Identification and estimation by penalization in Nonparametric Instrumental Regression*

Jean-Pierre Florens
Toulouse School of Economics

Jan Johannes
Universität Heidelberg

Sébastien Van Bellegem
Toulouse School of Economics

This revision: August 18, 2009

Abstract

The nonparametric estimation of a regression function from conditional moment restrictions involving instrumental variables is considered. The rate of convergence of penalized estimators is studied in the case where the regression function is not identified from the conditional moment restriction. We also study the gain of modifying the penalty in the estimation, considering derivatives in penalty. We analyze the effect of this modification on the identification of the regression function and the rate of convergence of its estimator.

Keywords: Instrumental variable, Nonparametric estimation, Ill-posed inverse problem, Identification, Penalized estimator, Tikhonov regularization, Sobolev norm

JEL classifications: Primary C14; secondary C30

*We are grateful to X. Chen, É. Renault, O. Scaillet and an anonymous referee for helpful comments on a preliminary version of this work

1 Introduction

Among the inverse problems arising in econometrics, more and more attention has been recently paid on nonparametric instrumental regression. In that problem we analyse an economic relationship between a response Y and a vector Z of p explanatory variables that are endogenous. It is also assumed that a set of q instruments denoted by the vector W is given and such that[1]

$$Y = \varphi(Z) + U \; , \qquad Z \in [0,1]^p \tag{1.1a}$$

and

$$\mathbb{E}(U|W) = 0 \; , \qquad W \in [0,1]^q \; . \tag{1.1b}$$

In this model, φ is a nonparametric function that defines the relationship of interest, and is solution of the conditional moment equation

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

The identification of the function φ from (1.2) is of course not straightforward and sufficient conditions can be found in Carrasco, Florens, and Renault (2008).

The above setting is the core of many econometric studies, see e.g. Darolles, Florens, and Renault (2002), Newey and Powell (2003), Matzkin (2003), Chernozhukov and Hansen (2005), Das (2005) or Hall and Horowitz (2005) to name but a few. Among the recent developments in econometric theory, we mention the test of exogeneity of Blundell and Horowitz (2007), the nonparametric instrumental quantile regression of Horowitz and Lee (2007) and the semi-nonparametric estimation of Engel curve with shape-invariant specification of Blundell, Chen, and Kristensen (2007).

Solving the moment equation (1.2) is an *ill-posed inverse problem* because the solution φ does not depend continuously on the regression function $\mathbb{E}(Y|W)$ in the L^2 norm. Therefore, even if we estimate F consistently and get a consistent estimator of the conditional expectations in equation (1.2), it is not guaranteed that the resulting estimator $\hat{\varphi}$ converges in probability to φ.

Note that, if the problem was *well-posed*, a simple least square estimator of φ would be legitimated. If $\widehat{\mathbb{E}}(Y|W)$ and $\widehat{\mathbb{E}}(\varphi(Z)|W)$ denote some consistent estimators of the conditional expectations (nonparametric kernel estimators for instance), the least square estimator finds the function $\hat{\varphi}_{LS}$ that minimizes the L^2 norm $\|\widehat{\mathbb{E}}(\varphi(Z)|W) - \widehat{\mathbb{E}}(Y|W)\|$ over $\varphi \in L^2([0,1]^p)$. However because the inversion of (1.2) is *ill-posed*, that procedure does not lead to a consistent estimator and needs to be modified. One popular recommendation in that situation is, instead of minimizing the above L^2 norm, to minimize the following biased, or penalized, L^2 norm:

$$\|\widehat{\mathbb{E}}(\varphi(Z)|W) - \widehat{\mathbb{E}}(Y|W)\|^2 + \alpha\|\varphi\|^2$$

[1] In this paper we consider the support of Z and W to be compact and included in $[0,1]^p$ and $[0,1]^q$ respectively. All results extend to non compact support in a straightforward way. The simplification considered here allows to use a discrete decompositon of the operators appearing in the proof section.

for some strictly positive α.

Adding a small bias in order to stabilize the inversion is in the line of the well-known paradigm of shrinkage estimators, among which the ridge regression is one standard example. In inverse problem theory, that stabilization procedure is called the "Tikhonov regularization" and has been largely considered in econometric theory, see the above mentioned papers. Of course, the choice of α is crucial in this procedure and in theory α decreases to zero as the sample size growths. The rate of decreasing of α results from a balance between the bias due to the regularization and the variance due to the instability of the inversion.

Another important question is the choice of the norm used in the penalty. The choice of an L^2 norm has been extensively studied in econometrics and leads to the so-called Tikhonov regularization. However, the literature in nonparametric statistics sometimes recommends the use of another penalty function (Wahba (1977), Craven and Wahba (1979), White and Wooldridge (1991), Gagliardini and Scaillet (2006), Chen (2008)). For instance, we could ask what would be the advantages if that norm was replaced by a norm that includes the derivatives of φ such as the Sobolev norm.

It is an open question to analyse what would be the gain when using another penalty norm in the context of nonparametric instrumental regression. The goal of this paper is to give a precise answer and recommendation to that question. Moreover, we link that question to the question of *identification* of the function φ from (1.2). That issue is addressed below at three levels:

(A) Suppose that the function φ is *not* identified from (1.2), but a minimal norm least square solution exists (we give existence conditions below). In that case, we derive the rate of convergence of the Tikhonov-regularized estimator to this solution. We show that in some situations the rate of convergence is not optimal. This is what we call the *saturation effect*;

(B) Suppose we replace the above L^2 penalty by a stronger penalty. By "stronger penalty", we think on a Sobolev penalty that includes the L^2 norm of the first m derivatives of φ for instance. We give conditions on the instruments W and the joint distribution F such that the function φ is identified and under which the *optimal rate of convergence can now be recovered* (no saturation effect);

(C) If the instruments do not fulfill the conditions in (B) and the solution φ is *not* identified, a minimal Sobolev norm least square solution can be defined. Then, if we use a Sobolev-type penalty, we also derive the rate of convergence to this solution, and show that this rate is not always optimal. In other words, in that general situation, we again discover a *saturation effect*.

In the following, each of these three steps is addressed in a separate section. One important aspect of the results below is the structural assumption we impose on φ in order to derive the rates of convergence. These rates are driven by a *relative measure of regularity of φ with respect to the conditional expectation operation* $\mathbb{E}(\varphi(Z)|W)$, a condition that is called "source condition" and already motivated in Florens, Johannes, and Van Bellegem

(2005) and Johannes, Van Bellegem, and Vanhems (2007). It can be surprising that the rate of convergence is not related to the sole smoothness regularity of φ. The intuitive reason is that φ is only identified through the conditional moment equation (1.2) and therefore the conditional expectation operation (is an integral transform of φ) is also determinant in the rate. A precise discussion on that aspect is to be found in the following Section.

Finally, the appendix of the paper presents a unified framework for the proof of all results.

2 Convergence rate of Tikhonov estimator in the nonidentified case

2.1 Source conditions on the minimal norm least square solution

It is convenient to rewrite the moment equation (1.2) in terms of operators between Hilbert spaces. Assume that the vector (Z, W) has a joint density f_{ZW} and f_W, resp. f_Z, denote the marginal densities of W, resp. Z. The moment equation can then be written as

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

with the function

$$r(w) = \mathbb{E}(Y|W = w)f_W(w)$$

and the operator

$$T : L^2[0,1]^p \to L^2[0,1]^q : g \to Tg = \mathbb{E}(g(Z)|W = \cdot)f_W(\cdot).$$

Using this notation, it is easy to give conditions on T and r for the identification and the existence of a solution φ to the moment equation (2.1). A solution is identified if and only if the operator T is injective. If it is not the case that is if the solution is not identified, we may consider the minimal norm least square solution φ^+ that we will now define. We first consider the set of *least square solutions* of (2.1), that is the set of functions $\varphi \in L^2[0,1]^p$ that minimize the L^2 norm $\|r - T\varphi\|$. One can show[2] that φ is least square solution if and only if

$$T^\star T\varphi = T^\star r \tag{2.2}$$

where $T^\star$ is the adjoint operator of T that is given by

$$T^\star : L^2[0,1]^q \to L^2[0,1]^p : h \to T^\star h = \mathbb{E}(h(W)|Z = \cdot)f_Z(\cdot).$$

Among the set of least square solutions, the *minimal* L^2 *norm* (or minimal norm) solution is the function φ^+ such that $\|\varphi^+\| \leqslant \|\varphi\|$ for all least square solution φ. That solution is unique[3].

[2]See e.g. Theorem 2.6 of Engl, Hanke, and Neubauer (2000).
[3]See e.g. Theorem 2.5 of Engl, Hanke, and Neubauer (2000).

The existence conditions of the minimal norm least square solution φ^+ can also be characterized in terms of the operator T and the function r. Consider the range of the operator $\mathcal{R}(T)$. That range is not a closed subset of $L^2[0,1]^q$ in general.[4] Define the orthogonal subspace $\mathcal{R}(T)^\perp$, that is the set of all functions that are orthogonal to the functions in $\mathcal{R}(T)$. Then, the minimal norm least square solution φ^+ exists if and only if the function r belongs to the subset $\mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$ of $L^2[0,1]^q$.

We can now define the measure of regularity of the nonparametric problem that leads the rates the convergence of the estimator. To simplify the exposition, we assume that the operator T is compact but we note that, although that assumption is reasonable for most practical cases, it can be easily relaxed. The compactness of T allows to decompose the operator $(T^\star T)$ that appears in (2.2) using a discrete system of eigenfunctions. More precisely, there exists an orthogonal system $\{\phi_j\}$ in $L^2[0,1]^p$ and a strictly positive sequence $\lambda_1 \geqslant \lambda_2 \geqslant \ldots > 0$ such that

$$(T^\star T)g = \sum_{j=1}^{\infty} \lambda_j^2 \langle g, \phi_j \rangle \phi_j \qquad \text{for all } g \in L^2[0,1]^p$$

where $\langle \cdot, \cdot \rangle$ represents the inner product in L^2. The sequence (λ_j^2, ϕ_j) is called the *eigenvalue decomposition* of $T^\star T$. Because the problem is ill-posed, the sequence of eigenvalues λ_j tends to zero as j tends to infinity, and the degree of ill-posedness is characterized by the rate of decreasing of λ_j^2 to zero.

Once the eigenvalue decomposition of a selfadjoint operator is defined, we can easily define any exponent of the operator: for all $\gamma \in \mathbb{R}$ we set

$$(T^\star T)^\gamma g := \sum_{j=1}^{\infty} \lambda_j^{2\gamma} \langle g, \phi_j \rangle \phi_j$$

which is defined for all $g \in L^2[0,1]^p$ such that the series is a well-defined function in $L^2[0,1]^p$ (that is for all g such that $\sum_{j=1}^{\infty} \lambda_j^{4\gamma} \langle g, \phi_j \rangle^2 < \infty$). That definition is used in the following assumption.

