Electronic Journal of Statistics
Vol. 15 (2021) 6151–6190
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1938

Adaptive estimation for some nonparametric instrumental variable models with full independence

Fabian Dunker

School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand
e-mail: fabian.dunker@canterbury.ac.nz

Abstract: The problem of endogeneity in statistics and econometrics is often handled by introducing instrumental variables (IV) which fulfill the mean independence assumption, i.e. the unobservable is mean independent of the instruments. When full independence of IV's and the unobservable is assumed, nonparametric IV regression models and nonparametric demand models lead to nonlinear integral equations with unknown integral kernels. We prove convergence rates for the mean integrated square error of the iteratively regularized Newton method applied to these problems. Compared to related results we derive stronger convergence results that rely on weaker nonlinearity restrictions. We demonstrate in numerical simulations for a nonparametric IV regression that the method produces better results than the standard model.

MSC2020 subject classifications: Primary 62G08; secondary 62G20.
Keywords and phrases: Instrumental variables, nonparametric regression, quantile regression, inverse problem, regularization.

Received April 2020.

1. Introduction

Dependence of an unobservable error term and covariates is a frequent problem in statistical and econometrical modeling known as endogeneity. An efficient way to deal with endogeneity is to use instrumental variables (IV) in the estimation. These are additional variables which can assumed to be independent or mean independent of the unobservable. In the context of nonparametric estimation the IV approach usually leads to ill-posed problems with an unknown operator that needs to be estimated. The solution φ of the nonparametric IV problem can be characterized by a possibly nonlinear operator equation

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

In some regression models $\psi = 0$, in others ψ 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 Banach or Hilbert spaces $\mathbb{X}$ and $\mathbb{Y}$ which is unknown 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 arbitrarily small variance in $\widehat{\psi}$ and $\widehat{\mathcal{F}}$ we usually have $\mathbb{V}\mathrm{ar}(\|\widehat{\mathcal{F}}^{-1}\widehat{\psi}\|_{\mathbb{X}}) = \infty$ and the straightforward estimator $\widehat{\varphi} = \widehat{\mathcal{F}}^{-1}\widehat{\psi}$ is typically inconsistent. We discuss specific examples for nonparametric IV models and the related operators together with the respective literature in Section 2.

In this paper we describe and analyze a consistent estimator for this type of problem, when $\mathcal{F}$ is an operator between Hilbert spaces. The estimator is based on the iteratively regularized Gauß-Newton method (IRGNM) with iterated Tikhonov regularization defined below in (11). Details about the method will be given in Section 3.

This method was suggested by Bakushinskiĭ (1992). Important monographs on this topic are Bakushinskiĭ and Kokurin (2004) and Kaltenbacher, Neubauer and Scherzer (2008). These contributions consider only problems with known operators and deterministic right hand side in equation (1). The use of IRGNM for nonparametric IV problems was proposed and analyzed by Dunker et al. (2014). They derived rates for convergence in probability with a priori parameter choice using variational methods.

The novelty of this paper is that we prove significantly faster convergence rates for the mean integrated squared error (MISE) rather than convergence in probability under a different set of assumptions. In addition, we propose adaptive estimation with Lepskiĭ's principle and prove rates for this case. Furthermore, we assume a significantly weaker nonlinearity condition for the operator $\mathcal{F}$ which has a clear interpretation and is reasonable for most applications while the nonlinearity condition in Dunker et al. (2014) is difficult to interpret and to check. We also prove faster rates of convergence when the regression function is smooth enough. Our proofs do not use variational methods. Instead we rely on spectral methods as in Bauer, Hohage and Munk (2009). We also use a modification of Hoeffding's inequality from McDiarmid (1989).

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. The estimator is introduced in Section 3. Section 4 contains convergence rate theorems. Finally, we present some numerical simulations in Section 5. All proofs are in the Appendix.

