frac{1}{128D\eta}\frac{n\upsilon_{M_n^+ + 1}}{(M_n^+ + 1)^2} + 3\log(M_n^+ + 1) + 5\log n\right) \\ &\leqslant C(h, \upsilon, \eta, D) \end{aligned} \tag{A.36}$$

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press

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

$$n^4 (M_n^h)^4 P(\mathcal{G}^c) \leqslant n^5 P\big(|\widehat{\sigma}_Y^2 - \sigma_Y^2| > \sigma_Y^2/2\big) \leqslant 1024 \sigma_Y^{-20} n^5 \mathbb{E}\,\big|n^{-1} \sum_{i=1}^n Y_i^2 - \sigma_Y^2\big|^{10}$$
$$\leqslant 1024 \sigma_Y^{-20} \mathbb{E}\,|Y^2 - \sigma_Y^2|^{10} \leqslant C(\gamma, \rho, \sigma, \eta). \tag{A.37}$$

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

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

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

https://doi.org/10.1017/S0266466614000966 Published online by Cambridge University Press