Convergence of the risk for nonparametric IV quantile regression and nonparametric IV regression with full independence

Fabian Dunker[*]

November 12, 2015

Abstract: In econometrics some nonparametric instrumental regression models and nonparametric demand models with endogeneity lead to nonlinear integral equations with unknown integral kernels. We prove convergence rates of the risk for the iteratively regularized Newton method applied to these problems. Compared to related results we relay on a weaker non-linearity condition and have stronger convergence results. We demonstrate by numerical simulations for a nonparametric IV regression problem with continuous instrument and regressor that the method produces better results than the standard method.

JEL classification: C13, C14, C31, C36
Keywords and phrases: Nonparametric regression, instrumental variables, nonlinear inverse problems, iterative regularization

[*] Courant Research Centre: Poverty, Equity and Growth in Developing Countries, University of Göttingen, Humboldtallee 3, 37073 Göttingen, Germany, *Email:* fdunker@uni-goettingen.de, *Phone:* +49 551 39 21674; Financial support by the DFG in the project SFB 823 and by the Ministry of Education and Cultural Affairs of Lower Saxony in the project Reducing Poverty Risk is gratefully acknowledged.

1 Introduction

Endogeneity of an unobservable and covariates is a frequent problem in econometric modeling. An efficient way to deal with endogeneity is to use instrumental variables (IV) in the estimation. These are variables which are independent or mean independent of the unobservable. In the context of nonparametric estimation the IV approach usually leads to ill-posed problems with unknown operator. That means the solution φ of the nonparametric IV problem can be characterized by a possible nonlinear operator equation

$$\mathcal{F}(\varphi) = \psi. \tag{1}$$

In some regression models $\psi = 0$. In other models it is a function that has to be estimated from observations by some estimator $\widehat{\psi}$. The operator $\mathcal{F} : \mathbb{X} \to \mathbb{Y}$ is an integral operator between some function spaces $\mathbb{X}$ and $\mathbb{Y}$. This operator is not exactly known in applications. Only an estimator $\widehat{\mathcal{F}}$ is available. The inverse of the operators $\mathcal{F}$ or $\widehat{\mathcal{F}}$ is usually not continuous. Even with an arbitrary small variance in $\widehat{\psi}$ and $\widehat{\mathcal{F}}$ we usually have $\mathbb{V}\text{ar}(\|\widehat{\mathcal{F}}^{-1}\widehat{\psi}\|_{\mathbb{X}}) = \infty$. Hence the straightforward estimator $\widehat{\varphi} = \widehat{\mathcal{F}}^{-1}\widehat{\psi}$ is inconsistent. We discuss specific examples for nonparametric IV models and the related operators in Section 2.

In this paper we describe and analyze a consistent estimator for this type of problems, when $\mathcal{F}$ is an operator between Hilbert spaces. The estimator is based on the iteratively regularized Gauß-Newton method (IRGNM) with iterated Tykhonov regularization defined below in (12). Details about that method will be given in Section 2.3

This method was suggested by Bakushinskiĭ (1992a). Important monographs on this topic are Bakushinskiĭ and Kokurin (2004) and Kaltenbacher et al. (2008). To use the IRGNM for nonparametric IV problems was proposed and analyzed by Dunker et al. (2014a) with rates for convergence in probability.

In contrast to Dunker et al. (2014a), we prove in this paper convergence rates of the risk under a significantly different set of assumptions. Instead of variational methods as in Dunker et al. (2014a) we relay on spectral methods and a modification of Hoeffding's inequality. Furthermore, we use a significantly weaker non-linearity condition for the operator $\mathcal{F}$ and prove faster rates when the regression function is smooth enough.

In the framework of Dunker et al. (2014a) the non-linearity is constrained by the so called tangential cone condition. This condition is hard to check and difficult to interpret. It can not be reduced to primitive conditions in the context of nonparametric IV. In this paper a Lipschitz condition (18) restricts the non-linearity of $\mathcal{F}$. We will give an easy interpretation of this condition.

Dunker et al. (2014a) derived rates in terms of $\delta := \|\mathcal{F}(\varphi^\dagger) - \widehat{\mathcal{F}}(\varphi^\dagger)\|$ where $\varphi^\dagger$ is the exact solution to (1). In that framework no convergence faster then $O_p(\delta^{-3/4})$ was proven even if $\varphi^\dagger$ is arbitrarily smooth. In this paper we show the rate $\delta^{-3/4}$ under similar smoothness conditions and provide results for higher rates for the risk when $\varphi^\dagger$ is smooth enough. Hence, results in this paper are complementary to results in Dunker et al. (2014a). In addition, to rates in δ we derive rates in the sample size n and analyze adaptive estimation.

The paper is organized as follows. We discuss in Section 2 some IV models which fit into the framework of this paper and explain the estimator. In Section 3 we analyze the error of the Gauss-Newton method applied to IV models. Using this analysis we give convergence rate results in Section 4. Finally, we present some numerical simulations in Section 5. All proofs are in the Appendix A.

2 Application: nonparametric instrumental variables

2.1 Nonparametric instrumental regression

Mean independence The first econometric model that combined nonparametric specification and instrumental variables was a nonparametric regression model with separable error term U and mean independent instruments Z

$$Y = \varphi(X) + U \qquad \text{with } \mathbb{E}[U|Z] = 0. \tag{2}$$

Here and in all following models Y and U are one-dimensional random variables. Whereas, X and Z can be a random vectors or one-dimensional. The dimensions of X and Z do not have to coincide. This model was proposed by Newey and Powell (2003) and Florens (2003). It was further investigated and applied in Hall and Horowitz (2005), Blundell et al. (2007), Chen and Reiss (2011), Horowitz

(2011), Florens et al. (2011), Horowitz (2014), Chen and Christensen (2015), as well as in Breunig and Johannes (2015) among others.

We can write model (2) equivalently by $\mathbb{E}[\varphi(X)|Z] = \mathbb{E}[Y|Z]$. The left hand side of this equation is the so called conditional expectation operator. It maps a function in X to a function in Z. If we assume that conditional on Z the variables X and Y have densities $f_{X|Z}$ and $f_{Y|Z}$, this leads to the integral equation

$$\int f_{X|Z}(x|z)\varphi(x)dx = \int f_{Y|Z}(y|z)dy \qquad \text{for all } z \in \operatorname{supp}(Z). \tag{3}$$

The conditional expectation operator is given as

$$(\mathcal{F}_{ce}\varphi)(z) := \int f_{X|Z}(x|z)\varphi(x)dx.$$

It is a linear integral operator with integral kernel $f_{X|Z}(x|z)$. We denote the right hand side of the equation by $\psi(z) := \int f_{Y|Z}(y|z)dy$. Then model (2) is equivalent to the operator equation

$$(\mathcal{F}_{ce}\varphi)(z) = \psi(z). \tag{4}$$

The integral kernel $f_{X|Z}$ and thereby $\mathcal{F}_{ce}$ as well as the function ψ are not known exactly in practice. They have to be estimated from a sample of Y, X, Z. An estimator $\widehat{f}_{X|Z}$ gives an estimator for the operator in a natural way $(\widehat{\mathcal{F}}_{ce}\psi)(z) := \int \widehat{f}_{X|Z}(x|z)\psi(x)dx$.

The main focus of this paper is on non-linear operator equations. Hence, we are not particularly interested in this problem. However, for a linear operator like $\mathcal{F}_{ce}$ the IRGNM with iterated Tykhonov regularization defined in (12) reduces to the usual iterated Tykhonov regularization

$$\overline{\varphi}_{i+1} := \operatorname*{argmin}_{\varphi \in \mathbb{X}} \left(\|\widehat{\mathcal{F}}_{ce}(\varphi) - \widehat{\psi}\|^2 + \alpha\|\varphi - \overline{\varphi}_i\|^2 \right).$$

This iteration starts at some initial guess $\overline{\varphi}_0$ and is stopped when $i = m$ for some number m. In this sens our results for the IRGNM hold for (2) with iterated Tykhonov regularization as well. Tykhonov regularization for model (2) is well understood and our results for the IRGNM applied to this simpler case are not new. We use model (2) together with iterated Tykhonov regularization only as a benchmark for the IRGNM applied to model (5) below.

The operator formulation (4) gives rise to an identification analysis of model (2). The model identifies the regression function φ if and only if $\mathcal{F}_{ce}$ is injective. This property is usually called completeness. Detailed discussions of the identification can be found in D'Haultfoeuille (2011), D'Haultfoeuille and Fevrier (2011), and Andrews (2011).

Full independence In practical applications econometricians claim that some variable Z is a good instrument when they have compelling reasons that Z is independent of the unobservable U. It is usually not supposed that Z might be mean independent but not fully independent of U. This motivates the regression model

$$Y = \varphi(X) + U \qquad \text{with } U \perp\!\!\!\perp Z \text{ and } \mathbb{E}[U] = 0. \tag{5}$$

Here $\mathbb{E}[U|Z] = 0$ is replaced by full independence $U \perp\!\!\!\perp Z$ and $\mathbb{E}[U] = 0$. This model was proposed in Dunker et al. (2014a). Since the new assumptions imply $\mathbb{E}[U|Z] = 0$ but not vice versa the model (5) makes stronger assumptions than model (2). Thereby, it uses more information than model (2). Consequently, when ever (2) identifies the solution so does (5). Furthermore, there are cases in which (5) can identify a solution, while (2) fails. This is for example the case with discrete instruments and continuous regressors as discussed in D'Haultfoeuille and Fevrier (2011) and Dunker et al. (2014a).

We can translate model (5) into an operator equation as above by defining the operator

$$(\widetilde{\mathcal{F}}_{ind}(\varphi))(u,z) := \begin{pmatrix} \mathbb{P}[Y - \varphi(X) \leq u] - \mathbb{P}[Y - \varphi(X) \leq u | Z = z] \\ \mathbb{E}[Y - \varphi(X)] \end{pmatrix}. \tag{6}$$

When Y, X, Z have a joint density f_{YXZ} an alternative operator is

$$(\mathcal{F}_{ind}(\varphi))(u,z) := \begin{pmatrix} \int f_{YXZ}(u + \varphi(x), x, z) - f_{YX}(u + \varphi(x), x) f_Z(z)\, dx \\ \int \varphi(x) f_x(x)\, dx - \int y f_Y(y)\, dy \end{pmatrix}. \tag{7}$$

Model (5) is equivalent to the operator equations

$$\widetilde{\mathcal{F}}_{ind}(\varphi) = 0 \qquad \text{or} \qquad \mathcal{F}_{ind}(\varphi) = 0.$$

Note that the operators are nonlinear due to the first line of (6) or (7). Furthermore, the operators are not known exactly in practice and have to be estimated. In particular, an estimator $\widehat{f}_{YXZ}$ for the joint density gives an estimator $\widehat{\mathcal{F}}_{ind}$ for $\mathcal{F}_{ind}$ just by replacing f_{YXZ} by $\widehat{f}_{YXZ}$.

By the curse of dimensionality the first line of the operator will dominate the convergence rates. Hence, when we discus this example below we will focus on the first component of the operator. Let us denote the integral kernel of the first line of the operator (7) by

$$\begin{aligned} k_{ind}(y,x,z) &:= f_{YXZ}(y,x,z) - f_{YX}(y,x)f_Z(z) \text{ or} \\ \widehat{k}_{ind}(y,x,z) &:= \widehat{f}_{YXZ}(y,x,z) - \widehat{f}_{YX}(y,x)\widehat{f}_Z(z) \end{aligned}$$

respectively. The first component of the operator then reads $(\mathcal{F}(\varphi))(z) = \int k_{ind}(\varphi(x),x,z)dx$. We will express some properties of $\mathcal{F}_{ind}$ and $\widehat{\mathcal{F}}_{ind}$ in terms of k_{ind} and $\widehat{k}_{ind}$.

Demand models We want to point out a related application in industrial organization. Some recent approaches to model demand in differentiated product markets nonparametrically lead to similar operator equations as above. These models characterize a structural function that explains the demand for every product in the market by observed (and quantified) properties of the products. Similarities to IV regression appear when Y corresponds to a specific observed characteristic of products. X corresponds to a vector of the demand of products and of other observed product characteristics including the price. U is an unobserved product characteristic. In this interpretation the function φ is not a regression function. Instead, it is the demand function inverted in the unobservable U. Nevertheless, the mathematical structure of this problem is the same as in a regression model. The price and the unobservable U are usually dependent. This endogeneity can be treated with instrumental variables. An approach with mean independence similar to (2) was proposed in Berry and Haile (2011) and Berry and Haile (2014). While Dunker et al. (2014b) suggest to assume full independence and use a model similar to (5). For more information on this application we additionally refer to Berry et al. (2013).

2.2 Nonparametric instrumental quantile regression and non-separable models

Nonparametric instrumental quantile regression Another regression model that leads to a nonlinear operator equation is the nonparametric instrumental quantile regression proposed by Horowitz and Lee (2007). For $q \in [0,1]$ the q-th quantile regression function φ_q is characterized by

$$Y = \varphi_q(X) + U \qquad \mathbb{P}(U \leq 0 | Z = z) = q \qquad \text{for all } z. \tag{8}$$

With the assumption that a joint density f_{YXZ} exists, the model is equivalent to the an operator equation $\mathcal{F}_q(\varphi_q) = 0$ with

$$(\mathcal{F}_q(\varphi))(z) := \int F_{YXZ}(\varphi(x), x, z)\, dx - q f_Z(z). \tag{9}$$

and $F_{YXZ}(y,x,z) := \int_{-\infty}^{y} f_{YXZ}(\tilde{y}, x, z)\, d\tilde{y}$. This operator is again nonlinear. Different estimation procedures for this model were proposed and analyzed in Horowitz and Lee (2007), Chen and Pouzo (2012), Dunker et al. (2014a), and Breunig (2015). Local identification properties of this and related models are discussed in Chen et al. (2014).

In order to get a consistent notation with Section 2.1 we define an integral kernel for the operator above by

$$k_q(y,x,z) := F_{YXZ}(y,x,z) - q f_{XZ}(x,z)$$

and denote an estimate by $\widehat{k}_q$. Then $(\mathcal{F}_q(\varphi))(z) = \int k_q(\varphi(x), x, z) dx$. To replace $q f_Z(z)$ by $\int q f_{XZ}(x,z) dx$ is clearly impractical for applications. But this way of writing the operators makes it easier to discuss (7) and (9) in a unified framework.

Non-separable model The forgoing models have in common that the unobservable U enters in a separable way. One model that falls in our framework which allows for an unseparable error term was proposed in Chernozhukov et al.

(2007). See also Chernozhukov and Hansen (2005).

$$\begin{aligned} Y = \phi(X, U) \qquad & \text{with } U \perp\!\!\!\perp Z \text{ and} \\ & \qquad \phi(x, u) \text{ strictly monotonic increasing in } u. \end{aligned} \tag{10}$$

It is pointed out in Horowitz and Lee (2007), and Chernozhukov et al. (2007) that this model is already contained in model (8). Let F_U be the cumulative distribution function of U. Renormalize $\widetilde{U} := F_u^{-1}(U)$ and $\widetilde{\phi}(x, u) = \phi(x, F_U(u))$. Then $\widetilde{U}$ is uniformly distributed on $[0, 1]$. The value of $\widetilde{U}$ corresponds to a quantile in model (8). This reduces (10) to model (8) with $\varphi_q(x) = \widetilde{\phi}(x, q)$.

2.3 The estimator

In this section we introduce an abstract setup which comprises the examples above. For this setup an estimator based on the iteratively regularized Gauß Newton-method is introduced. In the general setup we assume that a function $\varphi^\dagger$ is characterized by the possibly nonlinear operator equation

$$\mathcal{F}(\varphi^\dagger) = 0. \tag{11}$$

Where $\mathcal{F} : B_{2R}(\varphi_0) \subseteq \mathbb{X} \to \mathbb{Y}$ is an operator between Hilbert spaces. A ball $B_{2R}(\varphi_0)$ with radius $2R$ around an initial guess φ_0 must be contained in the domain of $\mathcal{F}$. In practice, large values of R are usually possible. The operator equation is allowed to be ill-posed in the sens that $\mathcal{F}^{-1}$ is not continuous. Furthermore, the operator $\mathcal{F}$ is not known exactly in applications. Only a series of estimators $\widehat{\mathcal{F}}_n : B_{2R}(\varphi_0) \subseteq \mathbb{X} \to \mathbb{Y}_n$ are available where n usually corresponds to a sample size. We can allow for image spaces $\mathbb{Y}_n$ that depend on the estimator and might not be contained in $\mathbb{Y}$. This could be a finite dimensional approximation space. The method we want to propose is based on linearizing $\mathcal{F}$. Therefore, we make the following assumption.

Assumption 1.

1. $\|\varphi^\dagger - \varphi_0\| < R$
2. $\mathcal{F}$ and all $\widehat{\mathcal{F}}_n$ are well defined on $B_{2R}(\varphi_0)$.
3. $\mathcal{F}$ and all $\widehat{\mathcal{F}}_n$ are Fréchet differentiable on $B_{2R}(\varphi_0)$.