2. Nonparametric instrumental variable models

In our general framework a function $\varphi^\dagger$ is characterized by the possibly nonlinear operator equation

$$\mathcal{F}(\varphi^\dagger) = 0, \tag{2}$$

i.e. $\varphi^\dagger$ is the true solution. Here $\mathcal{F} : B_{2R}(\varphi_0) \subseteq \mathbb{X} \to \mathbb{Y}$ is an operator between Hilbert spaces with norms $\|\cdot\|_{\mathbb{X}}$ and $\|\cdot\|_{\mathbb{Y}}$ respectively. Tyical examples for $\mathbb{X}$ and $\mathbb{Y}$ are L^2 and L^2 based Sobolev spaces H^i for $i = 1, 2, \ldots$. A ball $B_{2R}(\varphi_0)$ with radius $2R$ around an initial guess φ_0 is contained in the domain of $\mathcal{F}$. In practice, large values of R are possible. The operator equation is allowed to

be ill-posed, i.e. $\mathcal{F}^{-1}$ may not be continuous. Furthermore, the operator $\mathcal{F}$ is not known in applications. Only a series of estimators $\widehat{\mathcal{F}}_n : B_{2R}(\varphi_0) \subseteq \mathbb{X} \to \mathbb{Y}$ are available where n denotes the sample size. We assume that $\varphi^\dagger$ is a unique solution to (2) in B_{2R}, i.e. the problem is locally identified. In the following we discuss econometric examples for this setup.

2.1. Quantile regression and non-separable models

Nonparametric IV quantile regression The first example is nonparametric IV quantile regression proposed by Horowitz and Lee (2007). For $q \in [0,1]$ the q-th quantile regression function φ_q is characterized by

$$Y = \varphi_q(X) + U \qquad \mathbb{P}(U \leq 0 | Z = z) = q \qquad \text{for all } z. \tag{3}$$

Here and in all following models Y and U are univariate random variables, while X and Z can be multivariate and their dimensions do not have to coincide. The regressor X, the instrument Z and the response Y are observed, while the error term U is unobservable. If the joint density f_{YXZ} exists, the model is equivalent to an operator equation $\mathcal{F}_q(\varphi_q) = 0$ with

$$(\mathcal{F}_q(\varphi))(z) := \int F_{YXZ}(\varphi(x), x, z)\, dx - q f_Z(z) \tag{4}$$

where $F_{YXZ}(y,x,z) := \int_{-\infty}^{y} f_{YXZ}(\tilde{y}, x, z)\, d\tilde{y}$. Different estimation procedures for this model were proposed and analyzed in Horowitz and Lee (2007), Chen and Pouzo (2012), Gagliardini and Scaillet (2012a) Dunker et al. (2014), Breunig (2015), and Kaido and Wüthrich (2021). Local identification properties of this and related models are discussed in Chen et al. (2014).

We can write $(\mathcal{F}_q(\varphi))(z) = \int k_q(\varphi(x), x, z) dx$ with integral kernel

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

Replacing $q f_Z(z)$ by $\int q f_{XZ}(x,z) dx$ is impractical in applications but makes it easier to discuss properties of (4) in this paper. $\mathcal{F}_q$ and k_q are unknown and have to be estimated. If we plug-in a density estimator $\widehat{f}_{YXZ}$, we get straight forward estimators $\widehat{k}_q$ and $\widehat{\mathcal{F}}_q$.

Non-separable model A related example that falls in our framework is nonparametric IV regression with unseparable error, wich was proposed in Chernozhukov, Imbens and Newey (2007). See also Chernozhukov and Hansen (2005). The model is

$$\begin{aligned} Y = \phi(X, U) \qquad & \text{with } U \perp\!\!\!\perp Z \text{ and} \\ & \phi(x,u) \text{ strictly monotonic increasing in } u. \end{aligned} \tag{5}$$

It was pointed out in Horowitz and Lee (2007), and Chernozhukov, Imbens and Newey (2007) that this model is already contained in model (3). Let F_U

be the cumulative distribution function of U. Normalize $\widetilde{U} := F_U(U)$ and $\widetilde{\phi}(x,\tilde{u}) := \phi(x, F_U^{-1}(\tilde{u}))$. Then $\widetilde{U}$ is uniformly distributed on $[0,1]$. The value of $\widetilde{U}$ corresponds to a quantile in model (3). This reduces (5) to model (3) with $\varphi_q(x) = \widetilde{\phi}(x,q)$.