Assumption 2.1 (Source condition). *There exists an exponent $\beta > 0$ and function $\psi \in L^2[0,1]^p$ such that the minimal norm least square solution fulfills $\varphi^+ = (T^\star T)^{\beta/2}\psi$.*

Equivalently, the source condition assumes that the function $(T^\star T)^{-\beta/2}\varphi^+$ belongs to $L^2[0,1]^p$. Note that, for all appropriate g,

$$(T^\star T)^{-\beta/2} g = \sum_{j=1}^{\infty} \frac{\langle g, \phi_j \rangle}{\lambda_j^\beta} \phi_j$$

and, because the eigenvalues λ_j tend to zero, the index β that appears in the source condition is one measure of the degree of ill-posedness of the problem. As we will see below, that index is driving the rate of convergence of the nonparametric estimator. Using a converse result,

[4] The fact that $\mathcal{R}(T)$ is not a closed set is another way to tell that the problem is ill-posed. For a complete exposition, see chapter 2 in Engl, Hanke, and Neubauer (2000). Note that in the parametric model $\mathcal{R}(T)$ is a closed space.

we also show below why the source condition is the natural assumption in the context of nonparametric instrumental variables.

To close this section, we illustrate the source condition in two meaningful examples where φ is univariate.

EXAMPLE 2.1. Some results of Hall and Horowitz (2005) are based under the assumption that there exists two constants $0 < c_1 \leqslant c_2 < \infty$ and $\gamma > 0$ such that $c_1 j^{-2\gamma} \leqslant \lambda_j^2 \leqslant c_2 j^{-2\gamma}$, and assuming that the eigenvectors ϕ_j are the trigonometric functions. In that setting, the source condition is equivalent to assume

$$\sum_{j=1}^{\infty} \frac{\langle \varphi^+, \phi_j \rangle^2}{j^{-2\gamma\beta}} < \infty$$

which is equivalent to assume that φ^+ is $(\gamma\beta)$ times weakly differentiable.

EXAMPLE 2.2. In the case where (Y, Z, W) is jointly Normal, the joint density f_{YZW} is infinitely differentiable. In that situation one can show that the eigenvalues are exponentially decreasing (e.g. Hille and Tamarkin (1931)), that is there exists two constants $0 < c_1 \leqslant c_2 < \infty$ and $\gamma > 0$ such that $c_1 \exp(-2j^\gamma) \leqslant \lambda_j^2 \leqslant c_2 \exp(-2j^\gamma)$. The source condition is here equivalent to assume

$$\sum_{j=1}^{\infty} \frac{\langle \varphi^+, \phi_j \rangle^2}{\exp(-2\beta j^\gamma)} < \infty \, .$$

In that setting, if ϕ_j are the trigonometric functions, one can show (e.g. Kawata (1972)) that the source condition implies that φ^+ is infinitely differentiable (whatever the value of β is).

These examples also show that usual regularity conditions such as the number of derivatives of the solution can be recovered by the source condition.

2.2 Tikhonov regularization with known conditional expectation

The function r and the operator T in the moment equation (2.1) are unknown and must be estimated in practice. It is not the goal of this paper to analyse the statistical properties of estimators of r and T. In subsection 2.4 below, we recall usual nonparametric estimators of these quantities.

Our goal is instead to relate the rate of convergence of penalized L^2 norm estimators of φ^+ to the rate of convergence of the estimators of r and T. To start with, let us first suppose that the operator T is known and only r is estimated from a sample of (Y, Z, W). Denote by $\hat{r}$ an estimator of r. The penalized L^2 norm estimator of φ^+ is defined as the minimizer of

$$\hat{\varphi}_\alpha = \underset{\phi : \|\phi\| < \infty}{\arg\inf} \{ \|T\phi - \hat{r}\|^2 + \alpha \|\phi\|^2 \} \tag{2.3}$$

for a given regularization parameter $\alpha > 0$. The minimizer is unique and given by:

$$\hat{\varphi}_\alpha = \left(\alpha I + T^\star T \right)^{-1} T^\star \hat{r}$$

where I denotes the identity operator. The next result gives the rate of convergence of this regularized estimator.

Theorem 2.1. *Suppose there exists a function $\psi \in L^2[0,1]^p$ such that $\varphi^+ = (T^\star T)^{\beta/2}\psi$ for some $\beta > 0$ (Assumption 2.1). Let $\hat{\varphi}_\alpha$ be the minimizer of the penalized quadratic error (2.3).*

If $\beta \leqslant 2$ and if the regularization parameter α is chosen as

$$\alpha = C\{\mathbb{E}\|\hat{r} - r\|^2\}^{\frac{1}{1+\beta}}$$

for any strictly positive constant C, then the rate of convergence of $\hat{\varphi}_\alpha$ is given by

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi^+\|^2 = O\left(\{\mathbb{E}\|\hat{r} - r\|^2\}^{\frac{\beta}{\beta+1}}\right).$$

If $\beta > 2$ and if the regularization parameter α is chosen as

$$\alpha = C\{\mathbb{E}\|\hat{r} - r\|^2\}^{1/3}$$

for any strictly positive constant C, then the rate of convergence of $\hat{\varphi}_\alpha$ is given by

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi^+\|^2 = O\left(\{\mathbb{E}\|\hat{r} - r\|^2\}^{2/3}\right).$$

Remark 2.1. **(i)** The quality of the convergence of the penalized estimator is different for $\beta \leqslant 2$ or $\beta > 2$. That phenomenon is called the "saturation effect" of the Tikhonov regularization. It is a limitation of the Tikhonov regularization method which is not observed with other regularization methods (see e.g. Johannes, Van Bellegem, and Vanhems (2007)). To understand the technical reason of this limitation, we refer to the appendix, in particular inequality (A.1) of Lemma A.1.

(ii) When $\beta \leqslant 2$, the result is known to be optimal over the class of functions that fulfill the source condition (a related reference is Mair and Ruymgaart (1996)).

(iii) When $\beta > 2$, then the rate of convergence is not optimal in the minimax sense. In a purely deterministic setting, Proposition 5.3 of Engl, Hanke, and Neubauer (2000) shows that for any choice of the regularization parameter α, the bound

$$\sup\left\{\mathbb{E}\|\hat{\varphi}_\alpha - \varphi^+\|^2 \text{ such that } \mathbb{E}\|\hat{r} - r\|^2 \leq \delta\right\} = o(\delta^{2/3})$$

can only hold if $\varphi^+ = 0$. We conjecture that this result holds also in the random setting, meaning that the Tikhonov regularization for the ill-posed linear problem (1.2) never yields a convergence rate faster than $O(\{\mathbb{E}\|\hat{r} - r\|^2\}^{2/3})$.

(iv) The result also shows that the index β of the source condition is the parameter that leads to the rate of convergence. This shows that the source condition is a natural sufficient condition in the context of nonparametric instrumental variables. A natural question to ask is whether a necessary condition can be proved. A partial answer is given in Section 5.1 (p. 120) of Engl, Hanke, and Neubauer (2000): If the estimator $\hat{\varphi}_\alpha$ is such that $O(\{\mathbb{E}\|\hat{r} - r\|^2\}^{\frac{\beta}{\beta+1}})$ for a $\beta \leqslant 2$, then $(T^\star T)^{-\nu/2}\varphi^+$ belongs to $L^2[0,1]^p$

for all $\nu < \beta$. Because that converse result does not hold for $\nu = \beta$, we cannot say that the source condition with β is a necessary condition, but each solution that has this rate of convergence also fulfills the source condition with $\nu < \beta$. In that sense, the source condition is the natural assumption to impose on φ in order to derive the rate of convergence in nonparametric instrumental regression.

2.3 Convergence of the general Tikhonov regularized estimator

If moreover the operator T is unknown, we need to estimate it from a sample of (Y, Z, W) and we denote by $\hat{T}$ the resulting estimator. The penalized estimator of φ^+ is thus defined as the minimizer

$$\hat{\varphi}_\alpha = \arg\inf_{\phi:\|\phi\|<\infty} \{\|\hat{T}\phi - \hat{r}\|^2 + \alpha\|\phi\|^2\}. \tag{2.4}$$

Again, the minimizer is unique and given by

$$\hat{\varphi}_\alpha = \left(\alpha I + \hat{T}^\star \hat{T}\right)^{-1} \hat{T}^\star \hat{r}.$$

The next theorem generalizes the previous one and related the rate of convergence of $\hat{\varphi}^+$ to the one of $\hat{r}$ and $\hat{T}$. In order to state that result, we first need to define the norm for the operator T. By definition, for any operator T,

$$\|T\| = \sup_{\phi \in L^2[0,1]^p} \|T\phi\|$$

where $\|T\phi\|$ in this formula is the $L^2[0,1]^q$ norm of the function $T\phi$.

Theorem 2.2. *Suppose there exists a function $\psi \in L^2[0,1]^p$ such that $\varphi^+ = (T^\star T)^{\beta/2}\psi$ for some $\beta > 0$. Let $\hat{\varphi}_\alpha$ be the minimizer of the penalized quadratic error (2.4). Assume that there exists a constant K such that $\mathbb{E}\|\hat{T} - T\|^4 \leqslant K(\mathbb{E}\|\hat{T} - T\|^2)^2$.*

If $0 < \beta \leqslant 2$ and if the regularization parameter α is chosen as

$$\alpha = C\{\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\}^{\frac{1}{1+\beta}}$$

for any strictly positive constant C, then the rate of convergence of $\hat{\varphi}_\alpha$ is given by

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi^+\|^2 = O\left(\{\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\}^{\frac{\beta}{1+\beta}}\right).$$

If $\beta > 2$ and if the regularization parameter α is chosen as

$$\alpha = C\{\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\}^{1/3}$$

for any strictly positive constant C, then the rate of convergence of $\hat{\varphi}_\alpha$ is given by

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi^+\|^2 = O\left(\{\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\}^{2/3}\right).$$

Remark 2.2. **(i)** The remarks of the previous result still holds for this theorem. For the minimax optimality of the rate of convergence we refer to Chen and Reiss (2007). Note also that the saturation effect is still present.

(ii) In many econometric studies a preconditionning of the moment equation is considered, that is we preliminary apply the dual operator $T^\star$ to (2.1) and start the analysis from the moment equation $T^\star r = T^\star T\varphi$. If we denote $\tilde{r} = T^\star r$ and $\tilde{T} = T^\star T$, then the new moment equation can be written as $\tilde{r} = \tilde{T}\varphi$ and the above results hold with $\tilde{r}$, $\tilde{T}$ instead of r, T. To illustrate that point, consider Theorem 2.2. The source condition assumed in that theorem is $\varphi = (T^\star T)^{\beta/2}\psi$, that is $\varphi = (\tilde{T}^\star\tilde{T})^{\frac{\beta/2}{2}}$. Therefore the rate of convergence of the estimator is

$$\mathbb{E}\|\hat{\varphi}_\alpha - \varphi^+\|^2 = O\left(\{\mathbb{E}\|\widehat{T^\star r} - T^\star r\|^2 + \mathbb{E}\|\widehat{T^\star T} - T^\star T\|^2\}^{\frac{\beta}{2+\beta}}\right).$$

when $\beta < 2$ (and similarly for $\beta > 2$). Preconditionning is considered for instance in the seminal work of Darolles, Florens, and Renault (2002).

