Oracle Inequality for Instrumental Variable Regression

Jean-Michel Loubes, Clément Marteau

Oracle Inequality for Instrumental Variable Regression

J-M. Loubes & C. Marteau

January 27, 2009

Abstract

We tackle the problem of estimating a regression function observed in an instrumental regression framework. This model is an inverse problem with unknown operator. We provide a spectral cut-off estimation procedure which enables to derive oracle inequalities which warrants that our estimate, built without any prior knowledge, behaves as well as, up to log term, if the best model were known.

Keywords: Inverse Problems, Instrumental Variables, Model Selection, Econometrics .
Subject Class. MSC-2000: 62G05, 62G20 .

Introduction

An economic relationship between a response variable Y and a vector of explanatory variables X is often represented by an equation

$$Y = \varphi(X) + U,$$

where φ is the parameter of interest which models the relationship while U is an error term. Contrary to usual statistical regression models, the error term is correlated with the explanatory variables X, hence $\mathbf{E}(U|X) \neq 0$, preventing direct estimation of φ. To overcome the endogeneity of X, we assume that there exists an observed random variable W, called the instrument, which decorrelates the effects of the two variables X and Y in the sense that $\mathbf{E}(U|W) = 0$. It is often the case in economics, where the practical construction of instrumental variables play an important part. For instance [CIN07] present practical situations where prices of goods and quantity in goods can be explained using an instrument. This situation is also encountered when dealing with simultaneous equations, error-in-variable models, treatment model with endogenous effects. It defines the so-called instrumental variable regression model which has received a growing interest among the last decade and turned to be a challenging issue in statistics. In particular, we refer to [HN91], [NP03] [Flo03] for general references on the use of instrumental variables in economics while [HH05], [DFR03] and [FJvB07] deal with the statistical estimation problem.

More precisely, we aim at estimating a function φ observed in the following observation model

$$Y = \varphi(X) + U, \quad \begin{cases} \mathbf{E}(U|X) & \neq 0 \\ \mathbf{E}(U|W) & = 0 \end{cases} \tag{1}$$

Hence, the model (1) can be rewritten as an inverse problem using the expectation conditional operator with respect to W, which will be denoted T, as follows :

$$r := \mathbf{E}(Y|W) = \mathbf{E}(\varphi(X)|W) = T\varphi. \tag{2}$$

The function r is not known and only an observation $\hat{r}$ is available, leading to the inverse problem $\hat{r} = T\varphi + \delta$, 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, but contrary to most of the problems tackled in the literature on inverse problems (see [EHN96], [MR96], [CGPT02], [CHR03], [LL08] and [O'S86] for general references), the operator T is unknown either, which transforms the model into an inverse problem with unknown operator. Few results exist in this settings and only very recently new methods have arised. In particular [CH05], [Mar06, Mar08], or [EK01] and [HR08] in a more general case, construct estimators which enable to estimate inverse problem with *unobserved* 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 attention has been paid to this estimation issue, with different kinds of technics such as kernel based Tikhonov regularization [DFR03] or [HH05], regularization in Hilbert scales [FJvB07], finite dimensional sieve minimum distance estimator [NP03], with different rates and different smoothness assumptions, providing sometimes minimax rates of convergence. But, to our knowledge, all 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 [CR08] for general comments and insightful comments on such assumptions).

Hence, in this paper, we provide under some conditions, an adaptive estimation procedure of the function φ which converges, without prior regularity assumption, at the optimal rate of convergence, up to a logarithmic term. Moreover, we derive an oracle inequality which ensures optimality among the different choices of estimators.

The article falls into the following parts. Section 1 is devoted to the mathematical presentation of the instrumental variable framework and the building of the estimator. Section 2 provides the asymptotic behaviour of this adaptive estimate as well as an oracle inequality, while technical Lemmas and proofs are gathered in Section 3.

1 Inverse Problem for IV regression

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

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

with the corresponding scalar product $< .,. >_X$ and $< .,. >_W$. Then the conditional expectation operator of X with respect to W is defined as an operator T

$$\begin{aligned} T: \quad & L_X^2 \to L_W^2 \\ & g \to \mathbf{E}(g(X)|W). \end{aligned}$$

The model (1) can be written, as discussed in [CR08], as

$$\begin{aligned} Y_i &= \varphi(X_i) + \mathbf{E}[\varphi(X_i)|W_i] - \mathbf{E}[\varphi(X_i)|W_i] + U_i \\ &= \mathbf{E}[\varphi(X_i)|W_i] + V_i \\ &= T\varphi(W_i) + V_i, \end{aligned} \tag{3}$$

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 with 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 and depends on the unknown distribution of the observed variables $f_{(Y,X,Z)}$. Estimation of an 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.

Assume that T is compact and admits a singular value decomposition (SVD) $(\lambda_j, \phi_j, \psi_j)_{j \geqslant 1}$, which provides a natural basis adapted to the operator for representing the function φ, see for instance [EHN96]. More precisely, let T^* be the adjoint operator of T, then T^*T is a compact operator on L_X^2 with eigenvalues λ_j^2, $j \geqslant 1$ 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

$$r(w) = \mathbf{E}(Y|W = w) = T\varphi(w) = \sum_{j \geqslant 1} \lambda_j < \varphi, \phi_j > \psi_j(w), \tag{4}$$

$$\text{and} \quad r(w) = \sum_{j \geqslant 1} r_j \psi_j(w), \tag{5}$$

with $r_j =< Y, \psi_j >$ that can be estimated by

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

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

In a very general framework, full estimation of an operator is a hard task hence 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 known.