2.2. Nonparametric IV regression

Mean independence The simplest nonparametric IV regression model has a separable error term and a mean independence condition

$$Y = \varphi(X) + U \qquad \text{with } \mathbb{E}[U|Z] = 0. \tag{6}$$

This model was proposed by Newey and Powell (2003) and Florens (2003). It was further studied and applied in Hall and Horowitz (2005), Blundell, Chen and Kristensen (2007), Chen and Reiss (2011), Darolles et al. (2011) Florens, Johannes and Van Bellegem (2011), Horowitz (2011), Johannes, Van Bellegem and Vanhems (2011), Gagliardini and Scaillet (2012b), Chen and Pouzo (2012), Horowitz (2014a), Chen and Christensen (2015), Breunig and Johannes (2016), Chen and Christensen (2018), as well as Babii (2020) among others. For an overview see Horowitz (2014b).

We can write (6) equivalently as $\mathbb{E}[\varphi(X)|Z] = \mathbb{E}[Y|Z]$ and if the conditional densities $f_{X|Z}$ and $f_{Y|Z}$ exist, as

$$\int f_{X|Z}(x|z)\varphi(x)dx = \int y f_{Y|Z}(y|z)dy \qquad \text{for all } z \in \operatorname{supp}(Z). \tag{7}$$

We define the linear integral operator $(\mathcal{F}_{ce}\varphi)(z) := \int f_{X|Z}(x|z)\varphi(x)dx$ with integral kernel $f_{X|Z}(x|z)$ and the function $\psi(z) := \int y f_{Y|Z}(y|z)dy$. Model (6) can be given in operator form $(\mathcal{F}_{ce}\varphi)(z) = \psi(z)$. The integral kernel $f_{X|Z}$ and thereby $\mathcal{F}_{ce}$ as well as the function ψ are unknown and have to be estimated from a sample of Y, X, Z. An Density estimators $\widehat{f}_{X|Z}$ and $\widehat{f}_{Y|Z}$ give estimators $\widehat{\mathcal{F}}_{ce}$ and $\widehat{\psi}$ in a natural way. While the main focus of this paper is on nonlinear operator equations, we use model (6) as a benchmark for the IRGNM applied to model (8) below.

The model identifies the regression function φ if and only if $\mathcal{F}_{ce}$ is injective. This property is called completeness, see D'Haultfoeuille (2011), D'Haultfoeuille and Février (2015), Andrews (2017), and Babii and Florens (2020).

Full independence In many applications the error term can be assumed to be independent of the instrument. Hence, mean independence of the instrument can be replace by full independence as proposed in Dunker et al. (2014)

$$Y = \varphi(X) + U \qquad \text{with } U \perp\!\!\!\perp Z \text{ and } \mathbb{E}[U] = 0. \tag{8}$$

Since the new assumptions $U \perp\!\!\!\perp Z$ and $\mathbb{E}[U] = 0$ imply $\mathbb{E}[U|Z] = 0$ but not vice versa model (8) makes stronger assumptions than model (6). Consequently,

whenever (6) identifies the solution so does (8). Furthermore, there are cases in which (8) can identify a solution, while (6) fails. This is for example the case with discrete instruments and continuous regressors as discussed in Dunker et al. (2014), Torgovitsky (2015), D'Haultfoeuille and Février (2015), and Loh (2019).