2.4 Examples

Even if this paper does not study the statistical properties of $\hat{r}$ and $\hat{T}$, we find useful to give two examples of estimators that fulfill the assumptions of our results. To write these estimators, we suppose that we observe n vectors (Y_i, Z_i, W_i) identically distributed as (Y, Z, W).

Example 2.3 (Kernel estimator). Kernel estimators of r and T have often been used, see e.g. Darolles, Florens, and Renault (2002), Hall and Horowitz (2005), Carrasco, Florens, and Renault (2008) among others. Let $K(\cdot)$ be a multivariate kernel function. Estimators of r, T and $T^\star$ are given by

$$\hat{r}(\cdot) = \frac{1}{nh_W^q}\sum_{i=1}^n Y_i K\left(\frac{W_i - \cdot}{h_W}\right)$$

$$\hat{T}\phi(\cdot) = \frac{1}{nh_W^q h_Z^p}\sum_{i=1}^n K\left(\frac{W_i - \cdot}{h_W}\right)\int K\left(\frac{Z_i - z}{h_Z}\right)\phi(z)dz \qquad \text{for all } \phi \in L^2[0,1]^p$$

$$\hat{T}^\star g(\cdot) = \frac{1}{nh_W^q h_Z^p}\sum_{i=1}^n K\left(\frac{Z_i - \cdot}{h_Z}\right)\int K\left(\frac{W_i - w}{h_W}\right)g(w)dw \qquad \text{for all } g \in L^2[0,1]^q$$

Rates of convergence of these estimators under sampling assumptions and smoothness restrictions on the joint density $f_{Y,W,Z}$ can be found e.g. in Lemma A.3 of Florens, Johannes, and Van Bellegem (2005).

Example 2.4 (Series expansion). Another popular nonparametric estimator is given by series expansion or sieve estimators, see e.g. Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007) or Chen (2008). Let $\boldsymbol{\phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_{m_Z}(\cdot))'$ be a vector of functions that form an orthonormal a basis of $\Phi_{m_Z} \subseteq L^2[0,1]^p$ and $\boldsymbol{\psi}(\cdot) = (\psi_1(\cdot), \ldots, \psi_{m_W}(\cdot))'$ be a vector of functions that form an orthonormal basis of $\Psi_{m_W} \subseteq L^2[0,1]^q$. Note that the number of elements in these bases depends on the parameters m_Z, m_W. In order to derive the series estimator of r and T, we first define the vector and matrix

$$\hat{v} = \frac{1}{n}\sum_{i=1}^n Y_i\boldsymbol{\psi}(W_i)\ , \qquad \widehat{M} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(W_i)\boldsymbol{\phi}(Z_i)'$$

Therefore the series estimator of T is $\widehat{T}g(\cdot) := \boldsymbol{\psi}(\cdot)'\widehat{M}\langle g, \boldsymbol{\phi}\rangle$ where $\langle g, \boldsymbol{\phi}\rangle$ denotes the column vector $(\langle g, \phi_1\rangle, \ldots, \langle g, \phi_{m_Z}\rangle)'$. The estimator of $T^\star$ is the dual of $\widehat{T}$, that is $\widehat{T}^\star h(\cdot) := \boldsymbol{\phi}(\cdot)'\widehat{M}'\langle h, \tilde{\boldsymbol{\psi}}\rangle$ where $\langle h, \boldsymbol{\psi}\rangle$ analogously denotes the column vector $(\langle h, \psi_1\rangle, \ldots, \langle h, \psi_{m_W}\rangle)'$. Finally the series estimator of r is $\widehat{r}(\cdot) = \boldsymbol{\psi}(\cdot)'\widehat{v}$. The estimator of the least square norm solution φ^+ is $\hat{\varphi} = \sum_{j=1}^{m_Z} \hat{a}_j \phi_j$ where the coefficients $\hat{a} = (\hat{a}_j)$ are given by $\hat{a} = (\alpha I + \hat{M}'\hat{M})^{-1}\hat{M}'\hat{v}$. Rates of convergence of these estimators under sampling assumptions and smoothness restrictions on the joint density $f_{Y,W,Z}$ can be found e.g. in Proposition 3.1 of Johannes, Van Bellegem, and Vanhems (2007), who also consider the most relevant case where $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are not orthonormal.

3 Recovering the optimal rate of convergence by stronger penalization

In this section, we start to study the situation where the norm of the penalty is no longer the L^2 norm. We give conditions on the penalty function under which the optimal rate of convergence is always recovered, that is there is no saturation effect.

The core of the results in this section is to define an appropriate connection between the operator T and the norm considered in the penalty. We start with the simplest (but unrealistic) situation where the penalty involves T via the source condition. Then we consider the realistic case where the penalty norm is a Sobolev-type of norm that includes the derivatives of the function.

3.1 Convergence with a penalty adapted to T

To simplify the exposition, we assume in this subsection that the solution φ is identified from the moment equation (2.1) (e.g. the operator T is injective).

Suppose first that T is known and the source condition is satisfied for some $\beta > 0$ (Assumption 2.1). In the result below we consider the estimator of φ that is given by the unique minimizer

$$\hat{\varphi}_\alpha^\beta = \arg\inf_{\phi}\{\|T\phi - \hat{r}\|^2 + \alpha\|(T^\star T)^{-\beta/2}\phi\|^2\}, \tag{3.1}$$

for a given regularization parameter $\alpha > 0$. This estimator is unfeasible as long as T is unknown, but it can be considered as an ideal estimator and it is worth studying its properties. Due to the particular penalty function considered here, the solution is forced to satisfy the source condition. We therefore say that the norm is "stronger" than the L^2 norm because it reflects an a priori knowledge about the solution φ. To illustrate this, we can consider the situation of Example 2.1. In that example, the modified penalty imposes that the solution is $(\gamma\beta)$ times weakly differentiable.

The resulting estimator is equivalently given by:

$$\hat{\varphi}_\alpha^\beta = \left(\alpha I + (T^\star T)^{\beta+1}\right)^{-1}(T^\star T)^\beta T^\star \hat{r}.$$

The following result derives the rate of convergence of this estimator, which is found to be the optimal minimax rate of convergence.

Theorem 3.1. *Suppose there exists a function $\psi \in L^2[0,1]^p$ such that $\varphi = (T^\star T)^{\beta/2}\psi$ for some $\beta > 0$. Let $\hat\varphi_\alpha^\beta$ be the minimizer of the penalized functional (3.1). If the regularization parameter α is chosen as*

$$\alpha = C\,\mathbb{E}\|\hat r - r\|^2,$$

for any constant C, then we obtain

$$\mathbb{E}\|\hat\varphi_\alpha^\beta - \varphi\|^2 = O\left(\{\mathbb{E}\|\hat r - r\|^2\}^{\frac{\beta}{\beta+1}}\right) .$$

That result shows that the saturation effect does not occur when the penalty is adapted to the source condition. Of course, a drawback is that this penalty function depends on the operator T which is unknown in practice. The next result considers the feasible situation where the penalty does not involve T.

3.2 Convergence with a Sobolev penalty

Consider the Sobolev space of periodic functions in $L^2[0,1]$, that are defined for all $\mathfrak{s} > 0$ as

$$H^{\mathfrak{s}} = \left\{ g \in L^2[0,1] \text{ such that } \|g\|_{\mathfrak{s}}^2 := \sum_{k=1}^{\infty} k^{2\mathfrak{s}} \langle g, e_k\rangle^2 < \infty \right\}$$

where

$$e_k = \sqrt{2}\sin\left[\pi x(k-1/2)\right] \ . \tag{3.2}$$

For integer m an equivalent definition of H^m is provided in terms of the weak derivatives of its elements (e.g. Mair and Ruymgaart (1996)):

$$H^m = \Big\{ g \in L^2[0,1] \text{ such that } f^{(m-1)} \text{ is absolutely continuous}, f^{(m)} \in L^2[0,1], \\ f^{(2j)}(0) = f^{(2j)}(1) = 0, \quad j = 0,1,\dots,\lfloor (m-1)/2 \rfloor \Big\}.$$

In terms of the weak derivatives of g, the Sobolev norm of g is equivalently defined as $\|g\|_{\mathfrak{s}} = \pi^{-2\mathfrak{s}}\|g^{(\mathfrak{s})}\|$. The Sobolev space in $L^2[0,1]^p$ can be defined analogously for multivariate functions. Its definition however involves a multi-index notation that complicates the presentation of the result without adding any essential benefit. Therefore, in what follows, we only present the results in the case of an univariate function φ. Note further that in case $\mathfrak{s} < 0$ we define the Hilbert space $H^{\mathfrak{s}}$ by completion with respect to the Hilbert space norm $\|\cdot\|_{\mathfrak{s}}$.

In the following, we consider a penalty given by the Sobolev norm. We derive the optimal rate of convergence under the following assumption that determines a particular connection between the operator T and the Sobolev norm.

Assumption 3.1. *The operator T in the moment equation (2.1) is* adapted to the Sobolev spaces $(H^{\mathfrak{s}})$, *that is there exists two constants $0 < d \leqslant D < \infty$ and a number $\mathfrak{a} > 0$ such that*

$$d\|g\|_{-\mathfrak{a}} \leqslant \|Tg\| \leqslant D\|g\|_{-\mathfrak{a}}$$

for every $g \in L^2[0,1]$.

This formal constraint should be seen as an assumption on the joint distribution F of (Y, Z, W). We illustrate that assumption on the above examples. In Example 2.1 the explicit form of the eigenvectors and eigenvalues actually impose a strong constraint on the distribution F itself. In that example, we can see after some algebra that the operator T is adapted to the Sobolev spaces with $\mathfrak{a} = \gamma$. Example 2.2 is an example where T is never adapted, therefore the Normal case is not covered by the results of this section (see also Remark 3.1 (iii) below).

Note also that Assumption 3.1 implies that the operator T is injective, that is the solution φ is identified from the moment equation (2.1).

For a given $\mathfrak{p} > 0$, the estimator of φ with $H^{\mathfrak{p}}$ penalty is now defined as the minimizer

$$\hat{\varphi}_\alpha^{\mathfrak{p}} = \arg\inf_{\phi}\{\|T\phi - \hat{r}\|^2 + \alpha\|\phi\|_{\mathfrak{p}}^2\}. \tag{3.3}$$

where the penalty $\|\phi\|_{\mathfrak{p}}^2$ is now the norm in the Sobolev space $H^{\mathfrak{p}}$. In order to give an explicit expression of this estimator we denote by $T_{|\mathfrak{p}}$ the operator T restricted on the functions in $H^{\mathfrak{p}}$. By definition the operator $T_{|\mathfrak{p}}$ maps $H^{\mathfrak{p}}[0,1]$ to $L^2[0,1]$. Its adjoint is denoted by $T_{|\mathfrak{p}}^{\star}$ and is defined with respect to the inner product in $H^{\mathfrak{p}}$, that is

$$\langle g, h\rangle_{\mathfrak{p}} := \sum_{k=1}^{\infty} k^{2\mathfrak{p}}\langle g, e_k\rangle\langle h, e_k\rangle.$$

Then the explicit solution to the minimization problem (3.3) is given by

$$\hat{\varphi}_\alpha^{\mathfrak{p}} = \left(\alpha I + T_{|\mathfrak{p}}^{\star}T_{|\mathfrak{p}}\right)^{-1} T_{|\mathfrak{p}}^{\star}\hat{r}\ .$$

We can now derive the rate of convergence of $\hat{\varphi}_\alpha^{\mathfrak{p}}$.