Note that this assumption is often met for the special case of deconvolution. Consider

the case where the unknown function φ reduces to the identity. Hence model (1) reduces to the usual deconvolution model

$$Y = X + U.$$

Set f_U the unknown density of the noise U and assume that $f_U \in L^2(\mathbf{R})$ is a 1 periodic function. Let also T_U be the convolution operator defined by $T_U g = g \star f_U$. In this special case, the spectral decomposition of the operator T_U is known, given by the unitary Fourier transform and the usual real trigonometric basis on $[0, 1]$ are the eigenvectors .

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 [EHN96])

$$\hat{\varphi}_m^0 = \sum_{j=1}^{m} \frac{\hat{r}_j}{\lambda_j} \phi_j \tag{6}$$

Since the λ_j's are unknown, we first build an estimator of the eigenvalues. For this, using the decomposition (4), we obtain

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

So the eigenvalue λ_j can be estimated by

$$\hat{\lambda}_j = \frac{1}{n} \sum_{j=1}^{n} \psi_j(W_i)\phi_j(X_i). \tag{7}$$

As studied in [CH05], replacing directly the eigenvalues by their estimates in (6) does not yield a consistent estimator, hence using their same strategy we define an upper bound for the resolution level

$$M = \inf \left\{ k \leqslant N : |\hat{\lambda}_k| \leqslant \frac{1}{\sqrt{n}} \log n \right\} - 1, \tag{8}$$

for N to be chosen later. The parameter N provides an upper bound for M in order to ensure that M is not too large. 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, define two deterministic lower and upper bounds M_0, M_1 as

$$M_0 = \inf \left\{ k : |\lambda_k| \leqslant \frac{1}{\sqrt{n}} \log^2 n \right\} - 1, \tag{9}$$

and

$$M_1 = \inf \left\{ k : |\lambda_k| \leqslant \frac{1}{\sqrt{n}} \log^{3/4} n \right\}, \tag{10}$$

we will show in Section 3, that with high probability $M_0 \leqslant M < M_1$.

Now, thresholding the spectral decomposition in (6) leads to the following estimator

$$\hat{\varphi}_m = \sum_{j=1}^{m} \frac{\hat{r}_j}{\hat{\lambda}_j} 1_{j \leqslant M} \phi_j. \tag{11}$$

The asymptotic behaviour 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 Main result

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

Bounded SVD functions: There exists a finite constant C_1 such that

$$\forall j \geqslant 1, \quad \|\phi_j\|_\infty < C_1, \quad \|\psi_j\|_\infty < C_1 \tag{12}$$

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

$$\forall j \geqslant 1, \forall k \geqslant 2, \quad \mathbf{E}(Y_j^k) < \frac{k!}{2} v c^{k-2}. \tag{13}$$

These two conditions are required in order to obtain concentration bounds using first Hoeffding type inequality, then Bernstein inequality, see for instance [vdG00]. 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.

IP: 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

$$\lambda_L k^{-t} \leqslant \lambda_k \leqslant \lambda_U k^{-t}, \forall k \geqslant 1 \tag{14}$$

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 [CH05], is needed when comparing the residual error of the estimator with the risk in order to obtain the oracle inequality.

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

$$\forall j \geqslant 1, \quad \sigma_L^2 \leqslant \sigma_j^2 \leqslant \sigma_U^2. \tag{15}$$

Note that Condition (13) implies the upper bound of Condition (15). The lower bound is similar to the variance condition in Assumption 3.1 in [CR08]. We 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.

2.1 Oracle inequality

First, let $R_0(m,\varphi)$ be the quadratic estimation risk for the naive estimator $\hat{\varphi}_m^0$ (6), defined by

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

The best model would be obtained by choosing a minimizer of this quantity, namely

$$m_0 = \arg\min_m R_0(m,\varphi). \tag{16}$$

This risk depends on the unknown function φ hence m_0 is 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 $\hat{\lambda}_k^{-1}\hat{r}_k$ and take for estimator of σ_k^2, $\hat{\sigma}_k^2$, defined by

$$\begin{aligned} \hat{\sigma}_k^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\psi_k(W_i) - \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_k(W_i)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\psi_k(W_i) - \hat{r}_k\right)^2. \end{aligned}$$

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

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

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

$$m_1 = \arg\min_{m\leqslant M} R_0(m,\varphi). \tag{17}$$

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

$$R(m,\varphi) = \sum_{k>m}\varphi_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m}\lambda_k^{-2}\sigma_k^2, \ \forall m \in \mathbb{N}. \tag{18}$$

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

$$U(m,r,\lambda) = -\sum_{k=1}^{m}\hat{\lambda}_k^{-2}\hat{r}_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m}\hat{\lambda}_k^{-2}\hat{\sigma}_k^2, \ \forall m \in \mathbb{N}. \tag{19}$$

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

$$m^\star := \arg\min_{m \leqslant M} U(m, r, \lambda), \tag{20}$$

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

$$\varphi^\star = \sum_{k=1}^{m^\star} \hat{\lambda}_k^{-1} \hat{r}_k \phi_k. \tag{21}$$

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

Theorem 2.1. *Let $\varphi^\star$ the projection estimator defined in (21). Then, there exists B_0, B_1, B_2 and τ positive constants independent of n such that:*

$$\begin{aligned}\mathbb{E}\|\varphi^\star - \varphi\|^2 \;\leqslant\; & B_0 \log^2(n).\left[\inf_m R(m,\varphi)\right] + \frac{B_1}{n}\left(\log(n).\|\varphi\|^2\right)^{2t} \\ & + \Omega + \log^2(n).\Gamma(\varphi),\end{aligned}$$

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

$$\Gamma(\varphi) = \sum_{k=\min(M_0, m_0)}^{m_0} \left[\varphi_k^2 + \frac{1}{n}\lambda_k^{-2}\sigma_k^2\right], \tag{22}$$

with the convention $\sum_a^b = 0$ if $a = b$.

