Statistics & Risk Modeling 29, 215–242 (2012) / **DOI** 10.1524/strm.2012.1044

Adaptive estimation for an inverse regression model with unknown operator

Jean-Michel Loubes, Clément Marteau

Received: November 11, 2009; Accepted: May 26, 2012

Summary: We are interested in the problem of estimating a regression function φ observed with a correlated noise $Y = \varphi(X) + U$. Contrary to the usual regression model, U is not centered conditionaly on X but rather on an observed variable W. Hence this model turns to be a difficult inverse problem where the corresponding operator is unknown since it is related to the joint distribution of (X, W). We focus on the case where the eigenvalues of the corresponding operator are observed with small perturbations and, using a well adapted spectral cut-off estimation procedure, we build a data driven estimates and derive an oracle inequality.

1 Introduction

The problem of estimating a regression function from noisy and pointwise observations is at the heart of modern statistical research. The observations are usually an i.i.d. sample $(Y_i, X_i)_{i=1}^n$, associated to the model

$$Y = \varphi(X) + U, \tag{1.1}$$

where φ is the unknown function to be estimated. The design $(X_i)_{i=1}^n$ can either be deterministic or random, according to the considered setting while the term U corresponds to some centered noise independent of X. We may refer for instance to [29] for an introduction to the model (1.1) where a penalized least square estimator is proved to reach the minimax rate of convergence over a wide range of functional spaces. Model selection approaches leading to an oracle inequality are tackled in [2]. We may also mention [1] for a study of the model (1.1) in a heteroscedastic setting. Kernel methods have been widely investigated, see for instance [5] and references therein while projections methods have been extensively developed over the past decades, see for instance [12] for a pioneer work.

In this article, contrary to previous statistical regression models, the error term is correlated with the explanatory variables X. In particular, $\mathbf{E}(U\,|\,X) \neq 0$ preventing a direct estimation of φ. To overcome the endogeneity of X, we assume that there exists an

AMS 2000 subject classification: Primary: 62G05, 62G20
Key words and phrases: Inverse problems, model selection, unknown operator

observed random variable W, which decorrelates the effects of the two variables X and Y in the sense that $\mathbf{E}(U|W) = 0$. Hence we aim at estimating a function φ from i.i.d. observations of (Y, X, W) satisfying the following condition

$$Y = \varphi(X) + U, \quad \begin{cases} \mathbf{E}(U|X) & \neq 0 \\ \mathbf{E}(U|W) & = 0 \end{cases} \tag{1.2}$$

The model (1.2) is often encountered when dealing with simultaneous equations, error-in-variable models, treatment model with endogenous effects. In econometrics, it defines the so-called instrumental variable regression model which has received a growing interest among the last decade. In particular, we refer to [26] for general references on the use of instrumental variables in economics.

We will see in Section 2 that the model (1.2) can be rewritten as an inverse problem using the expectation conditional operator with respect to W, as follows:

$$r := \mathbf{E}(Y|W) = \mathbf{E}(\varphi(X)|W) := T\varphi(W). \tag{1.3}$$

The function r is unknown and only an observation $\widehat{r}$ is available, leading to the inverse problem $\widehat{r} = T\varphi + \delta$, where φ is defined as the solution of a noisy Fredholm equation of the first order which may generate an ill-posed inverse problem. The literature on inverse problems in statistics is large, see for instance [14], [23], [7], [11] or [22] for general references. However, contrary to most of the problems tackled in this literature, the operator T is unknown since it depends on the joint distribution of X and W. Hence, the problem is turned into an inverse problem with unknown operator. Few results exist in this settings and only very recently new methods have arisen. In particular [8], [24, 25] or [13] and [18] in a more general case, construct estimators which enable to estimate inverse problem with unknown operators in an adaptive way, i.e. getting optimal rates of convergence without prior knowledge of the regularity of the functional parameter of interest.

In this work, we are facing an even more difficult situation since both r and the operator T have to be estimated from the same sample. Some papers tackle this topic, see for instance [15] for a complete introduction to the model, or [4], [17], [16], but all the proposed estimators rely on the prior knowledge of the regularity of the function φ. The objective of this work is to extend previous adaptive estimation procedures to the particular case where the operator is partially unknown. Our estimator is based on a spectral cut-off procedure. It requires the knowledge of the eigenvectors of $T^\star T$ but the eigenvalues are estimated from the observation sample. In this setting, we provide under some conditions, an oracle inequality to control the estimation error of the adaptive estimate, built in this paper. In particular, we prove that the risk of our estimator can be compared, up to a log term, to the risk of the best possible estimator (in a sense which will be precised later on).

The article falls into the following parts. Section 2 is devoted to the mathematical presentation of the instrumental variable framework and the building of the estimator. Section 3 provides the asymptotic behavior of this adaptive estimate as well as an oracle inequality, while technical Lemmas and proofs are gathered in Section 4.

2 Inverse problem formulation

2.1 A statistical framework

We observe an i.i.d. sample (Y_i, X_i, W_i) for $i = 1, \dots, n$ with unknown distribution $f(Y, X, W)$. Define the following Hilbert spaces

$$L^2_X = \{h : \mathbb{R} \to \mathbb{R},\ \|h\|^2_X := \mathbf{E}(h^2(X)) < +\infty\}$$
$$L^2_W = \{g : \mathbb{R} \to \mathbb{R},\ \|g\|^2_W := \mathbf{E}(g^2(W)) < +\infty\},$$

with the corresponding scalar product $\langle ., . \rangle_X$ and $\langle ., . \rangle_W$. For the sake of simplicity, we only consider in this paper the case where φ is univariate. Nevertheless the approach presented in this paper may be extended to the multivariate case (i.e. with a variable X of dimension $d > 1$).

Then the conditional expectation operator of X with respect to W is defined as an operator T

$$\begin{aligned} T : \quad & L^2_X \to L^2_W \\ & g \to \mathbf{E}[g(X)|W = .]\,. \end{aligned}$$

Following for instance [10], the model (1.2) can be written as

$$\begin{aligned} Y_i &= \varphi(X_i) + \mathbf{E}[\varphi(X_i)|W_i] - \mathbf{E}[\varphi(X_i)|W_i] + U_i \\ &= \mathbf{E}[\varphi(X_i)|W_i] + V_i \\ &= T\varphi(W_i) + V_i, \end{aligned} \tag{2.1}$$

where $V_i = \varphi(X_i) - \mathbf{E}[\varphi(X_i)|W_i] + U_i$, is such that $\mathbf{E}(V|W) = 0$. The parameter of interest is the unknown function φ. Hence, the observation model turns to be an inverse problem with unknown operator T and a correlated noise V. Solving this issue amounts to deal with the estimation of the operator and then controlling the correlation with respect to the noise.

The operator T is unknown since it depends on the unknown distribution $f_{(Y,X,Z)}$ of the observed variables. The estimation of this operator can be performed either by directly using an estimate of $f_{(Y,X,Z)}$, or if exists, by estimating the spectral value decomposition of the operator.

In the following, we assume that T is compact and admits a singular value decomposition (SVD) $(\lambda_j, \phi_j, \psi_j)_{j \geq 1}$. Such a decomposition provides a natural basis adapted to the operator for representing the function φ, see for instance [14]. More precisely, let T^* be the adjoint operator of T. Then T^*T is a compact operator on L^2_X with eigenvalues λ_j^2, $j \geq 1$ associated to the corresponding eigenfunctions ϕ_j, while ψ_j are defined by $\psi_j = \frac{T\phi_j}{\|T\phi_j\|}$. So we obtain

$$T\phi_j = \lambda_j \psi_j, \quad T^*\psi_j = \lambda_j \phi_j.$$

We can write the following decompositions

$$r(w) = \mathbf{E}(Y|W = w) = T\varphi(w) = \sum_{j \geq 1} \lambda_j \langle \varphi, \phi_j \rangle_X \psi_j(w), \tag{2.2}$$

$$\text{and} \quad r(w) = \sum_{j \geq 1} r_j \psi_j(w), \tag{2.3}$$

with $r_j = \langle Y, \psi_j \rangle_W$ that can be estimated by

$$\widehat{r}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_j(W_i).$$

Hence the noisy observations are the $\widehat{r}_j$'s and will be used to estimate the regression function φ in the inverse problem framework.

In a very general framework, full estimation of an operator is a hard task. Some attention has been paid to this estimation issue, with different kinds of technics such as kernel based Tikhonov regularization [15] or [17], regularization in Hilbert scales [16], finite dimensional sieve minimum distance estimator [26], with different rates and different smoothness assumptions, providing sometimes minimax rates of convergence. But, to our knowledge, most of the proposed estimators rely on prior knowledge on the regularity of the function φ expressed through an embedding condition into a smoothness space or an Hilbert scale, or a condition linking the regularity of φ to the regularity of the operator, namely a link condition or source condition (see [10] for general comments and insightful comments on such assumptions).

Yet, such general methods depend on the choice of a regularization parameter which has to be well chosen. In the following, to provide an automatic data driven choice, we restrict ourselves to the case where the SVD of the operator is partially known in the sense that the eigenvalues λ_j's are unknown but the eigenvectors ϕ_j's and ψ_j's are available. This assumption is restrictive for practical applications but, as discussed at the end of this section, some convolution issues can still be handled that way. In addition, this assumption is commonly encountered in the inverse problem literature. Actually, in many inverse problems, the regularization parameter depends on the ill-posedness of the problem. This index is generally expressed through the mere decay of the eigenvalues or through the decay of the eigenvalues compared to the decay of the coefficients of the function to be estimated (Source Condition assumption), which surely also requires some knowledge of the SVD decomposition of the operator.