Theorem 3.2. *Assume that the solution φ of the moment equation (2.1) satisfies $\|\varphi\|_{\mathfrak{p}} < \infty$ for some $\mathfrak{p} > 0$, and assume that the operator T is adapted to the Sobolev spaces with $\mathfrak{a} > 0$ (Assumption 3.1). Let $\hat{\varphi}_\alpha^{\mathfrak{p}}$ be the minimizer of the penalized functional (3.3). If the regularization parameter α is chosen as*

$$\alpha = C\ \mathbb{E}\|\hat{r} - r\|^2$$

for any strictly positive constant C and if $-\mathfrak{a} \le \mathfrak{s} \le \mathfrak{p}$ then

$$\mathbb{E}\|\hat{\varphi}_\alpha^{\mathfrak{p}} - \varphi\|_{\mathfrak{s}}^2 = O\left(\{\mathbb{E}\|\hat{r} - r\|^2\}^{\frac{\mathfrak{p}-\mathfrak{s}}{\mathfrak{p}+\mathfrak{a}}}\right).$$

Remark 3.1. **(i)** The two conditions of the theorem are $\|\varphi\|_{\mathfrak{p}} < \infty$ and the adaptation of T to the Sobolev spaces for some $\mathfrak{a} > 0$. When the operator T is adapted, the condition $\|\varphi\|_{\mathfrak{p}} < \infty$ and the source condition (Assumption 2.1) are equivalent with $\beta = \mathfrak{p}/\mathfrak{a}$.

(ii) The result of the theorem is about the asymptotic convergence in the $H^{\mathfrak{s}}$ norm. In particular, if $\mathfrak{s}$ is an integer, Theorem 3.2 also give the rate of convergence of the $\mathfrak{s}$th derivative of φ. Convergence in the L^2 norm corresponds to $\mathfrak{s} = 0$. As $\beta = \mathfrak{p}/\mathfrak{a}$, the optimal minimax rate of convergence is found for every $\beta > 0$ (no saturation effect).

(iii) The fact that the Normal situation of Example 2.2 is not covered by this result has important consequences in practice. If the variables (Y, Z, W) are normaly distributed, the operator is not adapted and the polynomial rate of convergence derived above cannot be reached. It is known that the minimax-optimal rate of convergence over the class $H^{\mathfrak{p}}$ is logarithmic in that situation, and is reached by a simple penalized estimator with L^2 penalty (without any saturation effect). The conclusion of this example is that a Sobolev penalty does not improve the convergence in the Normal model when φ is in the class $H^{\mathfrak{p}}$.

(iv) Consider again the Normal situation of Example 2.2. For these variables, the source condition $(T^\star T)^{-\beta/2}\varphi \in L^2[0,1]$ for some $\beta > 0$ implies that φ is infinitely differentiable and therefore φ satisfies the moment equation $r = T_{|\mathfrak{p}}\varphi$ for all $\mathfrak{p} > 0$. Moreover we can see after some algebra that if the source condition $(T^\star T)^{-\beta/2}\varphi \in L^2[0,1]$ holds for some $\beta > 0$, then $(T^\star_{|\mathfrak{p}} T_{|\mathfrak{p}})^{-\nu/2}\varphi \in L^2[0,1]$ for any $\mathfrak{p} > 0$ and $\nu < \beta$. In that situation for any $\mathfrak{p} > 0$ with $(T^\star_{|\mathfrak{p}} T_{|\mathfrak{p}})^{-\beta/2}\varphi \in L^2[0,1]$, Theorem 2.1 gives the optimal rate of convergence $\mathbb{E}\|\hat{\varphi}^{\mathfrak{p}}_\alpha - \varphi\|^2_{\mathfrak{p}} = O\left(\{\mathbb{E}\|\hat{r} - r\|^2\}^{\beta/(\beta+1)}\right)$ when $\beta \leq 2$, while if $\beta > 2$ the saturation effect occurs, i.e., $\mathbb{E}\|\hat{\varphi}^{\mathfrak{p}}_\alpha - \varphi\|^2_{\mathfrak{p}} = O\left(\{\mathbb{E}\|\hat{r} - r\|^2\}^{2/3}\right)$. The conclusion of this example is that a Sobolev penalty does not allow to overcome the saturation effect in the Normal setting.

In the case where T is unknown, let $\hat{T}$ be an estimator from a sample of observations. In the next general result, we still assume that T is adapted to the Sobolev spaces. The estimator is given by

$$\hat{\varphi}^{\mathfrak{p}}_\alpha = \underset{\phi:\|\phi\|_{\mathfrak{p}}<\infty}{\arg\inf}\ \{\|\hat{T}\phi - \hat{r}\|^2 + \alpha\|\phi\|^2_{\mathfrak{p}}\}. \tag{3.4}$$

and is explicitly given by

$$\hat{\varphi}^{\mathfrak{p}}_\alpha = \left(\alpha I + \hat{T}^\star_{|\mathfrak{p}}\hat{T}_{|\mathfrak{p}}\right)\hat{T}^\star_{|\mathfrak{p}}\hat{r}$$

The following result extends the previous theorem.

Theorem 3.3. *Assume that the minimal norm solution φ^+ satisfies $\|\varphi^+\|_{\mathfrak{p}} < \infty$ for some $\mathfrak{p} > 0$, and assume that the operator T is adapted to the Sobolev spaces with $\mathfrak{a} > 0$ (Assumption 3.1). Moreover assume there exists two constants C_1 and C_2 such that the estimators $\hat{r}$ and $\hat{T}$ fulfill $\mathbb{E}\|\hat{r} - r\|^4 \leqslant C_1(\mathbb{E}\|\hat{r} - r\|^2)^2$ and $\mathbb{E}\|\hat{T} - T\|^4 \leqslant C_2(\mathbb{E}\|\hat{T} - T\|^2)^2$. Let $\hat{\varphi}^{\mathfrak{p}}_\alpha$ be the minimizer of the penalized functional (3.4). If the regularization parameter α is chosen as*

$$\alpha = C\ \left(\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\right)$$

for any strictly positive constant C and if $-\mathfrak{a} \leqslant \mathfrak{s} \leqslant \mathfrak{p}$, then

$$\mathbb{E}\|\hat{\varphi}^{\mathfrak{s}}_\alpha - \varphi\|^2_{\mathfrak{s}} = O\left(\{\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\}^{\frac{\mathfrak{p}-\mathfrak{s}}{\mathfrak{p}+\mathfrak{a}}}\right).$$

As we have already mentioned, although the assumption of adaptation of T to Sobolev space is convenient and allows to derive fast optimal rates of convergence, it is an strong constraint for many practical situations. In particular, when T is not injective, that assumption is not satisfied. It is also not fulfilled in the Normal setting of Example 2.2 (cf. Remark 3.1 (iii)).

The next section derives rates of convergence when that condition is relaxed.

4 The effect of a stronger penalization in the nonidentified case

Even if the operator T is not injective, we might want to use a Sobolev-type of penalty if our economic interest is for instance to focus on the mth derivative, if it exists.

For some $\mathfrak{s} > 0$, we therefore define the *minimal $H^{\mathfrak{s}}$ norm solution* among the least-squares solutions as the function $\varphi_{\mathfrak{s}}^{+}$ such that $\|\varphi^{+}\|_{\mathfrak{s}} \leqslant \|\varphi\|_{\mathfrak{s}}$ for all least-square solution φ of the moment equation (2.1).

Here again, existence of this solution can be characterized in terms of the operator T. This solution exists if and only if $r \in \mathcal{R}(T_{|\mathfrak{s}}) \oplus \mathcal{R}(T)^{\perp}$, where $T_{|\mathfrak{s}}$ denotes the restriction of T on the Sobolev space $H^{\mathfrak{s}}$.

We again consider the functional (3.4) with Sobolev penalty. However, note that it is not obvious that the minimizer defined by

$$\varphi_{\alpha}^{\mathfrak{s}} = \arg\inf_{\phi}\{\|T\phi - r\|^2 + \alpha\|\phi\|_{\mathfrak{s}}^2\}$$

converges to a solution. Locker and Prenter (1980) showed that the minimizer $\varphi_{\alpha}^{\mathfrak{s}}$ actually converges to $\varphi_{\mathfrak{s}}^{+} \in H_{\mathfrak{s}}$, as the regularization parameter α tends to zero. Moreover, they proved that $\|\varphi_{\alpha}^{\mathfrak{s}}\| \to \infty$ if there does not exist a least-squares solution in $H^{\mathfrak{s}}$.

Theorem 4.1. *Suppose there exists a function $\psi \in H^{\mathfrak{s}}$ such that the minimal $H^{\mathfrak{s}}$ norm least-squares solution satisfies the source condition $\varphi_{\mathfrak{s}}^{+} = (T_{|\mathfrak{s}}^{\star}T_{|\mathfrak{s}})^{\beta/2}\psi$ for some $\beta > 0$. Suppose also there exists a positive, finite constant K such that $\mathbb{E}\|\hat{T} - T\|^4 \leqslant K(\mathbb{E}\|\hat{T} - T\|^2)^2$. Let $\hat{\varphi}_{\alpha}^{\mathfrak{s}}$ be the minimizer of the penalized functional (3.4).*

If $0 \leqslant \beta \leqslant 2$ and if the regularization parameter α is chosen as

$$\alpha = C\left(\{\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\}^{\frac{1}{1+\beta}}\right)$$

for any strictly positive constant C, then the rate of convergence of $\hat{\varphi}_{\alpha}^{\mathfrak{s}}$ is

$$\mathbb{E}\|\hat{\varphi}_{\alpha}^{\mathfrak{s}} - \varphi_{\mathfrak{s}}^{+}\|_{\mathfrak{s}}^2 = O(\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2)^{\frac{\beta}{\beta+1}}.$$

If $\beta > 2$ and if the regularization parameter α is chosen as

$$\alpha = C\left(\{\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2\}^{\frac{1}{3}}\right)$$

for any strictly positive constant C, then the rate of convergence of $\hat{\varphi}_{\alpha}^{\mathfrak{s}}$ is

$$\mathbb{E}\|\hat{\varphi}_{\alpha}^{\mathfrak{s}} - \varphi_{\mathfrak{s}}^{+}\|_{\mathfrak{s}}^2 = O(\mathbb{E}\|\hat{r} - r\|^2 + \mathbb{E}\|\hat{T} - T\|^2)^{\frac{2}{3}}.$$

To understand that result consider e.g. the case $\mathfrak{s} = 1$. There we are interested to estimate the solution φ_1^{+} whose first derivative has minimal L^2 norm. Theorem 4.1 derives the rate of convergence of the first derivative of the estimator, to the first derivative of φ_1^{+}.