This gives all components we need to define the iteratively regularized Gauß-Newton method (IRGNM) with iterated Tykhonov regularization. This method consists of two nested iterations. The outer iteration is a Newton method. It starts with an initial guess φ_0 and produces in the $j-1$-the step the estimate $\widehat{\varphi}_j$. In the j-th Newton iteration step the operator is linearized by

$$\widehat{\mathcal{F}}(\varphi) \approx \widehat{\mathcal{F}}'_n[\widehat{\varphi}_j](\varphi - \widehat{\varphi}_j) + \widehat{\mathcal{F}}_n(\widehat{\varphi}_j).$$

This linearizion is used in an m-times iterated Tykhonov regularization – the inner iteration of the method. In the following scheme the Newton iteration is counted be j and the Tykhonov iteration by i

$$\begin{aligned}
&\overline{\varphi}_{j+1,0} && := \widehat{\varphi}_j \\
&\overline{\varphi}_{j+1,i+1} && := \operatorname*{argmin}_{\varphi \in \mathbb{X}} \left(\|\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j](\varphi - \widehat{\varphi}_j) + \widehat{\mathcal{F}}_n(\widehat{\varphi}_j)\|^2_{\mathbb{Y}_n} + \alpha_j \|\varphi - \overline{\varphi}_{j+1,i}\|^2 \right) \\
&\widehat{\varphi}_{j+1} && := \overline{\varphi}_{j+1,m} \\
&\text{stop if } \|\widehat{\varphi}_j - \varphi_0\| > 2R \text{ and set } \widehat{\varphi}_{j+1} = \varphi_0.
\end{aligned} \tag{12}$$

Here $\alpha_j > 0$ is a regularization parameter. With a small α_j the method has a large variance. With a larger α_j the variance can be controlled while some bias is added to the estimates. We choose α_0 large enough to stabilize the problem. In every Newton step α_j decays by

$$\alpha_{j+1} = q_\alpha \alpha_j \quad \text{with some fixed} \quad 0 < q_\alpha < 1 \tag{13}$$

to reduce the bias. A second parameter that has to be chosen is the number of inner iterations m. A large m is of advantage when $\varphi^\dagger$ is very smooth. When m is chosen to large or for less smooth $\varphi^\dagger$ there is no significant effect on the estimator. Since the inner iteration is numerically cheap a large value of m can be taken without high costs.

An alternative formulation of the method can be obtained by using the functional calculus. We denote by $\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j]^*$ the adjoint operator of $\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j]$ and we set

$$g_\alpha(\lambda) := \frac{(\lambda + \alpha)^m - \alpha^m}{\lambda(\lambda + \alpha)^m.} \tag{14}$$

Then

$$\begin{aligned}&\widehat{\varphi}_{j+1} = \varphi_0 + g_{\alpha_j}\left(\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j]^*\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j]\right)\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j]^*\left(\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j](\widehat{\varphi}_j - \varphi_0) - \widehat{\mathcal{F}}_n(\widehat{\varphi}_j)\right)\\ &\text{stop if } \|\widehat{\varphi}_j - \varphi_0\| > 2R \text{ and set } \widehat{\varphi}_{j+1} = \varphi_0.\end{aligned} \tag{15}$$

is equivalent to (12). The function g_{α_j} is applied to the self-adjoint operator $\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j]^*\widehat{\mathcal{F}}'_n[\widehat{\varphi}_j]$ in the sens of the functional calculus. Compare Bakushinskiĭ and Kokurin (2004) p.23-24.

One crucial parameter choice has to be made for this method. The Newton iteration has to stop at an appropriate iteration step. The size of the regularization parameter is linked to the number of steps. Hence, the number of steps corresponds to an bias variance trade of. In addition, we added some kind of emergency stop. The iteration always ends when $\|\widehat{\varphi}_j - \varphi_0\| > 2R$. Finding the right step to stop the iteration plays an important role in the convergence analysis which is presented in the next section.

Example 2 (Fréchet differentiability)**.** Assumption 1 is usually fulfilled in our examples. The operators are well defined and Fréchet differentiable on the whole space under mild integrability conditions on the joint density f_{YXZ}. The Fréchet derivative of the operator in (7) exists when f_{YXZ} is partially differentiable in the first variable. The operator in (9) is differentiable without further assumptions.

$$(\mathcal{F}'_{ind}[\varphi]\psi)(u,z) = \begin{pmatrix} \int\left[\frac{\partial}{\partial y}f_{YXZ}(u+\varphi(x),x,z) - \frac{\partial}{\partial y}f_{YX}(u+\varphi(x),x)f_Z(z)\right]\psi(x)\,dx \\ \int\psi(x)f_x(x)\,dx\end{pmatrix},$$

$$(\mathcal{F}'_q[\varphi](\psi))(z) = \int f_{YXZ}(\varphi(x),x,z)\psi(x)\,dx.$$

Note that the derivatives are linear integral operators with kernel $\frac{\partial}{\partial y}k(\varphi(x),x,z)$.

3 Error analysis

To abbreviate the formulas in this section we introduce the following notations

$$T_\dagger := \mathcal{F}'[\varphi^\dagger] \qquad \widehat{T}_{n,j} := \widehat{\mathcal{F}}'_n[\widehat{\varphi}_j] \qquad \widehat{T}_{n\dagger} := \widehat{\mathcal{F}}'_n[\varphi^\dagger].$$

This brings the iteration scheme in a more compact form

$$\begin{aligned}&\widehat{\varphi}_{j+1} = \varphi_0 + g_{\alpha_j}(\widehat{T}^*_{n,j}\widehat{T}_{n,j})\widehat{T}^*_{n,j}\left(\widehat{T}_{n,j}(\widehat{\varphi}_j - \varphi_0) - \widehat{\mathcal{F}}_n(\widehat{\varphi}_j)\right)\\ &\text{stop if } \|\widehat{\varphi}_j - \varphi_0\| > 2R \text{ and set } \widehat{\varphi}_{j+1} = \varphi_0.\end{aligned} \tag{16}$$

The error analysis starts with a discussion of smoothness assumptions in form of source conditions. Then we will decompose the error $e_{j+1} := \widehat{\varphi}_{j+1} - \varphi^\dagger$ into different components derive estimates for each component.

3.1 Source conditions

As usual for nonparametric methods a smoothness assumption has to be imposed on the true solution $\varphi^\dagger$ to get convergence rates. In our setup with an ill-posed operator equation (1) it is necessary to link the smoothness of $\varphi^\dagger$ to the smoothing properties of the operator $\mathcal{F}$. An efficient and popular way to formulate this is a source condition. The following definition uses the functional calculus.

Definition 3. Let $\Lambda : [0,\infty) \to [0,\infty)$ be continuous, strictly increasing with $\Lambda(0) = 0$. A representation of the initial error

$$\varphi_0 - \varphi^\dagger = \Lambda(T^*_\dagger T_\dagger)\omega\ , \qquad \omega \in \mathbb{X} \tag{17}$$

is called spectral source condition and Λ is called an index function.

When $T_\dagger$ is a linear integral operator with kernel $\frac{\partial}{\partial y}k(\varphi^\dagger(x), x, z)$ as in Example 2 this definition can be interpreted in the following way. We assume for simplicity that $T_\dagger$ is compact. This for example the case if $\frac{\partial}{\partial y}k(\varphi^\dagger(x), x, z)$ is continuous. It is shown in Reade (1984) and Little and Reade (1984) that the singular values of such an operator decay at least polynomially if $\frac{\partial}{\partial y}k(\varphi^\dagger(x), x, z)$ belongs to a Sobolev space, and exponentially if $\frac{\partial}{\partial y}k(\varphi^\dagger(x), x, z)$ is analytic.

Let $(\sigma_t,\, u_t,\, v_t)$ be the singular system of $T_\dagger$. The source condition (17) implies for $e_0 = \varphi_0 - \varphi^\dagger$

$$\omega = \sum_{t\in\mathbb{N}} \frac{\langle e_0,\, v_t\rangle}{\Lambda(\sigma_t^2)} u_t \in \mathbb{X} \qquad \text{and thereby} \qquad \sum_{t=1}^{\infty}\left(\frac{\langle e_0,\, v_t\rangle}{\Lambda(\sigma_t^2)}\right)^2 < \infty.$$

Hence, a w fulfilling (17) only exists if Λ compensates the decay of the singular values in a way that $\Lambda(\sigma_t^2)^{-1} = \mathcal{O}(\langle e_0,\, v_t\rangle)$. The decay of singular values describes the smoothing properties of the $T_\dagger$ with respect to the singular vectors. While the decay of $\langle e_0, v_t\rangle$ describes the smoothness of e_0 with respect to the singular vectors. Thus, the rate of decay for $\Lambda(x)$ when $x \searrow 0$ compares these two degrees of smoothness. For the examples above the source condition compares the smoothness of f_{YXZ} with the smoothness of the regression function $\varphi^\dagger$.

When σ_t and $\langle e_0, v_t\rangle$ both decay polynomially or both decay exponentially, i.e.

$$\sigma_t \lesssim \exp(-c_\sigma t), \qquad \langle e_0, v_t\rangle \lesssim \exp(-c_{e_0} t)$$

with some constants c_σ and c_{e_0}, the source condition is fulfilled with $\Lambda(x) = x^\mu$. Where $\mu > 0$ is a sufficiently small constant. This is called a Hölder source condition. This concept goes back to Lavrent′ev (1962) and Morozov (1968). For exponential decay of σ_t but only polynomial decay of $\langle e_0, v_t\rangle$ the source condition is true when the operator is rescaled to $\|T_\dagger\| < 1$ and $\Lambda(x) = (-\ln(x))^{-p}$ with some $0 < p$. In this case the smoothing properties of $T_\dagger$ are much stronger then the smoothness of e_0. This choice of Λ is called a logarithmic source condition. It was proposed by Mair (1994) and Hohage (1997).

Despite the word "condition" in the name "source condition" it is rather a relation that selects an index function. For any compact injective operator $T_\dagger$ and any e_0 there is always an index function Λ such that a source condition is fulfilled.

Theorem 4. *Let $H : \mathbb{X} \to \mathbb{X}$ be a compact, injective, self adjoint, nonnegative linear operator. For every $e_0 \in \mathbb{X}$ there is a $\omega \in X$ and a concave index function Λ such that* (17) *is true.*

Proof. Mathé and Hofmann (2008) Corollary 2.

□

In this paper we focus on Hölder source conditions with $\mu \geq 1/2$. Notice, that this implies $e_0 \in \text{Range}(\mathcal{F}'[\varphi^\dagger]^*)$. The case of $\mu \leq 1/2$ and exponential source conditions was analyzed in Dunker et al. (2014a). We make the formal assumption:

Assumption 5. The true solution $\varphi^\dagger$ fulfills a source condition (17) with an index function that satisfying $\Lambda(x)x^{-\mu} = \mathcal{O}(1)$ for $x \searrow 0$ with $0 \leq \mu \leq \frac{1}{2}$.

3.2 Lipschitz condition

Another important assumption is a restriction on the nonlinearity of the operator. We use for this purpose the following Lipschitz condition on the Fréchet derivative.

Assumption 6. There exists $L > 0$ such that

$$\|\widehat{\mathcal{F}}'_n[\xi_1] - \widehat{\mathcal{F}}'_n[\xi_2]\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} \le L\|\xi_1 - \xi_2\|_{\mathbb{X}} \tag{18}$$

for all $\xi_1, \xi_2 \in B_R(\varphi^\dagger)$ and large n.

The norm on the left hand side of the inequality is the operator norm for linear operators

$$\|T\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} := \sup_{f \in \mathbb{X} \text{ with } \|f\|_{\mathbb{X}} \le 1} \|Tf\|_{\mathbb{Y}}.$$

Example 7. We discuss sufficient but not necessary conditions for the operators (7) and (9) to fulfill (18). Assume $\widehat{k}_n(y, x, z)$ is twice differentiable in y and the support of the instrument has finite measure $\mu(\operatorname{supp}(Z)) < \infty$. Then condition (18) is implied by the boundedness of the second partial derivative

$$\sup_{y,z,w} \left| \frac{\partial^2}{\partial y^2} \widehat{k}_n(y, z, w) \right| < \infty, \quad \text{for all sufficiently large } n.$$

To show this, we use the fact that the operator norm of a linear integral operator is bounded by the Hilbert-Schmidt norm. This is the L^2 norm of the integral kernel.

$$\begin{aligned}
\|\widehat{\mathcal{F}}'[\xi_1] - \widehat{\mathcal{F}}'_n[\xi_2]\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} &\le \|\widehat{\mathcal{F}}'[\xi_1] - \widehat{\mathcal{F}}'_n[\xi_2]\|_{HS} \\
&= \sqrt{\iint \left(\frac{\partial}{\partial y} \widehat{k}_n(\xi_1(x), x, z) - \frac{\partial}{\partial y} \widehat{k}_n(\xi_2(x), x, z) \right)^2 dx\, dz} \\
&\le \sqrt{\iint \left(\sup_{y,\widetilde{x}} \left| \frac{\partial^2}{\partial y^2} \widehat{k}_n(y, \widetilde{x}, z) \right| (\xi_1(x) - \xi_2(x)) \right)^2 dx\, dz} \\
&= \mu(\operatorname{supp}(Z)) \sup_{y,x,z} \left| \frac{\partial^2}{\partial y^2} \widehat{k}_n(y, x, z) \right| \|\xi_1 - \xi_2\|_{\mathbb{X}}
\end{aligned}$$

Most density estimators are strongly consistent. If $\widehat{k}_n$ is estimated by a strongly consistent estimator, there is for every constant $c \ge 0$ a $N > 0$ such that for all

$n > N$

$$\sup_{y,x,z} \left| \frac{\partial^2}{\partial y^2} \widehat{k}_n(y,x,z) \right| \leq \sup_{y,x,z} \left| \frac{\partial^2}{\partial y^2} k(y,x,z) \right| + c.$$

Hence, we can set with some sufficiently large constant c

$$L := \sup_{y,x,z} \left| \frac{\partial^2}{\partial y^2} k(y,x,z) \right| + c.$$

Then (18) holds as long as $\sup_{y,x,z} \left| \frac{\partial^2}{\partial y^2} k(y,x,z) \right| < \infty$. For the operators (7) and (9) this is implied by

$$\sup_{y,x,z} \left| \frac{\partial^2}{\partial y^2} f_{YXZ}(y,x,z) \right| < \infty \qquad \text{or} \qquad \sup_{y,x,z} \left| \frac{\partial}{\partial y} f_{YXZ}(y,x,z) \right| < \infty$$

respectively.

3.3 Error decomposition

The error in the $j+1$-th Newton step is

$$\begin{aligned} e_{j+1} &= \widehat{\varphi}_{j+1} - \varphi^\dagger \\ &= \varphi_0 - \varphi^\dagger + g_{\alpha_j}(\widehat{T}_{n,j}^* \widehat{T}_{n,j}) \widehat{T}_{n,j}^* \left(\widehat{T}_{n,j}(\widehat{\varphi}_j - \varphi_0) - \widehat{\mathcal{F}}_n(\widehat{\varphi}_j) \right). \end{aligned}$$

We decompose the error into four parts. These are an approximation error, a propagated noise error, an error due to noise in the derivative, and a nonlinearity error

$$e_{j+1} = e_{j+1}^{app} + e_{j+1}^{noi} + e_{j+1}^{der} + e_{j+1}^{nl}$$

with the following structure.

approximation error $e_{j+1}^{app} := r_{\alpha_j}(\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger}) \Lambda(\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger}) \omega$

propagated noise error $e_{j+1}^{noi} := g_{\alpha_j}(\widehat{T}_{n,j}^* \widehat{T}_{n,j}) \widehat{T}_{n,j}^* [-\widehat{\mathcal{F}}_n(\varphi^\dagger)]$

derivative noise error $e_{j+1}^{der} := r_{\alpha_j}(\widehat{T}_{n,j}^* \widehat{T}_{n,j}) [\Lambda(T_\dagger^* T_\dagger) - \Lambda(\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger})] \omega$

nonlinearity error $e_{j+1}^{nl} := g_{\alpha_j}(\widehat{T}_{n,j}^* \widehat{T}_{n,j}) \widehat{T}_{n,j}^* [\widehat{\mathcal{F}}_n(\varphi^\dagger) - \widehat{\mathcal{F}}_n(\widehat{\varphi}_j) + \widehat{T}_{n,j}(\widehat{\varphi}_j - \varphi^\dagger)]$

$$+ [r_{\alpha_j}(\widehat{T}_{n,j}^* \widehat{T}_{n,j}) - r_{\alpha_j}(\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger})] \Lambda(\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger}) \omega$$

The function r_α is defined as $r_\alpha(\lambda) := 1 - \lambda g_\alpha(\lambda)$. In our case with g_α as in (14) we have: $r_\alpha(\lambda) = \left(\frac{\alpha}{\lambda+\alpha}\right)^m$. A similar related decomposition without e^{der}_{j+1} was proposed in Bakushinskiĭ (1992b) for the case of exactly known operators. In the rest of the section we will analyze each error component in detail.