2.2 A general estimation approach

If the operator were known we could provide an estimator using the spectral decomposition of the function φ as follows. For a given decomposition level m, define the projection estimator (also called spectral cut-off [14])

$$\widehat{\varphi}_m^0 = \sum_{j=1}^{m} \frac{\widehat{r}_j}{\lambda_j} \phi_j \tag{2.4}$$

Since the λ_j's are unknown, we first build an estimator of the eigenvalues. For this, using the decomposition (2.2), we obtain

$$\begin{aligned} \lambda_j &= \langle T\phi_j, \psi_j \rangle_W \\ &= \mathbf{E}[T\phi_j(W)\psi_j(W)] \\ &= \mathbf{E}[\mathbf{E}[\phi_j(X)|W]\psi_j(W)] \\ &= \mathbf{E}[\phi_j(X)\psi_j(W)]. \end{aligned}$$

So the eigenvalue λ_j can be estimated by

$$\widehat{\lambda}_j = \frac{1}{n}\sum_{i=1}^{n} \psi_j(W_i)\phi_j(X_i). \tag{2.5}$$

As studied in [8], replacing directly the eigenvalues by their estimates in (2.4) does not yield a consistent estimator, hence using their same strategy we define an upper bound for the resolution level

$$M = \inf\left\{k \leq N : |\widehat{\lambda}_k| \leq \frac{1}{\sqrt{n}}\log n\right\} - 1. \tag{2.6}$$

The parameter N provides an upper bound for M in order to ensure that M is not too large. Typically, N can be chosen of order n^ρ with $\rho > 1$. The main idea behind this definition is that when the estimates of the eigenvalues are too small with respect to the observation noise, trying to still provide an estimation of the inverse λ_k^{-1} only amplificates the estimation error. To avoid this trouble, we truncate the sequence of the estimated eigenvalues when their estimate is too small, i.e. smaller than the noise level. We point out that this parameter M is a random variable which we will have to control. More precisely, if we define two deterministic lower and upper bounds M_0, M_1 as

$$M_0 = \inf\left\{k : |\lambda_k| \leq \frac{1}{\sqrt{n}}\log^2 n\right\} - 1, \tag{2.7}$$

and

$$M_1 = \inf\left\{k : |\lambda_k| \leq \frac{1}{\sqrt{n}}\log^{3/4} n\right\}, \tag{2.8}$$

then, we will show in Section 4, that with high probability $M_0 \leq M < M_1$. Note that if in the definition (2.6) the set is empty, we set $M = 0$. However, from the remark above, this case happens with very small probability.

Now, thresholding the spectral decomposition in (2.4) leads to the following estimator

$$\widehat{\varphi}_m = \sum_{j=1}^{m} \frac{\widehat{r}_j}{\widehat{\lambda}_j} 1_{j\leq M}\phi_j. \tag{2.9}$$

The asymptotic behavior of this estimate depends on the choice of m. In the next section, we provide an optimal procedure to select the parameter m that gives rise to an adaptive estimator $\varphi^\star$ and an oracle inequality.

2.3 Examples

In this section, we present a brief discussion concerning the knowledge of the eigenvectors $(\phi_j)_{j\in\mathbb{N}}$ and $(\psi_j)_{j\in\mathbb{N}}$ of the unknown operator T. Assume that the link between X and the instrument W is of the form $X = \mathcal{L}(W, Z)$ with Z an independent random variable with distribution $\mathbf{P}_Z$. Then the operator has the following form

$$T\varphi(w) = \int \varphi \circ \mathcal{L}(w, Z) d\mathbf{P}_Z(Z) = \int \varphi(x) K_{\mathcal{L}}(x, w) dx$$

with a change of variable under some differentiability conditions on $\mathcal{L}$. Under technical assumptions, the operator defines a Fredholm integral operator with kernel $K_{\mathcal{L}}$ depending on the the link function $\mathcal{L}$ and the distribution of Z. Such operators are well studied in [20] or [27] for instance and, in many cases, the SVD decomposition will be available, which enables to use the estimation procedure developed in this paper. We point out that the knowledge of the operator implies that the distribution of X should also be known.

As a practical example, one may be interested in the following particular case. Assume that the function $\varphi : \mathbb{R} \to \mathbb{R}$ of interest is periodic with period 1 and that we observe

$$Y_i = \varphi(X_i) + U_i, \ \forall \in \{1, \dots, n\},$$

where the X_i are i.i.d. uniform random variables on $[0,1]$. In this particular case, $L^2(X) = L^2([0,1])$. We moreover assume that

$$W_i = X_i + Z_i, \ \forall i \in \{1, \dots, n\},$$

where the Z_i are i.i.d. random variables with unknown density $g : \mathbb{R} \to \mathbb{R}$.

In this case, for all $w \in \mathbb{R}$, we have

$$Tf(w) = \mathbf{E}(f(X)|W = w) = \mathbf{E}(f(w - Z)) = \int_{-\infty}^{+\infty} f(w-z)g(z)dz,$$

with adjoint

$$T^\star h(x) = \mathbf{E}(h(W)|X = x) = \int_{-\infty}^{+\infty} h(z+x)g(z)dz, \ \forall x \in \mathbb{R},$$

for all periodic functions f, h belonging respectively in $L^2(X)$ and $L^2(W)$. Hence, T is a convolution type operator. Let $(\phi_k)_{k\in\mathbb{Z}}$ be the usual complex trigonometric basis on $[0,1]$. Since X is uniform on $[0,1]$, $(\phi_z)_{z\in\mathbb{Z}}$ is an orthonormal basis of $L^2(X)$. With simple algebra, it is possible to prove that this sequence corresponds to the eigenvectors of $T^\star T$. The corresponding eigenvalues are related to the Fourier coefficients of the density g_Z. The eigenvalues are obviously unknown but may be easily estimated using the procedure presented above.

3 Main result

Consider the following assumptions on both the data Y_i, $i = 1, \dots, n$ and the eigenfunctions ϕ_k and ψ_k for $k \geq 1$.

Assumption 3.1 (Bounded SVD functions) *There exists a finite constant C_1 such that*

$$\forall j \geq 1, \quad \|\phi_j\|_\infty < C_1, \quad \|\psi_j\|_\infty < C_1. \tag{3.1}$$

Assumption 3.2 (Exponential moment conditions) *The observation Y satisfy to the following moment condition. There exists some positive numbers $v \geq \mathbf{E}(Y_j^2)$ and c such that*

$$\forall j \geq 1, \forall k \geq 2, \quad \mathbf{E}(Y_j^k) < \frac{k!}{2} v c^{k-2}. \tag{3.2}$$

These two conditions are required in order to obtain concentration bounds using first Hoeffding type inequality, then Bernstein inequality, see for instance [30]. Requiring bounded SVD functions may be seen as a restrictive condition. Yet it is met when the eigenvectors are trigonometric functions. However, this condition can be also be turned into a moment condition if we replace the concentration bound by a Bernstein type inequality. Note also that the moment conditions on Y amounts to require a bounded regression function φ and equivalent moment conditions on the errors U_j.

Assumption 3.3 (Degree of ill-posedness) *We assume that there exists t, called the degree of ill-posedness of the operator which controls the decay of the eigenvalues of the operator T. More precisely, there are constants λ_L, λ_U such that*

$$\lambda_L k^{-t} \leq \lambda_k \leq \lambda_U k^{-t}, \forall k \geq 1 \tag{3.3}$$

In this paper, we only consider the case of mildly ill-posed inverse problems, i.e. when the eigenvalues decay at a polynomial rate. This assumption, also required in [8], is needed when comparing the residual error of the estimator with the risk in order to obtain the oracle inequality.

Assumption 3.4 (Enough ill-posedness) *Let $\sigma_j^2 = \mathrm{Var}(Y\psi_j(W))$. We assume that there exist two positive constants σ_L^2 and σ_U^2 such that*

$$\forall\, j \geq 1, \quad \sigma_L^2 \leq \sigma_j^2 \leq \sigma_U^2. \tag{3.4}$$

Note that Condition (3.2) implies the upper bound of Condition (3.4). We also point out that this condition is not needed when building an estimator for the regression function. However it turns necessary when obtaining the lower bound to get a minimax result, or when obtaining an oracle inequality.

3.1 Oracle inequality

All the estimation errors will be given with respect to the L_X^2 norm which is a natural choice for this kind of problems. First, let $R_0(m,\varphi)$ be the quadratic estimation risk for the naive estimator $\widehat{\varphi}_m^0$ (2.4), defined for all $m \in \mathbb{N}$, by

$$\begin{aligned} R_0(m,\varphi) &= \mathbf{E}\|\widehat{\varphi}_m^0 - \varphi\|_X^2 \\ &= \sum_{k>m} \varphi_k^2 + \frac{1}{n}\sum_{k=1}^{m} \lambda_k^{-2}\sigma_k^2, \ \forall m \in \mathbb{N}, \end{aligned}$$

with $\varphi_k = \langle \varphi, \phi_k \rangle_X$. The best model would be obtained by choosing a minimizer of this quantity, namely

$$m_0 = \arg\min_m R_0(m,\varphi). \tag{3.5}$$

This risk depends on the unknown function φ hence m_0 is referred to as the oracle. We aim at constructing an estimator of $R_0(m,\varphi)$ which, by minimization, could give rise

to a convenient choice for m, i.e. as close as possible to m_0. The first step would be to replace φ_k by their estimates $\widehat{\lambda}_k^{-1}\widehat{r}_k$ and take for estimator of σ_k^2, $\widehat{\sigma}_k^2$, defined by

$$\widehat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\psi_k(W_i) - \frac{1}{n}\sum_{j=1}^{n}Y_j\psi_k(W_j)\right)^2$$
$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i\psi_k(W_i) - \widehat{r}_k)^2.$$

This would lead us to consider the empirical risk for any $m \leq M$, the cut-off which warrants a good behavior for the $\widehat{\lambda}_j$'s

$$U_0(m,r,\lambda) = -\sum_{k=1}^{m}\widehat{\lambda}_k^{-2}\widehat{r}_k^2 + \frac{c}{n}\sum_{k=1}^{m}\widehat{\lambda}_k^{-2}\widehat{\sigma}_k^2,\ \forall m \in \mathbb{N},$$

for a well chosen constant c. The corresponding random oracle within the range of models which are considered would be

$$m_1 = \arg\min_{m\leq M} R_0(m,\varphi). \tag{3.6}$$

Unfortunately, the correlation between the errors V_i and the observations Y_i prevents an estimator defined as a minimizer of $U_0(m,r,\lambda)$ to achieve the quadratic risk $R_0(m,\varphi)$. Indeed, we have to use a stronger penalty, leading to an extra error in the estimation that shall be discussed later in the paper. More precisely, c in the penalty is not a constant anymore but is allowed to depend on the number of observations n.