Note also that we rediscover a saturation effect in this theorem. Therefore the rate of convergence is optimal in the minimax sense only if $\beta \leqslant 2$.

5 Discussion

Considering derivatives in the penalty of the Tikhonov estimator is a well studied technique in the nonparametric literature. In nonparametric instrumental regression the operator to be inverted is the conditional expectation and has to be estimated from data. Considering derivatives in the penalty when the operator is estimated is theoreticaly less clear and is an open topic of econometrics.

The first main result of this paper is to show that *if the operator of the inverse problem is adapted to the space of differentiable functions* (Sobolev spaces) then the Tikhonov estimator penalized by derivatives can reach optimal rates of convergence (Theorem 3.3). These rates of convergence are considered in the L_2 norm as well as the L_2 norm of the derivatives. Moreover, the adaptation of the operator is a sufficient condition of identification (Assumption 3.1).

This paper also argue that the adaptivity assumption on the operator is restrictive. In particular, it does not allow the endogeneous variables and instruments to be normally distributed. Therefore, we also study the convergence of the estimator when this adaptivity condition does not hold. Because the solution is not necessarily identified, we study the convergence to the minimal norm solution, and we show that the penalized estimator is optimal only if the regression function is not too regular compared with the smoothness of the operator (Theorem 4.1 with $\beta < 2$).

A Proofs

The first lemma collects a set of useful inequalities on the operator norm that are used in our proofs. Note also that, in what follows, we write $A \lesssim B$ when there exists a positive, finite constant c that does not depend on A, B and is such that $A \leqslant cB$.

Lemma A.1. *Let $K : \mathbb{H} \to \mathbb{G}$ be a linear operator defined between the two Hilbert spaces $\mathbb{H}$ and $\mathbb{G}$, and let $K^\star$ be the adjoint operator of K. Then, for all $\alpha > 0$, the following bounds on the operator norm holds true:*

$$\|\alpha(\alpha I + K^\star K)(K^\star K)^\gamma\| \leqslant \begin{cases} \alpha^\gamma & \text{if } 0 < \gamma \leqslant 1 \\ \|K^\star K\|^{\gamma-1}\alpha & \text{if } \gamma > 1 \end{cases} , \tag{A.1}$$

$$\|(\alpha I + K^\star K)^{-1} K^\star\| = \|K(\alpha I + K^\star K)^{-1}\| \lesssim 1/\sqrt{\alpha} \, , \tag{A.2}$$

$$\|\alpha I + K^\star K\| \leqslant 1/\alpha \, , \tag{A.3}$$

$$\|K(\alpha I + K^\star K)^{-1} K^\star\| \leqslant 1 \, , \tag{A.4}$$

$$\|K[I - (\alpha I + K^\star K)^{-1} K^\star K]\| \lesssim \sqrt{\alpha} \, , \tag{A.5}$$

$$\|I - (\alpha I + K^\star K)^{-1} K^\star K\| \leqslant 1 \, . \tag{A.6}$$

Proof. By definition of the operator norm for selfadjoint operators, and by straightforward algebra, we can write for $\gamma \leqslant 1$:

$$\|\alpha(\alpha I + K^\star K)(K^\star K)^\gamma\| = \sup_{\lambda \in \sigma(K^\star K)} \left(\frac{\alpha \lambda^\gamma}{\alpha + \lambda} \right) \leqslant \alpha^\gamma$$

where $\sigma(K^\star K)$ denotes the spectrum of the operator $K^\star K$. If $\gamma > 1$, we have

$$\|\alpha(\alpha I + K^\star K)(K^\star K)^\gamma\| = \sup_{\lambda \in \sigma(K^\star K)} \left(\lambda^{\gamma-1} \frac{\alpha\lambda}{\alpha + \lambda} \right) \leqslant \alpha \sup_{\lambda \in \sigma(K^\star K)} \left(\lambda^{\gamma-1} \right) = \alpha \|K^\star K\|^{\gamma-1}$$

for all $\alpha > 0$. This proves the first inequality. To prove the second inequality, we first notice that $\|K(\alpha I + K^\star K)^{-1}\| = \|(K^\star K)^{1/2}(\alpha I + K^\star K)^{-1}\|$ and then proceed similarly:

$$\|(K^\star K)^{1/2}(\alpha I + K^\star K)^{-1}\| = \sup_{\lambda\in\sigma(K^\star K)} \left(\frac{\sqrt{\lambda}}{\alpha+\lambda}\right) \leqslant 1/\sqrt{\alpha}$$

for all $\alpha > 0$. The other inequalities are directly derived from (A.1). □

Proof of Theorem 2.1. The proof is a straightforward generalization of the standard proof when φ is identified, and we give its main lines for the sake of completeness. We start from the standard decomposition of the loss between the bias term and the stochastic term:

$$\|\varphi^+ - \hat{\varphi}_\alpha\|^2 \lesssim \|\varphi^+ - \varphi_\alpha\|^2 + \|\varphi_\alpha - \hat{\varphi}_\alpha\|^2$$

where $\varphi_\alpha = (\alpha I + T^\star T)^{-1}T^\star r$ is called the regularized solution. Using also $T^\star r = T^\star T\varphi^+$ and the source condition ($\varphi^+ = (T^\star T)^{\beta/2}\psi$) the bias term is bounded as follows:

$$\begin{aligned}\|\varphi^+ - \varphi_\alpha\|^2 &= \|(T^\star T)^{\beta/2}\psi - (\alpha I + T^\star T)^{-1}T^\star T(T^\star T)^{\beta/2}\psi\|^2\\ &\leqslant \|\psi\|^2\|\{I - (\alpha I + T^\star T)^{-1}T^\star T\}(T^\star T)^{\beta/2}\|^2\end{aligned}$$

where the last norm is now the norm on operators. Note that $I-(\alpha I+T^\star T)^{-1}T^\star T = \alpha(\alpha I+T^\star T)^{-1}$. Therefore, Lemma A.1, (A.1), implies that the bias term is bounded by $\alpha^{\min(\beta,2)}$.

An expansion of $\hat{\varphi}_\alpha$ and φ_α in the stochastic term leads to the bound

$$\|\hat{\varphi}_\alpha - \varphi_\alpha\|^2 \lesssim \|(\alpha I + T^\star T)^{-1}T^\star\|^2 \cdot \|r - \hat{r}\|^2 \tag{A.7}$$

where the first factor is a norm over operators. This norm is equal to the norm of the adjoint operator, that is $\|(\alpha I + T^\star T)^{-1}T^\star\|^2 = \|T(\alpha I + T^\star T)^{-1}\|^2 \leqslant 1/\alpha$ by Lemma A.1, inequality (A.2). Therefore the stochastic term (A.7) is bounded by $\|r - \hat{r}\|^2/\alpha$. The final result follows. □

Lemma A.2. *Let χ be a random function and K be a linear, compact operator between Hilbert spaces.*

1. *If $\mathbb{E}\|\chi\|^2 \lesssim (\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{1/2}$ with $0 < \beta \leqslant 1$, then*

$$(\mathbb{E}\|\chi\|^2)^{1/\beta} \lesssim \mathbb{E}\|K\chi\|^2(\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{-1/2}.$$

2. *If $\mathbb{E}\|(K^\star K)^{-\beta/2}\chi\|^2 \lesssim 1$ with $\beta > 0$, then $\mathbb{E}\|(K^\star K)^{-\mathfrak{s}/2}\chi\|^2 \lesssim (\mathbb{E}\|K\chi\|^2)^{(\beta-\mathfrak{s})/(1+\beta)}$ for all $\mathfrak{s}$ such that $-1 \leqslant \mathfrak{s} \leqslant \beta$.*

Proof. Let $\{\lambda_i, g_i\}$ denote the eigenvalue decomposition of $K^\star K$, where $(\lambda_i)_i$ is a sequence of strictly positive eigenvalues of $K^\star K$. Define the function $\Phi(t) := t^{1/\beta}, t \geqslant 0$ which is convex for all $0 < \beta \leq 1$. By Jensen's inequality,

$$\begin{aligned}\Big(\frac{\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2}{\mathbb{E}\|\chi\|^2}\Big)^{\frac{1}{\beta}} &= \Phi\Big(\frac{\sum_i \lambda_i^\beta \mathbb{E}\langle\chi, g_i\rangle^2}{\sum_i \mathbb{E}\langle\chi, g_i\rangle^2}\Big)\\ &\leqslant \frac{\sum_i \Phi(\lambda_i^\beta)\mathbb{E}\langle\chi, g_i\rangle^2}{\sum_i \mathbb{E}\langle\chi, g_i\rangle^2} = \frac{\sum_i \lambda_i\mathbb{E}\langle\chi, g_i\rangle^2}{\sum_i \mathbb{E}\langle\chi, g_i\rangle^2} = \frac{\mathbb{E}\|K\chi\|^2}{\mathbb{E}\|\chi\|^2}.\end{aligned}$$

We multiply this inequality by $\mathbb{E}\|\chi\|^2/\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2$ and obtain

$$\Big(\frac{\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2}{\mathbb{E}\|\chi\|^2}\Big)^{\frac{1-\beta}{\beta}} \leqslant \frac{\mathbb{E}\|K\chi\|^2}{\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2}.$$

Moreover, we have $(\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{1/4} \lesssim (\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{1/2}/(\mathbb{E}\|\chi\|^2)^{1/2}$ by using that $\mathbb{E}\|\chi\|^2 \lesssim (\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{1/2}$. Consequently, it follows

$$\Big(\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2\Big)^{\frac{1-\beta}{2\beta}} \lesssim \frac{\mathbb{E}\|K\chi\|^2}{\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2}$$

and hence by multiplying with $(\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{1/2}$ it follows

$$(\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{\frac{1}{2\beta}} \lesssim \frac{\mathbb{E}\|K\chi\|^2}{(\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{1/2}},$$

which by using again $\mathbb{E}\|\chi\|^2 \lesssim (\mathbb{E}\|(K^\star K)^{\beta/2}\chi\|^2)^{1/2}$ implies the first result.