3.4 Approximation error

The terminology "approximation error" suggests that it is independent of the noise and measures only the effect of the regularization on the approximation. While this is true in the theory for non-random operators it is violated in our case because $e^{app}_{j+1} := r_{\alpha_j}(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})\Lambda(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})\omega$ obviously depends on the estimator $\widehat{T}_{n\dagger}$. Nevertheless, the standard way to bound the norm of this term leads to a bound that does not dependent on $\widehat{T}_{n\dagger}$ and measures only effects of the regularization. This bound relays on the following assumption.

Assumption 8.

1. The number of iterations m of the Tykhonov regularization is large enough such that
$$\Lambda(x)^{-1}x^m = \mathcal{O}(1) \quad \text{for } x \searrow 0.$$
2. The initial regularization parameter α_0 is large enough such that
$$\alpha_0 \geq \frac{\|\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger}\|}{1 - q_\alpha}.$$

If Assumption 8 holds, there exists a constant C_Λ with:

$$\|r_\alpha\Lambda\|_\infty \leq C_\Lambda\Lambda(\alpha) \qquad \text{for all } \alpha \geq 0.$$

Hence, with some constant $\rho \geq \|\omega\|$ the approximation error is bounded by

$$\|e^{app}_{j+1}\| \leq C_\Lambda\Lambda(\alpha_j)\rho. \tag{19}$$

Furthermore, in our setting with $\alpha_j := q_\alpha\alpha_{j-1}$ the following inequalities hold with

$\gamma_{app} := q_\alpha^{-m}$

$$
\begin{aligned}
&\|e_{j+1}^{app}\| \le \|e_j^{app}\| \le \gamma_{app}\|e_{j+1}^{app}\| && \text{for } j \ge 1 \\
&\text{and } \|e_0^{app}\| \le \gamma_{app}\|e_1^{app}\| && \text{since } \alpha_0 \ge \frac{\|\widehat{T}_{n\dagger}^*\widehat{T}_{n\dagger}\|}{1-q_\alpha}.
\end{aligned} \tag{20}
$$

Note that the approximation error tends to 0 with increasing j because α_j is decreasing while Λ is strictly increasing and $\Lambda(0) = 0$.

3.5 Propagated noise error

The propagated noise error $e_{j+1}^{noi} := g_{\alpha_j}(\widehat{T}_{n,j}^*\widehat{T}_{n,j})\widehat{T}_{n,j}^*[-\widehat{\mathcal{F}}_n(\varphi^\dagger)]$ can be bounded by using some standard estimates and the functional calculus. Note that there exists a constant C_g only depending on the function g_α such that for any linear bounded operator $T : \mathbb{X} \to \mathbb{Y}$ and $\psi \in \mathbb{Y}$

$$
\begin{aligned}
\|g_\alpha(TT^*)\|_{\mathcal{L}(\mathbb{Y},\mathbb{Y})} &\le \|g_\alpha\|_\infty = \sup_{x\ge 0}\left(\frac{(x+\alpha)^m - \alpha^m}{x(x+\alpha)^m}\right) \le \frac{C_g}{\alpha}\ , \\
\|g_\alpha(TT^*)TT^*\|_{\mathcal{L}(\mathbb{Y},\mathbb{Y})} &\le \sup_{x\ge 0}|g_\alpha(x)x| = \sup_{x\ge 0}\left(\frac{(x+\alpha)^m - \alpha^m}{(x+\alpha)^m}\right) = 1\ , \text{ and} \\
\|g_\alpha(T^*T)T^*\psi\|_{\mathbb{X}}^2 &= \langle g_\alpha(TT^*)\psi,\, g_\alpha(TT^*)TT^*\psi\rangle_{\mathbb{Y}} \\
&\le \sup_{x\ge 0}|xg_\alpha(x)|\|g_\alpha\|_\infty\|\psi\|_{\mathbb{Y}}^2 \le \frac{C_g}{\alpha}\|\psi\|_{\mathbb{Y}}^2.
\end{aligned} \tag{21}
$$

Hence,

$$
\begin{aligned}
&\|e_{j+1}^{noi}\|_{\mathbb{X}} = \|g_{\alpha_j}(\widehat{T}_{n,j}^*\widehat{T}_{n,j})\widehat{T}_{n,j}^*\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{X}} \le \sqrt{\frac{C_g}{\alpha_j}}\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}} \quad \text{and} \\
&\mathbb{E}\left(\|e_{j+1}^{noi}\|_{\mathbb{X}}^2\right) \le \frac{C_g}{\alpha_j}\mathbb{E}\left(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}^2\right).
\end{aligned} \tag{22}
$$

This bound does not depend on the noise in the derivative $\widehat{T}_{n,j}$ but only on the error in the operator $\widehat{\mathcal{F}}_n$ at $\varphi^\dagger$. Note that the bound grows with decreasing regularization parameter. In contrast to the approximation error it grows with the number of Newton steps. Thus, we have to find the right step J where $\|e_J^{app}\|$ and $\|e_J^{noi}\|$ are balanced such that non of them becomes large.

In addition to the risk bound of e_j^{noi} a concentration inequality is needed to bound the risk of the Newton method. This is formulated as an assumption on the estimator $\widehat{\mathcal{F}}_n$. Lemma 11 at the end of this section shows that this assumption holds for the operators (7) and (9) of the regression models (5) and (8).

Assumption 9. There are constants $c_1, c_2 \geq 0$ such that for all $n \in \mathbb{N}$

$$\mathbb{P}\left\{\left|\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}} - \mathbb{E}\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}\right| \geq \sqrt{\tau\,\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}\right)}\right\} \leq c_1 e^{-c_2\tau}. \tag{23}$$

The following example illustrates the asymptotic behavior of the risk bound (22) for (7) and (9).

Example 10. Let $\mathcal{F}_{ind}(\varphi)(u,z) = \int k_{ind}(\varphi(x)+u, x, z)dx$ be a nonlinear integral operator as in (7) or $\mathcal{F}_q(\varphi)(z) = \int k_q(\varphi(x), x, z)dx$ as in (9). We consider the case where these operators are maps between L^2-spaces and where $\operatorname{supp}(X) \subset \mathbb{R}^{d_X}$, $\operatorname{supp}(Z) \subset \mathbb{R}^{d_Z}$. Assume for $\mathcal{F}_{ind}$ that all derivatives of degree r of the density f_{YXZ} exist and are bounded. For the operator in (9) we assume less smoothness. Derivatives of degree r of F_{YXZ} should exist and be bounded. Let the joint density f_{YXZ} be estimated by a kernel density estimator $\widehat{f}_{YXZ}$. With a kernel of sufficiently high order and with a common bandwidth h. This gives naturally estimators $\widehat{k}_{ind}$ and $\widehat{\mathcal{F}}_{ind}$ for the regression with full independence and $\widehat{k}_q$ and $\widehat{\mathcal{F}}_q$ for the quantile regression.

With these smoothness assumptions, sample size n, and bandwidth h the estimators $\widehat{k}_{ind}$ and $\widehat{k}_q$ converge in both cases with the rate

$$\mathbb{E}(\|k - \widehat{k}\|^2_{L^2}) = \mathcal{O}(n^{-1}h^{-d_X-d_Z-1} + h^{2r}).$$

The operators $\widehat{\mathcal{F}}_{ind}$ and $\widehat{\mathcal{F}}_q$ integrate $\widehat{k}$ over x. So the dimension of X should not play a roll in the asymptotic convergence of the MISE of $\widehat{\mathcal{F}}_{ind}(\varphi)$ and $\widehat{\mathcal{F}}_q(\varphi)$. Corollary 25 in the Appendix proves this fact and shows the rates

$$\begin{aligned}\mathbb{E}(\|\widehat{\mathcal{F}}_{ind}(\varphi) - \mathcal{F}_{ind}(\varphi)\|^2_{L^2}) &= \mathcal{O}(n^{-1}h^{-d_Z-1} + h^{2r})\\ &= \mathcal{O}(n^{-\frac{2r}{2r+d_Z+1}}) \quad \text{when } h \sim n^{-\frac{1}{2r+d_Z+1}} \qquad \text{and}\\ \mathbb{E}(\|\widehat{\mathcal{F}}_q(\varphi) - \mathcal{F}_q(\varphi)\|^2_{L^2}) &= \mathcal{O}(n^{-1}h^{-d_Z} + h^{2r}) = \mathcal{O}(n^{-\frac{2r}{2r+d_Z}}) \quad \text{when } h \sim n^{-\frac{1}{2r+d_Z}}.\end{aligned}$$

With the bound in (22) the rate of the MISE of the propagated noise error is

$$\mathbb{E}\left(\|e^{noi}_{j+1}|^2_{L^2}\right) \leq \frac{C_g}{\alpha_j}\mathbb{E}\left(\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|^2_{L^2}\right) = \mathcal{O}\left(\alpha_j^{-1}(n^{-1}h^{-d_Z-1} + h^{2r})\right) \text{ for } \mathcal{F}_{ind},$$
$$\mathbb{E}\left(\|e^{noi}_{j+1}|^2_{L^2}\right) \leq \frac{C_g}{\alpha_j}\mathbb{E}\left(\|\widehat{\mathcal{F}}_{q}(\varphi^\dagger)\|^2_{L^2}\right) = \mathcal{O}\left(\alpha_j^{-1}(n^{-1}h^{-d_Z} + h^{2r})\right) \text{ for } \mathcal{F}_{q}.$$

Lemma 11. *Consider the operators* (7) *and* (9) *as maps into* $L^2(U, Z)$ *or* $L^2(Z)$ *respectively. Assume that* f_{YXZ} *is estimated by a kernel density estimator with a product kernel composed of a one-dimensional kernel* K_Y *and two multivariate kernels* K_X *and* K_Z *corresponding to the dimensions* $\dim(X) = d_X$ *and* $\dim(Z) = d_Z$ *with joint bandwidth* h*. Then the following exponential inequality holds for* $\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2}$

$$\mathbb{P}\left\{\left|\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2} - \mathbb{E}\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2}\right| \geq \sqrt{\tau\,\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2}\right)}\right\} \leq 2e^{-c_2\tau}.$$

The Lemma shows that Assumption 9 is true under these assumptions. Higher order kernels are not ruled by the Lemma.

3.6 Derivative noise error

The third component in the error decomposition is the error in the approximation of the derivative

$$e^{der}_{j+1} := r_{\alpha_j}(\widehat{T}^*_{n,j}\widehat{T}_{n,j})[\Lambda(T^*_\dagger T_\dagger) - \Lambda(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})]\omega.$$

The simple observation that $r_\alpha(x) = \left(\frac{\alpha}{x+\alpha}\right)^m \leq 1$ for $x \in [0, \infty)$ independent of α or m leads to the estimate

$$\|e^{der}_{j+1}\|_{\mathbb{X}} \leq \rho\|\Lambda(T^*_\dagger T_\dagger) - \Lambda(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})\|.$$

Where the norm on the right hand side of the inequality is the usual operator norm. A way to simplify the term $\|\Lambda(T^*_\dagger T_\dagger) - \Lambda(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})\|$ is provided by the following lemma.

Lemma 12 (Egger (2005) Lemma 3.2.)**.** *For two linear bounded operators between*

Hilbert spaces A and B and $\mu \geq 0$ exists a constant c_μ such that

$$\|(A^*A)^\mu - (B^*B)^\mu\| \leq c_\mu \begin{cases} \|A-B\| \ (1 + \|A\| + \|B\| + |\ln(\|A-B\|)|) & \text{for } \mu = \frac{1}{2} \\ \|A-B\| \left|\|A\| - \|B\|\right|^\mu & \text{for } \mu > \frac{1}{2}\,. \end{cases}$$

Hence, with some constant C_d

$$\|e^{der}_{j+1}\|_{\mathbb{X}} \leq C_d \rho \begin{cases} \|\widehat{T}_{n\dagger} - T_\dagger\|_D \,|\ln(\|\widehat{T}_\dagger - T_\dagger\|_D)| & \text{for } \mu = \frac{1}{2} \\ \|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu} & \text{for } \mu > \frac{1}{2} \end{cases} \tag{24}$$

The norm $\|\cdot\|_D$ on the right hand side which measures the deviation in the derivative is either the operator norm or some norm dominating the operator norm. In Example 14 and Lemma 15 we will choose the Hilber-Schmidt norm as $\|\cdot\|_D$ for technical reasons.

The bounds in (24) are independent of the regularization parameter α and of the number of Newton steps j. They depend only on the noise in the Fréchet derivative of $\mathcal{F}$ at $\varphi^\dagger$. One advantage is that this error does not need to be balanced with the approximation error and the propagated noise error.

In addition to this estimate of $\|e^{der}_{j+1}\|$ an exponential inequality similar to (23) is needed to bound the tail behavior of $\|e^{der}_{j+1}\|$. Convergence of the expected square error will be proved later only for the case $\mu > 1/2$. The case $\mu = 1/2$ just allows to show convergence in probability. Therefore we make the following assumption just for the case $\mu > 1/2$.

Assumption 13. There are constants c_3 and c_4 such that for all $n \in \mathbb{N}$

$$\mathbb{P}\left[\left|\|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu}\right)\right| \geq \sqrt{\tau \operatorname{\mathbb{V}ar}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu}\right)}\right] \leq c_3 e^{-c_4 \tau}. \tag{25}$$

Where $\|\cdot\|_D$ is the operator norm $\|\cdot\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$ or some norm that dominates the operator norm.

In the rest of the section an example illustrates how $\|e^{der}_{j+1}\|_{L_2}$ converges for the operators (7) and (9) of the regression problems. Afterwards, Lemma 15 show that Assumption 13 holds for these operators with the Hilbert-Schmidt norm. This norms dominates the operator norm, when T is a map between L^2 spaces.

Example 14. We adopt the assumptions and constructions of $\widehat{\mathcal{F}}_{ind}$, $\widehat{\mathcal{F}}_q$, $\widehat{k}_{ind}$ and $\widehat{k}_q$ from Example 10. When Assumption 1 holds, the Fréchet derivatives have the form

$$\begin{aligned}
\widehat{\mathcal{F}}'_{ind}[\varphi]\psi(u,z) &= \int \frac{\partial}{\partial y}\widehat{k}_{ind}(\varphi(x)+u,\, x,\, z)\psi(z)\,dz,\\
\widehat{\mathcal{F}}'_q[\varphi]\psi(z) &= \int \frac{\partial}{\partial y}\widehat{k}_q(\varphi(x),\, x,\, z)\psi(z)\,dz.
\end{aligned}$$

The Hilbert-Schmidt norm bounds the operator norm from above and is in our case the L^2 norm of the integral kernels $\frac{\partial}{\partial y}\widehat{k}_{ind}(\varphi(x)+u,x,z)$ and $\frac{\partial}{\partial y}\widehat{k}_q(\varphi(x),x,z)$. A more explicit representation of the kernel density estimator is needed in the formula. Therefor, we introduce the notation $\widehat{\kappa}_{n,h}(u,x,z) := \frac{\partial}{\partial y}\widehat{k}_{ind}(u,x,z)$ when a sample of size n and the bandwidth h are used to estimate $\widehat{k}_{ind}$. Accordingly, $\widehat{\kappa}_{1,1}$ stands for the partial derivative of the unscaled kernel and $\kappa(u,x,z) := \frac{\partial}{\partial y}k_{ind}(u,x,z)$.

$$\begin{aligned}
&\mathbb{E}\left(\|e^{der}_{j+1}\|^2_{\mathbb{X}}\right) \leq C_d\rho\mathbb{E}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{2(1+\mu)}\right)\\
&\quad = C_d\rho\mathbb{E}\left(\int\left(\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z) - \kappa(\varphi(x)+u,x,z)\right)^2 d(u,x,z)\right)^{1+\mu}\\
&\quad = C_d\rho\mathbb{E}\Bigg(\int\left(\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z) - \mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z)\right)^2\\
&\qquad\qquad + \left(\mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z) - \kappa(\varphi(x)+u,x,z)\right)^2 d(u,x,z)\Bigg)^{1+\mu}\\
&\quad \leq 2^{1+\mu}C_d\rho\int\mathbb{E}\left|\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z) - \mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z)\right|^{2(1+\mu)}d(u,x,z)\\
&\qquad\qquad + 2^{1+\mu}C_d\rho\left(\int\left(\mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z) - \kappa(\varphi(x)+u,x,z)\right)^2 d(u,x,z)\right)^{1+\mu}
\end{aligned}$$

Where Jensen's inequality is used in the last inequality. We analyze the second term first. Here $\mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z) - \kappa(\varphi(x)+u,x,z)$ is the bias of a partial derivative of a $1+d_X+d_Z$-dimensional kernel density estimator. Hence,

$$\left(\int\left(\mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z) - \kappa(\varphi(x)+u,x,z)\right)^2 d(u,x,z)\right)^{1+\mu} = \mathcal{O}\left(h^{2(r-1)(1+\mu)}\right).$$

The expectation in the first term can be analyzed with the usual change in variables

$$
\begin{aligned}
&\mathbb{E}\Big|\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z)-\mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z)\Big|^{2(1+\mu)}\\
&\quad=\int\Big|\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z)-\mathbb{E}\widehat{\kappa}_{n,h}(\varphi(x)+u,x,z)\Big|^{2(1+\mu)}f_{YXZ}(\tilde{y},\tilde{x},\tilde{z})d(\tilde{y},\tilde{x},\tilde{z})\\
&\quad=\frac{h^{(d_X+d_Z+1)}}{n^{1+\mu}h^{2(1+\mu)(d_X+d_Z+2)}}\int\Big|\widehat{\kappa}_{1,1}(\bar{u},\bar{x},\bar{z})-\mathbb{E}\widehat{\kappa}_{1,1}(\bar{u},\bar{x},\bar{z})\Big|^{2(1+\mu)}\\
&\qquad\qquad\qquad\qquad f_{YXZ}(y-h(\varphi(x)+\bar{u}),x+h\bar{x},z+h\bar{z})d(\bar{y},\bar{x},\bar{z})\\
&\quad=n^{-1-\mu}h^{-((1+2\mu)(d_X+d_Z+2)+1)}\big(Cf_{YXZ}(y,x,z)+\mathcal{O}(h)\big)+\mathcal{O}(n^{-1-\mu}).
\end{aligned}
$$

The constant C in the last line does not depend on n or h. Combining the analysis of both terms yields

$$
\mathbb{E}\left(\|e^{der}_{j+1}\|^2_{\mathbb{X}}\right)=\mathcal{O}\left(n^{-1-\mu}h^{-((1+2\mu)(d_X+d_Z+2)+1)}+h^{2(r-1)(1+\mu)}\right).
$$

A similar computation can be carried out for the quantile regression problem with (9). We can conclude that a large value of μ has a positive impact on the convergence of $\mathbb{E}\left(\|e^{der}_{j+1}\|^2_{\mathbb{X}}\right)$. Larger values of μ correspond to more smoothness in the solution, i.e. a less ill-posed problem. We like to point out that in the special case of a linear operator always $\|e^{der}_{j+1}\|^2_{\mathbb{X}}=0$ for all j.