We can translate model (8) into an operator equation by defining the operator

$$(\widetilde{\mathcal{F}}_{ind}(\varphi))(u,z) := \begin{pmatrix} \mathbb{P}[Y-\varphi(X)\le u] - \mathbb{P}[Y-\varphi(X)\le u|Z=z] \\ \mathbb{E}[Y-\varphi(X)] \end{pmatrix}. \tag{9}$$

When Y, X, Z have a joint density f_{YXZ}, taking the derivative with respect to u yields the alternative operator

$$(\mathcal{F}_{ind}(\varphi))(u,z) := \begin{pmatrix} \int f_{YXZ}(u+\varphi(x),x,z) - f_{YX}(u+\varphi(x),x) f_Z(z)\,dx \\ \int \varphi(x) f_X(x)\,dx - \int y f_Y(y)\,dy \end{pmatrix}. \tag{10}$$

Model (8) is equivalent to the operator equations $\widetilde{\mathcal{F}}_{ind}(\varphi) = 0$ or $\mathcal{F}_{ind}(\varphi) = 0$. Note that the operators are nonlinear due to the first line of (9) or (10). Furthermore, the operators are not known and have to be estimated. A density estimator $\widehat{f}_{YXZ}$ gives a straight forward estimator $\widehat{\mathcal{F}}_{ind}$.

For any φ that sets the first line of the operator $\mathcal{F}_{ind}$ to 0 also $c+\varphi$ with $c \in \mathbb{R}$ sets it to 0. In addition, for any φ, the second line of $\mathcal{F}_{ind}$ is set to 0 by $\varphi - \mathbb{E}[Y-\varphi(X)]$. Hence, for any solution φ of the first line of the operator we have $\mathcal{F}_{ind}\big(\varphi - \mathbb{E}[Y-\varphi(X)]\big) = 0$. The nonlinear inverse problem is to find a φ that solves the first line of $\mathcal{F}_{ind}$. The second line is a parametric problem that can be estimated with the parametric rate. When we discuss this example below we will only consider the first line of the operator as this is dominating the convergence rate.

Let us denote the integral kernel of the first line of the operator (10) and its estimator by $k_{ind}(y,x,z) := f_{YXZ}(y,x,z) - f_{YX}(y,x)f_Z(z)$ and $\widehat{k}_{ind}(y,x,z) := \widehat{f}_{YXZ}(y,x,z) - \widehat{f}_{YX}(y,x)\widehat{f}_Z(z)$ respectively. Then the first component of the operator reads $(\mathcal{F}_{ind}(\varphi))(u,z) = \int k_{ind}(u+\varphi(x),x,z)dx$.

Further examples We briefly comment on further econometric models that fall into the framwork of this paper. A problem that has a similar mathematical structure as IV regression appears in some nonparametric demand models for differentiated products. It was considered with mean independence assumption similar to (6) in Berry and Haile (2011), Berry and Haile (2014) and with full independence similar to (8) in Dunker, Hoderlein and Kaido (2014). Some models for games of incomplete information lead to a nonlinear inverse problem with deterministic operator, see for example Florens and Sbaï (2010). Nonlinear inverse problems with deterministic operators also occur in functional linear quantile regression (without instrumental variables) as in Kato (2012). The estimator in this paper can be applied to these type of problems. However, the error analysis would be different since there is no randomness in the operator. Also related are nonparametric ARCH(∞) models which can be treated as linear inverse problem, see Linton and Mammen (2005). Further linear inverse problems in econometrics are discussed in Carrasco, Florens and Renault (2007).

3. Estimation

3.1. The estimator

Remember that $\varphi^\dagger$ denotes the true solution and let φ_0 be an initial guess. Our method is based on linearizing $\mathcal{F}$ which motivates the following assumption.

Assumption 1. 1. $\|\varphi^\dagger - \varphi_0\|_{\mathbb{X}} < R$
2. $\mathcal{F}$ and all $\widehat{\mathcal{F}}_n$ are Fréchet differentiable on $B_{2R}(\varphi_0)$ with Fréchet derivatives $\mathcal{F}'$ and $\widehat{\mathcal{F}}_n'$ respectively.