The proof of the second result is similar, we only give a sketch. The function $\Psi(t) := t^{(1+\beta)/(\beta-\mathfrak{s})}, t \geqslant 0$ which is convex for all $-1 \leqslant \mathfrak{s} < \beta$ and, by similar arguments, we have that

$$\Big(\frac{\mathbb{E}\|(K^\star K)^{-\mathfrak{s}/2}\chi\|^2}{\mathbb{E}\|(K^\star K)^{-\beta/2}\chi\|^2}\Big)^{\frac{1+\beta}{\beta-\mathfrak{s}}} = \frac{\mathbb{E}\|K\chi\|^2}{\mathbb{E}\|(K^\star K)^{-\beta/2}\chi\|^2}.$$

The same manipulations as above gives the second result □

Proof of Theorem 2.2. If we denote $\varphi_\alpha^+ = (\alpha I + T^\star T)^{-1}T^\star T\varphi^+$ the "regularized minimal norm solution", we use the decomposition $\hat\varphi_\alpha - \varphi^+ = E_1 + E_2 + E_3$ where

$$\begin{aligned} E_1 &= (\alpha I + \hat T^\star \hat T)^{-1}\left(\hat T^\star \hat r - \hat T^\star \hat T\varphi^+\right) , \\ E_2 &= \varphi_\alpha^+ - \varphi^+ , \\ E_3 &= (\alpha I + \hat T^\star \hat T)^{-1}\hat T^\star \hat T\varphi^+ - (\alpha I + T^\star T)^{-1}T^\star T\varphi^+ . \end{aligned}$$

The term $\mathbb{E}\|E_1\|^2$ is bounded using the inequality (A.2) of Lemma A.1:

$$\begin{aligned} \|E_1\|^2 &\lesssim \|(\alpha I + \hat T^\star \hat T)^{-1}\hat T^\star\|^2 \cdot \|\hat r - \hat T\varphi^+\|^2 \\ &\lesssim (\|\hat r - r\|^2 + \|\hat T - T\|^2)/\alpha. \end{aligned}$$

In order to bound the two other terms, we need to consider the two cases where $\beta > 1$ and $0 < \beta \leqslant 1$. In both cases we will show that

$$\mathbb{E}\|E_2 + E_3\|^2 \lesssim \alpha^{\min(\beta,2)} + \mathbb{E}\|\hat T - T\|^2/\alpha + (\mathbb{E}\|T - \hat T\|^2)^\beta. \tag{A.8}$$

Choosing α as in the statement of the theorem will lead to the result.

(i) Consider the case $\beta > 1$. The term E_2 is the bias term, that we bound as in the proof of Theorem 2.1:

$$\|E_2\|^2 \lesssim \begin{cases} \alpha^\beta & \text{if } \beta < 2 \\ \alpha^2 & \text{if } \beta \geqslant 2. \end{cases}$$

In order to control the last term, we first note that $\varphi_\alpha^+ - \varphi^+ = -\alpha(\alpha I + T^\star T)^{-1}\varphi^+$ and therefore

$$\begin{aligned} E_3 &= \alpha\left\{(\alpha I + \hat T^\star \hat T)^{-1} - (\alpha I + T^\star T)^{-1}\right\}\varphi^+ \\ &= (\alpha I + \hat T^\star \hat T)^{-1}\left(T^\star T - \hat T^\star \hat T\right)(\varphi_\alpha^+ - \varphi^+). \end{aligned}$$

Note that in the last factor the regularization bias E_2 appears. Using the preceeding bound on E_2 together with inequality (A.3), we get $\mathbb{E}\|E_3\|^2 \lesssim \frac{1}{\alpha^2} \cdot \mathbb{E}\|T^\star T - \hat T^\star \hat T\|^2 \cdot \alpha^{\min(\beta,2)}$. Note that for $\beta > 1$, $\alpha^{\min(\beta,2)}/\alpha = o(1)$ if $\alpha = o(1)$ for all $\beta > 1$. The assumption $\mathbb{E}\|\hat T - T\|^4 \lesssim (\mathbb{E}\|\hat T - T\|^2)^2$ implies that

$\mathbb{E}\|\hat{T}^\star\hat{T}-T^\star T\|^2\lesssim\mathbb{E}\|\hat{T}-T\|^2(1+\mathbb{E}\|\hat{T}-T\|^2)=O(\mathbb{E}\|\hat{T}-T\|^2)$ provided that $\mathbb{E}\|\hat{T}-T\|^2=o(1)$. The inequality (A.8) follows now by combination of the last bounds.

(ii) Let $0<\beta\leqslant 1$ and write $E_{23}=E_2+E_3=-\widehat{R}_\alpha\varphi^+$ where $\widehat{R}_\alpha:=[I-(\alpha I+\widehat{T}^\star\widehat{T})^{-1}\widehat{T}^\star\widehat{T}]$. The idea of the proof is to apply the first result of Lemma A.2, which implies

$$(\mathbb{E}\|E_{23}\|^2)^{1/\beta}\lesssim\mathbb{E}\|TE_{23}\|^2(\mathbb{E}\|(T^\star T)^{\beta/2}E_{23}\|^2)^{-1/2}\tag{A.9}$$

provided that

$$\mathbb{E}\|E_{23}\|^2\lesssim(\mathbb{E}\|(T^\star T)^{\beta/2}E_{23}\|^2)^{1/2}.\tag{A.10}$$

We first check (A.10) . Following inequality (A.6) of Lemma A.1 we note that $\|\widehat{R}_\alpha^{1/2}\|\leqslant 1$. Therefore we can write

$$\|E_{23}\|^2=\|\widehat{R}_\alpha\varphi^+\|^2\leqslant\|\widehat{R}_\alpha^{1/2}\varphi^+\|^2=\langle\widehat{R}_\alpha\varphi^+,\varphi^+\rangle=\langle(T^\star T)^{\beta/2}E_{23},(T^\star T)^{-\beta/2}\varphi^+\rangle.\tag{A.11}$$

By the source condition ($\varphi^+=(T^\star T)^{\beta/2}\psi$) and the Cauchy-Schwarz inequality, the last term is bounded by $\|(T^\star T)^{\beta/2}E_{23}\|$ and thus we get (A.10).

It remains to evaluate the right hand side of (A.9). From (A.11) together with inequality (A.2) of Lemma A.1 it follows

$$\|\widehat{T}E_{23}\|^2=\|\widehat{T}\widehat{R}_\alpha\varphi^+\|^2=\|(\widehat{T}^\star\widehat{T})^{1/2}\widehat{R}_\alpha\varphi^+\|^2\leqslant\alpha\|\widehat{R}_\alpha^{1/2}\varphi^+\|^2\lesssim\alpha\|(T^\star T)^{\beta/2}E_{23}\|$$

and hence $\mathbb{E}\|\widehat{T}E_2\|^2\lesssim\alpha(\mathbb{E}\|(T^\star T)^{\beta/2}E_2\|^2)^{1/2}$. Using (A.10) and again the Cauchy Schwarz inequality, we get

$$\mathbb{E}\|(T-\widehat{T})E_{23}\|^2\leqslant\mathbb{E}\|(T-\widehat{T})\|^2\|E_{23}\|^2\lesssim(\mathbb{E}\|T-\widehat{T}\|^4)^{1/2}(\mathbb{E}\|(T^\star T)^{\beta/2}E_{23}\|^2)^{1/2}.\tag{A.12}$$

Therefore

$$\begin{aligned}\mathbb{E}\|TE_{23}\|^2&\lesssim\mathbb{E}\|(T-\widehat{T})E_{23}\|^2+\mathbb{E}\|\widehat{T}E_{23}\|^2\\&\lesssim\Big\{(\mathbb{E}\|T-\widehat{T}\|^4)^{1/2}+\alpha\Big\}(\mathbb{E}\|(T^\star T)^{\beta/2}E_{23}\|^2)^{1/2}.\end{aligned}$$

that provides the necessary bound for (A.9) implying the expected inequality (A.8). □

Proof of Theorem 3.1. If we define the operator $T_\beta=T(T^\star T)^{\beta/2}$, we can rewrite the estimator as

$$\widehat{\varphi}_\alpha^\beta=(\alpha I+T_\beta^\star T_\beta)^{-1}(T^\star T)^{\beta/2}T_\beta^\star\widehat{r}=(T^\star T)^{\beta/2}(\alpha I+T_\beta^\star T_\beta)^{-1}T_\beta^\star\widehat{r}$$

and moreover we define

$$\varphi_\alpha^\beta=(T^\star T)^{\beta/2}(\alpha I+T_\beta^\star T_\beta)^{-1}T_\beta^\star T\varphi.$$

Then using the source condition $\varphi=(T^\star T)^{\beta/2}\psi$, we easily verify that $\widehat{\varphi}_\alpha^\beta$ and φ_α^β satisfy

$$\begin{aligned}&T(\widehat{\varphi}_\alpha^\beta-\varphi_\alpha^\beta)=T_\beta(\alpha I+T_\beta^\star T_\beta)^{-1}T_\beta^\star(\widehat{r}-r),\\&T(\varphi-\varphi_\alpha^\beta)=T_\beta[I-(\alpha I+T_\beta^\star T_\beta)^{-1}T_\beta^\star T_\beta]\psi,\\&(T^\star T)^{-\beta/2}(\widehat{\varphi}_\alpha^\beta-\varphi_\alpha^\beta)=(\alpha I+T_\beta^\star T_\beta)^{-1}T_\beta^\star(\widehat{r}-r),\\&(T^\star T)^{-\beta/2}(\varphi-\varphi_\alpha^\beta)=[I-(\alpha I+T_\beta^\star T_\beta)^{-1}T_\beta^\star T_\beta]\psi.\end{aligned}$$

Using Lemma A.1 together with the triangle inequality we obtain

$$\begin{aligned}&\mathbb{E}\|T(\widehat{\varphi}_\alpha^\beta-\varphi)\|^2\lesssim\mathbb{E}\|\widehat{r}-r\|^2+\alpha\|\psi\|^2=(\alpha^{-1}\mathbb{E}\|\widehat{r}-r\|^2+\|\psi\|^2)\alpha,\\&\mathbb{E}\|(T^\star T)^{-\beta/2}(\widehat{\varphi}_\alpha^\beta-\varphi)\|^2\lesssim\alpha^{-1}\mathbb{E}\|\widehat{r}-r\|^2+\|\psi\|^2.\end{aligned}$$

Using $\alpha = C\,\mathbb{E}\|\widehat{r}-r\|^2$ the last bounds simplify to $\mathbb{E}\|T(\widehat{\varphi}_\alpha^\beta-\varphi)\|^2 = O(\mathbb{E}\|\widehat{r}-r\|^2)$ and $\mathbb{E}\|(T^\star T)^{-\beta/2}(\widehat{\varphi}_\alpha^\beta-\varphi)\|^2 = O(1)$. With Lemma A.2 we finally obtain

$$\mathbb{E}\|(\widehat{\varphi}_\alpha^\beta - \varphi)\|^2 = O\left(\{\mathbb{E}\|T(\widehat{\varphi}_\alpha^\beta - \varphi)\|^2\}^{\frac{\beta}{\beta+1}}\right) = O\left(\{\mathbb{E}\|\widehat{r} - r\|^2\}^{\frac{\beta}{1+\beta}}\right)$$

which proves the Theorem. □

In order to prove the next results, we introduce the operator L that is defined by $Lf := \sum_{k=1}^\infty k^2\langle f, e_k\rangle e_k$, where $\{e_k\}$ is the orthonormal system of trigonometric functions (cf. (3.2)). The operator L is a densely defined, unbounded, self-adjoint, strictly positive operator in $L^2[0,1]$. We denote by $\mathcal{D}(L)$ the domain of L and $\mathcal{R}(L)$ denotes its range. The operator $L^{\mathfrak{s}/2}$ is defined by $L^{\mathfrak{s}/2}f := \sum_{k=1}^\infty k^{\mathfrak{s}}\langle f, e_k\rangle e_k$ and it generates the Sobolev spaces $H^{\mathfrak{s}}$, that is $H^{\mathfrak{s}} = \mathcal{D}(L^{\mathfrak{s}/2})$. Moreover, $H^{\mathfrak{s}}$ endowed with the inner product $\langle f, g\rangle_{\mathfrak{s}} := \langle L^{\mathfrak{s}/2}f, L^{\mathfrak{s}/2}g\rangle$ is a Hilbert space. The associated norm is given by $\|f\|_{\mathfrak{s}} = \|L^{\mathfrak{s}/2}f\|$ and satisfies $\|f\|_m = \pi^{-2m}\|f^{(m)}\|$ for every integer m.