Hence, now define $R(m,\varphi)$ the penalized estimation risk as

$$R(m,\varphi) = \sum_{k>m}\varphi_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m}\lambda_k^{-2}\sigma_k^2,\ \forall m \in \mathbb{N}. \tag{3.7}$$

The best choice for m would be a minimizer of this quantity, which yet depends on the unknown regression function φ. Hence, to mimic this risk, define the following empirical criterion

$$U(m,r,\lambda) = -\sum_{k=1}^{m}\widehat{\lambda}_k^{-2}\widehat{r}_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m}\widehat{\lambda}_k^{-2}\widehat{\sigma}_k^2,\ \forall m \in \mathbb{N}. \tag{3.8}$$

Then, the best estimator is selected by minimizing this quantity as follows

$$m^\star := \arg\min_{m\leq M} U(m,r,\lambda), \tag{3.9}$$

Finally, the corresponding adaptive estimator $\varphi^\star$ is defined as:

$$\varphi^\star = \sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-1}\widehat{r}_k\phi_k. \tag{3.10}$$

The performances of $\varphi^\star$ are presented in the following theorem.

Theorem 3.5 *Let $\varphi^\star$ the projection estimator defined in (3.10). Then, under Assumptions (2.1) to (2.4), there exists B_0, B_1, B_2 and τ positive constants independent of n such that:*

$$\mathbb{E}\|\varphi^\star - \varphi\|_X^2 \;\le\; B_0 \log^2(n) \cdot \Big[\inf_m R(m,\varphi)\Big] + \frac{B_1}{n}\left(\log(n)\cdot\|\varphi\|_X^2\right)^{2t} + \Omega + \log^2(n)\cdot\Gamma(\varphi),$$

where $\Omega \le B_2(1+\|\varphi\|_X^2)\exp\left\{-\log^{1+\tau} n\right\}$, m_0 denotes the oracle bandwidth and

$$\Gamma(\varphi) = \begin{cases} \sum_{k=M_0}^{m_0}\left[\varphi_k^2 + \frac{1}{n}\lambda_k^{-2}\sigma_k^2\right], & \text{if } M_0 \le m_0, \\ 0, & \text{if } M_0 > m_0. \end{cases} \tag{3.11}$$

We obtain a non asymptotic inequality which guarantees a pertinent and adaptive choice for the bandwidth parameter m. In particular, the risk $R(m^\star,\varphi)$ of the corresponding estimator can be compared, up to a logarithmic factor, to the best possible risk $\inf_m R(m,\varphi)$ among all the projection estimators that could be constructed. We point out that we lose a $\log^2(n)$ factor when compared with the bound obtained in [8]. This loss comes partly from the fact that the error on the operator is not deterministic nor even due to a independent noisy observation of the eigenvalues. Here, the λ_k's have to be estimated using the available data by $\widehat{\lambda}_k$. In the econometric model, both the operator and the regression function are estimated on the same sample, which leads to high correlation effects that are made explicit in Model (2.1), hampering the rate of convergence of the corresponding estimator.

An oracle inequality only provides some information on the asymptotic behavior of the estimator if the remainder term $\Gamma(\varphi)$ is of smaller order than the risk of the oracle. This remainder term models the error made when truncating the eigenvalues, i.e. the error of selecting a model close to the random oracle $m_1 \le M$ and not close to the true oracle m_0. In the next section, we prove that, under some assumptions, this extra term is smaller than the risk of the estimator.

3.2 Rate of convergence

To get a rate of convergence for the estimator, we need to specify the regularity of the unknown function φ and compare it with the degree of ill-posedness of the operator T, following the usual conditions in the statistical literature on inverse problems, see for example [23] or [3] for some examples.

Assumption 3.6 (Regularity condition) *Assume that the function φ is such that there exists s and a constant C such that*

$$\varphi \in H_s(C) = \left\{ \nu = (\nu_k)_k,\ \text{s.t.}\ \sum_{k\ge 1} k^{2s}\nu_k^2 < C \right\}. \tag{3.12}$$

This assumption corresponds to functions whose regularity is governed by the smoothness index s. This parameter is unknown and yet governs the rate of convergence.

In the special cases where the eigenfunctions are the Fourier basis, this set corresponds to Sobolev classes. We provide in the following corollary a rate of convergence for our estimator.

Corollary 3.7 *Let $\varphi^\star$ be the model selection estimator defined in* (3.10)*. Then, we get the following rate of convergence*

$$\sup_{\varphi\in H_s(C)} \mathbb{E}\|\varphi^\star-\varphi\|_X^2 = O\left(\left(\frac{n}{\log^{2\gamma} n}\right)^{\frac{-2s}{2s+2t+1}}\right),$$

with $\gamma = 2+2s+2t$.

We point out that $\varphi^\star$ is constructed without prior knowledge of the unknown regularity s of φ. The rate of convergence that we obtain corresponds, up to some logarithmic terms, to the one given in [10]. Note that this rate corresponds to the minimax rate of convergence under the additional assumption that the **error term** V in the model (2.1) follows a Gaussian distribution and under the assumption (3.4) for the variance of this noise. This bound is the usual bound when estimating a function in an inverse model with known operator, see for instance in [9] for a review. In this sense, our estimator is said to be almost asymptotically adaptive. Following [10], we point out that Hall and Horowitz in [17] also obtain another minimax optimal rate of convergence in a similar settings but under different regularity assumptions. More recently a lower bound for the minimax rate of convergence in a closely related setting has been given in [19] under different weaker assumptions than in [10].

Remark 3.8 In an equivalent way, we could have imposed a supersmooth assumption, on the function φ, i.e. assuming that for given γ, t and constant C,

$$\sum_{k=1}^{\infty} \exp(2\gamma k^t)\varphi_k^2 < C.$$

Following the guidelines of the proof of Corollary 3.7 and Theorem 3.5, we obtain that $M_0 > m_0 \sim (a\ 2\gamma \log n)^{1/t}$ with $2a\gamma > 1$, leading to the optimal recovery rate for supersmooth functions in inverse problems.

3.3 Conclusion

In this work, we provide some new paths in order to build adaptive estimators for an inverse regression problem with unknown operator. We restrict ourselves to the framework where the eigenvectors are known and only the eigenvalues must be estimated. In this case, we prove that for smooth functions φ, estimating the eigenvalues and using a threshold enables to get a good estimator of the regression function and to build an adaptive procedure. The price to pay for not knowing the operator is only an extra $\log^2 n$ with respect to usual inverse problems and is mainly due to the correlation induced by the V_i's. We do not claim that we achieved optimality of the estimation procedure. Yet we provide a general way to get oracle inequalities for a class of estimators in this setting, which highlight the mathematical problems related to the adaptation in this instrumental variable problem.

4 Technical lemmas

First of all, we point out that, throughout all the paper, C denotes some generic constant that may vary from line to line.

Lemma 4.1 *Set $\mathcal{M} = \{M_0 \leq M < M_1\}$, where M, M_0, M_1 are respectively defined in (2.6), (2.7) and (2.8). Then, for all $n \geq 1$*

$$P(\mathcal{M}^c) \leq C M_0 e^{-\log^{1+\tau} n},$$

where C and τ denote positive constants independent of n.

Proof: It is easy to see that

$$P(\mathcal{M}^c) = P\left(\{M < M_0\} \cup \{M \geq M_1\}\right) \leq P(M < M_0) + P(M \geq M_1).$$

Using (2.6) and (2.8)

$$P(M \geq M_1) = P\left(\bigcap_{k=1}^{M_1} \left\{|\widehat{\lambda}_k| \geq \frac{1}{\sqrt{n}} \log n\right\}\right) \leq P\left(|\widehat{\lambda}_{M_1}| \geq \frac{1}{\sqrt{n}} \log n\right).$$

The definition of $\widehat{\lambda}_{M_1}$ yields

$$\begin{aligned} P(M \geq M_1) &\leq P\left(\left|\widehat{\lambda}_{M_1} - \lambda_{M_1} + \lambda_{M_1}\right| \geq \frac{1}{\sqrt{n}} \log n\right) \\ &\leq P\left(\left|\widehat{\lambda}_{M_1} - \lambda_{M_1}\right| \geq \frac{1}{\sqrt{n}} \log n - |\lambda_{M_1}|\right) \\ &\leq P\left(\left|\frac{1}{n} \sum_{i=1}^{n} \phi_{M_1}(X_i)\psi_{M_1}(W_i) - \mathbf{E}[\phi_{M_1}(X)\psi_{M_1}(W)]\right| \geq b_n\right), \end{aligned}$$

where $b_n = n^{-1/2} \log n - |\lambda_{M_1}|$ for all $n \in \mathbb{N}$. Let $k \in \mathbb{N}$ and $x \in [0,1]$ be fixed. Assumption (3.1) and Hoeffding inequality yield

$$\begin{aligned} P(|\widehat{\lambda}_k - \lambda_k| > x) &\leq 2\exp\left\{-\frac{(nx)^2}{2\sum_{i=1}^{n} \mathrm{Var}(\phi_{M_1}(X_i)\psi_{M_1}(W_i)) + 2nCx/3}\right\} \\ &= 2\exp\left\{-\frac{nx^2}{2\mathrm{Var}(\phi_{M_1}(X)\psi_{M_1}(W)) + 2Cx/3}\right\}. \end{aligned}$$

Using again the Assumption (3.1) on the bases $(\phi_k)_{k\in\mathbb{N}}$ and $(\psi_k)_{k\in\mathbb{N}}$

$$\mathrm{Var}(\phi_{M_1}(X)\psi_{M_1}(W)) \leq \mathbb{E}[\phi^2_{M_1}(X)\psi^2_{M_1}(W)] \leq C_1^4.$$

Hence

$$P(|\widehat{\lambda}_k - \lambda_k| > x) \leq 2\exp\left(-Cnx^2\right), \ \forall x \in [0,1], \tag{4.1}$$

for some constant C depending on C_1 but independent of n. Using (2.8), $1 > b_n > 0$ for all $n \in \mathbb{N}$. Therefore, using (4.1) with $x = b_n$, we obtain

$$P(M \geq M_1) \leq 2\exp\left\{-Cnb_n^2\right\} \leq 2\exp\left\{-C(\log n - \log^{3/4} n)^2\right\} \leq C\exp\left\{-\log^{1+\tau} n\right\},$$

where C and τ denote positive constants independent of n.