We obtain a non asymptotic inequality which guarantees that the estimator achieves the optimal bound, up to a logarithmic factor, among all the estimators that could be constructed. We point out that we loss a $\log^2(n)$ factor when compared with the bound obtained in [CH05]. The explanation of this loss comes 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 $\hat{\lambda}_k$. 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 (3), hampering the rate of convergence of the corresponding estimator.

An oracle inequality only provides some information on the asymptotic behaviour 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 when selecting a model close to the random oracle $m_1 \leqslant M$ and not 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.

2.2 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 [MR96] or [CT02], [BHMR07] for some examples.

Regularity Condition Assume that the function φ is such that there exists s and a constant C such that

$$\sum_{k \geqslant 1} k^{2s} \varphi_k^2 < C \tag{23}$$

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 prove that our estimator achieves the optimal rate of convergence without prior assumption on s.

Corollary 2.2. *Let $\varphi^\star$ be the model selection estimator defined in* (21)*. Then, under the Sobolev embedding assumption* (23)*, we get the following rate of convergence*

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

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

We point out that $\varphi^\star$ is constructed without prior knowledge of the unknown regularity s of φ, yet achieving the optimal rate of convergence, up to some logarithmic terms. In this sense, our estimator is said to be asymptotically adaptive.

Remark 2.3. 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 proof of Corollary 2.2, we obtain that $M_0 > m_0 \sim (a2\gamma \log n)^{1/t}$ with $2a\gamma > 1$, leading to the optimal recovery rate for supersmooth functions in inverse problems.

In conclusion, this work shows that provided the eigenvectors are known, for smooth functions φ, estimating the eigenvalues and using a threshold suffices to get a good estimator of the regression function in the instrumental variable framework. The price to pay for not knowing the operator is only an extra $\log^2 n$ with respect to usual inverse problems and is only due to the correlation induced by the V_i's. One could object that when dealing with unknown operators, the knowledge of the eigenvectors is a huge hint and some papers have considered the case of completely unknown operators, using functional approach, see for instance [DFR03], [FJvB07], but their estimate clearly rely on smoothness assumptions for the regression. Hence the two approaches are complementary since we provide more refined adaptive result with the sake of stronger assumptions. Nevertheless, using similar techniques to develop a fully adaptive estimation procedure would be the last step towards a full understanding of the IV regression model.

3 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.

Recall that we have introduced

$$M = \inf\left\{k \leqslant N : |\hat{\lambda}_k| \leqslant \frac{1}{\sqrt{n}} \log n\right\} - 1,$$

The term N provides a deterministic upper bound for M and ensures that M is not too large. Typically, choose $N = n^4$. The following lemma provides a control of the bandwidth M by M_0 and M_1 respectively defined in (9) and (10).

Lemma 3.1. *Set $\mathcal{M} = \{M_0 \leqslant M < M_1\}$. Then, for all $n \geqslant 1$,*

$$P(\mathcal{M}^c) \leqslant CM_0 e^{-\log^{1+\tau} n},$$

where C and τ denote positive constants independent of n.

PROOF. It is easy to see that:

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

Using (8) and (10),

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

Thanks to the definition of $\hat{\lambda}_{M_1}$:

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

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 (12) and Hoeffding inequality yields

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

Using again the assumption (12) on the bases $(\phi_k)_{k\in\mathbb{N}}$ and $(\psi_k)_{k\in\mathbb{N}}$,

$$\mathrm{Var}(\phi_{M_1}(X)\psi_{M_1}(W)) \leqslant C_1^4 \mathbb{E}[\phi_{M_1}^2(X)\psi_{M_1}^2(W)] \leqslant 1.$$

Hence,

$$P(|\hat{\lambda}_k - \lambda_k| > x) \leqslant 2\exp\left(-C\frac{x^2}{3}\right), \ \forall x \in [0,1], \tag{24}$$

with C depending on C_1.

Using (10), $1 > b_n > 0$ for all $n \in \mathbb{N}$. Therefore, using (24) with $x = b_n$, we obtain:

$$\begin{aligned} P(M \geqslant M_1) \leqslant 2\exp\left\{-\frac{nb_n^2}{3}\right\} &\leqslant 2\exp\left\{-\frac{1}{3}(\log n - \log^{3/4} n)^2\right\}, \\ &\leqslant C\exp\left\{-\log^{1+\tau} n\right\}, \end{aligned}$$

where C and τ denote positive constants independent of n.