The iteratively regularized Gauß-Newton method with iterated Tikhonov regularization consists of two nested iterations. The outer iteration is a Newton method. It starts at φ_0 and produces in the j-th step the estimate $\widehat{\varphi}_{j+1}$. In the j-th step the operator is linearized as $\widehat{\mathcal{F}}(\varphi) \approx \widehat{\mathcal{F}}_n'[\widehat{\varphi}_j](\varphi - \widehat{\varphi}_j) + \widehat{\mathcal{F}}_n(\widehat{\varphi}_j)$. A regular Newton method would invert the linear operator $\widehat{\mathcal{F}}_n'[\widehat{\varphi}_j]$ to compute the next step. Due to the ill-posedness, this would be unstable and we use a regularized inverse instead. The regularized inverse is computed by m-times iterated Tikhonov regularization which is the inner iteration of the method. In the following scheme the Newton iteration is indexed by j and the Tikhonov iteration by i, and $\alpha_j > 0$ is a regularization parameter

$$
\begin{aligned}
&\text{for } j = 0 \text{ to } J\\
&\quad \overline{\varphi}_{j+1,0} := \widehat{\varphi}_j\\
&\quad \text{for } i = 0 \text{ to } m-1\\
&\qquad \overline{\varphi}_{j+1,i+1} := \operatorname*{argmin}_{\varphi \in \mathbb{X}} \left(\|\widehat{\mathcal{F}}_n'[\widehat{\varphi}_j](\varphi - \widehat{\varphi}_j) + \widehat{\mathcal{F}}_n(\widehat{\varphi}_j)\|_{\mathbb{Y}}^2 + \alpha_j \|\varphi - \overline{\varphi}_{j+1,i}\|_{\mathbb{X}}^2 \right)\\
&\quad \text{end}\\
&\quad \widehat{\varphi}_{j+1} := \overline{\varphi}_{j+1,m}\\
&\quad \text{stop if } \|\widehat{\varphi}_j - \varphi_0\|_{\mathbb{X}} > 2R \text{ and set } \widehat{\varphi}_{j+1} = \varphi_0\\
&\text{end.}
\end{aligned}
\tag{11}
$$

As usual for Newton methods, convergence can fail if the initial guess $\widehat{\varphi}_0$ is too far from the true solution $\varphi^\dagger$. In practice and in simulations the method proves to be quite robust to the choice of $\widehat{\varphi}_0$. If no a priori information about $\varphi^\dagger$ is available, $\widehat{\varphi}_0 = 0$ is usually a good choice.

With a small α_j the method has a large variance due to the ill-posedness of $\mathcal{F}$. While a larger α_j controls the variance but adds some bias. We choose α_0 large enough to stabilize the problem and let α_j decay in every Newton step by

$$
\alpha_{j+1} = q_\alpha \alpha_j \quad \text{with some fixed} \quad 0 < q_\alpha < 1 \tag{12}
$$

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 for very smooth $\varphi^\dagger$. We will address the choice of α_0 and m in Section 4.1.1 and Assumption 3. The Newton

iteration needs to be stopped 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 a bias variance trade-off. We will investigate parameter choice with a priori knowledge in Section 4.2 and fully data driven in Section 4.4.

We introduce the following notations for shorter formulas

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

An alternative formulation of the method can be obtained by using the functional calculus. Let $\widehat{T}^*_{n,j}$ denote the adjoint operator of $\widehat{T}_{n,j}$ and set

$$g_\alpha(\lambda) := \frac{(\lambda+\alpha)^m - \alpha^m}{\lambda(\lambda+\alpha)^m}\ . \tag{13}$$