The bound of $P(M < M_0)$ follows the same lines

$$P(M < M_0) = P\left(\bigcup_{j=1}^{M_0}\left\{|\widehat{\lambda}_j| \leq \frac{\log n}{\sqrt{n}}\right\}\right) \leq \sum_{j=1}^{M_0} P\left(|\widehat{\lambda}_j| \leq \frac{\log n}{\sqrt{n}}\right) \leq \sum_{j=1}^{M_0} P\left(\widehat{\lambda}_j \leq \frac{\log n}{\sqrt{n}}\right).$$

Let $j \in \{1, \ldots, M_0\}$ be fixed.

$$P\left(\widehat{\lambda}_j \leq \frac{\log n}{\sqrt{n}}\right) = P\left(\widehat{\lambda}_j - \lambda_j \leq \tilde{b}_{n,j}\right),$$

where $\tilde{b}_{n,j} = n^{-1/2}\log n - \lambda_j$ for all $n \in \mathbb{N}$. Thanks to (2.7), $\tilde{b}_{n,j} < 0$ for all $n \in \mathbb{N}$. Using (4.1) with $x = -\tilde{b}_{n,j}$, we get

$$P\left(\widehat{\lambda}_j \leq \frac{\log n}{\sqrt{n}}\right) \leq \exp\left\{-Cn\tilde{b}_{n,j}^2\right\} \leq C\exp\left\{-\log^{1+\tau} n\right\},$$

for some $C, \tau > 0$. This concludes the proof of Lemma 4.1. □

Lemma 4.2 *Let $\mathcal{B}$ the event defined by*

$$\mathcal{B} = \bigcap_{k=1}^{M}\left\{|\lambda_k^{-1}\mu_k| \leq \frac{1}{2}\right\}, \text{ where } \mu_k = \widehat{\lambda}_k - \lambda_k, \ \forall k \in \mathbb{N}^*.$$

Then

$$P(\mathcal{B}^c) \leq CM_1 e^{-\log^{1+\tau} n},$$

for some $\tau > 0$ and positive constant C.

Proof: Using simple algebra and Lemma 4.1

$$P(\mathcal{B}^c) = P(\mathcal{B}^c \cap \mathcal{M}) + P(\mathcal{B}^c \cap \mathcal{M}^c) \leq P(\mathcal{B}^c \cap \mathcal{M}) + P(\mathcal{M}^c) \leq P(\mathcal{B}^c \cap \mathcal{M}) + CM_0 e^{-\log^{1+\tau} n}.$$

Then

$$P(\mathcal{B}^c \cap \mathcal{M}) = P\left(\bigcup_{k=1}^{M}\left\{|\lambda_k^{-1}\mu_k| > \frac{1}{2}\right\} \cap \mathcal{M}\right) \leq P\left(\bigcup_{k=1}^{M_1-1}\left\{|\lambda_k^{-1}\mu_k| \geq \frac{1}{2}\right\}\right).$$

Let $k \in \{1, \ldots, M_1 - 1\}$ be fixed. Remark that

$$P\left(|\lambda_k^{-1}\mu_k| \geq \frac{1}{2}\right) = P\left(|\mu_k| \geq \frac{|\lambda_k|}{2}\right) \leq P\left(|\widehat{\lambda}_k - \lambda_k| \geq \frac{1}{2\sqrt{n}}\log^{3/4} n\right).$$

Then, using (4.1) with $x = 2n^{-1/2}\log^{3/4} n$

$$P\left(|\widehat{\lambda}_k - \lambda_k| \geq \frac{1}{2\sqrt{n}}\log^{3/4} n\right) \leq C e^{-\log^{1+\tau} n}, \tag{4.2}$$

for some $\tau > 0$ and a positive constant C. This concludes the proof of Lemma 4.2. □

The following lemma provides some tools for the control of the ratio $\widehat{\lambda}_k^{-1}\lambda_k$ on the event $\mathcal{B}$.

Lemma 4.3 *For all $k \leq M$, we have*

$$\left(\frac{\lambda_k}{\widehat{\lambda}_k} - 1\right)^2 \mathbf{1}_{\mathcal{B}} \leq \frac{2}{3}\lambda_k^{-2}(\widehat{\lambda}_k - \lambda_k)^2 \mathbf{1}_{\mathcal{B}}.$$

Moreover, we have the following expansion

$$\left(\frac{\lambda_k}{\widehat{\lambda}_k}\right)^2 = 1 - 2\lambda_k^{-1}(\widehat{\lambda}_k - \lambda_k) + \lambda_k^{-2}(\widehat{\lambda}_k - \lambda_k)^2 v_k,$$

where v_k is uniformly bounded on the event $\mathcal{B}$.

Proof: Let $k \leq M$ be fixed. Then

$$\left(\frac{\lambda_k}{\widehat{\lambda}_k} - 1\right)^2 \mathbf{1}_{\mathcal{B}} = \left(\frac{\mu_k}{\widehat{\lambda}_k}\right)^2 \mathbf{1}_{\mathcal{B}} = \left(\frac{\mu_k}{\lambda_k + \mu_k}\right)^2 \mathbf{1}_{\mathcal{B}} \leq \frac{2}{3}\lambda_k^{-2}(\widehat{\lambda}_k - \lambda_k)^2 \mathbf{1}_{\mathcal{B}},$$

where the μ_k are defined in Lemma 4.2. The end of the proof is based on a Taylor expansion of the ratio $(\widehat{\lambda}_k^{-1}\lambda_k)^2 = (1 + \lambda_k^{-1}\mu_k)^{-2}$. The variable v_k depends on $\lambda_k^{-1}\mu_k$ and can be easily bounded on the event $\mathcal{B}$. Remark that a similar expansion holds for $\widehat{\lambda}_k^{-1}\lambda_k$. □

Lemma 4.4 *Let $\bar{m}$ a random variable measurable with respect to $(Y_i, X_i, W_i)_{i=1,\ldots,n}$ such that $\bar{m} \leq M$. Then, for all $K > 1$ and $\gamma > 0$*

$$(i)\ \mathbf{E}\left[\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}(\widehat r_k-r_k)^2\right]\le\frac{\log^K(n)}{n}\mathbf{E}\left[\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}\sigma_k^2\right]+CNe^{-\log^K n},$$

$$\begin{aligned}(ii)\ \mathbf{E}\left[\sum_{k=1}^{\bar m}\lambda_k^{-2}(\widehat r_k-r_k)r_k\right]\le{}&\gamma^{-1}\frac{\log^K(n)}{n}\mathbf{E}\left[\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}\sigma_k^2\right]\\&+C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\gamma^{-1}R(m_0,\varphi)\\&+\gamma\mathbf{E}\sum_{k>\bar m}\varphi_k^2,\end{aligned}$$

where $C>0$ is a positive constant independent of n, m_0 denotes the oracle bandwidth and N has been introduced in (2.6).

Proof: Let $Q>0$ a positive term which will be chosen later. With simple algebra

$$\begin{aligned}&\mathbf{E}\left[\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}(\widehat r_k-r_k)^2\right]\\&=\mathbf{E}\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}(\widehat r_k-r_k)^2\mathbf{1}_{\left\{(\widehat r_k-r_k)^2<\frac{Q\sigma_k^2}{n}\right\}}+\mathbf{E}\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}(\widehat r_k-r_k)^2\mathbf{1}_{\left\{(\widehat r_k-r_k)^2\ge\frac{Q\sigma_k^2}{n}\right\}}\\&\le\frac{Q}{n}\mathbf{E}\left[\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}\sigma_k^2\right]+\mathbf{E}\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}\left((\widehat r_k-r_k)^2-\frac{Q\sigma_k^2}{n}\right)\mathbf{1}_{\left\{(\widehat r_k-r_k)^2\ge\frac{Q\sigma_k^2}{n}\right\}}.\end{aligned}\tag{4.3}$$

In the sequel, we are interested in the behavior of the second term in the right hand side of (4.3). Since $\widehat\lambda_k^{-2}\le n\log^{-2}n$ for all $k\le M$ and $\bar m\le N$

$$\begin{aligned}&\mathbf{E}\sum_{k=1}^{\bar m}\widehat\lambda_k^{-2}\left((\widehat r_k-r_k)^2-\frac{Q\sigma_k^2}{n}\right)\mathbf{1}_{\left\{(\widehat r_k-r_k)^2\ge\frac{Q\sigma_k^2}{n}\right\}}\\&\le\frac{n}{\log^2 n}\sum_{k=1}^{N}\mathbf{E}\left((\widehat r_k-r_k)^2-\frac{Q\sigma_k^2}{n}\right)\mathbf{1}_{\left\{(\widehat r_k-r_k)^2\ge\frac{Q\sigma_k^2}{n}\right\}}.\end{aligned}\tag{4.4}$$