Lemma A.3. *Let K be a linear operator between Hilbert space such that $\|h\|_{-\mathfrak{a}} \lesssim \|Kh\|$ for some $\mathfrak{a} > 0$ and for every function $h \in L^2[0,1]$. If χ is a random function such that $\mathbb{E}\|\chi\|_{\mathfrak{p}}^2 \lesssim 1$ with $\mathfrak{p} > 0$, then for all $-\mathfrak{a} \leqslant \mathfrak{s} \leqslant \mathfrak{p}$ the inequality $\mathbb{E}\|\chi\|_{\mathfrak{s}}^2 \lesssim (\mathbb{E}\|K\chi\|^2)^{(\mathfrak{p}-\mathfrak{s})/(\mathfrak{a}+\mathfrak{p})}$ holds true.*

Proof. Define the function $\Psi(t) := t^{(\mathfrak{a}+\mathfrak{p})/(\mathfrak{p}-\mathfrak{s})}, t \geqslant 0$ which is convex for all $-\mathfrak{a} \leqslant \mathfrak{s} < \mathfrak{p}$. Due to Jensen's inequality we obtain

$$\begin{aligned}\Big(\frac{\mathbb{E}\|\chi\|_{\mathfrak{s}}^2}{\mathbb{E}\|\chi\|_{\mathfrak{p}}^2}\Big)^{\frac{\mathfrak{a}+\mathfrak{p}}{\mathfrak{p}-\mathfrak{s}}} &= \Psi\Big(\frac{\sum_k k^{-2(\mathfrak{p}-\mathfrak{s})}k^{2\mathfrak{p}}\mathbb{E}\langle\chi, e_k\rangle^2}{\sum_k k^{2\mathfrak{p}}\mathbb{E}\langle\chi, e_i\rangle^2}\Big) \leqslant \frac{\sum_k \Psi(k^{-2(\mathfrak{p}-\mathfrak{s})})k^{2\mathfrak{p}}\mathbb{E}\langle\chi, e_k\rangle^2}{\sum_k k^{2\mathfrak{p}}\mathbb{E}\langle\chi, e_k\rangle^2}\\ &= \frac{\mathbb{E}\|\chi\|_{-\mathfrak{a}}^2}{\mathbb{E}\|\chi\|_{\mathfrak{p}}^2} \lesssim \frac{\mathbb{E}\|K\chi\|^2}{\mathbb{E}\|\chi\|_{\mathfrak{p}}^2}.\end{aligned}$$

We multiply this inequality by $\mathbb{E}\|\chi\|_{\mathfrak{p}}^2/\mathbb{E}\|\chi\|_{\mathfrak{s}}^2$ and obtain

$$\Big(\frac{\mathbb{E}\|\chi\|_{\mathfrak{s}}^2}{\mathbb{E}\|\chi\|_{\mathfrak{p}}^2}\Big)^{\frac{\mathfrak{a}+\mathfrak{s}}{\mathfrak{p}-\mathfrak{s}}} \leqslant \frac{\mathbb{E}\|K\chi\|^2}{\mathbb{E}\|\chi\|_{\mathfrak{s}}^2}.$$

From $\mathbb{E}\|\chi\|_{\mathfrak{p}}^2 \lesssim 1$ we have $\mathbb{E}\|\chi\|_{\mathfrak{s}}^2 \lesssim \mathbb{E}\|\chi\|_{\mathfrak{s}}^2/\mathbb{E}\|\chi\|_{\mathfrak{p}}^2$. Consequently, we obtain

$$(\mathbb{E}\|\chi\|_{\mathfrak{s}}^2)^{\frac{\mathfrak{a}+\mathfrak{p}}{\mathfrak{p}-\mathfrak{s}}} \lesssim \mathbb{E}\|\chi\|_{\mathfrak{s}}^2\Big(\frac{\mathbb{E}\|\chi\|_{\mathfrak{s}}^2}{\mathbb{E}\|\chi\|_{\mathfrak{p}}^2}\Big)^{\frac{\mathfrak{a}+\mathfrak{s}}{\mathfrak{p}-\mathfrak{s}}} \lesssim \mathbb{E}\|K\chi\|^2,$$

which proves the result. □

Proof of Theorem 3.2. If we define the operator $T_{\mathfrak{p}} = TL^{-\mathfrak{p}/2}$, we can rewrite the estimator as $\widehat{\varphi}_\alpha^{\mathfrak{p}} = L^{-\mathfrak{p}/2}(\alpha I + T_{\mathfrak{p}}^\star T_{\mathfrak{p}})^{-1}T_{\mathfrak{p}}^\star\widehat{r}$. Moreover, define $\varphi_\alpha^{\mathfrak{p}} := L^{-\mathfrak{p}/2}(\alpha I + T_{\mathfrak{p}}^\star T_{\mathfrak{p}})^{-1}T_{\mathfrak{p}}^\star T\varphi$. We easily check the following identities:

$$\begin{aligned}&T(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi_\alpha^{\mathfrak{p}}) = T_{\mathfrak{p}}(\alpha I + T_{\mathfrak{p}}^\star T_{\mathfrak{p}})^{-1}T_{\mathfrak{p}}^\star(\widehat{r} - r),\\ &T(\varphi - \varphi_\alpha^{\mathfrak{p}}) = T_{\mathfrak{p}}[I - (\alpha I + T_{\mathfrak{p}}^\star T_{\mathfrak{p}})^{-1}T_{\mathfrak{p}}^\star T_{\mathfrak{p}}]L^{\mathfrak{p}/2}\varphi,\\ &L^{\mathfrak{p}/2}(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi_\alpha^{\mathfrak{p}}) = (\alpha I + T_{\mathfrak{p}}^\star T_{\mathfrak{p}})^{-1}T_{\mathfrak{p}}^\star(\widehat{r} - r),\\ &L^{\mathfrak{p}/2}(\varphi - \varphi_\alpha^{\mathfrak{p}}) = [I - (\alpha I + T_{\mathfrak{p}}^\star T_{\mathfrak{p}})^{-1}T_{\mathfrak{p}}^\star T_{\mathfrak{p}}]L^{\mathfrak{p}/2}\varphi.\end{aligned}$$

Using Lemma A.1 together with the triangle inequality we obtain

$$\begin{aligned}&\mathbb{E}\|T(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2 \lesssim \mathbb{E}\|\widehat{r} - r\|^2 + \alpha\|\varphi\|_{\mathfrak{p}}^2 = (\alpha^{-1}\mathbb{E}\|\widehat{r} - r\|^2 + \|\varphi\|_{\mathfrak{p}}^2)\alpha,\\ &\mathbb{E}\|\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi\|_{\mathfrak{p}}^2 \lesssim \alpha^{-1}\mathbb{E}\|\widehat{r} - r\|^2 + \|\varphi\|_{\mathfrak{p}}^2.\end{aligned}$$

Using $\alpha = C\mathbb{E}\|\widehat{r}-r\|^2$ the last bounds simplify to $\mathbb{E}\|T(\widehat{\varphi}_\alpha^{\mathfrak{p}}-\varphi)\|^2 = O(\mathbb{E}\|\widehat{r}-r\|^2)$ and $\mathbb{E}\|\widehat{\varphi}_\alpha^{\mathfrak{p}}-\varphi\|_{\mathfrak{p}}^2 = O(1)$. With Lemma A.3 we finally obtain

$$\mathbb{E}\|\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi\|_{\mathfrak{s}}^2 = O\left(\{\mathbb{E}\|T(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2\}^{\frac{\mathfrak{p}-\mathfrak{s}}{\mathfrak{p}+\mathfrak{a}}}\right) = O\left(\{\mathbb{E}\|\widehat{r} - r\|^2\}^{\frac{\mathfrak{p}-\mathfrak{s}}{\mathfrak{a}+\mathfrak{p}}}\right)$$

which proves the Theorem. □

Proof of Theorem 3.3. Define the operators $T_{\mathfrak{p}} = TL^{-\mathfrak{p}/2}$ and $\widehat{T}_{\mathfrak{p}} = \widehat{T}L^{-\mathfrak{p}/2}$, and rewrite the estimator as $\widehat{\varphi}_\alpha^{\mathfrak{p}} = L^{-\mathfrak{p}/2}(\alpha I + \widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}})^{-1}\widehat{T}_{\mathfrak{p}}^\star \widehat{r}$. We also define $\varphi_\alpha^{\mathfrak{p}} := L^{-\mathfrak{p}/2}(\alpha I + \widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}})^{-1}\widehat{T}_{\mathfrak{p}}^\star \widehat{T}\varphi$. Therefore we directly check

$$\begin{aligned}
&T(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi_\alpha^{\mathfrak{p}}) = T_{\mathfrak{p}}(\alpha I + \widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}})^{-1}\widehat{T}_{\mathfrak{p}}^\star(\widehat{r} - \widehat{T}\varphi),\\
&T(\varphi - \varphi_\alpha^{\mathfrak{p}}) = T_{\mathfrak{p}}[I - (\alpha I + \widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}})^{-1}\widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}}]L^{\mathfrak{p}/2}\varphi,\\
&L^{\mathfrak{p}/2}(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi_\alpha^{\mathfrak{p}}) = (\alpha I + \widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}})^{-1}\widehat{T}_{\mathfrak{p}}^\star(\widehat{r} - \widehat{T}\varphi),\\
&L^{\mathfrak{p}/2}(\varphi - \varphi_\alpha^{\mathfrak{p}}) = [I - (\alpha I + \widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}})^{-1}\widehat{T}_{\mathfrak{p}}^\star \widehat{T}_{\mathfrak{p}}]L^{\mathfrak{p}/2}\varphi.
\end{aligned}$$