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

$$\begin{aligned} P(M < M_0) = P\left(\bigcup_{j=1}^{M_0}\left\{|\hat{\lambda}_j| \leqslant \frac{\log n}{\sqrt{n}}\right\}\right) &\leqslant \sum_{j=1}^{M_0} P\left(|\hat{\lambda}_j| \leqslant \frac{\log n}{\sqrt{n}}\right), \\ &\leqslant \sum_{j=1}^{M_0} P\left(\hat{\lambda}_j \leqslant \frac{\log n}{\sqrt{n}}\right). \end{aligned}$$

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

$$\begin{aligned} P\left(\hat{\lambda}_j \leqslant \frac{\log n}{\sqrt{n}}\right) &= P\left(\hat{\lambda}_j - \lambda_j \leqslant \frac{\log n}{\sqrt{n}} - \lambda_j\right), \\ &= P\left(\frac{1}{n}\sum_{i=1}^{n}\{\phi_j(X_i)\psi_j(X_i) - \mathbb{E}[\phi_j(X_i)\psi_j(X_i)]\} \leqslant \tilde{b}_n\right), \end{aligned}$$

where $\tilde{b}_n = n^{-1/2}\log n - \lambda_j$ for all $n \in \mathbb{N}$. Thanks to (9), $\tilde{b}_n < 0$ for all $n \in \mathbb{N}$. Using Hoeffding inequality and Assumption (12) :

$$P\left(\hat{\lambda}_j \leqslant \frac{\log n}{\sqrt{n}}\right) \leqslant \exp\left\{-\frac{n\tilde{b}_n^2}{2 + 2/3|\tilde{b}_n|}\right\} \leqslant C\exp\left\{-\log^{1+\tau} n\right\},$$

for some $C, \tau > 0$. This concludes the proof of Lemma 3.1.

□

Lemma 3.2. *Let $\mathcal{B}$ the event defined by:*

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

Then,

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

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

PROOF. Using simple algebra and Lemma 3.1

$$\begin{aligned} P(\mathcal{B}^c) &= P(\mathcal{B}^c \cap \mathcal{M}) + P(\mathcal{B}^c \cap \mathcal{M}^c), \\ &\leqslant P(\mathcal{B}^c \cap \mathcal{M}) + P(\mathcal{M}^c), \\ &\leqslant P(\mathcal{B}^c \cap \mathcal{M}) + CM_0 e^{-\log^{1+\tau} n}. \end{aligned}$$

Then,

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

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

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

Then, using (24) with $x = 2n^{-1/2}\log^{3/4} n$:

$$P\left(|\hat{\lambda}_k - \lambda_k| \geqslant \frac{1}{2\sqrt{n}}\log^{3/4} n\right) \leqslant Ce^{-\log^{1+\tau} n}, \tag{25}$$

for some $\tau > 0$ and a positive constant C. This concludes the proof of Lemma 3.2.

□

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

Lemma 3.3. *For all $k \leqslant M$, we have:*

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

Moreover, we have the following expansion:

$$\frac{\hat{\lambda}_k^{-1}}{\lambda_k} = 1 - \lambda_k^{-1}(\hat{\lambda}_k - \lambda_k) + \lambda_k^{-2}(\hat{\lambda}_k - \lambda_k)^2 \nu_k,$$

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

PROOF. Let $k \leqslant M$ be fixed. Then

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

where the μ_k are defined in Lemma 3.2. The end of the proof is based on a Taylor expansion of the ratio $\hat{\lambda}_k^{-1}\lambda_k = (1 + \lambda_k^{-1}\mu_k)^{-1}$. The variable ν_k depends on $\lambda_k^{-1}\mu_k$ and can be easily bounded on the event $\mathcal{B}$.

□

Lemma 3.4. *Let $\bar{m}$ a random variable measurable with respect to $(Y_i, X_i, W_i)_{i=1\ldots n}$ such that $\bar{m} \leqslant M$. Then, for all $K > 1$ and $\gamma > 0$,*

$$(i)\ \mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(\hat{r}_k - r_k)^2\right] \leqslant \frac{\log^K(n)}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}\sigma_k^2\right] + CNne^{-\log^K n},$$

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

where $C > 0$ is a positive constant independent of n, m_0 denotes the oracle bandwidth and N has been introduced in (8).

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

$$\begin{aligned}&\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(\hat{r}_k - r_k)^2\right] \\ &= \mathbf{E}\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 < \frac{Q\sigma_k^2}{n}\right\}} + \mathbf{E}\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 \geqslant \frac{Q\sigma_k^2}{n}\right\}}, \\ &\leqslant \frac{Q}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}\sigma_k^2\right] + \mathbf{E}\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 \geqslant \frac{Q\sigma_k^2}{n}\right\}}. \qquad (26)\end{aligned}$$

In the sequel, we are interested in the behavior of the second term in the right hand side of (26). Since $\hat{\lambda}_k^{-2} \leqslant n \log^{-2} n$ for all $k \leqslant M$ and $\bar{m} \leqslant N$, we obtain:

$$\mathbf{E} \sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2} (\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 \geqslant \frac{Q\sigma_k^2}{n}\right\}} \leqslant \frac{n}{\log^2 n} \sum_{k=1}^{N} \mathbf{E}(\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 \geqslant \frac{Q\sigma_k^2}{n}\right\}}. \tag{27}$$

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

$$\mathbf{E}(\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 \geqslant \frac{Q\sigma_k^2}{n}\right\}} \leqslant \int_{\frac{Q\sigma_k^2}{n}}^{+\infty} P\left((\hat{r}_k - r_k)^2 > x\right) dx.$$

Then,

$$P\left((\hat{r}_k - r_k)^2 \geqslant x\right) = P\left(|\hat{r}_k - r_k| \geqslant \sqrt{x}\right).$$