Then (11) is equivalent to

$$\begin{aligned}
&\text{for } j = 0 \text{ to } J \\
&\quad \widehat{\varphi}_{j+1} = \varphi_0 + g_{\alpha_j}(\widehat{T}^*_{n,j}\widehat{T}_{n,j})\widehat{T}^*_{n,j}\left(\widehat{T}_{n,j}(\widehat{\varphi}_j - \varphi_0) - \widehat{\mathcal{F}}_n(\widehat{\varphi}_j)\right) \\
&\quad \text{stop if } \|\widehat{\varphi}_j - \varphi_0\|_{\mathbb{X}} > 2R \text{ and set } \widehat{\varphi}_{j+1} = \varphi_0 \\
&\text{end.}
\end{aligned} \tag{14}$$

Example 1 (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 (10) exists when f_{YXZ} is partially differentiable in the first variable. The operator in (4) is differentiable without further assumptions.

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

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

The derivatives are linear integral operators with kernels $\frac{\partial}{\partial y}k_{ind}(\varphi(x),x,z)$ and $\frac{\partial}{\partial y}k_q(\varphi(x),x,z)$ respectively.

4. Convergence Rates

The convergence theory is presented in four steps. We start by introducing assumptions for the general operator equation (2) as well as for the IV regression models (10) and (4). Afterwards, we state convergence rate result for the MISE with a priori choice of the stopping parameter j. Then we compare the result to Horowitz and Lee (2007). The last step is a theorem with data driven choice of j by Lepskiĭ's principle.

4.1. Assumptions

4.1.1. Smoothness condition

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

$$\varphi_0 - \varphi^\dagger = \Lambda(T_\dagger^* T_\dagger)\omega\,, \qquad \omega \in \mathbb{X} \text{ with } \rho := \|\omega\|_{\mathbb{X}} \tag{15}$$

is called a 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 1, this definition can be interpreted in the following way. We assume for simplicity that $T_\dagger$ is compact which is for example the case if $\frac{\partial}{\partial y} k(\varphi^\dagger(x), x, z)$ is continuous. It was 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 a singular system for $T_\dagger$. The source condition (15) 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 ω fulfilling (15) only exists if Λ compensates the decay of the singular values in a way that $\langle e_0,\, v_t \rangle \Lambda(\sigma_t^2)^{-1}$ is square summable. 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. $\sigma_t \lesssim \exp(-c_\sigma t)$ and $\langle e_0,\, v_t \rangle \lesssim \exp(-c_{e_0} t)$ 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. A source condition with polynomial Λ is called a Hölder source condition, which is a concept that 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 holds when the operator is rescaled to $\|T_\dagger\| < 1$ and $\Lambda(x) = (-\ln(x))^{-p}$ with some $0 < p$. This choice of Λ was proposed by Mair (1994) and Hohage (1997) and is called logarithmic source condition.

Despite the word "condition" in the name "source condition" it is rather a relation that selects an index function. Corollary 2 in Mathé and Hofmann (2008) shows that for any compact injective operator $T_\dagger$ and any e_0 exists an index function Λ such that a source condition is fulfilled.

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

Assumption 2. The true solution $\varphi^\dagger$ fulfills a source condition (15) with sufficiently small ρ and with an index function that satisfies $\Lambda(x) = \mathcal{O}(x^\mu)$ for $x \searrow 0$ with $\mu > 1/2$.

Example 2. For the nonparametric IV examples (10) and (4) Assumption 2 implies that $\varphi_0 - \varphi^\dagger$ is in some Sobolev class if $\frac{\partial}{\partial y} k_{ind}(\varphi(x), x, z)$ or $\frac{\partial}{\partial y} k_q(\varphi(x), x, z)$ have Sobolev smoothness. If $\frac{\partial}{\partial y} k_{ind}(\varphi(x), x, z)$ or $\frac{\partial}{\partial y} k_q(\varphi(x), x, z)$ are analytic, $\varphi_0 - \varphi^\dagger$ must be infinitely smooth.