Let $k\in\{1,\dots,N\}$ be fixed. It follows from integration by part that

$$\mathbf{E}\left((\widehat r_k-r_k)^2-\frac{Q\sigma_k^2}{n}\right)\mathbf{1}_{\left\{(\widehat r_k-r_k)^2\ge\frac{Q\sigma_k^2}{n}\right\}}=\int_{\frac{Q\sigma_k^2}{n}}^{+\infty}P\left((\widehat r_k-r_k)^2>x\right)dx.$$

Then

$$P\left((\widehat r_k-r_k)^2\ge x\right)=P\left(|\widehat r_k-r_k|\ge\sqrt{x}\right).$$

Assumption (3.2) together with Bernstein inequality entails that

$$\begin{aligned} P\left(|\widehat{r}_k - r_k| \geq \sqrt{x}\right) &= P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Y_i\psi_k(W_i) - \mathbf{E}[Y_i\psi_k(W_i)])\right| \geq \sqrt{x}\right) \\ &\leq \exp\left\{-\frac{n^2 x}{2\sum_{i=1}^{n}\mathrm{Var}(Y_i\psi_k(W_i)) + Cn\sqrt{x}}\right\} \\ &= \exp\left\{-\frac{nx}{2\sigma_k^2 + C\sqrt{x}}\right\}, \end{aligned}$$

for some $C > 0$. Set $D = (2\sigma_k^2 C^{-1})^2$. We obtain

$$\begin{aligned} &\mathbf{E}\left((\widehat{r}_k - r_k)^2 - \frac{Q\sigma_k^2}{n}\right)\mathbf{1}_{\left\{(\widehat{r}_k - r_k)^2 \geq \frac{Q\sigma_k^2}{n}\right\}} \\ &\quad\leq \int_{\frac{Q\sigma_k^2}{n}}^{D} \exp\left\{-\frac{nx}{4\sigma_k^2}\right\}dx + \int_{D}^{+\infty} \exp\left\{-\frac{nx}{C\sqrt{x}}\right\}dx \\ &\quad\leq \left[-\frac{4\sigma_k^2}{n}e^{-\frac{nx}{4\sigma_k^2}}\right]_{Q\sigma_k^2/n}^{+\infty} + \int_{D}^{+\infty} \exp\left\{-Cn\sqrt{x}\right\}dx \\ &\quad\leq \frac{4\sigma_k^2}{n}e^{-Q/4} + e^{-Cn}. \end{aligned}$$

Hence, we have

$$\mathbf{E}\left((\widehat{r}_k - r_k)^2 - \frac{Q\sigma_k^2}{n}\right)\mathbf{1}_{\left\{(\widehat{r}_k - r_k)^2 \geq \frac{Q\sigma_k^2}{n}\right\}} \leq \frac{C\sigma_k^2}{n}e^{-Q/4} + e^{-Cn}, \tag{4.5}$$

for some $C > 0$. Using (4.4) and (4.5)

$$\mathbf{E}\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}(\widehat{r}_k - r_k)^2\mathbf{1}_{\left\{(\widehat{r}_k - r_k)^2 \geq \frac{Q\sigma_k^2}{n}\right\}} \leq \frac{CN}{\log^2 n}e^{-Q/4} + \frac{nNe^{-Cn}}{\log^2 n}.$$

From (4.3), we eventually obtain

$$\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}(\widehat{r}_k - r_k)^2\right] \leq \frac{Q}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}\sigma_k^2\right] + \frac{CN}{\log^2 n}e^{-Q/4} + \frac{nNe^{-Cn}}{\log^2 n}.$$

Choose $Q = \log^K(n)$ in order to conclude the proof of (i).

Now, consider the bound of (ii). Let m_0 be the oracle bandwidth defined in (3.5). With the convention $\sum_a^b = -\sum_b^a$ if $b < a$

$$\begin{aligned}
\mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)r_k &= \mathbf{E}\sum_{k=m_0}^{\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)r_k \\
&\leq \mathbf{E}\left|\sum_{k=m_0}^{\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)r_k\right| \\
&\leq \mathbf{E}\sum_{k=1}^{+\infty}\left|(\mathbf{1}_{\{k\leq\bar{m}\}} - \mathbf{1}_{\{k\leq m_0\}})\lambda_k^{-2}(\widehat{r}_k - r_k)r_k\right|. \qquad (4.6)
\end{aligned}$$

Indeed, $\mathbf{E}[\widehat{r}_k] = r_k$ for all $k \in \mathbb{N}$. Then remark that

$$\begin{aligned}
\left|\mathbf{1}_{\{k\leq\bar{m}\}} - \mathbf{1}_{\{k\leq m_0\}}\right| &= \left|(\mathbf{1}_{\{k\leq\bar{m}\}} + \mathbf{1}_{\{k\leq m_0\}})(\mathbf{1}_{\{k\leq\bar{m}\}} - \mathbf{1}_{\{k\leq m_0\}})\right| \\
&= (\mathbf{1}_{\{k\leq\bar{m}\}} + \mathbf{1}_{\{k\leq m_0\}})\left|\mathbf{1}_{\{k>\bar{m}\}} - \mathbf{1}_{\{k>m_0\}}\right| \\
&\leq \mathbf{1}_{\{k>\bar{m}\}}\mathbf{1}_{\{k\leq m_0\}} + \mathbf{1}_{\{k>m_0\}}\mathbf{1}_{\{k\leq\bar{m}\}}. \qquad (4.7)
\end{aligned}$$

Using the Cauchy–Schwarz inequality and using that for all a, b and $1 > \gamma > 0$, $2ab \leq \gamma a^2 + \gamma^{-1}b^2$

$$\begin{aligned}
&\mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)r_k \\
&\leq \left(\mathbf{E}\sum_{k>\bar{m}} \lambda_k^{-2}r_k^2\right)^{\frac{1}{2}}\left(\mathbf{E}\sum_{k\leq m_0} \lambda_k^{-2}(\widehat{r}_k - r_k)^2\right)^{\frac{1}{2}} \\
&\qquad + \left(\mathbf{E}\sum_{k>m_0} \lambda_k^{-2}r_k^2\right)^{\frac{1}{2}}\left(\mathbf{E}\sum_{k\leq\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)^2\right)^{\frac{1}{2}} \\
&\leq \gamma\left\{\mathbf{E}\sum_{k>\bar{m}} \varphi_k^2 + \sum_{k>m_0} \varphi_k^2\right\} + \gamma^{-1}\left\{\mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)^2 + \mathbf{E}\sum_{k=1}^{m_0} \lambda_k^{-2}(\widehat{r}_k - r_k)^2\right\}.
\end{aligned}$$

We eventually obtain

$$\mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)r_k \leq \gamma^{-1}R(m_0, \varphi) + \gamma\mathbf{E}\sum_{k>\bar{m}} \varphi_k^2 + \gamma^{-1}\left\{\mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}(\widehat{r}_k - r_k)^2\right\}.$$

We conclude the proof using a string of inequalities similar to (i). In particular, using Assumption (3.3), we obtain the bound $\lambda_k^{-2} \leq CN^{2t}$ for all $k \leq M$. □

Lemma 4.5 *Let $\bar{m}$ a random variable measurable with respect to $(Y_i, X_i, W_i)_{i=1,\dots,n}$ such that $\bar{m} \leq M$. Then, for all $\gamma \in (0,1)$*

$$\mathbf{E}\sum_{k=1}^{\bar{m}}(\widehat{\lambda}_k^{-2}-\lambda_k^{-2})r_k^2 \;\leq\; \frac{\gamma+\gamma^{-1}\log^{3/2}n}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\lambda_k^{-2}\sigma_k^2\right]+\frac{1}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|^2}{\gamma}\right)^{2t} + \log^2(n)\cdot R(m_0,\varphi)+\Omega\,.$$

Proof: The term in the left hand side can be written as

$$\mathbf{E}\sum_{k=1}^{\bar{m}}(\widehat{\lambda}_k^{-2}-\lambda_k^{-2})r_k^2 = \mathbf{E}\sum_{k=1}^{\bar{m}}\left(\frac{\lambda_k^2}{\widehat{\lambda}_k^2}-1\right)\lambda_k^{-2}r_k^2 = \mathbf{E}\sum_{k=1}^{\bar{m}}\left(\frac{\lambda_k^2}{\widehat{\lambda}_k^2}-1\right)\varphi_k^2.$$

Using Lemma 4.3, we obtain

$$\mathbf{E}\sum_{k=1}^{\bar{m}}(\widehat{\lambda}_k^{-2}-\lambda_k^{-2})r_k^2 = -2\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-1}\mu_k\right]+\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\mu_k^2\nu_k\right] = S_1+S_2,$$

where the μ_k are defined in Lemma 4.2 and ν_k denotes a variable uniformly bounded on $\mathcal{B}$. First consider the bound on S_2. Using (4.1) with $x = n^{-1/2}\log n$, we obtain

$$\begin{aligned} S_2 \;&=\; \mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\mu_k^2\nu_k\right] \leq C\,\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\mu_k^2\right]+\Omega \\ &\leq\; C\,\frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\right]+C\,\|\varphi\|_X^2 e^{-\log^{1+\tau}n}, \end{aligned} \tag{4.8}$$

where C, τ denote positive constants independent of n and Ω is defined in Theorem 3.5. Thanks to our assumptions on the sequence $(\lambda_k)_{k\in\mathbb{N}}$, for all $\gamma > 0$

$$\begin{aligned} S_2 \;&\leq\; \frac{\log^2 n}{n}\|\varphi\|_X^2\mathbf{E}\sup_{k\leq\bar{m}}\lambda_k^{-2}+C\,\|\varphi\|_X^2 e^{-\log^{1+\tau}n} \\ &\leq\; \frac{\gamma}{n}\sum_{k=1}^{\bar{m}}\lambda_k^{-2}\sigma_k^2+\frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|^2}{\gamma}\right)^{2t}+\Omega, \end{aligned} \tag{4.9}$$

where for the last inequality, we have used (3.3), (3.4) and the bound

$$\sup_{k\leq\bar{m}}\lambda_k^{-2} \leq \frac{1}{x}\sum_{k=1}^{\bar{m}}\lambda_k^{-2}+Cx^{2t},$$

with $x = \gamma^{-1}\log^2(n)\cdot\|\varphi\|_X^2$. More details on this bound can be found in [7].