Assumption (13) together with Bernstein inequality entails that:

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

Now remark that:

$$2\sigma_k^2 = C\sqrt{x} \Leftrightarrow x = D, \text{ with } D = \left(\frac{2\sigma_k^2}{C}\right)^2.$$

We obtain:

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

Hence, we have

$$\mathbf{E}(\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 \geqslant \frac{Q\sigma_k^2}{n}\right\}} \leqslant \frac{C\sigma_k^2}{n} e^{-Q/4} + e^{-Cn}, \tag{28}$$

for some $C > 0$. Using (28) and (27),

$$\mathbf{E} \sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2} (\hat{r}_k - r_k)^2 \mathbf{1}_{\left\{(\hat{r}_k - r_k)^2 \geqslant \frac{Q\sigma_k^2}{n}\right\}} \leqslant \frac{CNn}{\log^2 n} e^{-Q/4} + e^{-Cn}.$$

From (26), we eventually obtain:

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

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

Now, consider the bound of (ii). Let m_0 the oracle bandwidth defined in (16). With the convention $\sum_a^b = -\sum_b^a$ if $b < a$,

$$\begin{aligned}
\mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}(\hat{r}_k - r_k)r_k &= \mathbf{E}\sum_{k=m_0}^{\bar{m}} \lambda_k^{-2}(\hat{r}_k - r_k)r_k, \\
&\leqslant \mathbf{E}\left|\sum_{k=m_0}^{\bar{m}} \lambda_k^{-2}(\hat{r}_k - r_k)r_k\right|, \\
&\leqslant \mathbf{E}\sum_{k=1}^{+\infty}\left|(\mathbf{1}_{\{k\leqslant\bar{m}\}} - \mathbf{1}_{\{k\leqslant m_0\}})\lambda_k^{-2}(\hat{r}_k - r_k)r_k\right|. \qquad (29)
\end{aligned}$$

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

$$\begin{aligned}
\left|\mathbf{1}_{\{k\leqslant\bar{m}\}} - \mathbf{1}_{\{k\leqslant m_0\}}\right| &= \left|(\mathbf{1}_{\{k\leqslant\bar{m}\}} + \mathbf{1}_{\{k\leqslant m_0\}})(\mathbf{1}_{\{k\leqslant\bar{m}\}} - \mathbf{1}_{\{k\leqslant m_0\}})\right|, \\
&= (\mathbf{1}_{\{k\leqslant\bar{m}\}} + \mathbf{1}_{\{k\leqslant m_0\}})\left|\mathbf{1}_{\{k>\bar{m}\}} - \mathbf{1}_{\{k>m_0\}}\right|, \\
&\leqslant \mathbf{1}_{\{k>\bar{m}\}}\mathbf{1}_{\{k\leqslant m_0\}} + \mathbf{1}_{\{k>m_0\}}\mathbf{1}_{\{k\leqslant\bar{m}\}}. \qquad (30)
\end{aligned}$$

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

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

We eventually obtain:

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

We conclude the proof using a similar to (i) string of inequalities. In particular, using Assumption (14), we obtain the bound $\lambda_k^{-2} \leqslant CN^{2t}$ for all $k \leqslant M$.

□

Lemma 3.5. *Let $\bar{m}$ a random variable measurable with respect to $(Y_i, X_i, W_i)_{i=1\ldots n}$ such that $\bar{m} \leqslant M$. Then, for all $\gamma \in (0, 1)$,*

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

PROOF. The term in the left hand side can be rewritten as:

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

Using Lemma 3.3, we obtain:

$$\mathbf{E}\sum_{k=1}^{\bar{m}}(\hat{\lambda}_k^{-2}-\lambda_k^{-2})r_k^2=-\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-1}\mu_k\right]+\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\mu_k^2\nu_k\right]=W_1+W_2,$$

where the μ_k are defined in Lemma 3.2. First consider the bound of W_2. Using (24) with $x=n^{-1/2}\log n$, we obtain:

$$\begin{aligned}W_2 &= \mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\mu_k^2\nu_k\right]\leqslant C\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\mu_k^2\right]+\Omega,\\ &\leqslant C\frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\right]+C\|\varphi\|^2e^{-\log^{1+\tau}n},\end{aligned}\tag{31}$$

where C denotes a positive constant independent of n. Thanks to our assumptions on the sequence $(\lambda_k)_{k\in\mathbb{N}}$, for all $\gamma>0$

$$\begin{aligned}W_2 &\leqslant \frac{\log^2 n}{n}\|\varphi\|^2\mathbf{E}\sup_{k\leqslant\bar{m}}\lambda_k^{-2}+C\|\varphi\|^2e^{-\log^{1+\tau}n},\\ &\leqslant \frac{\gamma}{n}\sum_{k=1}^{\bar{m}}\lambda_k^{-2}\sigma_k^2+\frac{C}{n}\left(\frac{\log^2(n).\|\varphi\|^2}{\gamma}\right)^{2t}+\Omega,\end{aligned}\tag{32}$$

where for the last inequality, we have used (14) and the bound:

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

with $x=\gamma^{-1}\log^2(n).\|\varphi\|^2$. More details on this bound can be found in [CGPT02].