Lemma 15. *Consider the operators* (7) *and* (9) *as maps into* $L^2(U,Z)$ *or* $L^2(Z)$ *respectively. Assume that* $\widehat{k}_n$ *is estimated by a kernel density estimator with a product kernel composed of* K_Y*,* K_X*, and* K_Z *with joint bandwidth* h *as in Lemma 15. Then the following exponential inequality holds for* $\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2}$

$$
\mathbb{P}\left[\left|\|\widehat{T}_{n\dagger}-T_\dagger\|^{1+\mu}_{HS}-\mathbb{E}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|^{1+\mu}_{HS}\right)\right|\geq\sqrt{\tau\,\mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|^{1+\mu}_{HS}\right)}\right]\leq 2e^{-c_4\tau}.
$$

3.7 Nonlinearity error

It is easy to check that the last part of the error decomposition

$$
\begin{aligned}
e^{nl}_{j+1}:=g_{\alpha_j}(\widehat{T}^*_{n,j}\widehat{T}_{n,j})\widehat{T}^*_{n,j}[\widehat{\mathcal{F}}_n(\varphi^\dagger)-\widehat{\mathcal{F}}_n(\widehat{\varphi}_j)+\widehat{T}_{n,j}(\widehat{\varphi}_j-\varphi^\dagger)]&\\
+[r_{\alpha_j}(\widehat{T}^*_{n,j}\widehat{T}_{n,j})-r_{\alpha_j}(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})]\Lambda(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})\omega.&
\end{aligned}
$$

vanishes if the operator $\widehat{\mathcal{F}}_n$ is linear. In contrast, it can become arbitrary large when $\widehat{\mathcal{F}}_n$ is nonlinear and no constraints on its nonlinearity are imposed. This is why we call e^{nl}_{j+1} the nonlinearity error.

To control $\|e^{nl}\|$ a restriction on the nonlinearity of $\widehat{\mathcal{F}}_n$ is necessary. We already introduced a suitable constraint in Assumption 6. The Lipschitz condition (18) in Assumption 6 allows to bound the Taylor reminder of the first term in the nonlinearity error by

$$\|\widehat{\mathcal{F}}_n(\varphi^\dagger) - \widehat{\mathcal{F}}_n(\widehat{\varphi}_j) + \widehat{T}_{n,j}(\widehat{\varphi}_j - \varphi^\dagger)\| \leq \frac{L}{2}\|\widehat{\varphi}_j - \varphi^\dagger\|^2 = \frac{L}{2}\|e_j\|^2.$$

For the norm of the second term an additional inequality is needed. It was shown in Bakushinskiĭ and Kokurin (2004) Chapter 4.1 that for every $\mu \geq \frac{1}{2}$ there is a constant C_μ, such that for two linear operators $A, B : \mathbb{X} \to \mathbb{Y}$ between Hilbert spaces

$$\|[r_\alpha(A^*A) - r_\alpha(B^*B)](B^*B)^\mu\| \leq C_\mu\|A - B\|.$$

This yields in our case

$$\begin{aligned}\|[r_{\alpha_j}(\widehat{T}^*_{n,j}\widehat{T}_{n,j}) - r_{\alpha_j}(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})]\Lambda(\widehat{T}^*_{n\dagger}\widehat{T}_{n\dagger})\omega\| &\leq C_\mu\|\widehat{T}_{n,j} - \widehat{T}_{n\dagger}\|\rho \\ &\leq C_\mu\rho L\|\widehat{\varphi}_j - \varphi^\dagger\| = C_\mu\rho L\|e_j\|.\end{aligned}$$

Putting both estimates together and use (21) gives

$$\|e^{nl}_{j+1}\| \leq \frac{L\sqrt{C_g}}{2\sqrt{\alpha_j}}\|e_j\|^2 + C_\mu\rho L\|e_j\|. \tag{26}$$

This error bound grows quadratically and so does $\|e^{nl}_{j+1}\|$ in many cases. The Newton iteration has to be stopped at a sufficiently small step j to control the nonlinearity error which is similar to the propagated noise error. In fact, the nonlinearity error is bounded by the other three error components for some Newton steps. The next Lemma computes an appropriate stopping parameter J_{max} such that $\|e^{nl}_j\|$ is dominated for all $j \leq J_{max}$.

Lemma 16. *Let Assumptions 1, 5, 6, 8 hold true with a sufficiently small ρ in Assumption 5. Assume that $B_{2R}(\varphi_0) \subset \mathrm{dom}(\mathcal{F})$ and that $\varphi^\dagger \in B_R(\varphi_0)$. Choose a monotonically increasing function Φ such that $\|e^{noi}_j + e^{der}_j\| \leq \Phi(j)$ for all $j \geq 0$.*

Define

$$J_{max} := \max\left\{ j \in \mathbb{N} : \frac{\Phi(j)}{\sqrt{\alpha_j}} \leq C_{stop} \right\} \textit{ with } 0 < C_{stop} \leq min\left\{ \frac{1}{8L\sqrt{C_g}}, \frac{R}{4\sqrt{\alpha_0}} \right\}. \tag{27}$$

Then it holds for all $j := 1,\ 2, \dots, J_{max}$ that

$$\|e_j^{nl}\| \leq \gamma_{nl}\left(\|e_j^{app}\| + \Phi(j)\right) \textit{ and } \widehat{\varphi}_j \in B_R(\varphi^\dagger),$$

with $\gamma_{nl} := 8L\sqrt{C_g}C_{stop} \leq 1$.

The assumption that ρ is sufficiently small means that the initial guess must be close enough to the true solution. As always for Newton type methods we get only local convergence. In practice, the convergence radius seems to be quite large and does usually not restrict the applicability of the method.

The lemma is formulated for a deterministic setting. For random errors the assumption that a function $\Phi(j)$ exists with $\|e_j^{noi} + e_j^{der}\| \leq \Phi(j)$ usually holds only with a certain probability. In the next section it will be important to control this probability. We will use the following construction for Φ.

Choose sequences δ_n^{noi}, σ_n^{noi}, δ_n^{der}, and σ_n^{der} with

$$\begin{aligned} &\delta_n^{noi} \geq E(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|), \quad (\sigma_n^{noi})^2 \geq Var(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|), \quad \text{and} \\ &\delta_n^{der} \geq E(\|\widehat{T}_{n\dagger} - T_\dagger\|^{1+\mu}), \quad (\sigma_n^{der})^2 \geq Var(\|\widehat{T}_{n\dagger} - T_\dagger\|^{1+\mu}). \end{aligned}$$

In addition, we choose a weight function $\tau(j)$ with

$$\tau(j+1)q_\alpha \geq \tau(j) \tag{28}$$

and q_α as in (13) for all j. We define

$$\Phi_n^{noi}(\tau, j) := \sqrt{\frac{C_g}{\alpha_j}}\delta_n^{noi} + C_d\rho\delta_n^{der} + \sqrt{\tau(j)}\left(\sqrt{\frac{C_g}{\alpha_j}}\sigma_n^{noi} + C_d\rho\sigma_n^{der}\right). \tag{29}$$

By construction this $\Phi_n^{noi}(\tau, j)$ is monotonically increasing in j. Hence, $\Phi_n^{noi}(\tau, j)$ fulfills the assumptions of the last lemma. The probability that $\|e_j^{noi} + e_j^{der}\| \leq \Phi_n^{noi}(\tau, j)$ holds can be estimated with the concentration inequalities (23) and

(25).

4 Convergence rates

4.1 Convergence rates with a priory parameter choice

Throughout this section it is assumed that a priory knowledge about the approximation error $\|e_j^{app}\|$ is available. This information is used to choose a Newton step where the iteration is stopped. This implicitly balances approximation error and noise error. It is similar to a bias variance trade-off. Knowing or at least estimating $\|e_j^{app}\|$ involves knowledge about the true solution $\varphi^\dagger$ which is usually not available in applications. A purely data driven choice of an admissible Newton step to stop the iteration is discussed in the next subsection.

The following lemma gives convergence rates in a deterministic setting, i.e. $0 = \mathbb{V}\mathrm{ar}(\|\widehat{\mathcal{F}}(\varphi^\dagger)\|) = \mathbb{V}\mathrm{ar}(\|\widehat{T}_{n\dagger}\|)$. The crucial point is to show that the maximal stopping parameter J_{max} computed in Lemma 16 is larger or equal to a suitable stopping parameter.

Lemma 17. *Suppose that the Assumptions 1, 5, 6, and 8 are fulfilled. Assume that $B_{2R}(\varphi_0) \subset \mathrm{dom}(\mathcal{F})$, and that ρ is small enough as in Lemma 16. Let $\widetilde{\delta}_n^{noi}$ and $\widetilde{\delta}_n^{der}$ be a sequences such that $\widetilde{\delta}_n^{noi} \geq \|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|$ and*

$$\widetilde{\delta}_n^{der} \geq \begin{cases} \|\widehat{T}_{n\dagger} - T_\dagger\| \, |\ln(\|\widehat{T}_\dagger - T_\dagger\|)| & \text{if } \mu = \frac{1}{2} \\ \|\widehat{T}_{n\dagger} - T_\dagger\|^{1+\mu} & \text{if } \mu > \frac{1}{2} \end{cases}$$

Set

$$\widetilde{J} := \operatorname*{argmin}_{j\in\mathbb{N}} \left(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}} \widetilde{\delta}_n^{noi} \right) \quad \text{and} \quad J := \min\{J_{max}, \widetilde{J}\}.$$

Then there exists a constant C such that

$$\|\widehat{\varphi}_J - \varphi^\dagger\| \leq C \inf_{j\in\mathbb{N}} \left(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}} \widetilde{\delta}_n^{noi} + C_d \rho \widetilde{\delta}_n^{der} \right).$$

This lemma implies convergence in probability of the estimator with the same rate. It thereby compares to the results in (Dunker et al., 2014a). However, the sets of assumption for the results in this paper and in (Dunker et al., 2014a) differ significantly. The next theorem improves Lemma 17 by proving rates for the risk of the estimator.

Theorem 18. *Let the Assumptions 9 and 13 and the conditions of Lemma 17 hold with $\mu > 1/2$ in Assumption 5. Choose sequences δ_n^{noi}, σ_n^{noi}, δ_n^{der}, and σ_n^{der} such that*

$$\begin{aligned} \delta_n^{noi} &\geq E(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|), & (\sigma_n^{noi})^2 &\geq Var(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|), \\ \delta_n^{der} &\geq E(\|\widehat{T}_{n\dagger} - T_\dagger\|^{1+\mu}), & (\sigma_n^{der})^2 &\geq Var(\|\widehat{T}_{n\dagger} - T_\dagger\|^{1+\mu}). \end{aligned}$$

Define the stopping index

$$J := \operatorname*{argmin}_{j\in\mathbb{N}} \left(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}(\delta_n^{noi} + \sigma_n^{noi})} \right)$$

and set

$$J^* := \begin{cases} J & \text{if } \widehat{\varphi}_j \in B_{2R}(\varphi_0) \text{ for } j = 1, \dots, J \\ 0 & \text{else.} \end{cases}$$

Then, there exist constants $C > 1$ and $\bar{\delta}^{noi}$, $\bar{\sigma}^{noi}$, $\bar{\delta}^{der}$ and $\bar{\sigma}^{der}$ such that

$$\sqrt{E(\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|^2)} \leq C \min_{j\in\mathbb{N}} \left(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}(\delta_n^{noi} + \sigma_n^{noi}) + C_d\rho(\delta_n^{der} + \sigma_n^{der})} \right)$$

for all $\delta_n^{noi} \in (0, \bar{\delta}^{noi}]$, $\sigma_n^{noi} \in (0, \bar{\sigma}^{noi}]$, $\delta_n^{der} \in (0, \bar{\delta}^{der}]$ and $\sigma_n^{der} \in (0, \bar{\sigma}^{der}]$.

Corollary 19. *If the assumptions of Theorem 18 hold true, the risk of the estimator achieves the rate*

$$\sqrt{E(\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|^2)} = O\left(\rho^{\frac{1}{2\mu+1}} (\delta_n^{noi} + \sigma_n^{noi})^{\frac{2\mu}{2\mu+1}} + \delta_n^{der} + \sigma_n^{der} \right) \tag{30}$$

or similarly

$$E\left(\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|^2\right) = O\left(\rho^{\frac{2}{2\mu+1}} \left(E(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|^2) \right)^{\frac{2\mu}{2\mu+1}} + E\left(\|\widehat{T}_{n\dagger} - T_\dagger\|^{2+2\mu} \right) \right).$$

On first sight, δ_n^{der} and σ_n^{der} do not seem to play a role for the rates. The exponent $\frac{2\mu}{2\mu+1}$ slows down the convergence with respect to δ_n^{noi} and σ_n^{noi}. It looks as if they dominate the convergence rate. But δ_n^{der} and σ_n^{der} usually correspond to the convergence of an estimator for the derivative of a density. In contrast, $\delta_n^{noi} + \sigma_n^{noi} = E(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|^2)$ usually corresponds to the estimation of the density itself. Hence, $\delta_n^{der} + \sigma_n^{der}$ often decay slower than $\delta_n^{noi} + \sigma_n^{noi}$. Which of these terms dominates the convergence depends on specific properties of the application, e.g. smoothness of the density of the observables, smoothness of the true solution, estimation procedures for $\widehat{\mathcal{F}}_n$, and the numbers of covariates and instruments. This will become apparent in the next example. Furthermore, we want to mention that the rate in (30) is known to be optimal in the very special case that $\mathcal{F}$ is linear with separable noise. See for example Tautenhahn (1998).

Example 20. Let the assumptions of Examples 10 and 14 be true with $r \geq 2$. Combining the results of Corollary 19 and Examples 10 and 14 we get the rate

$$\begin{aligned} &E(\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|^2) \\ &\quad = O\left(\rho^{\frac{2}{2\mu+1}}\left(n^{-1}h^{-(d_Z+1)}\right)^{\frac{2\mu}{2\mu+1}} + n^{-1-\mu}h^{-((1+2\mu)(d_X+d_Z+2)+1)} + h^{\frac{4\mu r}{2\mu+1}}\right). \end{aligned}$$

4.2 Convergence rates for adaptive estimation with Lepskiĭ's principle

The convergence rates presented so far relay on a priori information about $\|e_j^{app}\|$ that is usually not available in practice. One approach to this problem is to use an oracle for the true solution and estimate the error by an oracle inequality. In the context of statistical inverse problems Lepskiĭ's principle is a popular way to do this. We refer to Tsybakov (2000), Bauer and Hohage (2005), Mathé (2006), and Bauer et al. (2009).