Now, we are interested in the bound on S_1. Using (4.7) and a string of inequalities similar to (4.6), we obtain

$$\begin{aligned} S_1 &= \mathbf{E}\sum_{k=1}^{\bar{m}} \varphi_k^2 \lambda_k^{-2} \mu_k \\ &\leq \mathbf{E}\sum_{k=1}^{+\infty} \mathbf{1}_{\{k>\bar{m}\}} \mathbf{1}_{\{k\leq m_0\}} \varphi_k^2 |\lambda_k^{-1} \mu_k| + \mathbf{E}\sum_{k=1}^{+\infty} \mathbf{1}_{\{k>m_0\}} \mathbf{1}_{\{k\leq \bar{m}\}} \varphi_k^2 |\lambda_k^{-1} \mu_k| \\ &\leq \left(\mathbf{E}\sum_{k>\bar{m}} \varphi_k^2\right)^{\frac{1}{2}} \left(\mathbf{E}\sum_{k\leq m_0} \lambda_k^{-2}(\widehat{\lambda}_k - \lambda_k)^2\right)^{\frac{1}{2}} \\ &\qquad + \left(\mathbf{E}\sum_{k>m_0} \varphi_k^2\right)^{\frac{1}{2}} \left(\mathbf{E}\sum_{k\leq \bar{m}} \lambda_k^{-2}(\widehat{\lambda}_k - \lambda_k)^2\right)^{\frac{1}{2}}. \end{aligned}$$

Hence, for all $\gamma > 0$

$$\begin{aligned} S_1 &\leq \gamma \left\{ \mathbf{E}\sum_{k>\bar{m}} \varphi_k^2 + \sum_{k>m_0} \varphi_k^2 \right\} \\ &\qquad + \gamma^{-1} \left\{ \mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}(\widehat{\lambda}_k - \lambda_k)^2 + \mathbf{E}\sum_{k=1}^{m_0} \lambda_k^{-2}(\widehat{\lambda}_k - \lambda_k)^2 \right\}. \end{aligned}$$

Using (4.1) once again with $x = n^{-1/2}\log^{3/4} n$, we obtain for all $\gamma > 0$

$$\begin{aligned} S_1 &\leq \gamma \left\{ \mathbf{E}\sum_{k>\bar{m}} \varphi_k^2 + \sum_{k>m_0} \varphi_k^2 \right\} \\ &\qquad + \frac{\gamma^{-1}\log^{3/2} n}{n} \left\{ \mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}\sigma_k^2 + \sum_{k=1}^{m_0} \lambda_k^{-2}\sigma_k^2 \right\} + C e^{-\log^{1+\tau} n}. \end{aligned}$$

This concludes the proof of Lemma 4.5. □

Lemma 4.6 *Let $\bar{m}$ a random variable measurable with respect to $(Y_i, X_i, W_i)_{i=1,\ldots,n}$ such that $\bar{m} \leq M$. Then*

$$\begin{aligned} &\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \widehat{\lambda}_k^{-2}(\widehat{\sigma}_k^2 - \sigma_k^2)\right] \\ &\quad \leq C\frac{\log n}{n^{3/2}}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \widehat{\lambda}_k^{-2}\sigma_k^2\right] + \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \widehat{\lambda}_k^{-2}(r_k^2 - \widehat{r}_k^2)\right] + C e^{-\log^2 n}, \end{aligned}$$

for some $C > 0$ independent of n.

Proof: First remark that, for all $k \geq 1$

$$\begin{aligned}\widehat{\sigma}_k^2 - \sigma_k^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i\psi_k(W_i) - \widehat{r}_k)^2 - \sigma_k^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2\psi_k^2(W_i) + \widehat{r}_k^2 - 2\widehat{r}_k^2 - \left(\mathbf{E}[Y^2\psi_k^2(W)] - \mathbf{E}[Y\psi_k(W)]^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i^2\psi_k^2(W_i) - \mathbf{E}[Y^2\psi_k^2(W)]\right\} + (r_k^2 - \widehat{r}_k^2).\end{aligned}$$

Hence, we obtain

$$\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}(\widehat{\sigma}_k^2 - \sigma_k^2)\right] = \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}\rho_k\right] + \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}(r_k^2 - \widehat{r}_k^2)\right], \qquad (4.10)$$

where for all $k \in \mathbb{N}$

$$\rho_k = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i^2\psi_k^2(W_i) - \mathbf{E}[Y^2\psi_k(W)]\right\}.$$

We are interested in the first term in the right hand side of (4.10). Let $\delta > 0$ a positive constant which will be chosen later

$$\begin{aligned}\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}\rho_k\right] &= \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}\rho_k\mathbf{1}_{\{\rho_k\leq\delta\}}\right] + \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}\rho_k\mathbf{1}_{\{\rho_k>\delta\}}\right] \\ &\leq \frac{\delta}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}\right] + \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}(\rho_k - \delta)\mathbf{1}_{\{\rho_k>\delta\}}\right].\end{aligned}$$

Since $\bar{m} \leq M$, from integration by part

$$\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\widehat{\lambda}_k^{-2}(\rho_k - \delta)\mathbf{1}_{\{\rho_k>\delta\}}\right] \leq \frac{1}{\log^2 n}\sum_{k=1}^{N}\int_{\delta}^{+\infty} P(\rho_k \geq x)dx.$$

Let $k \in \mathbb{N}$ and $x \geq \delta$ be fixed. Using Bernstein inequality

$$\begin{aligned}P(\rho_k \geq x) &= P\left(\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i^2\psi_k^2(W_i) - \mathbf{E}[Y^2\psi_k(W)]\right\} \geq x\right) \\ &\leq \exp\left\{-\frac{n^2x^2}{2\sum_{i=1}^{n}\mathrm{Var}(Y_i^2\psi_k^2(W_i)) + Cxn/3}\right\} \\ &\leq \exp\left\{-\frac{nx^2}{2D_0 + D_1x}\right\},\end{aligned}$$

with the hypotheses (3.2) and (3.1) on Y and $(\psi_k)_k$. The constants D_0 and D_1 are positive and independent of n. Therefore, for all $k \leq N$

$$\begin{aligned}
&\int_{\delta}^{+\infty} P(\rho_k \geq x)dx \\
&\leq \int_{\delta}^{2D_0/D_1} \exp\{-Cnx^2\}dx + \int_{2D_0/D_1}^{+\infty} \exp\{-nx\}dx \\
&\leq \int_{\delta}^{+\infty} \exp\{-Cn\delta x\}dx + \frac{1}{n}e^{-Cn} \\
&\leq \frac{C}{n\delta}\exp\{-n\delta^2\} + n^{-1}e^{-Cn},
\end{aligned}$$

for some $C > 0$. Choosing $\delta = n^{-1/2}\log n$ and using Assumption (3.4), we obtain

$$\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\tilde{m}} \widehat{\lambda}_k^{-2}\rho_k\right] \leq C\frac{\log n}{n^{3/2}}\mathbf{E}\left[\sum_{k=1}^{\tilde{m}} \widehat{\lambda}_k^{-2}\sigma_k^2\right] + Ce^{-\log^2 n}.$$

We use (4.10) in order to conclude the proof. □

5 Proofs

Proof of Theorem 3.5: The proof of our main result can be decomposed into four steps. In a first time, we prove that the quadratic risk of $\varphi^\star$ is close, up to some residual terms, to $\mathbf{E}\bar{R}(m^\star, \varphi)$ where

$$\bar{R}(m,\varphi) = \sum_{k>m} \varphi_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m} \widehat{\lambda}_k^{-2}\sigma_k^2, \ \forall m \in \mathbb{N}. \tag{5.1}$$

This result is uniform in m and justifies our choice of $\bar{R}(m,\varphi)$ as a criterion for the bandwidth selection.

In a second time, we show that $\mathbf{E}\bar{R}(m^\star,\varphi)$ and $\mathbf{E}U(m^\star, r, \varphi)$ are in some sense comparable. Then, according to the definition of $m^\star$ in (3.9)

$$U(m^\star, r, \varphi) \leq U(m, r, \varphi), \forall m \leq M.$$

We will conclude the proof by proving that for all $m \leq M$, $\mathbf{E}U(m,r,\varphi) = \mathbf{E}\|\widehat{\varphi}_m - \varphi\|^2$, up to a log term and some residual terms.