Using Lemma A.1 together with the triangle inequality we obtain

$$\mathbb{E}\|\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi\|_{\mathfrak{p}}^2 \lesssim \alpha^{-1}\mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^2 + \|\varphi\|_{\mathfrak{p}}^2.$$

and moreover we have

$$\mathbb{E}\|\widehat{T}(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2 \lesssim \mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^2 + \alpha\|\varphi\|_{\mathfrak{p}}^2$$

and

$$\begin{aligned}
\mathbb{E}\|(T - \widehat{T})(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2 &\lesssim (\mathbb{E}\|(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})^\star(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})\|^2)^{1/2}(\mathbb{E}\|\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi\|_{\mathfrak{p}}^4)^{1/2}\\
&\lesssim (\mathbb{E}\|(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})^\star(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})\|^2)^{1/2}\{\alpha^{-1}(\mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^4)^{1/2} + \|\varphi\|_{\mathfrak{p}}^2\}.
\end{aligned}$$

Combining all these bounds we obtain

$$\begin{aligned}
\mathbb{E}\|T(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2 &\lesssim \mathbb{E}\|(T - \widehat{T})(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2 + \mathbb{E}\|\widehat{T}(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2\\
&\lesssim \Big\{\alpha^{-1}(\mathbb{E}\|(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})^\star(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})\|^2)^{1/2}\{\alpha^{-1}(\mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^4)^{1/2} + \|\varphi\|_{\mathfrak{p}}^2\}\\
&\qquad + \alpha^{-1}\mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^2 + \|\varphi\|_{\mathfrak{p}}^2\Big\} \cdot \alpha
\end{aligned}$$

Using $\alpha = C\max((\mathbb{E}\|(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})^\star(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})\|^2)^{1/2}, (\mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^4)^{1/2}, \mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^2)$ the last bounds simplify to

$$\begin{aligned}
&\mathbb{E}\|T(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2 \lesssim \max((\mathbb{E}\|(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})^\star(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})\|^2)^{1/2}, (\mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^4)^{1/2}, \mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^2),\\
&\mathbb{E}\|\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi\|_{\mathfrak{p}}^2 \lesssim 1.
\end{aligned}$$

Applying now Lemma A.3 we obtain

$$\begin{aligned}
&\mathbb{E}\|\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi\|_{\mathfrak{s}}^2 \lesssim (\mathbb{E}\|T(\widehat{\varphi}_\alpha^{\mathfrak{p}} - \varphi)\|^2)^{(\mathfrak{p}-\mathfrak{s})/(\mathfrak{p}+\mathfrak{a})}\\
&\lesssim \max((\mathbb{E}\|(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})^\star(\widehat{T}_{\mathfrak{p}} - T_{\mathfrak{p}})\|^2)^{1/2}, (\mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^4)^{1/2}, \mathbb{E}\|\widehat{r} - \widehat{T}\varphi\|^2)^{(\mathfrak{p}-\mathfrak{s})/(\mathfrak{p}+\mathfrak{a})},
\end{aligned}$$

which proves the result under the moment conditions stated in the Theorem. □

Proof of Theorem 4.1. Using the definition of the operator L, it is easy to show that $T_{|\mathfrak{s}}^\star = L^{-\mathfrak{s}}T^\star$. We proceed now as in the proof of Theorem 2.2. If we denote $\varphi_\alpha^{\mathfrak{s}} = (\alpha I + T_{|\mathfrak{s}}^\star T_{|\mathfrak{s}})^{-1}T_{|\mathfrak{s}}^\star T_{|\mathfrak{s}}\varphi_{\mathfrak{s}}^+$ the "regularized minimal $H^{\mathfrak{s}}$ norm solution", we use the decomposition $\hat{\varphi}^{\mathfrak{s}}_\alpha - \varphi^+_{\mathfrak{s}} = E_1 + E_2 + E_3$ where

$$E_1 = (\alpha I + \hat{T}^\star_{|\mathfrak{s}}\hat{T}_{|\mathfrak{s}})^{-1}\left(\hat{T}^\star_{|\mathfrak{s}}\hat{r} - \hat{T}^\star_{|\mathfrak{s}}\hat{T}_{|\mathfrak{s}}\varphi^+_{\mathfrak{s}}\right) ,$$
$$E_2 = \varphi^{\mathfrak{s}}_\alpha - \varphi^+_{\mathfrak{s}} ,$$
$$E_3 = (\alpha I + \hat{T}^\star_{|\mathfrak{s}}\hat{T}_{|\mathfrak{s}})^{-1}\hat{T}^\star_{|\mathfrak{s}}\hat{T}_{|\mathfrak{s}}\varphi^+ - (\alpha I + T^\star_{|\mathfrak{s}}T_{|\mathfrak{s}})^{-1}T^\star_{|\mathfrak{s}}T_{|\mathfrak{s}}\varphi^+_{\mathfrak{s}} .$$

The term $\mathbb{E}\|E_1\|^2$ is bounded using the inequality (A.4) of Lemma A.1:

$$\begin{aligned}\mathbb{E}\|E_1\|^2_{\mathfrak{s}} &\lesssim \sup_{\substack{\phi\in L^2[0,1]^q\\ \|\phi\|\le 1}} \|(\alpha I + \hat{T}^\star_{|\mathfrak{s}}\hat{T}_{|\mathfrak{s}})^{-1}\hat{T}^\star_{|\mathfrak{s}}\phi\|^2_{\mathfrak{s}} \cdot \|\hat{r} - \hat{T}_{|\mathfrak{s}}\varphi^+_{\mathfrak{s}}\|^2 \\ &\lesssim (\|\hat{r} - r\|^2 + \|\hat{T} - T\|^2)/\alpha.\end{aligned}$$

We now show that

$$\mathbb{E}\|E_2 + E_3\|^2_{\mathfrak{s}} \lesssim \alpha^{\min(\beta,2)} + \mathbb{E}\|\hat{T} - T\|^2/\alpha + (\mathbb{E}\|T - \hat{T}\|^2)^\beta. \tag{A.13}$$

which will prove the result from the appropriate choice of α.

The proof of (A.13) follows the lines of the proof of Theorem 2.2. The L^2 norms are essentialy replaces by norms in $H^{\mathfrak{s}}$ and we only give below the missing arguments in order to conclude the proof.

(i) When $\beta > 1$, we similarly get the bounds $\|E_2\|^2_{\mathfrak{s}} \lesssim \alpha^{\min(\beta,2)}$ and

$$\|E_3\|^2_{\mathfrak{s}} \lesssim \frac{1}{\alpha^2} \cdot \sup_{\substack{\phi\in H^{\mathfrak{s}}\\ \|\phi\|_{\mathfrak{s}}\le 1}} \|(T^\star_{|\mathfrak{s}}T_{|\mathfrak{s}} - \hat{T}^\star_{|\mathfrak{s}}\hat{T}_{|\mathfrak{s}})\phi\|^2_{\mathfrak{s}} \cdot \alpha^{\min(\beta,2)}$$

Because $T^\star_{|\mathfrak{s}} = L^{-\mathfrak{s}}T^\star$, we have that $\sup_{\|\phi\|_{\mathfrak{s}}\le 1} \|(T^\star_{|\mathfrak{s}}T_{|\mathfrak{s}} - \hat{T}^\star_{|\mathfrak{s}}\hat{T}_{|\mathfrak{s}})\phi\|^2_{\mathfrak{s}} \lesssim \|T^\star T - \hat{T}^\star\hat{T}\|^2$ and we conclude as in the proof of Theorem 2.2.

(ii) When $0 < \beta \leqslant 1$, only the inequalities (A.12) have to be modified as follows:

$$\|(T - \widehat{T})E_{23}\|^2_{\mathfrak{s}} \leqslant \|E_{23}\|^2_{\mathfrak{s}} \cdot \sup_{\substack{\phi\in H^{\mathfrak{s}}\\ \|\phi\|_{\mathfrak{s}}\le 1}} \|(T_{|\mathfrak{s}} - \widehat{T}_{|\mathfrak{s}})\phi\|^2_s \lesssim \|E_{23}\|^2_{\mathfrak{s}} \cdot \|T - \widehat{T}\|^2$$

that implies

$$\mathbb{E}\|(T - \widehat{T})E_{23}\|^2_{\mathfrak{s}} \lesssim \left(\|(T^\star_{|\mathfrak{s}}T_{|\mathfrak{s}})^{\beta/2}E_{23}\|^2_{\mathfrak{s}}\right)^{1/2}\left(\mathbb{E}\|T - \widehat{T}\|^4\right)^{1/2}$$

and the conclusion follows. □

References

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-nonparametric IV estimation of shape-invariant Engel curves," *Econometrica*, 75, 1613–1669.

Blundell, R., and J. Horowitz (2007): "A Nonparametric Test of Exogeneity," *Rev. Econ. Stud.*, 74, 1035–1058.

Carrasco, M., J.-P. Florens, and E. Renault (2008): "Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6. Elsevier, to appear.

Chen, X. (2008): "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6. Elsevier, to appear.

Chen, X., and M. Reiss (2007): "On rate optimality for ill-posed inverse problems in econometrics," Working Paper.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quatile Treatment Effects," *Econometrica*, 73, 245–261.

Craven, P., and G. Wahba (1979): "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation.," *Numer.Math.*, 31, 377–403.

Darolles, S., J.-P. Florens, and E. Renault (2002): "Nonparametric Instrumental Regression," Working Paper # 228, IDEI, Université de Toulouse I.

Das, M. (2005): "Instrumental variables estimators of nonparametric models with discrete endogenous regressors," *Journal of Econometrics*, 124, 335–361.

Engl, H. W., M. Hanke, and A. Neubauer (2000): *Regularization of Inverse Problems.* Kluwer Academic, Dordrecht.

Florens, J.-P., J. Johannes, and S. Van Bellegem (2005): "Instrumental Regression in Partially Linear Models," Discussion Paper # 0537, Institut de statistique, Université catholique de Louvain.

Gagliardini, P., and O. Scaillet (2006): "Tikhonov Regularization for Functional Minimum Distance Estimators," Swiss Finance Institute Research Paper No. 06-30.

Hall, P., and J. L. Horowitz (2005): "Nonparametric methods for inference in the presence of instrumental variables," *Ann. Statist.*, 33, 2904–2929.

Hille, E., and J. D. Tamarkin (1931): "On the characteristic values of linear integral equations," *Acta Mathematica*, 57, 1–76.

Horowitz, J. L., and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75, 1191–1208.

Johannes, J., S. Van Bellegem, and A. Vanhems (2007): "Projection estimation in nonparametric instrumental regression," Working paper.

Kawata, T. (1972): *Fourier Analysis in Probability Theory.* Academic Press, New York.

Locker, J., and P. Prenter (1980): "Regularization with differential operators I: General theory," *J. Math. Anal. Appl.*, 74, 504–529.

Mair, B. A., and F. H. Ruymgaart (1996): "Statistical inverse estimation in Hilbert scales," *SIAM J. Appl. Math.*, 56(5), 1424–1444.

Matzkin, R. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71, 1339–1375.

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

Wahba, G. (1977): "Practical aproximate solutions to linear operator equations when the data are noisy," *SIAM J. Numer. Anal.*, 14, 651–667.

White, H., and J. Wooldridge (1991): "Some results on Sieve estimation with dependent observations," in *Nonparametric and Semi-Parametric Methods in Econometrics and Statistics*, ed. by W. Barnett, J. Powell, and G. Tauchen, New York. Cambridge University Press.