The next lemma shows that Lepskiĭ's principle achieves the same rate of convergence in a deterministic setting as the a priori parameter choice. The convergence is only slowed down by a constant factor.

Theorem 21. *Let the assumptions of Lemma 17 hold. Define the Lepskiĭ stopping parameter by*

$$\begin{aligned} J_{Lep} := \min\Bigg\{ j \leq J_{max} \,\Bigg|\, &\|\widehat{\varphi}_i - \widehat{\varphi}_j\| \leq 4(1+\gamma_{nl})\left[\sqrt{\frac{C_g}{\alpha_i}}\widetilde{\delta}_n^{noi} + C_d\rho\widetilde{\delta}_n^{der}\right] \\ &\textit{for all } i = 1, \ldots, J_{max}\Bigg\}. \end{aligned}$$

Then there exists a constant $\tilde{C}$, such that

$$\|\widehat{\varphi}_{J_{Lep}} - \varphi^\dagger\| \le \tilde{C} \inf_{j\in\mathbb{N}} \left(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}} \widetilde{\delta}_n^{noi} + C_d \rho \widetilde{\delta}_n^{der} \right).$$

This result implies convergence in probability of $\|\widehat{\varphi}_{J_{Lep}} - \varphi^\dagger\|$ with the same rate as in Lemma 17. However, the risk of the estimator with Lepskiĭ's principle does not always achieve the same rate as the estimator with a priori parameter choice. Tsybakov (2000) showed for a class of linear ill-posed problems that a Lepskiĭ type adaptive estimation loses a logarithmic factor in the convergence rate compared to the minimax rates. Nevertheless, Cavalier et al. (2002) showed that an adaptive estimation which obtains the optimal rates, is possible for some linear mildly ill-posed problems. It is an open question whether there is an adaptive estimation procedure for nonlinear mildly ill-posed inverse problems which achieves the asymptotic rates of convergence proved for a priori parameter choice in Theorem 18. The following theorem, adapted from Bauer et al. (2009), we prove convergence of the expected square error with Lepskiĭ type parameter choice with the loss of a logarithmic factor.

Theorem 22. *Let the assumptions of Theorem 18 hold. Take in the definition of J_{max} in* (27) *as Φ the function*

$$\widetilde{\Phi}_n^{noi}(j) := \sqrt{\frac{C_g}{\alpha_j}} \left(\delta_n^{noi} + \ln((\sigma_n^{noi})^{-2})\sigma_n^{noi}\right) + C_d\rho(\delta_n^{der} + \ln((\sigma_n^{der})^{-2})\sigma_n^{der}).$$

Define the Lepskiĭ stopping parameter by

$$J_{Lep} := \min\left\{ j \le J_{max} \middle| \|\widehat{\varphi}_i - \widehat{\varphi}_j\| \le 4(1+\gamma_{nl})\widetilde{\Phi}_n^{noi}(j) \quad \text{for all } i = 1, \ldots, J_{max} \right\}$$

and set

$$J^* := \begin{cases} J_{Lep} & \text{if } \widehat{\varphi}_j \in B_{2R}(\varphi_0) \text{ for } j = 1, \ldots, J_{max} \\ 0 & \text{else.} \end{cases}$$

Then there exist constants $C > 1$ and $\bar{\delta}^{noi}$, $\bar{\sigma}^{noi}$, $\bar{\delta}^{der}$ and $\bar{\sigma}^{der}$ such that

$$\sqrt{E(\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|^2)} \le C \min_{j\in\mathbb{N}} \Bigg(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}} \left(\delta_n^{noi} + \ln((\sigma_n^{noi})^{-1})\sigma_n^{noi}\right)$$

$$+ C_d\rho \left(\delta_n^{der} + \ln((\sigma_n^{der})^{-1})\sigma_n^{der}\right) \Big)$$

for all $\delta_n^{der} \in (0, \bar{\delta}^{der}]$, $\sigma_n^{der} \in (0, \bar{\sigma}^{der}]$, $\delta_n^{noi} \in (0, \bar{\delta}^{noi}]$ *and* $\sigma_n^{noi} \in (0, \bar{\sigma}^{noi}]$.

Corollary 23. *Let the assumptions of Theorem 22 be fulfilled. The risk of the estimator with Lepskiĭ type parameter choice achieves the rate*

$$E\big(\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|^2\big) \\ = \mathcal{O}\left(\rho^{\frac{1}{2\mu+1}} \left(\delta_n^{noi} + \ln((\sigma_n^{noi})^{-1})\sigma_n^{noi}\right)^{\frac{2\mu}{2\mu+1}} + \delta_n^{der} + \ln((\sigma_n^{der})^{-1})\sigma_n^{der}\right).$$

5 Numerical examples

We implemented the estimator based on the IRGNM and tested it with simulated data. As a test problem we chose nonparametric IV regression in accordance to the models in (2) and (5). The covariate X and instrument Z where one dimensional. This allows to compare the estimator based on model (5), on operator (7), and on the IRGNM (12) with the estimator based on model (2), on the operator equation (4), and on iterated Tikhonov regularization.

The regressor of the test example was generated by some function g and a random variable V such that

$$X = g(Z) + V \qquad \text{and} \qquad V \perp\!\!\!\perp Z.$$

In addition, an exact solution $\varphi^\dagger$ and an error term U_V depending on V but not on Z are chosen. Then Y is defined as

$$Y := \varphi^\dagger(X) + U_V.$$

With this construction both models (2) and (5) identify the true solution. The functions and probability densities that were chosen for the test example are

$$\varphi^\dagger(x) = \frac{1}{6}\sin(2\pi(x + 0,25)),$$

$$f_Z(z) = \frac{9}{7}\sqrt{z} + \frac{1}{7} \qquad \text{on the interval } [0,\ 1],$$

$$g(z) = 0,8z + 0,1,$$

$$f_V(v) = \frac{1}{0,08\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{v}{0,08}\right)^2\right), \text{ and}$$

$$f_{U_V}(y,v) = \frac{1}{0,07\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-2v}{0,07}\right)^2\right).$$

The densities of V and U_V are constructed with Gaussians. The expectation of U_V depends on v. The problem is solved on the domain

$$\operatorname{supp}(Y) \times \operatorname{supp}(X) \times \operatorname{supp}(Z) = [-1/2, 1/2] \times [0,1] \times [0,1]$$

discretized by 100 × 100 × 100 nodes. Figure 1 shows the exact solution (blue curve) compared to the solution a nonparametric regression without instrumental variables would yield asymptotically (green curve).

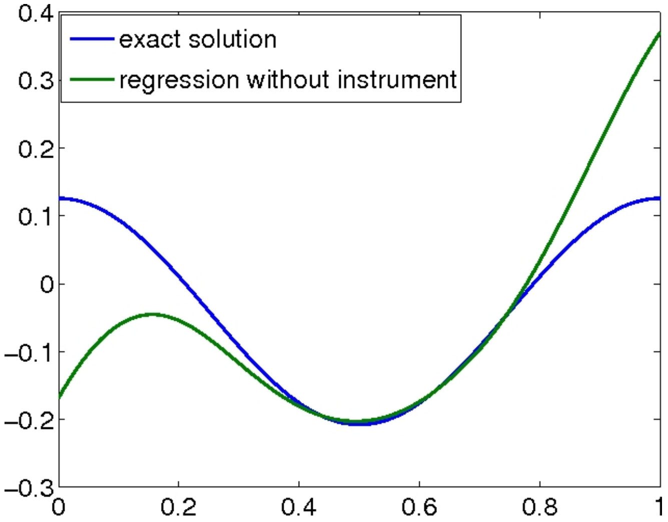

Figure 1: Necessity of the instrument: A standard nonparametric regression would asymptotically yield the green curve which is considerably different from the true curve $\varphi^\dagger$ in blue.

In a first step both methods were tested on this problem using the exact joint density f_{YXZ} instead of a density estimate. The initial guess for both methods was the constant function with the value $E[Y]$. The penalty functional was the squared H^1 norm and the regularization parameters were $\alpha_0 = 1$ and $\alpha_{n+1} = 0.9\alpha_n$. The reconstructions are plotted in Figure 2. The red curve shows the reconstruction with conditional mean assumption, the green curve the reconstruction with full independents. The exact solution is the blue curve.

Both methods converge to the exact solution apart from some deviations at the boundaries. They occur since V and U_V are constructed by Gaussians with unbounded support, while the computation is carried out on a compact domain. A small amount of probability mass gets lost at the boundaries which causes the deformations.

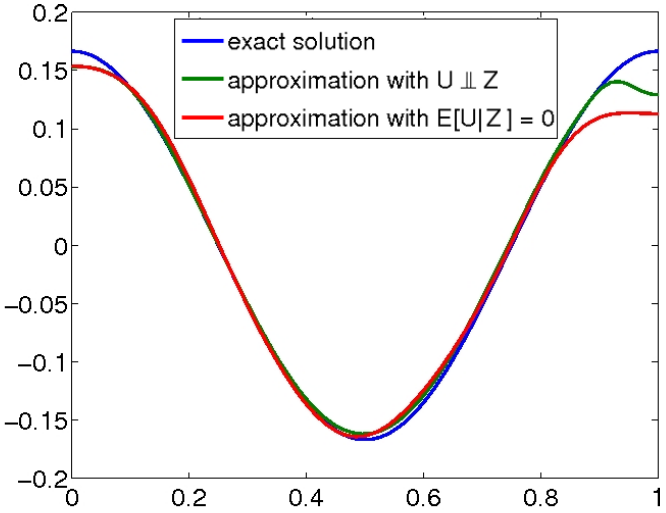

Figure 2: Reconstructions using the exact density f_{YZW}: The blue curve shows the exact solution $\varphi^\dagger$, the red curve the reconstruction with conditional mean assumption and the blue curve is the reconstruction with independent instrument.

In a second step both methods were tested on samples of 500 and 1000 data points. For each of the two sample sizes 1000 samples were generated. Then the joint density f_{YXZ} was estimated. For every sample both methods were evaluated on the same estimate of the density. The Lepskiĭ principle was used to find the stopping parameter of the Newton iteration. For the alternative approach with conditional mean assumption the regularization parameter α had to be chosen instead. This was done by Lepskiĭ's principle as well. I.e. the iterated Tykhonov regularization was computed for a large number of different α. Then one of these approximation was chosen by Lepskiĭ's principle. Hence, both methods are fully data driven.

The histograms in Figures 3–6 show the L^2 error of the reconstructions for both methods and different sample sizes. The values are normed by the initial error. I.e. on this scale the initial error becomes 1. A tabular below summarizes the means and some quantiles of the errors.

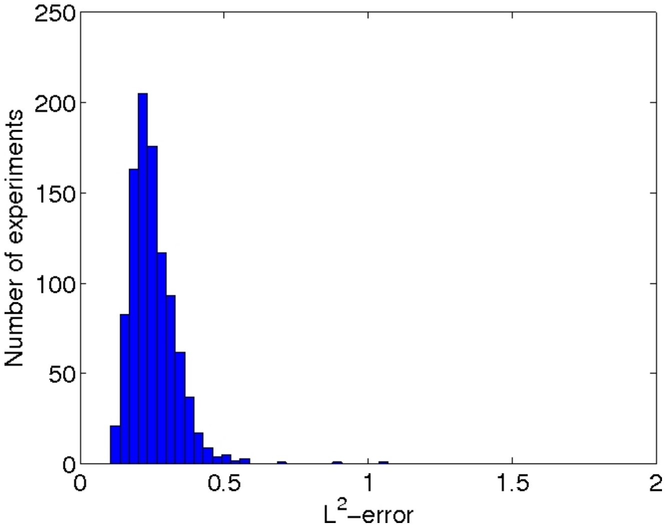

Figure 3: L^2 error of the IRGNM with the assumption $U \perp\!\!\!\perp Z$ and sample size $n = 500$

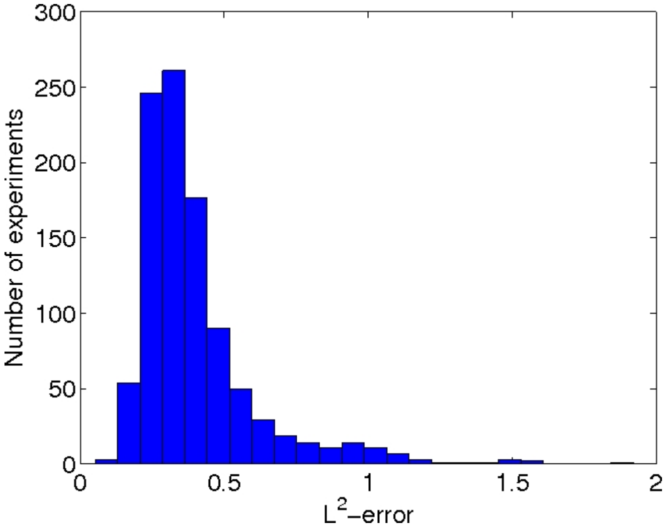

Figure 4: L^2 error of the iterated Tikhonov regularization with the assumption $E[U|Z] = 0$ and sample size $n = 500$

In Figure 3 and 4 we compare the errors of both methods for the sample size 500. Both methods produce acceptable results. The variance and the amount of outliers of the method with independent instrument is much smaller than for the method with the conditional mean assumption. The latter method produces a considerable number of outliers with the same or even larger errors than the initial guess. This can not be observed for the IRGNM. In addition, the mean error of the IRGNM is smaller.

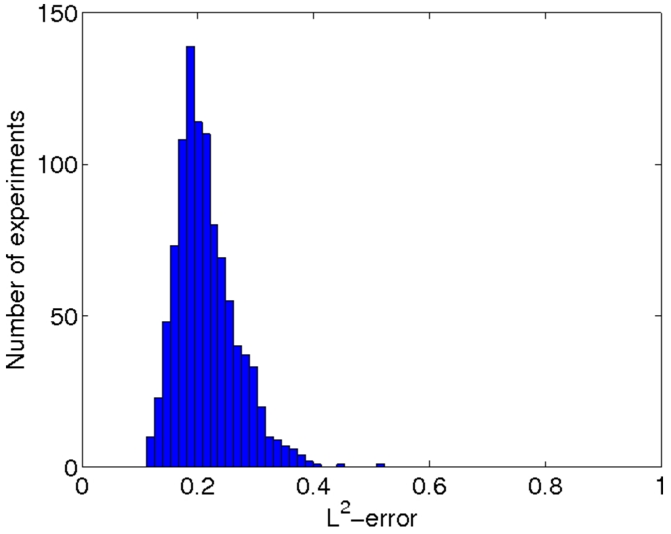

Figure 5: L^2 error of the IRGNM with the assumption $U \perp\!\!\!\perp Z$ and sample size $n = 1000$

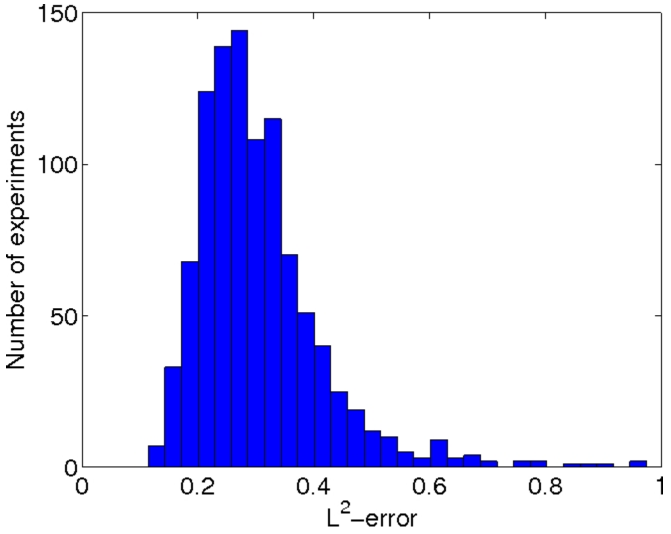

Figure 6: L^2 error of the Tikhonov regularization with the assumption $E[U|Z] = 0$ and sample size $n = 1000$

Similar histograms for samples of 1000 points are in Figures 5 and 6. Now both methods perform well. The advantages of the IRGNM with less outliers and smaller variance can be observed again. The difference in the mean error has become smaller for the larger samples but is still obvious. The following tabular provides the mean and some quantiles of the errors normed by the initial error.

sample size and method	mean	quantiles $q = 0.25$	$q = 0.5$	$q = 0.75$	$q = 0.9$
$n = 500$, $U \perp\!\!\!\perp W$	0.2535	0.2012	0.2398	0.2940	0.3495
$n = 500$, $E[U\|W] = 0$	0.4042	0.2738	0.3437	0.4475	0.6407
$n = 1000$, $U \perp\!\!\!\perp W$	0.2152	0.1780	0.2064	0.2439	0.2868
$n = 1000$, $E[U\|W] = 0$	0.3067	0.2339	0.2846	0.3482	0.4325

We close this section examples of median reconstructions for both sample sizes. They illustrate the advantage of the regression model with independent instruments solved with the IRGNM as well.