First note that

$$\mathbf{E}\|\varphi^\star - \varphi\|_X^2 = \mathbf{E}\sum_{k=1}^{+\infty}(\varphi_k^\star - \varphi_k)^2 = \mathbf{E}\sum_{k>m^\star}\varphi_k^2 + \mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-1}\widehat{r}_k - \varphi_k)^2.$$

This is the usual bias-variance decomposition. Then

$$\begin{aligned}
\mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-1}\widehat{r}_k - \varphi_k)^2 &= \mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-1}\widehat{r}_k - \widehat{\lambda}_k^{-1}r_k + \widehat{\lambda}_k^{-1}r_k - \varphi_k)^2 \\
&\leq 2\mathbf{E}\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\widehat{r}_k - r_k)^2 + 2\mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-1}r_k - \varphi_k)^2 = T_1 + T_2.
\end{aligned}$$

Concerning T_2, we use the following approach. For all $\gamma > 0$, using Lemma 4.3 and the bounds (4.8) and (4.9)

$$\begin{aligned} T_2 &= \mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-1} r_k - \varphi_k)^2 = \mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\widehat{\lambda}_k} - 1\right)^2 \varphi_k^2 \\ &= \mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\widehat{\lambda}_k} - 1\right)^2 \varphi_k^2 \mathbf{1}_{\mathcal{B}} + \mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\widehat{\lambda}_k} - 1\right)^2 \varphi_k^2 \mathbf{1}_{\mathcal{B}^c} \\ &\le \frac{2}{3}\mathbf{E}\left[\sum_{k=1}^{m^\star}\lambda_k^{-2}\mu_k^2\varphi_k^2\right] + \Omega \\ &\le \frac{\gamma}{n}\mathbf{E}\sum_{k=1}^{m^\star}\lambda_k^{-2}\sigma_k^2 + C\left(\frac{\|\varphi\|_X^2 \log^2(n)}{\gamma}\right)^{2t} + \Omega\,. \end{aligned} \tag{5.2}$$

where $\mu_k = \widehat{\lambda}_k - \lambda_k$ for all $k \in \mathbb{N}$. The term T_1 is bounded using Lemma 4.4 with $\bar{m} = m^\star$ and $K = 2$. Hence, for all $\gamma > 0$

$$\mathbf{E}\|\varphi^\star - \varphi\|_X^2 \le (1+\gamma)\mathbf{E}\bar{R}(m^\star, \varphi) + \frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t} + \Omega\,, \tag{5.3}$$

where $\bar{R}(m^\star, \varphi)$ is introduced in (5.1). This concludes the first step of our proof.

Now, our aim is to write $\mathbf{E}\bar{R}(m^\star, \varphi)$ in terms of $\mathbf{E}U(m^\star, r, \varphi)$

$$\begin{aligned} &\mathbf{E}U(m^\star, r, \varphi) \\ &= \mathbf{E}\left[-\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}\widehat{r}_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}\widehat{\sigma}_k^2\right] \\ &= \mathbf{E}\left[-\sum_{k=1}^{m^\star}\lambda_k^{-2} r_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}\sigma_k^2\right] - \mathbf{E}\left[\sum_{k=1}^{m^\star}\{\widehat{\lambda}_k^{-2}\widehat{r}_k^2 - \lambda_k^{-2} r_k^2\}\right] \\ &\qquad - \frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\sigma_k^2 - \widehat{\sigma}_k^2)\right] \\ &= \mathbf{E}\left[\sum_{k>m^\star}\varphi_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}\sigma_k^2\right] - \|\varphi\|_X^2 - \mathbf{E}\left[\sum_{k=1}^{m^\star}\{\widehat{\lambda}_k^{-2}\widehat{r}_k^2 - \lambda_k^{-2} r_k^2\}\right] \\ &\qquad - \frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\sigma_k^2 - \widehat{\sigma}_k^2)\right]. \end{aligned}$$

Hence

$$\mathbf{E}\bar{R}(m^\star,\varphi) = \mathbf{E}U(m^\star,r,\varphi) + \|\varphi\|_X^2 + \mathbf{E}\left[\sum_{k=1}^{m^\star}\{\widehat{\lambda}_k^{-2}\widehat{r}_k^2 - \lambda_k^{-2}r_k^2\}\right] + \frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\sigma_k^2 - \widehat{\sigma}_k^2)\right]. \tag{5.4}$$

Remark that

$$\begin{aligned}
&\mathbf{E}\left[\sum_{k=1}^{m^\star}\{\widehat{\lambda}_k^{-2}\widehat{r}_k^2 - \lambda_k^{-2}r_k^2\}\right] \\
&= \mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\widehat{r}_k^2 - r_k^2)\right] + \mathbf{E}\left[\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-2} - \lambda_k^{-2})r_k^2\right] \\
&= \mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}\{(\widehat{r}_k - r_k)^2 + 2(\widehat{r}_k - r_k)r_k\}\right] + \mathbf{E}\left[\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-2} - \lambda_k^{-2})r_k^2\right].
\end{aligned}$$

Using simple algebra

$$\begin{aligned}
&\mathbf{E}\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\widehat{r}_k - r_k)r_k \\
&= \mathbf{E}\sum_{k=1}^{m^\star}\lambda_k^{-2}(\widehat{r}_k - r_k)r_k + \mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-2} - \lambda_k^{-2})(\widehat{r}_k - r_k)r_k \\
&= \mathbf{E}\sum_{k=1}^{m^\star}\lambda_k^{-2}(\widehat{r}_k - r_k)r_k + \mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-1} - \lambda_k^{-1})r_k(\widehat{\lambda}_k^{-1} + \lambda_k^{-1})(\widehat{r}_k - r_k) \\
&\le \mathbf{E}\sum_{k=1}^{m^\star}\lambda_k^{-2}(\widehat{r}_k - r_k)r_k + \mathbf{E}\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-1} - \lambda_k^{-1})^2 r_k^2 + C\mathbf{E}\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\widehat{r}_k - r_k)^2 + \Omega.
\end{aligned}$$

Hence

$$\mathbf{E}\left[\sum_{k=1}^{m^\star}\{\widehat{\lambda}_k^{-2}\widehat{r}_k^2 - \lambda_k^{-2}r_k^2\}\right] \le C\mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}(\widehat{r}_k - r_k)^2\right] + 2\mathbf{E}\left[\sum_{k=1}^{m^\star}\lambda_k^{-2}(\widehat{r}_k - r_k)r_k\right] + \mathbf{E}\left[\sum_{k=1}^{m^\star}(\widehat{\lambda}_k^{-2} - \lambda_k^{-2})r_k^2\right] + \mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\widehat{\lambda}_k} - 1\right)^2\varphi_k^2.$$

Using Lemmata 4.4, 4.5 and (5.2), we obtain, for all $1 > \gamma > 0$ and $K > 1$

$$\begin{aligned}
&\mathbf{E}\left[\sum_{k=1}^{m^\star}\{\widehat{\lambda}_k^{-2}\widehat{r}_k^2 - \lambda_k^{-2}r_k^2\}\right] \qquad (5.5)\\
&\leq \left(2\gamma^{-1}\log^K n + C\gamma^{-1}\log^{3/2} n + \gamma\right)\cdot\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}\sigma_k^2\right] + \gamma^{-1}R(m_0,\varphi)\\
&\quad + \gamma\mathbf{E}\left[\sum_{k>m^\star}\varphi_k^2\right] + \Omega + C\gamma^{-1}N^{2t+1}e^{-\log^K n} + \frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}.
\end{aligned}$$

Remark that this result can be obtained for all $\bar{m}$ measurable with respect to the sample $(X_i, Y_i, W_i)_{i=1,\ldots,n}$. Then, from (5.4) and Lemma 4.6

$$\begin{aligned}
&\mathbf{E}\bar{R}(m^\star,\varphi)\\
&\leq \mathbf{E}U(m^\star,r,\varphi) + \|\varphi\|_X^2\\
&\quad + \left(2\gamma^{-1}\log^K n + C\gamma^{-1}\log^{3/2} n + C\frac{\log^2 n}{n^{1/2}}\right)\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{m^\star}\widehat{\lambda}_k^{-2}\sigma_k^2\right]\\
&\quad + \gamma^{-1}R(m_0,\varphi) + \gamma\mathbf{E}\left[\sum_{k>m^\star}\varphi_k^2\right] + C\gamma^{-1}N^{2t+1}e^{-\log^K n} + \Omega\\
&\quad + \frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t},
\end{aligned}$$

which can be rewritten

$$\begin{aligned}
&(1-\rho(\gamma,K,n))\mathbf{E}\bar{R}(m^\star,\varphi)\\
&\quad\leq \mathbf{E}U(m^\star,r,\varphi) + \|\varphi\|^2 + \gamma^{-1}R(m_0,\varphi)\\
&\qquad + C\gamma^{-1}N^{2t+1}e^{-\log^K n} + \Omega + \frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}, \qquad (5.6)
\end{aligned}$$

with

$$\rho(\gamma,K,n) = 2\gamma^{-1}\log^{K-2} n + \frac{C}{n^{1/2}} + \log^{-1/2} n + \gamma.$$

The third step of our proof can be easily derived from the definition of $m^\star$ and leads to the following result

$$\begin{aligned}
&(1-\rho(\gamma,K,n))\mathbf{E}\bar{R}(m^\star,\varphi)\\
&\quad\leq \mathbf{E}U(m_1,r,\varphi) + \|\varphi\|^2 + \gamma^{-1}R(m_0,\varphi)\\
&\qquad + C\gamma^{-1}N^{2t+1}e^{-\log^K n} + \Omega + \frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}, \qquad (5.7)
\end{aligned}$$

where m_1, defined in (3.6), denotes the oracle in the family $\{1,\dots,M\}$. In order to conclude the proof, we have to compute $\mathbf{E}U(m_1,r,\varphi)+\|\varphi\|^2$. To begin with, remark that

$$\begin{aligned}
&\mathbf{E}U(m_1,r,\varphi)+\|\varphi\|^2\\
&\quad=\mathbf{E}\left[-\sum_{k=1}^{m_1}\widehat{\lambda}_k^{-2}\widehat{r}_k^2+\frac{\log^2 n}{n}\sum_{k=1}^{m_1}\widehat{\lambda}_k^{-2}\widehat{\sigma}_k^2\right]+\|\varphi\|^2\\
&\quad=\mathbf{E}\left[-\sum_{k=1}^{m_1}\lambda_k^2 r_k^2\right]+\|\varphi\|_X^2+\frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m_1}\widehat{\lambda}_k^{-2}\sigma_k^2\right]\\
&\qquad+\mathbf{E}\left[\sum_{k=1}^{m_1}(\lambda_k^{-2}r_k^2-\widehat{\lambda}_k^{-2}\widehat{r}_k^2)\right]+\frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m_1}(\widehat{\lambda}_k^{-2}\widehat{\sigma}_k^2-\widehat{\lambda}_k^{-2}\sigma_k^2)\right].
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbf{E}U(m_1,r,\varphi)+\|\varphi\|^2\\
&\quad=\mathbf{E}\left[\sum_{k>m_1}\varphi_k^2+\frac{\log^2 n}{n}\sum_{k=1}^{m_1}\widehat{\lambda}_k^{-2}\sigma_k^2\right]+\mathbf{E}\left[\sum_{k=1}^{m_1}(\lambda_k^{-2}r_k^2-\widehat{\lambda}_k^{-2}\widehat{r}_k^2)\right]\\
&\qquad\qquad+\frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m_1}(\widehat{\lambda}_k^{-2}\widehat{\sigma}_k^2-\widehat{\lambda}_k^{-2}\sigma_k^2)\right]\\
&\quad=\mathbf{E}\bar{R}(m_1,\varphi)+F_1+F_2.
\end{aligned}$$