Closely related to the smoothness of the true solution is the choice of the parameters α_0 and m for the IRGNM. In the following assumption $\|T\|_{\mathcal{L}(\mathbb{X},\mathbb{X})} := \sup_\varphi \{\|T\varphi\|_\mathbb{X} \mid \|\varphi\|_\mathbb{X} = 1\}$ denotes the usual operator norm for linear operators.

Assumption 3.
1. The number of iterations of the Tikhonov regularization m is larger or equal to μ in the source conditions $m \geq \mu$, i.e. $\Lambda(x)^{-1} x^m = \mathcal{O}(1)$ for $x \searrow 0$.
2. The initial regularization parameter α_0 is large enough such that $\alpha_0 \geq \|\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger}\|_{\mathcal{L}(\mathbb{X},\mathbb{X})}/(1 - q_\alpha)$.

Both parameters need to be large enough but there is not much harm in choosing them larger than necessary. Any α_0 and m fulfilling Assumption 3 will lead to comparable estimates. However, increasing α_0 will lead to a few more Newton steps. The lower bound of α_0 depends on the derivative of the estimated operator and is thereby random but not unknown.

As usual for nonparametric methods the rate of convergence increases if the true solution $\varphi^\dagger$ is smoother, i.e. if μ is larger. But this increase is only realized if $m \geq \mu$. However, m does not act as a regularization parameter. Since the inner iteration is numerically cheap it is save to chose a larger value for m without having a significant disadvantage.

4.1.2. Nonlinearity restriction

The non-linearity of $\mathcal{F}$ needs to be restricted for the algorithm to work. We use a Lipschitz condition on the derivative for this purpose.

Assumption 4. There exists $L > 0$ such that

$$\|\widehat{\mathcal{F}}_n'[\xi_1] - \widehat{\mathcal{F}}_n'[\xi_2]\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} \leq L\|\xi_1 - \xi_2\|_\mathbb{X} \tag{16}$$

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

The special structure of the of $\mathcal{F}_{ind}$ and $\mathcal{F}_q$ allows us to replace Assumption 4 for the IV regression examples by the following alternative. Lemma 2 in the appendix shows that Assumption 5 implies Assumption 4.

Assumption 5. For the operators (10) and (4) the integral kernels k_{ind}, k_q, and their estimates are twice differentiable with respect to y with bounded derivative and the support of the instrument has finite measure $\mu(\operatorname{supp}(Z)) < \infty$. Furthermore, the integral kernels are estimated by an estimator which is strongly consistent for the second derivative. There exists $L > 0$ such that

$$\sqrt{\mu(\operatorname{supp}(Z))} \sup_{y,z,w} \left| \frac{\partial^2}{\partial y^2} k(y,z,w) \right| < L$$

with $k = k_{ind}$ or $k = k_q$ respectively.

Common nonparametric density estimators are strongly consistent. Assumption 5 implies for the operators (10) and (4)

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

respectively.

4.1.3. Concentration inequalities

The estimation error in the operator and its derivative needs to be bounded by exponential inequalities.

Assumption 6. There are constants $c_1, c_2, c_3, c_4 \geq 0$ such that for all $n \in \mathbb{N}$ and all $\tau \geq 0$

$$\mathbb{P}\left\{ \left| \|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}} - \mathbb{E}\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}} \right| \geq \sqrt{\tau \,\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}} \right)} \right\} \leq c_1 e^{-c_2 \tau} \text{ and } \tag{17}$$

$$\mathbb{P}\Bigg\{ \left| \|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu} \right) \right| \geq \sqrt{\tau \,\mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu} \right)} \Bigg\} \leq c_3 e^{-c_4 \tau}. \tag{18}$$

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

The following lemma shows that Assumption 6 holds for the IV regression applications (10) and (4) under mild conditions when $\mathbb{Y}$ is a L^2 space and $\|\cdot\|_D$ is the Hilbert-Schmidt norm. The Hilbert-Schmidt norm bounds the operator norm from above and is denoted by $\|\cdot\|_{HS}$. For linear integral operators it coincides with the L^2 norm of the integral kernel. We denote the space $L^2(\operatorname{supp}(U), \operatorname{supp}(Z))$ by $L^2(U,Z)$ and the space $L^2(\operatorname{supp}(Z))$ by $L^2(Z)$.