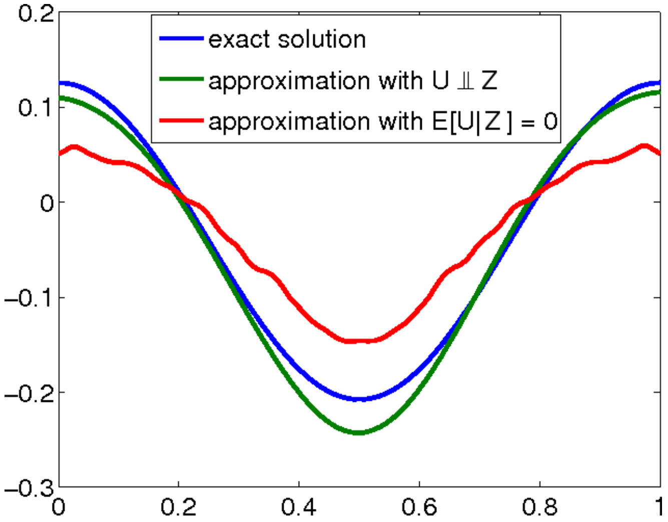

Figure 7: Example for reconstructions with sample size $n = 500$. The blue curve shows the exact solution, the red curve the reconstruction with the conditional mean assumption and the green curve the reconstruction with independent instrument.

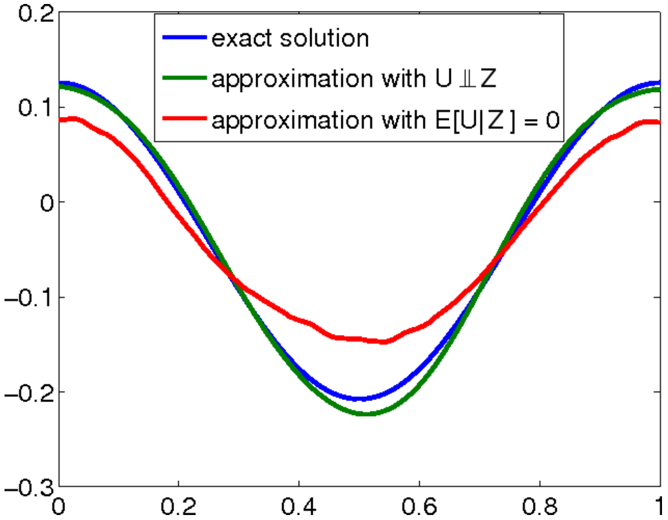

Figure 8: Example for reconstructions with sample size $n = 1000$. The blue curve shows the exact solution, the red curve the reconstruction with the conditional mean assumption and the green curve the reconstruction with independent instrument.

These results suggest that both methods give consistent estimators for the nonparametric instrumental regression with clear advantages for the regression model with independent instruments (5).

Acknowledgment

The author wants to thank Thorsten Hohage and Johannes Schmidt-Hieber for interesting and fruitful discussions on this topic. Financial support by the DFG in the project SFB 823 and by the Ministry of Education and Cultural Affairs of Lower Saxony in the project Reducing Poverty Risk is gratefully acknowledged.

References

Andrews, D. W. K. 2011. Examples of l^2-complete and boundedly-complete distributions. *Preprint*.

Bakushinskiĭ, A. B. 1992a. On a convergence problem of the iterative-regularized Gauss-Newton method. *Zhurnal Vychislitelnoi Matematiki i Matematicheskoi Fiziki*, 32(9):1503–1509.

Bakushinskiĭ, A. B. 1992b. On a convergence problem of the iterative-regularized Gauss-Newton method. *Zh. Vychisl. Mat. i Mat. Fiz.*, 32(9):1503–1509.

Bakushinskiĭ, A. B. and Kokurin, M. Y. 2004. *Iterative Methods for Approximate Solution of Inverse Problems*. Springer, Dordrecht.

Bauer, F. and Hohage, T. 2005. A Lepskij-type stopping rule for regularized Newton methods. *Inverse Problems*, 21:1975–1991.

Bauer, F., Hohage, T., and Munk, A. 2009. Regularized Newton methods for nonlinear inverse problems with random noise. *SIAM Journal on Numerical Analysis*, 47:1827–1846.

Berry, S., Gandhi, A., and Haile, P. 2013. Connected substitutes and invertibility of demand. *Econometrica*, 81(5):2087–2111.

Berry, S. and Haile, P. 2011. Nonparametric identification of multinomial choice demand models with heterogeneous consumers. *Cowles Foundation Discussion Paper*, (1787).

Berry, S. T. and Haile, P. A. 2014. Identification in differentiated products markets using market level data. *Econometrica*, 82(5):1749–1797.

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

Breunig, C. 2015. Goodness-of-fit tests based on series estimators in nonparametric instrumental regression. *J. Econometrics*, 184(2):328–346.

Breunig, C. and Johannes, J. 2015. Adaptive estimation of functionals in nonparametric instrumental regression. *Econometric Theory*, pages 1–43.

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. 2002. Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874.

Chen, X., Chernozhukov, V., Lee, S., and Newey, W. K. 2014. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809.

Chen, X. and Christensen, T. M. 2015. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *J. Econometrics*, 188(2):447–465.

Chen, X. and Pouzo, D. 2012. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321.

Chen, X. and Reiss, M. 2011. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(5):497–521.

Chernozhukov, V. and Hansen, C. 2005. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261.

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

D'Haultfoeuille, X. 2011. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27:460–471.

D'Haultfoeuille, X. and Fevrier, P. 2011. Identification of nonseparable models with endogeneity and discrete instruments. *Preprint*.

Dunker, F., Florens, J.-P., Hohage, T., Johannes, J., and Mammen, E. 2014a. Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *J. Econometrics*, 178(part 2-3):444–455.

Dunker, F., Hoderlein, S., and Kaido, H. 2014b. Nonparametric identification of endogenous and heterogeneous aggregate demand models: complements, bundles and the market level. *cemmap Working Papers*, (CWP23/14).

Egger, H. 2005. Accelerated Newton-Landweber iterations for regularizing nonlinear inverse problems. *SFB-F013-Report*, 2005-03.

Florens, J.-P. 2003. Inverse problems and structural economics: The example of instrumental variables. In Dewatripont, M., Hansen, L. P., and Turnovsky, S., editors, *Advances in Economics and Econometrics: Theory and Applications*, pages 284–311. Cambridge Univ. Press.

Florens, J.-P., Johannes, J., and Van Bellegem, S. 2011. Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory*, 27:472–496.

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

Hohage, T. 1997. Logarithmic convergence rates of the iteratively regularized Gauss-Newton method for an inverse potential and an inverse scattering problem. *Inverse Problems*, 13:1279–1299.

Horowitz, J. L. 2011. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394.

Horowitz, J. L. 2014. Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics*, 180(2):158 – 173.

Horowitz, J. L. and Lee, S. 2007. Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, 75(4):1191–1208.

Kaltenbacher, B., Neubauer, A., and Scherzer, O. 2008. *Iterative Regularization Methods for Nonlinear ill-posed Problems.* Radon Series on Computational and Applied Mathematics. de Gruyter, Berlin.

Lavrent′ev, M. M. 1962. *O nekotorykh nekorrektnykh zadachakh matematicheskoifiziki.* Izdat. Sibirsk. Otdel. Akad. Nauk SSSR, Novosibirsk.

Little, G. and Reade, J. B. 1984. Eigenvalues of analytic kernels. *SIAM J. Math. Anal.*, 15(1):133–136.

Mair, B. A. 1994. Tikhonov regularization for finitely and infinitely smoothing operators. *SIAM J. Math. Anal.*, 25(1):135–147.

Mathé, P. 2006. The Lepskiĭ principle revisited. *Inverse Problems*, 22(3):L11–L15.

Mathé, P. and Hofmann, B. 2008. How general are general source conditions? *Inverse Problems*, 24(1):015009, 5.

McDiarmid, C. 1989. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge.

Morozov, V. A. 1968. The principle of disparity in solving operator equations by the method of regularization. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 8:295–309.

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

Reade, J. B. 1984. Eigenvalues of smooth kernels. *Math. Proc. Cambridge Philos. Soc.*, 95(1):135–140.

Tautenhahn, U. 1998. Optimality for ill-posed problems under general source conditions. *Numer. Funct. Anal. Optim.*, 19(3-4):377–398.

Tsybakov, A. 2000. On the best rate of adaptive estimation in some inverse problems. *C. R. Acad. Sci. Paris Sér. I Math.*, 330(9):835–840.

A Appendix

A.1 Concentration inequalities

We prove Lemmas 11 and 15 with McDiamid's extension of Hoeffding's inequality.

Theorem 24 (McDiarmid (1989))**.** *Let $W_1, \dots, W_n$ be independent random variables. If $f : \operatorname{supp}(W_1, \dots, W_n) \to \mathbb{R}$ satisfies for $1 \le i \le n$*

$$\sup_{\substack{(w_1,\dots,w_n),(w_1',\dots,w_n')\\ \in \operatorname{supp}(W_1,\dots,W_n)}} |f(w_1, \dots, w_n) - f(w_1, \dots, w_{i-1}, w_i', w_{i+1}, \dots, w_n)| \le c_i. \quad (31)$$

Then

$$\mathbb{P}\{|f(W_1, \dots, W_n) - \mathbb{E}f(W_1, \dots, W_n)| \ge \sqrt{\tau}\} \le 2\exp\left(\frac{-2\tau}{\sum_{i=1}^n c_i^2}\right).$$

Proof. (of Lemma 11)

The joint density f_{YXZ} is estimated by a kernel density estimator with kernels K_Y, K_X, and K_Z and common bandwidth h. We write as usual

$$K_{Y,h}(y) = \frac{1}{h}K_Y\left(\frac{y}{h}\right), \qquad K_{X,h}(x) = \frac{1}{h^{d_X}}K_X\left(\frac{x}{h}\right), \qquad K_{Z,h}(z) = \frac{1}{h^{d_Z}}K_Z\left(\frac{z}{h}\right).$$

We first consider the operator (7) and show (31) with $W = (Y, X, Z)$ and

$$\begin{aligned} &f\big((y_1, x_1, z_1), \dots, (y_n, x_n, z_n)\big) := \|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)(u, z)\|_{L_2(u,z)} \\ &= \left\| \int n^{-1} \sum_{i=1}^n K_{Y,h}(\varphi^\dagger(x) - u - y_i) K_{X,h}(x - x_i) K_{Z,h}(z - z_i) dx \right\|_{L_2(u,z)}. \end{aligned}$$

In this case we have

$$\begin{aligned} &|f(w_1, \dots, w_n) - f(w_1, \dots, w_{i-1}, w_i', w_{i+1}, \dots, w_n)| \\ &= n^{-1} \Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i) K_{X,h}(x - x_i) K_{Z,h}(z - z_i) dx \end{aligned}$$

$$
\begin{aligned}
&\quad - \int K_{Y,h}(\varphi^\dagger(x) - u - y_i')K_{X,h}(x - x_i')K_{Z,h}(z - z_i')dx\Big\|_{L_2(u,z)} \\
&\leq 2n^{-1}\Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i)K_{X,h}(x - x_i)K_{Z,h}(z - z_i)dx\Big\|_{L_2(u,z)} \\
&= 2n^{-1}\Big\| \int K_{Y,h}(\varphi^\dagger(hx + x_i) - u - y_i)K_X(x)K_{Z,h}(z - z_i)dx\Big\|_{L_2(u,z)} \\
&= 2n^{-1}\bigg(\int \Big(\int K_{Y,h}(\varphi^\dagger(hx + x_i) - u - y_i)K_X(x)K_{Z,h}(z - z_i)dx\Big)^2 d(u,z)\bigg)^{1/2} \\
&\leq 2n^{-1}\bigg(\int K_X^2(x) \int K_{Y,h}^2(\varphi^\dagger(hx + x_i) - u - y_i)K_{Z,h}^2(z - z_i)d(u,z)\, dx\bigg)^{1/2} \\
&= 2n^{-1}\bigg(\int K_X^2(x)\|K_{Y,h}(u)K_{Z,h}(z)\|^2_{L^2(u,z)}dx\bigg)^{1/2} \\
&= 2n^{-1}\|K_X\|_{L^2}\|K_{Y,h}(u)K_{Z,h}(z)\|_{L^2(u,z)} \\
&= 2n^{-1}h^{-(d_Z+1)/2}\|K_X\|_{L^2}\|K_Y(u)K_Z(z)\|_{L^2(u,z)}.
\end{aligned}
$$

Together with Theorem 24 this proves

$$
\mathbb{P}\left\{ \Big| \|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2} - \mathbb{E}\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}\Big| \geq \sqrt{\tau}\right\} \leq 2\exp\left(\frac{-\tau n h^{d_Z+1}}{2\|K_X\|^2_{L^2}\|K_Y\|^2_{L^2}\|K_Z\|^2_{L_2}}\right). \tag{32}
$$

This shows subgaussianity of $\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}$. Hence, there exists a constant $c_2 > 0$ such that

$$
\mathbb{P}\left\{ \Big| \|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2} - \mathbb{E}\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}\Big| \geq \sqrt{\tau\, \mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}\right)}\right\} \leq 2\exp\left(-c_2\tau\right).
$$

This proves the assertion for the operator in (7).

A similar argument applies to the quantile regression operator in (9). We set $\bar{K}_{Y,h}(y) = \int_{-\infty}^{y} K_Y(\tilde{y})d\tilde{y}$ and $\bar{C}_Y := |\sup_y \bar{K}_{Y,h}(y) - \inf_t \bar{K}_{Y,h}(y)| = |\sup_y \bar{K}_{Y,1}(y) - \inf_y \bar{K}_{Y,1}(y)|$. Note that $\bar{C}_Y$ does not depend on h. Theorem 24 is now applied with

$$
\begin{aligned}
&f(w_1, \ldots, w_n) = \|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2} \\
&\quad = \Big\| n^{-1} \int \bar{K}_{Y,h}(\varphi^\dagger(x) - y_i)K_{X,h}(x - x_i)K_{Z,h}(z - z_i)dx - qK_{Z,h}(z - z_i)\Big\|_{L_2(z)}.
\end{aligned}
$$

The estimation

$$
\begin{aligned}
&|f(w_1,\dots,w_n)-f(w_1,\dots,w_{i-1},w_i',w_{i+1},\dots,w_n)|\\
&=n^{-1}\Big\|\int \bar{K}_{Y,h}(\varphi^\dagger(x)-y_i)K_{X,h}(x-x_i)K_{Z,h}(z-z_i)dx-qK_{Z,h}(z-z_i)\\
&\qquad\qquad -\int \bar{K}_{Y,h}(\varphi^\dagger(x)-y_i')K_{X,h}(x-x_i')K_{Z,h}(z-z_i')dx+qK_{Z,h}(z-z_i')\Big\|_{L_2}\\
&\le n^{-1}\Big\|\bar{C}_Y\int K_{X,h}(x-x_i)K_{Z,h}(z-z_i)dx\Big\|_{L_2}+2qn^{-1}\Big\|K_{Z,h}(z-z_i)\Big\|_{L_2}\\
&=\bar{C}_Y(1+2q)n^{-1}\|K_{Z,h}\|_{L_2}\\
&=\bar{C}_Y(1+2q)\|K_Z\|_{L_2}n^{-1}h^{d_Z/2}.
\end{aligned}
$$

proves together with Theorem 24

$$
\mathbb{P}\left\{\Big|\|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2}-\mathbb{E}\|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2}\Big|\ge\sqrt{\tau}\right\}\le 2\exp\left(\frac{-2\tau nh^{d_Z}}{\bar{C}_Y^2(1+2q)^2\|K_Z\|_{L_2}^2}\right). \quad (33)
$$

Hence, $\|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2}$ is subgaussian which proves the lemma.

□

Corollary 25. *Under the assumptions of Lemma 11*

$$
\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}\right)=\mathcal{O}(n^{-1}h^{-d_Z-1})\quad\textit{and}\quad \mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2}\right)=\mathcal{O}(n^{-1}h^{-d_Z}).
$$

Proof. This is a direct consequence of (32) and (33).