We are now interested in the bound of W_1. Using (30) and a similar to (29) string of inequalities, we obtain:

$$\begin{aligned}W_1 &= \mathbf{E}\sum_{k=1}^{\bar{m}}\varphi_k^2\lambda_k^{-2}\mu_k,\\ &\leqslant \mathbf{E}\sum_{k=1}^{+\infty}\mathbf{1}_{\{k>\bar{m}\}}\mathbf{1}_{\{k\leqslant m_0\}}\varphi_k^2|\lambda_k^{-1}\mu_k|+\mathbf{E}\sum_{k=1}^{+\infty}\mathbf{1}_{\{k>m_0\}}\mathbf{1}_{\{k\leqslant\bar{m}\}}\varphi_k^2|\lambda_k^{-1}\mu_k|,\\ &\leqslant \sqrt{\mathbf{E}\sum_{k>\bar{m}}\varphi_k^2}\sqrt{\mathbf{E}\sum_{k\leqslant m_0}\lambda_k^{-2}(\hat{\lambda}_k-\lambda_k)^2}+\sqrt{\mathbf{E}\sum_{k>m_0}\varphi_k^2}\sqrt{\mathbf{E}\sum_{k\leqslant\bar{m}}\lambda_k^{-2}(\hat{\lambda}_k-\lambda_k)^2}.\end{aligned}$$

Hence, for all $\gamma>0$,

$$W_1\leqslant\gamma\left\{\mathbf{E}\sum_{k>\bar{m}}\varphi_k^2+\sum_{k>m_0}\varphi_k^2\right\}+\gamma^{-1}\left\{\mathbf{E}\sum_{k=1}^{\bar{m}}\lambda_k^{-2}(\hat{\lambda}_k-\lambda_k)^2+\mathbf{E}\sum_{k=1}^{m_0}\lambda_k^{-2}(\hat{\lambda}_k-\lambda_k)^2\right\}.$$

Using (24) once again with $x = n^{-1/2}\log^{3/4} n$, we obtain for all $\gamma > 0$:

$$W_1 \leqslant \gamma \left\{ \mathbf{E}\sum_{k>\bar{m}} \varphi_k^2 + \sum_{k>m_0} \varphi_k^2 \right\} + \frac{\gamma^{-1}\log^2 n}{n} \left\{ \mathbf{E}\sum_{k=1}^{\bar{m}} \lambda_k^{-2}\sigma_k^2 + \sum_{k=1}^{m_0} \lambda_k^{-2}\sigma_k^2 \right\}.$$

This concludes the proof of Lemma 3.5.

□

Lemma 3.6. *Let $\bar{m}$ a random variable measurable with respect to $(Y_i, X_i, W_i)_{i=1\dots n}$ such that $\bar{m} \leqslant M$. Then,*

$$\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(\hat{\sigma}_k^2 - \sigma_k^2)\right] \leqslant C\frac{\log n}{n^{3/2}}.\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}\sigma_k^2\right] + \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(r_k^2 - \hat{r}_k^2)\right] + Ce^{-\log^2 n},$$

for some $C > 0$ independent of n.

PROOF. First remark that, for all $k \geqslant 1$,

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

Hence, we obtain

$$\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(\hat{\sigma}_k^2 - \sigma_k^2)\right] = \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}\rho_k\right] + \frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{\bar{m}} \hat{\lambda}_k^{-2}(r_k^2 - \hat{r}_k^2)\right], \tag{33}$$

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 (33). Let $\delta > 0$ a positive constant which will be chosen later:

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

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

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

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

$$\begin{aligned} P(\rho_k \geqslant x) &= P\left(\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i^2\psi_k^2(W_i) - \mathbf{E}[Y^2\psi_k(W)]\right\} \geqslant x\right), \\ &\leqslant \exp\left\{-\frac{n^2x^2}{2\sum_{i=1}^{n}\mathrm{Var}(Y_i^2\psi_k^2(W_i)) + Cxn/3}\right\}, \\ &\leqslant \exp\left\{-\frac{n^2x^2}{2nD_0 + D_1nx}\right\}, \\ &\leqslant \exp\left\{-\frac{nx^2}{2D_0 + D_1x}\right\}, \end{aligned}$$

with the hypotheses (13) and (12) on Y and $(\psi_k)_k$. The constants D_0 and D_1 are positive and independent of n. Therefore, for all $k \leqslant N$,

$$\begin{aligned} &\int_\delta^{+\infty} P(\rho_k \geqslant x)dx \\ &\leqslant \int_\delta^{2D_0/D_1} \exp\left\{-\frac{nx^2}{2D_0 + D_1x}\right\}dx + \int_{2D_0/D_1}^{+\infty} \exp\left\{-\frac{nx^2}{2D_0 + D_1x}\right\}dx, \\ &\leqslant \int_\delta^{2D_0/D_1} \exp\{-Cnx^2\}dx + \int_{2D_0/D_1}^{+\infty} \exp\{-nx\}dx, \\ &\leqslant \int_\delta^{+\infty} \exp\{-Cn\delta x\}dx + \frac{1}{n}e^{-Cn}, \\ &\leqslant \frac{C}{n\delta}\exp\{-n\delta^2\} + n^{-1}e^{-Cn}, \end{aligned}$$

for some $C > 0$. Choosing $\delta = n^{-1/2}\log n$ and using Assumption (15), we obtain:

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

We use (33) in order to conclude the proof.