The same bound as (5.5) occurs for F_1. By the same way, using Lemma 4.6

$$\begin{aligned}
F_2&=\frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m_1}(\widehat{\lambda}_k^{-2}\widehat{\sigma}_k^2-\lambda_k^{-2}\sigma_k^2)\right]\\
&\le C\frac{\log n}{n^{3/2}}\mathbf{E}\left[\sum_{k=1}^{m_1}\widehat{\lambda}_k^{-2}\sigma_k^2\right]+\frac{1}{n}\mathbf{E}\sum_{k=1}^{m_1}\widehat{\lambda}_k^{-2}(r_k^2-\widehat{r}_k^2)+Ce^{-\log^2 n}.
\end{aligned}$$

Therefore, for all $K\ge 1$

$$\begin{aligned}
&\mathbf{E}U(m_1,r,\varphi)+\|\varphi\|^2\\
&\quad\le\left(1+C\log^{K-2}n+\frac{C\log^{-1}n}{\sqrt{n}}\right)\mathbf{E}\bar{R}(m_1,\varphi)+R(m_0,\varphi)\\
&\qquad+C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}+\Omega. \qquad (5.8)
\end{aligned}$$

Using (5.7) and (5.8), we eventually obtain

$$\begin{aligned}
&(1-\rho(\gamma,K,N))\mathbf{E}\bar{R}(m^\star,\varphi)\\
&\le \left(1+\log^{K-2}n+\frac{C\log^{-1}n}{\sqrt{n}}\right)\mathbf{E}\bar{R}(m_1,\varphi)+C\gamma^{-1}\mathbf{E}R(m_0,\varphi)\\
&\quad +C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}+\Omega\\
&\le C\log^2(n)\cdot\mathbf{E}R(m_1,\varphi)+C\gamma^{-1}\mathbf{E}R(m_0,\varphi)\\
&\quad +C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}+\Omega\\
&\le C\log^2(n)\cdot R(m_0,\varphi)+\log^2(n)\cdot\Gamma(\varphi)+\frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}+\Omega,
\end{aligned}$$

for some positive constant C, where $\Gamma(\varphi)$ is introduced in Theorem 3.5. An appropriate choice of K and γ yields $\rho(\gamma,K,N)<1$, at least for n small enough. Hence, we get

$$\begin{aligned}
&\mathbf{E}\|\varphi^\star-\varphi\|^2\\
&\le C\log^2(n)\cdot R(m_1,\varphi)+\frac{C}{n}\left(\frac{\log^2(n)\cdot\|\varphi\|_X^2}{\gamma}\right)^{2t}+\Omega+\log^2(n)\cdot\Gamma(\varphi).
\end{aligned}$$

□

Proof of Corollary 3.7: We start by recalling the oracle inequality obtained for the estimator $\varphi^\star$.

$$\begin{aligned}
\mathbb{E}\|\varphi^\star-\varphi\|^2 \le\ & C_0\log^2(n)\cdot\left[\inf_m R(m,\varphi)\right]+\frac{C_1}{n}\left(\log(n)\cdot\|\varphi\|^2\right)^{2\beta}\\
&+\Omega+\log^2(n)\cdot\Gamma(\varphi).
\end{aligned}$$

We have to bound the risk under the regularity condition and the extra term $\log^2(n)\Gamma(\varphi)$. Recall that the risk is given by

$$R(m,\varphi)=\sum_{k>m}\varphi_k^2+\frac{\log^2 n}{n}\sum_{k=1}^{m}\lambda_k^{-2}\sigma_k^2.$$

Hence under (3.12), we obtain both upper bounds for two constants C_1 and C_2

$$\begin{aligned}
\sum_{k>m}\varphi_k^2 &\le m^{-2s}C_1,\\
\frac{\log^2 n}{n}\sum_{k=1}^{m}\lambda_k^{-2}\sigma_k^2 &\le C_2\frac{\log^2 n}{n}\sigma_U^2 m^{2t+1}.
\end{aligned}$$

An optimal choice is given by $m = [(n/\log n)^{\frac{1}{1+2s+2t}}]$, leading to the desired rate of convergence.

Now consider the remainder term $\Gamma(\varphi)$. Under Assumption 3.3, $M_0 \geq [n^{1/2s}/\log^2 n]$, but since $m_0 = [n^{\frac{1}{1+2s+2t}}]$ we get clearly that $m_0 \leq M_0$, which entails that $\Gamma(\varphi) = 0$. □

Acknowledgements. The authors would like to thank the two referees for their remarks that help to improve the presentation of the paper.

References

[1] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[2] Y. Baraud. Model selection for regression on a fixed design. *Probability Theory and Related Fields*, 117:467–493, 2000.

[3] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications, *SIAM J. Numerical Analysis*, 45:2610–2636, 2007.

[4] C. Breunig and J. Johannes, On rate optimal local estimation in nonparametric instrumental regression, *Arxiv:* 0902.2103, submitted, 2009.

[5] D. Bosq. *Nonparametric Statistics for Stochastic Processes. Estimation and Prediction*. Lecture Notes in Statistics 110, Springer-Verlag, New York, 1996.

[6] I. Castillo, and J.-M. Loubes. Estimation of the distribution of random shifts deformation. *Math. Methods Statist.*, 18:21–42, 2009.

[7] L. Cavalier, Y. Golubev, D. Picard and A. B. Tsybakov. Oracle inequalities for inverse problems. *Annals of Statistics*, 30:843–874, 2002.

[8] L. Cavalier and N. W. Hengartner. Adaptive estimation for inverse problems with noisy operators. *Inverse problems*, 21:1345–1361, 2005.

[9] L. Cavalier. Nonparametric statistical inverse problems. *Inverse Problems*, 24(3), Article ID 034004, 2008.

[10] X. Chen and M. Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Econom. Theory*, 27(3):497–521, 2011.

[11] A. Cohen, M. Hoffmann, and M. Reiss. Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.* 42(4):1479–1501, 2004.

[12] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81:425–455, 1994.

[13] S. Efromovich and V. Koltchinskii. On inverse problems with unknown operators. *IEEE Trans. Inform. Theory*, 47:2876–2894, 2001.

[14] H. W. Engl, M. Hank, and A. Neubauer, *Regularization of Inverse Problems*. Kluwer Academic Publishers Group, Dordrecht (1996).

[15] J.-P. Florens. *Inverse Problems and Structural Econometrics: the Example of Instrumental Variables*, volume 2 of *Advances in Economics and Econometrics: Theory and Applications*. Cambridge University Press, Cambridge, UK, 2003.

[16] J.-P. Florens, J. Johannes, and S. van Bellegem. Identification and estimation by penalization in nonparametric instrumental regression, *Econom. Theory*, 27:472–496, 2011.

[17] P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.*, 33(6):2904–2929, 2005.

[18] M. Hoffmann and M. Reiss. Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.*, 36:310–336, 2008.

[19] J. Johannes and M. Schwarz, Adaptive Gaussian inverse regression with partially unknown operator, *ArXiv 1204.1226*, 2012.

[20] R. Kress. *Linear Integral Equations*. Applied Mathematical Sciences 82, Springer-Verlag, New York, 1999.

[21] J. M. Loubes and C. Ludena. Penalized estimators for non linear inverse problems. *ESAIM Probab. Stat*, 14:173–191, 2010.

[22] J. M. Loubes and C. Ludena. Adaptive Complexity Regularization for inverse Problems. *Electronic Journal of Statistics*, 2:661–677, 2008.

[23] B. Mair and F. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56:1424–1444, 1996.

[24] C. Marteau. Regularization of inverse problems with unknown operator. *Math. Methods Statist.*, 15:415–443, 2006.

[25] C. Marteau. On the stability of the Risk Hull Method. *Journal of Statistical Planning and Inference*, 139:1821–1835, 2009.

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

[27] A. D. Polyanin and A. V. Manzhirov. *Handbook of Integral Equations*. Chapman & Hall/CRC, Boca Raton, FL, 2008.

[28] A. N. Tikhonov, A. S. Leonov, and A. S. Yagola. *Nonlinear Ill-Posed Problems*. Vol. 1, 2. Applied Mathematics and Mathematical Computation, 14, Translated from the Russian, Chapman & Hall, London, 1998.

[29] S. Van de Geer. Estimating a regression function. *Annals of Statistics*, 18:907–924, 1990.

[30] S. Van de Geer. *Applications of Empirical Process Theory*. Cambridge Series in Statistical and Probabilistic Mathematics 6, Cambridge University Press, Cambridge, 2000.

Jean-Michel Loubes
Institut de Mathématique de Toulouse
UMR5219, Université de Toulouse
31000 Toulouse
France
Jean-Michel.Loubes@math.univ-toulouse.fr

Clément Marteau
Institut de Mathématique de Toulouse
UMR5219, Université de Toulouse
31000 Toulouse
France
Clement.Marteau@math.univ-toulouse.fr