□

Proof. (of Lemma 15)

We follow the same strategy as in the proof of Lemma 11 and adopt the notation above. First, we prove the Lemma for IV regression with full independence. Theorem 24 is applied with $W=(Y,X,Z)$ to the operator (7) with

$$
\begin{aligned}
&f\big((y_1,x_1,z_1),\dots,(y_n,x_n,z_n)\big):=\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\\
&=\left\|n^{-1}\sum_{i=1}^n K_{Y,h}'(\varphi^\dagger(x)-u-y_i)K_{X,h}(x-x_i)K_{Z,h}(z-z_i)-\frac{\partial}{\partial y}f_{YXZ}(\varphi^\dagger(x)-u,x,z)\right\|_{L_2}^{1+\mu}
\end{aligned}
$$

where $K'_{Y,h}$ is the derivative of $K_{Y,h}$. This leads to the inequality

$$\begin{aligned}
&|f(w_1,\dots,w_n) - f(w_1,\dots,w_{i-1},w'_i,w_{i+1},\dots,w_n)| \\
&\le n^{-1-\mu}\big\|K'_{Y,h}(\varphi^\dagger(x)-u-y_i)K_{X,h}(x-x_i)K_{Z,h}(z-z_i) \\
&\qquad - K'_{Y,h}(\varphi^\dagger(x)-u-y'_i)K_{X,h}(x-x'_i)K_{Z,h}(z-z'_i)\big\|_{L_2}^{1+\mu} \\
&\le 2^{1+\mu}n^{-1-\mu}\big\|K'_{Y,h}(\varphi^\dagger(x)-u-y_i)K_{X,h}(x-x_i)K_{Z,h}(z-z_i)\big\|_{L_2}^{1+\mu} \\
&= 2^{1+\mu}n^{-1-\mu}\|K'_{Y,h}\|_{L_2}^{1+\mu}\,\|K_{X,h}\|_{L_2}^{1+\mu}\,\|K_{Z,h}\|_{L_2}^{1+\mu} \\
&= 2^{1+\mu}n^{-1-\mu}h^{\frac{(1+\mu)(d_X+d_Z+3)}{2}}\|K'_Y\|_{L_2}^{1+\mu}\,\|K_X\|_{L_2}^{1+\mu}\,\|K_Z\|_{L_2}^{1+\mu}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{P}\left\{\left|\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\right)\right| \ge \sqrt{\tau}\right\} \\
\le 2\exp\left(\frac{-\tau n^{1+2\mu}h^{(1+\mu)(d_X+d_Z+3)}}{2\|K_X\|_{L^2}^{2+2\mu}\|K_Y\|_{L^2}^{2+2\mu}\|K_Z\|_{L_2}^{2+2\mu}}\right).
\end{aligned}$$

This shows that $\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\right)$ is sub-Gaussian. Thus, there exist constant a c_4 such that

$$\begin{aligned}
\mathbb{P}\left\{\left|\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\right)\right| \ge \sqrt{\tau\,\mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\right)}\right\} \\
\le 2\exp\left(-c_4\tau\right).
\end{aligned}$$

A similar argument holds for the instrumental quantile regression problem. The kernel of the Fréchet derivative of the operator $\mathcal{F}_q$ in (9) at $\varphi^\dagger$ is simply $f_{YXZ}(\varphi^\dagger(x),x,z)$. So Theorem 24 is applied to

$$\begin{aligned}
&f\big((y_1,x_1,z_1),\dots,(y_n,x_n,z_n)\big) := \|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu} \\
&= \left\|n^{-1}\sum_{i=1}^{n} K_{Y,h}(\varphi^\dagger(x)-y_i)K_{X,h}(x-x_i)K_{Z,h}(z-z_i) - f_{YXZ}(\varphi^\dagger(x),x,z)\right\|_{L_2}^{1+\mu}.
\end{aligned}$$

Note that $\sup_y\big(K_{Y,h}(y)\big) - \inf_y\big(K_{Y,h}(y)\big) = h^{-1}\left[\sup_y\big(K_Y(y)\big) - \inf_y\big(K_Y(y)\big)\right]$ and set $C_Y = \left|\sup_y\big(K_Y(y)\big) - \inf_y\big(K_Y(y)\big)\right|$. This allows for the following estimation

$$\begin{aligned}
&|f(w_1, \ldots, w_n) - f(w_1, \ldots, w_{i-1}, w_i', w_{i+1}, \ldots, w_n)| \\
&\leq n^{-1-\mu} \big\| K_{Y,h}(\varphi^\dagger(x) - y_i) K_{X,h}(x - x_i) K_{Z,h}(z - z_i) \\
&\qquad\qquad - K_{Y,h}(\varphi^\dagger(x) - y_i') K_{X,h}(x - x_i') K_{Z,h}(z - z_i') \big\|_{L_2}^{1+\mu} \\
&\leq n^{-1-\mu} h^{-1-\mu} C_Y \big\| K_{X,h}(x - x_i') K_{Z,h}(z - z_i') \big\|_{L_2}^{1+\mu} \\
&= n^{-1-\mu} h^{-\frac{(1+\mu)(2+d_X+d_Z)}{2}} C_Y \|K_X\|_{L_2}^{1+\mu} \|K_Z\|_{L_2}^{1+\mu}.
\end{aligned}$$

This implies

$$\begin{aligned}
&\mathbb{P}\left\{ \left| \|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} \right) \right| \geq \sqrt{\tau} \right\} \\
&\qquad\qquad\qquad \leq 2 \exp\left(\frac{-2\tau n^{1+2\mu} h^{(1+\mu)(2+d_X+d_Z)}}{C_Y^2 \|K_X\|_{L_2}^{2+2\mu} \|K_Z\|_{L_2}^{2+2\mu}} \right).
\end{aligned}$$

and thereby the sub-Gaussianity of $\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu}$. Hence, there exist a constants c_4 such that

$$\begin{aligned}
&\mathbb{P}\left\{ \left| \|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} \right) \right| \geq \sqrt{\tau \, \mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} \right)} \right\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \leq 2 \exp\left(-c_4 \tau \right).
\end{aligned}$$

□

A.2 Nonlinearity error

Proof. (of Lemma 16)

We generalize the prove strategy of Lemma 2.2 in Bauer et al. (2009) to our setting. The proposition follows by induction on j. We start with the induction step. Assume that the proposition holds for $j-1$ with $2 \leq j \leq J_{max}$. Since Φ is increasing and by (20)

$$\begin{aligned}
\|e_{j-1}\| &\leq (1 + \gamma_{nl}) \left(\|e_{j-1}^{app}\| + \Phi(j-1) \right) \\
&\leq (1 + \gamma_{nl}) \left(\gamma_{app} \|e_j^{app}\| + \Phi(j) \right).
\end{aligned}$$

Combining this with inequality (26) and using $(a+b)^2 \le 2a^2 + 2b^2$ yields

$$\begin{aligned} \|e_j^{nl}\| \le\, & C_\mu \rho L(1+\gamma_{nl}) \left(\gamma_{app}\|e_j^{app}\| + \Phi(j)\right) \\ & + \frac{L\sqrt{C_g}}{\sqrt{\alpha_j}}(1+\gamma_{nl})^2 \left(\gamma_{app}^2 \|e_j^{app}\|^2 + \Phi(j)^2\right) . \end{aligned} \tag{34}$$

If $\rho \le \gamma_{nl}/(2C_\mu(1+\gamma_{nl})\gamma_{app})$, the first line on the right hand side is bounded by $1/2\gamma_{nl}\left(\|e_j^{app}\| + \Phi(j)\right)$. To bound the second line, we assume that $\rho \le \gamma_{nl}/(2C_\Lambda \alpha_0^{\mu-1/2} L\sqrt{C_g}(1+\gamma_{nl})^2\gamma_{app}^2)$. It follows from (19) that

$$\frac{\|e_j^{app}\|}{\sqrt{\alpha_j}} \le C_\Lambda \rho \alpha_j^{\mu-\frac{1}{2}} \le C_\Lambda \rho \alpha_0^{\mu-\frac{1}{2}} \le \frac{\gamma_{nl}}{2L(1+\gamma_{nl})^2\gamma_{app}^2} .$$

Thus, $L/\sqrt{\alpha_j}(1+\gamma_{nl})^2\gamma_{app}^2\|e_j^{app}\|^2 \le \frac{1}{2}\gamma_{nl}\|e_j^{app}\|$. By the definition of J_{max} the fact that $\gamma_{nl} \le 1$ we have

$$\frac{L\sqrt{C_g}}{\sqrt{\alpha_j}}(1+\gamma_{nl})^2\Phi^2(j) \le \frac{4L\sqrt{C_g}}{\sqrt{\alpha_j}}\Phi^2(j) \le 4L\sqrt{C_g}C_{stop}\Phi(j) \le \frac{\gamma_{nl}}{2}\Phi(j).$$

Therefore, the second line on the right hand side of (34) is also bounded by $\frac{1}{2}\gamma_{nl}(\|e_j^{app}\| + \Phi(j))$. Together with the estimation of the first line this gives

$$\|e_j^{nl}\| \le \gamma_{nl}\left(\|e_j^{app}\| + \Phi(j)\right) .$$

The base case $j = 1$ of the induction follows in exactly the same way, as long as α_0 is large enough. This is already contained in Assumption 8 and in (20).

Finally, we have to show that $\widehat{\varphi}_j \in B_R(\varphi^\dagger)$. If $\rho \le R/\left(2C_\Lambda \alpha_0^\mu(1+\gamma_{nl})\right)$, then

$$\|e_j^{app}\| \le C_\Lambda \rho \alpha_j^\mu \le C_\Lambda \rho \alpha_0^\mu \le \frac{R}{2(1+\gamma_{nl})}.$$

Moreover, the monotonicity of Φ and the definitions of J_{max}, C_{stop} and γ_{nl} imply:

$$\Phi(j) \le \Phi(J_{max}) \le C_{stop}\sqrt{\alpha_{J_{max}}} \le C_{stop}\sqrt{\alpha_0} \le \frac{R}{4} \le \frac{R}{2(1+\gamma_{nl})} .$$

This shows together with the first part of the proof that

$$\|e_j\| \le (1+\gamma_{nl})\left(\|e_j^{app}\| + \Phi(j)\right) \le R.$$

Hence, $\widehat{\varphi}_j \in B_R(\varphi^\dagger) \subset \mathrm{dom}(\mathcal{F})$.

$\square$

A.3 Convergence rates with a priory parameter choice

Proof. (of Lemma 17)

Notice that J also minimizes

$$\operatorname*{argmin}_{j\in\mathbb{N}} \left(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}}\widetilde{\delta}_n^{noi} + C_d\rho\widetilde{\delta}_n^{der} \right)$$

because $C_d\rho\widetilde{\delta}_n^{der}$ does not depend on j. Set $\Phi(j) := \sqrt{C_g/\alpha_j}\widetilde{\delta}_n^{noi} + C_d\rho\widetilde{\delta}_n^{der}$. If $\widetilde{J} \leq J_{max}$, the theorem is proven by Lemma 16 with $C = 1 + \gamma_{nl}$.

If $\widetilde{J} > J_{max}$, then $\widetilde{J} \geq J_{max} + 1$ and $\Phi(J_{max}+1)/C_{stop} \geq \sqrt{\alpha_{J_{max}+1}}$. Hence, by the monotonicity of Φ

$$\begin{aligned}
\left(1 + \frac{C_\Lambda\rho\alpha_0^{\mu-\frac{1}{2}}}{C_{stop}\sqrt{q_\alpha}}\right)\left(\|e_J^{app}\| + \Phi(\widetilde{J})\right) &\geq \left(1 + \frac{C_\Lambda\rho\alpha_0^{\mu-\frac{1}{2}}}{C_{stop}\sqrt{q_\alpha}}\right)\Phi(J_{max}+1) \\
&\geq \Phi(J_{max}) + C_\Lambda\rho\frac{\Phi(J_{max}+1)\alpha_0^{\mu-\frac{1}{2}}}{C_{stop}\sqrt{q_\alpha}} \\
&\geq \Phi(J_{max}) + C_\Lambda\rho\frac{\sqrt{\alpha_{J_{max}+1}}\alpha_0^{\mu-\frac{1}{2}}}{\sqrt{q_\alpha}} \\
&= \Phi(J_{max}) + C_\Lambda\rho\sqrt{\alpha_{J_{max}}}\alpha_0^{\mu-\frac{1}{2}} \\
&\geq \Phi(J_{max}) + C_\Lambda\rho\alpha_{J_{max}}^{\mu} \\
&\geq \Phi(J_{max}) + \|e_{J_{max}}^{app}\|.
\end{aligned}$$

This proves the lemma when $\widetilde{J} > J_{max}$ with

$$C = \left(1 + \frac{C_\Lambda\rho\alpha_0^{\mu-\frac{1}{2}}}{C_{stop}\sqrt{q_\alpha}}\right)(1 + \gamma_{nl}).$$

$\square$

Proof. (of Theorem 18)

Here and in the following two Lemmas we generalize the prove strategy of Theorem 3.2 and Lemmas 3.3, 3.4 in Bauer et al. (2009) to our setting. Similar to the last proof J is also a minimizer of

$$J = \operatorname*{argmin}_{j \in \mathbb{N}} \left(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}}(\delta_n^{noi} + \sigma_n^{noi}) + C_d \rho(\delta_n^{der} + \sigma_n^{der}) \right).$$

The proof uses a threshold argument. The key tool is the following construction. Define a chain of events with increasing noise level containing each other $A_1 \subset A_2 \subset \ldots \subset A_{k_{max}}$ by

$$A_k := \left\{ \widehat{\varphi}_j \in B_{2R}(\varphi_0) \text{ and } \|e_j^{noi} + e_j^{der}\| \leq \Phi_n^{noi}(\tau_k, j) \text{ for all } j = 1, \ldots, J \right\} \quad (35)$$

and

$$k_{max} := \max \left\{ \left\lfloor \frac{\ln\left((\sigma_n^{noi})^{-2}\right)}{c_2} \right\rfloor, \left\lfloor \frac{\ln\left((\sigma_n^{der})^{-2}\right)}{c_4} \right\rfloor \right\}.$$

with c_2 and c_4 from (23) and (25). The function $\Phi_n^{noi}(\tau_k, j)$ in (35) is defined as in (29). Set $\tau_k(j) := k + \frac{\ln(\kappa)}{c_2}(J - j)$ with some $\kappa > 1$ that is small enough to make inequality (28) true. Consequently, $\Phi_n^{noi}(\tau_k, j)$ is monotonically increasing in j as required for the application of Lemma 16. Notice that k_{max} is chosen in a way such that

$$\max\left\{e^{-c_2 k_{max}}, e^{-c_4 k_{max}}\right\} \leq \max\{(\sigma_n^{noi})^2, (\sigma_n^{der})^2\}.$$

Lemma 16 and Lemma 27 below show that $\|e_j^{noi} + e_j^{der}\| \leq \Phi_n^{noi}(\tau_k, j)$ implies $\widehat{\varphi}_j \in B_{2R}(\varphi_0)$ when σ_n^{noi} is sufficiently small. I.e. the second condition in the definition of A_k implies the first one.