□

4 Proofs

Proof of Theorem 1. The proof of our main result can be decomposed in 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

$$\bar{R}(m, \varphi) = \sum_{k>m}\varphi_k^2 + \frac{\log^2 n}{n}\sum_{k=1}^{m}\hat{\lambda}_k^{-2}\sigma_k^2, \ \forall m \in \mathbb{N}. \tag{34}$$

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 (20),

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

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

In order to begin the proof, remark that:

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

This is the usual bias-variance decomposition. Then

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

Concerning T_2, we use the following approach. For all $\gamma>0$, using Lemma 3.3 and the bounds (31) and (32):

$$\begin{aligned}T_2 &= \mathbf{E}\sum_{k=1}^{m^\star}(\hat{\lambda}_k^{-1}r_k-\varphi_k)^2=\mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\hat{\lambda}_k}-1\right)^2\varphi_k^2,\\ &= \mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\hat{\lambda}_k}-1\right)^2\varphi_k^2\mathbf{1}_{\mathcal{B}}+\mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\hat{\lambda}_k}-1\right)^2\varphi_k^2\mathbf{1}_{\mathcal{B}^c},\\ &\leqslant \frac{2}{3}\mathbf{E}\left[\sum_{k=1}^{m^\star}\lambda_k^{-2}\mu_k^2\varphi_k^2\right]+\Omega,\\ &\leqslant \frac{\gamma}{n}\mathbf{E}\sum_{k=1}^{m^\star}\lambda_k^{-2}\sigma_k^2+C\left(\frac{\|\varphi\|^2\log^2(n)}{\gamma}\right)^{2t}+\Omega.\end{aligned}\tag{35}$$

where $\mu_k=\hat{\lambda}_k-\lambda_k$ for all $k\in\mathbb{N}$. The term T_1 is bounded using Lemma 3.4 with $\bar{m}=m^\star$ and $K=2$. Hence, for all $\gamma>0$,

$$\mathbf{E}\|\varphi^\star-\varphi\|^2\leqslant(1+\gamma)\mathbf{E}\bar{R}(m^\star,\varphi)+\frac{C}{n}\left(\frac{\log^2(n).\|\varphi\|^2}{\gamma}\right)^{2t}+\Omega,\tag{36}$$

where $\bar{R}(m^\star,\varphi)$ is introduced in (34). This concludes the first step of our proof.

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

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

Hence,

$$\begin{aligned}
\mathbf{E}\bar{R}(m^\star,\varphi) &= \mathbf{E}U(m^\star,r,\varphi)+\|\varphi\|^2+\mathbf{E}\left[\sum_{k=1}^{m^\star}\{\hat{\lambda}_k^{-2}\hat{r}_k^2-\lambda_k^{-2}r_k^2\}\right]\\
&\qquad +\frac{\log^2 n}{n}\mathbf{E}\left[\sum_{k=1}^{m^\star}\hat{\lambda}_k^{-2}(\sigma_k^2-\hat{\sigma}_k^2)\right]. \qquad (37)
\end{aligned}$$

Remark that:

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

Using simple algebra:

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

Hence,

$$\begin{aligned}\mathbf{E}\left[\sum_{k=1}^{m^\star}\{\hat\lambda_k^{-2}\hat r_k^2-\lambda_k^{-2}r_k^2\}\right] &\leqslant C\mathbf{E}\left[\sum_{k=1}^{m^\star}\hat\lambda_k^{-2}(\hat r_k-r_k)^2\right]+2\mathbf{E}\left[\sum_{k=1}^{m^\star}\lambda_k^{-2}(\hat r_k-r_k)r_k\right]\\ &\quad+\mathbf{E}\left[\sum_{k=1}^{m^\star}(\hat\lambda_k^{-2}-\lambda_k^{-2})r_k^2\right]+\mathbf{E}\sum_{k=1}^{m^\star}\left(\frac{\lambda_k}{\hat\lambda_k}-1\right)^2\varphi_k^2.\end{aligned}$$

Using Lemmata 3.4, 3.5 and (35), we obtain, for all $1>\gamma>0$ and $K>1$:

$$\begin{aligned}&\mathbf{E}\left[\sum_{k=1}^{m^\star}\{\hat\lambda_k^{-2}\hat r_k^2-\lambda_k^{-2}r_k^2\}\right]\\ &\leqslant \left(2\gamma^{-1}\log^K n+C\gamma^{-1}\log^{3/2}n+\gamma\right).\frac{1}{n}\mathbf{E}\left[\sum_{k=1}^{m^\star}\hat\lambda_k^{-2}\sigma_k^2\right] \qquad (38)\\ &\quad+\gamma^{-1}R(m_0,\varphi)+\gamma\mathbf{E}\left[\sum_{k>m^\star}\varphi_k^2\right]+\Omega+C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\frac{C}{n}\left(\frac{\log^2(n).\|\varphi\|^2}{\gamma}\right)^{2t}.\end{aligned}$$

Remark that this result can be obtained for all $\bar m$ measurable with respect to the sample $(X_i,Y_i,W_i)_{i=1\dots n}$. Then, from (37) and Lemma 3.6,

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

which can be rewritten:

$$\begin{aligned}(1-\rho(\gamma,K,n))\mathbf{E}\bar R(m^\star,\varphi) &\leqslant \mathbf{E}U(m^\star,r,\varphi)+\|\varphi\|^2\\ &\quad+2\gamma^{-1}R(m_0,\varphi)+C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\Omega+\frac{C}{n}\left(\frac{\log^2(n).\|\varphi\|^2}{\gamma}\right)^{2t}, \qquad (39)\end{aligned}$$