To prepare the final step of the proof we estimate the probability of $A_k \backslash A_{k-1}$ and the probability of the event complementary to A_k. The following computation uses (18), (23), (25).

$$\begin{aligned} P(A_k \backslash A_{k-1}) &= P\left\{ \Phi_n^{noi}(\tau_{k-1}, j) < \|e_j^{noi} + e_j^{der}\| \leq \Phi_n^{noi}(\tau_k, j) \text{ for all } j = 1, \ldots, J \right\} \\ &\leq P\left\{ \Phi_n^{noi}(\tau_{k-1}, j) < \|e_j^{noi} + e_j^{der}\| \text{ for all } j = 1, \ldots, J \right\} \\ &\leq \sum_{j=1}^{J} c_1 e^{-c_2 \tau_k(j)} + c_3 e^{-c_4 \tau_k(j)} \leq (c_1 e^{-c_2 k} + c_3 e^{-c_4 k}) \sum_{j=1}^{J} \kappa^{j-J} \\ &\leq (c_1 e^{-c_2 k} + c_3 e^{-c_4 k}) \sum_{j=0}^{\infty} \kappa^j = \frac{c_1 e^{-c_2 k} + c_3 e^{-c_4 k}}{1 - \kappa^{-1}} \end{aligned}$$

$$\begin{aligned}P(\mathcal{C}A_k) &\le P\left\{\Phi_n^{noi}(\tau_{k-1},j) < \|e_j^{noi}+e_j^{der}\| \text{ for all } j=1,\dots,J\right\}\\ &\le (c_1e^{-c_2k}+c_3e^{-c_4k})\sum_{j=0}^{\infty}\kappa^j = \frac{c_1e^{-c_2k}+c_3e^{-c_4k}}{1-\kappa^{-1}}\ .\end{aligned}$$

In every event A_k we have $J=J^*$. The assumptions of Lemma 16 are fulfilled in A_k. This allows for the following error bound

$$\begin{aligned}\|\widehat{\varphi}_J-\varphi^\dagger\|^2 &\le \left[\|e_J^{app}\| + \sqrt{\frac{C_g}{\alpha_J}}\delta_n^{noi} + C_d\rho\delta_n^{der} + \sqrt{\tau_k(J)}\left(\sqrt{\frac{C_g}{\alpha_J}}\sigma_n^{noi} + C_d\rho\sigma_n^{der}\right)\right]^2\\ &\le 10\|e_J^{app}\|^2 + 10\frac{C_g}{\alpha_J}(\delta_n^{noi})^2 + 10(\delta_n^{der})^2C_d^2\rho^2 + 10k\frac{C_g}{\alpha_J}(\sigma_n^{noi})^2 + 10kC_d^2\rho^2(\sigma_n^{der})^2\\ &=: C_k.\end{aligned}$$

By the construction of the algorithm (12) the worst case error is $\|\widehat{\varphi}_{J^*}-\varphi^\dagger\| \le 3R$. This will serve as an error bound in the event $\mathcal{C}A_{k_{max}}$. Putting everything together yields

$$\begin{aligned}E(\|\widehat{\varphi}_{J^*}-\varphi^\dagger\|^2) &\le P(A_1)C_1 + \sum_{k=2}^{k_{max}} P(A_k\backslash A_{k-1})C_k + P(\mathcal{C}A_{k_{max}})9R^2\\ &\le 10\left(\|e_J^{app}\|^2 + \frac{C_g}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2C_d^2\rho^2\right)\\ &\quad + 10P(A_1)\left(\frac{C_g}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2C_d^2\rho^2\right)\\ &\quad + \sum_{k=2}^{k_{max}} P(A_k\backslash A_{k-1})\left(10k\frac{C_g}{\alpha_J}(\sigma_n^{noi})^2 + 10k(\sigma_n^{der})^2C_d^2\rho^2\right) + P(\mathcal{C}A_{k_{max}})9R^2\\ &\le 10\left(\|e_J^{app}\|^2 + \frac{C_g}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2C_d^2\rho^2\right) + P(\mathcal{C}A_{k_{max}})9R^2\\ &\quad + 10\left(\frac{C_g}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2C_d^2\rho^2\right)\left(2 + \sum_{k=3}^{k_{max}} kP(A_k\backslash A_{k-1})\right)\\ &\le 10\left(\|e_J^{app}\|^2 + \frac{C_g}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2C_d^2\rho^2\right) + \left(\frac{c_1e^{-c_2k_{max}}+c_3e^{-c_4k_{max}}}{1-\kappa^{-1}}\right)9R^2\\ &\quad + 10\left(\frac{C_g}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2C_d^2\rho^2\right)\left(2 + \sum_{k=2}^{k_{max}-1}(k+1)\frac{c_1e^{-c_2k}+c_3e^{-c_4k}}{1-\kappa^{-1}}\right)\end{aligned}$$

$$
\begin{aligned}
&\leq 10\left(\|e_J^{app}\|^2 + \frac{C_g}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2 C_d^2\rho^2\right) + (c' \max\{(\sigma_n^{noi})^2, (\sigma_n^{der})^2\})9R^2 \\
&\quad + 10\left(\frac{C_g}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2 C_d^2\rho^2\right)\left(2 + \sum_{k=2}^{\infty}(k+1)\frac{c_1 e^{-c_2 k} + c_3 e^{-c_4 k}}{1-\kappa^{-1}}\right) \\
&\leq 10\left(\|e_J^{app}\|^2 + \frac{C_g}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2 C_d^2\rho^2\right) + (c' \max\{(\sigma_n^{noi})^2, (\sigma_n^{der})^2\})9R^2 \\
&\quad + 10c''\left(\frac{C_g}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2 C_d^2\rho^2\right) \\
&\leq C\left(\|e_J^{app}\| + \sqrt{\frac{C_g}{\alpha_J}}(\delta_n^{noi} + \sigma_n^{noi}) + C_d\rho(\delta_n^{der} + \sigma_n^{der})\right)^2.
\end{aligned}
$$

We used that $P(A_1) + \sum_{k=2}^{k_{max}} P(A_k \backslash A_{k-1}) + P(\mathcal{C}A_{k_{max}}) = 1$ and $P(A_1) + P(A_2 \backslash A_1) \leq 1$. Furthermore, $c' > 0$, $c'' > 0$ and $C > 0$ are generic constants.

□

The following two lemmas are needed for the proof of Theorem 18 above.

Lemma 26. *Let the assumptions of Theorem 18 hold and define:*

$$
\begin{aligned}
\tilde{\Phi}(j) &:= \sqrt{\frac{C_g}{\alpha_j}}(\delta_n^{noi} + \sigma_n^{noi}) + C_d\rho(\delta_n^{der} + \sigma_n^{der}) \\
\underline{\Gamma}_{noi} &:= \frac{\sqrt{C_g/(q_\alpha\alpha_1)}(\delta_n^{noi} + \sigma_n^{noi}) + C_d\rho(\delta_n^{der} + \sigma_n^{der})}{\sqrt{C_g/\alpha_1}(\delta_n^{noi} + \sigma_n^{noi}) + C_d\rho(\delta_n^{der} + \sigma_n^{der})} \\
\overline{\Gamma}_{noi} &:= q_\alpha^{-\frac{1}{2}}.
\end{aligned}
$$

The following two bounds hold for the stopping index J in Theorem 18:

$$
(1 - \underline{\Gamma}_{noi}^{-1})\tilde{\Phi}(J) \leq (\gamma_{app} - 1)\|e_J^{app}\|, \tag{36}
$$

$$
J \geq \sup\left\{k \in \mathbb{N} \,\middle|\, \|e_1^{app}\|\gamma_{app}^{1-k} > \inf_{l\in\mathbb{N}}\left(C_\Lambda\rho\sqrt{\alpha_l} + \tilde{\Phi}(1)\overline{\Gamma}_{noi}^{l-1}\right)\right\}. \tag{37}
$$

Proof. Note that (20) implies

$$
1 < \underline{\Gamma}_{noi} \leq \frac{\tilde{\Phi}(j+1)}{\tilde{\Phi}(j)} \leq \overline{\Gamma}_{noi}, \quad \text{for all } j \in \mathbb{N}. \tag{38}
$$

We start with inequality (36). Assume the opposite holds true

$$(1-\underline{\Gamma}_{noi}^{-1})\tilde{\Phi}_n^{noi}(J) > (\gamma_{app}-1)\|e_J^{app}\|.$$

It would follow from (20) and (38) that

$$\|e_{J-1}^{app}\| + \tilde{\Phi}(J-1) \leq \gamma_{app}\|e_J^{app}\| + \underline{\Gamma}_{noi}^{-1}\tilde{\Phi}(J) < \|e_J^{app}\| + \tilde{\Phi}(J).$$

This is a contradiction to the definition of J and therefore proves (36).

In order to prove (37) assume that for some k, and some $l \geq 1$

$$\|e_1^{app}\|\gamma_{app}^{1-k} > C_\Lambda \rho\sqrt{\alpha_l} + \tilde{\Phi}(1)\overline{\Gamma}_{noi}^{l-1}.$$

It follows from (19), (20) and (38) that for all $j \leq k$

$$\begin{aligned}\|e_l^{app}\| + \tilde{\Phi}(l) \leq C_\Lambda\rho\sqrt{\alpha_l} + \tilde{\Phi}(1)\overline{\Gamma}_{noi}^{l-1} &< \|e_1^{app}\|\gamma_{app}^{1-k} \leq \|e_k^{app}\| \leq \|e_j^{app}\| \\ &\leq \|e_j^{app}\| + \tilde{\Phi}(j).\end{aligned}$$

As J is the minimizer for $\|e_j^{app}\| + \tilde{\Phi}(j)$ this implies $J > k$. Taking the infimum over l and the supremum over k gives the assertion.

□

Lemma 27. *Let the assumptions of Theorem 18 hold true. Define J_{max} as in Lemma 16 with $\tau_k(j) := k + \frac{\ln(\kappa)}{c_2}(J-j)$*

$$\begin{aligned}J_{max}(k) := \max\Bigg\{ j \in \mathbb{N} \Bigg| \Bigg[& \sqrt{\frac{C_g}{\alpha_j}}\delta_n^{noi} + C_d\rho\delta_n^{der} \\ & + \sqrt{\tau_k(j)}\left(\sqrt{\frac{C_g}{\alpha_j}}\sigma_n^{noi} + C_d\rho\sigma_n^{der}\right)\Bigg]\alpha_j^{-\frac{1}{2}} \leq C_{stop}\Bigg\}.\end{aligned}$$

There exist $\bar{\sigma}_n^{noi} > 0$ and $\bar{\sigma}_n^{der} > 0$ such that for all $\sigma_n^{noi} \leq \bar{\sigma}_n^{noi}$ and $\sigma_n^{der} \leq \bar{\sigma}_n^{der}$ and for all $k = 1, \ldots, k_{max}$ it holds that $J \leq J_{max}$.

Proof. Since $\tau_k(j)$ fulfills inequality (28) for $k \leq k_{max}$ and $j \leq J$,

$$\tau_k(J) \leq \tau_{k_{max}}(J) \leq \max\left\{\ln((\sigma_n^{noi})^{-2})/c_2, \ln((\sigma_n^{der})^{-2})/c_4\right\}.$$

Hence,

$$\begin{aligned}
&\left(\sqrt{\frac{C_g}{\alpha_j}}\delta_n^{noi} + C_d\rho\delta_n^{der} + \sqrt{\tau_k(j)}\left(\sqrt{\frac{C_g}{\alpha_j}}\sigma_n^{noi} + C_d\rho\sigma_n^{der}\right)\right)\alpha_j^{-\frac{1}{2}} \\
&\qquad\le \left(\sqrt{\frac{C_g}{\alpha_J}}\delta_n^{noi} + C_d\rho\delta_n^{der} + \sqrt{\tau_{k_{max}}(J)}\left(\sqrt{\frac{C_g}{\alpha_J}}\sigma_n^{noi} + C_d\rho\sigma_n^{der}\right)\right)\alpha_J^{-\frac{1}{2}} \\
&\qquad\le \max\left\{\sqrt{\frac{\ln((\sigma_n^{noi})^{-2})}{c_2}}, \sqrt{\frac{\ln((\sigma_n^{der})^{-2})}{c_4}}\right\}\tilde{\Phi}(J)\alpha_J^{-\frac{1}{2}} \\
&\qquad\le \max\left\{\sqrt{\frac{\ln((\sigma_n^{noi})^{-2})}{c_2}}, \sqrt{\frac{\ln((\sigma_n^{der})^{-2})}{c_4}}\right\}\frac{\gamma_{app}-1}{1-\underline{\Gamma}_{noi}^{-1}}\|e_J^{app}\|\alpha_J^{-\frac{1}{2}} \\
&\qquad\le C\max\left\{\sqrt{\ln((\sigma_n^{noi})^{-2})}, \sqrt{\ln((\sigma_n^{der})^{-2})}\right\}\alpha_J^{\mu-\frac{1}{2}}
\end{aligned}$$

with $C := \dfrac{\rho C_\Lambda(\gamma_{app}-1)}{\min\{c_2, c_4\}(1-\underline{\Gamma}_{noi}^{-1})}$.

Moreover, we have to take into account that in inequality (37)

$$\begin{aligned}
&\inf_{l\in\mathbb{N}}\left(C_\Lambda\rho\sqrt{\alpha_l} + \tilde{\Phi}(1)\overline{\Gamma}_{noi}^{l-1}\right) \\
&\qquad = \inf_{l\in\mathbb{N}}\left(C_\Lambda\rho\sqrt{\alpha_l} + \left(\alpha_1^{-\frac{1}{2}}(\delta_n^{noi}+\sigma_n^{noi}) + C_d\rho(\delta_n^{der}+\sigma_n^{der})\right)\overline{\Gamma}_{noi}^{l-1}\right)
\end{aligned}$$

decays with a polynomial rate in σ_n^{noi} and σ_n^{der}. Therefore, there exists a constant b for which $J \ge -b\max\{\ln(\sigma_n^{noi}), \ln(\sigma_n^{der})\}$, while $\lim\limits_{x\to\infty} xq_\alpha^{cx}$ goes to 0 for every c as $q_\alpha < 1$. Hence, there are $\bar{\sigma}_n^{noi}$ and $\bar{\sigma}_n^{der}$ such that for all $\sigma_n^{noi} \in\,]0, \bar{\sigma}_n^{noi}]$ and for all $\sigma_n^{der} \in\,]0, \bar{\sigma}_n^{der}]$ it holds:

$$\begin{aligned}
&C\max\left\{\sqrt{\ln((\sigma_n^{noi})^{-2})}, \sqrt{\ln((\sigma_n^{der})^{-2})}\right\}\alpha_J^{\mu-\frac{1}{2}} \\
&\le C\max\left\{\sqrt{\ln((\sigma_n^{noi})^{-2})}, \sqrt{\ln((\sigma_n^{der})^{-2})}\right\}\left(\frac{\alpha_0}{q_\alpha}\right)q_\alpha^{\max\left\{\sqrt{\ln((\sigma_n^{noi})^{-2})},\sqrt{\ln((\sigma_n^{der})^{-2})}\right\}\frac{b}{2}(\mu-\frac{1}{2})} \\
&\le C_{stop}.
\end{aligned}$$

Together with the first estimate this proves the assertion.

□

A.4 Convergence rates for adaptive estimation

Proof. (of Corollary 19) The rates follow from Theorem 18 together with the bound (19) of $\|e_j^{app}\|$.

□

Proof. (of Theorem 21)

The Theorem follows directly by applying Corollary 1 in Mathé (2006) to our problem.

□

Proof. (of Theorem 22)

When $\widetilde{\Phi}_n^{noi}$ is used in the definition of J_{max} in (27), it follows that $J_{max} = O(\ln((\sigma_n^{noi})^{-1}) + \ln((\sigma_n^{der})^{-1}))$. Consider the event A defined as in (35) with

$$\tau(j) := \max\left\{ \frac{\ln\left((\sigma_n^{noi})^{-2}\right)}{c_2}, \frac{\ln\left((\sigma_n^{der})^{-2}\right)}{c_4} \right\}.$$

Applying the Lepskiĭ principle (e.g. Corollary 1 in Mathé (2006)) in this event gives the estimate

$$\|\widehat{\varphi}_{Lep} - \varphi^\dagger\| \le 6 q_\alpha^{-\frac{1}{2}} (1 + \gamma_{nl}) \min_{j=1,\dots,J_{max}} \left(\|e^{app}\| + \widetilde{\Phi}_n^{noi} \right).$$

In Lemma 27 it was shown that for sufficiently small values of δ_n^{noi}, σ_n^{noi}, δ_n^{der} and σ_n^{der} the parameter J_{max} is large enough. Hence, in the asymptotics we can take the infimum over $\mathbb{N}$

$$\|\widehat{\varphi}_{Lep} - \varphi^\dagger\| \le 6 q_\alpha^{-\frac{1}{2}} (1 + \gamma_{nl}) \inf_{j\in\mathbb{N}} \left(\|e^{app}\| + \widetilde{\Phi}_n^{noi} \right).$$

In addition, we estimate the probability of the opposite event of A by

$$\begin{aligned} P(\mathcal{C}A) &\le \sum_{j=1}^{J_{max}} c_1 \exp(-\ln((\sigma_n^{noi})^{-2}) + c_3 \exp(-\ln((\sigma_n^{der})^{-2}) \\ &\le J_{max} \left(c_1 (\sigma_n^{noi})^2 + c_3 (\sigma_n^{der})^2 \right) \\ &\le C' \max\left\{ \ln((\sigma_n^{noi})^{-1})(\sigma_n^{noi})^2, \ln((\sigma_n^{der})^{-1})(\sigma_n^{der})^2 \right\} \\ &\le C'' \min_{j\in\mathbb{N}} \Bigg(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}} (\delta_n^{noi} + \ln((\sigma_n^{noi})^{-1}) \sigma_n^{noi}) \end{aligned}$$

$$+ C_d\rho \left(\delta_n^{der} + \ln((\sigma_n^{der})^{-1})\sigma_n^{der}\right) \Bigg)$$

with two constants C' and C''. We used in the third row that $J_{max} = O(\ln((\sigma_n^{noi})^{-1}) + \ln((\sigma_n^{der})^{-1}))$ and in the fourth row that $\alpha_j^{-\frac{1}{2}}$ is monotonically increasing in j.

We finish the proof with the estimation of the risk

$$\begin{aligned} E[\|\widehat{\varphi}_{Lep} - \varphi^\dagger\|^2] &\leq P(A)36q_\alpha^{-1}(1+\gamma_{nl})^2 \inf_{j\in\mathbb{N}} \left(\|e^{app}\| + \widetilde{\Phi}_n^{noi}\right)^2 + P(\mathcal{C}A)9R^2 \\ &\leq C \min_{j\in\mathbb{N}} \Bigg(\|e_j^{app}\| + \sqrt{\frac{C_g}{\alpha_j}}\left(\delta_n^{noi} + \ln((\sigma_n^{noi})^{-1})\sigma_n^{noi}\right) \\ &\qquad + C_d\rho\left(\delta_n^{der} + \ln((\sigma_n^{der})^{-1})\sigma_n^{der}\right)\Bigg). \end{aligned}$$

□

Proof. (of Corollary 23)

The rate follows from the last theorem and the bound (19) of $\|e_j^{app}\|$. □