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:

$$\begin{aligned}(1-\rho(\gamma,K,n))\mathbf{E}\bar R(m^\star,\varphi) &\leqslant \mathbf{E}U(m_1,r,\varphi)+\|\varphi\|^2+2\gamma^{-1}R(m_0,\varphi)\\ &\quad+C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\Omega+\frac{C}{n}\left(\frac{\log^2(n).\|\varphi\|^2}{\gamma}\right)^{2t}, \qquad (40)\end{aligned}$$

where m_1 is defined in (17) and denotes the oracle in the family $\{1,\dots,M\}$. In order to

conclude the proof, we have to compute $\mathbf{E}U(m_1,r,\varphi)+\|\varphi\|^2$. In a first time, remark that:

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

Hence,

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

The same bound as (38) occurs for F_1. By the same way, using Lemma 3.6:

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

Therefore, for all $K\geqslant 1$,

$$\begin{aligned}\mathbf{E}U(m_1,r,\varphi)+\|\varphi\|^2 &\leqslant \left(1+C\log^{K-2}n+\frac{C\log^{-1}n}{\sqrt{n}}\right)\mathbf{E}\bar{R}(m_1,\varphi)+R(m_0,\varphi)\\ &\quad+C\gamma^{-1}N^{2t+1}e^{-\log^K n}+\frac{C}{n}\left(\frac{\log^2(n).\|\varphi\|^2}{\gamma}\right)^{2t}+\Omega. \qquad (41)\end{aligned}$$

Using (40) and (41), we eventually obtain:

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

for some positive constant C, where $\Gamma(\varphi)$ is introduced in Theorem 1. With an appropriate choice of K, this leads to:

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

□

Proof of Corollay 2.2 We start by recalling the oracle inequality obtained for the estimator $\varphi^\star$.

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

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 (23), we obtain both upper bounds for two constants C_1 and C_2

$$\sum_{k>m}\varphi_k^2\leqslant m^{-2s}C_1,$$

$$\frac{\log^2 n}{n}\sum_{k=1}^{m}\lambda_k^{-2}\sigma_k^2\leqslant C_2\frac{\log^2 n}{n}\sigma_U^2 m^{2t+1}.$$

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 [IP], $M_0\geqslant[n^{1/2s}/\log^2 n]$, but since $m_0=[n^{\frac{1}{1+2s+2t}}]$ we get clearly that $m_0\leqslant M_0$, which entails that $\Gamma(\varphi)=0$.

References

[BHMR07] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636 (electronic), 2007.

[CGPT02] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874, 2002. Dedicated to the memory of Lucien Le Cam.

[CH05] L. Cavalier and N. W. Hengartner. Adaptive estimation for inverse problems with noisy operators. *Inverse Problems*, 21(4):1345–1361, 2005.

[CHR03] A. Cohen, M. Hoffmann, and M. Reiss. Adaptive wavelet galerkin methods for linear inverse problems. *SIAM*, 1(3):323–354, 2003.

[CIN07] V. Chernozhukov, G. W. Imbens, and W. K. Newey. Instrumental variable estimation of nonseparable models. *J. Econometrics*, 139(1):4–14, 2007.

[CR08] X. Chen and M. Reiss. On rate optimality for ill-posed inverse problems in econometrics. *Cowles Foundation Discussion Paper No. 1626*, 2008.

[CT02] L. Cavalier and A. Tsybakov. Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*, 123(3):323–354, 2002.

[DFR03] S. Darolles, J-P. Florens, and E. Renault. Nonparametric instrumental regression. *preprint*, 2003.

[EHN96] H. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[EK01] S. Efromovich and V. Koltchinskii. On inverse problems with unknown operators. *IEEE Trans. Inform. Theory*, 47(7):2876–2894, 2001.

[FJvB07] J-P. Florens, J. Johannes, and S. van Bellegem. Indentification and estimation by penalization in nonparametric instrumental regression, 2007.

[Flo03] 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.

[HH05] P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.*, 33(6):2904–2929, 2005.

[HN91] J. Hausman and W. Newey. Nonparametric estimation of exact consumers surplus and deadweight loss. *Econometrica*, 63:1445–1476, 1991.

[HR08] M. Hoffmann and M. Reiss. Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.*, 36(1):310–336, 2008.

[LL08] J-M. Loubes and C. Ludena. Adaptive complexity regularization for inverse problems. *Electronic Journal Of Statistics*, 2:661–677, 2008.

[Mar06] C. Marteau. Regularization of inverse problems with unknown operator. *Math. Methods Statist.*, 15(4):415–443 (2007), 2006.

[Mar08] C. Marteau. On the stability of the risk hull method. *Journal of Statistical Planning and Inference (to appear)*, 2008.

[MR96] B. Mair and F. Ruymgaart. Statistical inverse estimation in Hilbert scales. *SIAM J. Appl. Math.*, 56(5):1424–1444, 1996.

[NP03] W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

[O'S86] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4):502–527, 1986. With comments and a rejoinder by the author.

[vdG00] Sara A. van de Geer. *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2000.

Jean-Michel Loubes
Equipe de probabilites et statistique,
Insistut de Mathématique de Toulouse,
UMR5219, Université de Toulouse,
31000 Toulouse France
Jean-Michel.Loubes@math.univ-toulouse.fr

Clément Marteau
Insistut de Mathématique de Toulouse,
INSA Département GMM.
Clement.Marteau@insa-toulouse.fr