Lemma 1. *Consider the operators* (10) *and* (4) *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*. Assume for* (10) *that* $n^{-1}h^{-d_z-1} = O(1)$*, and* $n^{-1-2\mu}h^{-(1+\mu)(d_x+d_z+3)} = O(1)$*. Assume for* (4) *that* $n^{-1}h^{-d_z} = O(1)$*, and* $n^{-1-2\mu}h^{-(1+\mu)(d_x+d_z+2)} = O(1)$*. Then constants* $c_2, c_4 > 0$ *exist such that for all* $\tau \geq 0$

$$\mathbb{P}\left\{\left|\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2} - \mathbb{E}\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2}\right| \geq \sqrt{\tau\,\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{L^2}\right)}\right\} \leq 2e^{-c_2\tau} \tag{19}$$

and

$$\mathbb{P}\left\{\left|\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu}\right)\right| \geq \sqrt{\tau\,\mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu}\right)}\right\} \leq 2e^{-c_4\tau}. \tag{20}$$

4.2. Convergence rates with a priori parameter choice

Our first convergence rate theorem assumes that μ in Assumption 2 is known, i.e. the smoothness of the true solution is known. Adaptive estimation will be discussed in the next section.

Theorem 1. *Let Assumptions 1, 2, 3, 4, and 6 hold. Define the stopping index as*

$$J := \operatorname*{argmin}_{j\in\mathbb{N}} \left(\alpha_j^\mu + \alpha_j^{-1/2}\mathbb{E}\big[\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}^2\big]\right)$$

and set

$$J^* := \begin{cases} J & \text{if } \widehat{\varphi}_j \in B_{2R}(\varphi_0) \text{ for } j = 1,\dots,J \\ 0 & \text{else.} \end{cases} \tag{21}$$

Then,

$$\mathbb{E}\left[\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|_{\mathbb{X}}^2\right] = \mathcal{O}\left(\left(\mathbb{E}\big[\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}^2\big]\right)^{\frac{2\mu}{2\mu+1}} + \mathbb{E}\big[\|\widehat{T}_{n\dagger} - T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}^{2+2\mu}\big]\right).$$

In the special cases of the IV regression examples in L^2 spaces, the convergence rate can be given more explicitly.

Corollary 1. *Let* $\mathbb{X}$ *be an* L^2 *space and let Assumptions 1, 2, 3, 5 and the conditions of Lemma 1 hold. Assume that in the case of operator* (10) *the density* f_{YXZ} *and that in case of operator* (4) *the function* F_{YXZ} *is* r *times differentiable and is estimated with a kernel estimator where the kernel is of order at least* r*. If* J^* *is chosen as in Theorem 1, then*

$$\mathbb{E}\big[\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|_{L^2}^2\big] = \mathcal{O}\left((n^{-1}h^{-(d_Z+1)})^{\frac{2\mu}{2\mu+1}} + n^{-1-\mu}h^{-((1+2\mu)(d_X+d_Z+2)+1)} + h^{\frac{4\mu r}{2\mu+1}}\right).$$

4.3. Comparison to an alternative quantile regression estimator

We can compare our result to the rates for nonparametric quantile regression in Horowitz and Lee (2007). They use nonlinear Tikhonov regularization for the operator $\mathcal{F}_q$ and proved optimal rates under assumptions which are more restrictive than ours. In contrast to our rates, their rates do not depend on the derivative $\widehat{T}_{n\dagger}$. The main challenge for nonlinear Tikhonov regularization

$$\widehat{\varphi} = \operatorname*{argmin}_{\varphi} \|\mathcal{F}_q\varphi\|_{\mathbb{Y}}^2 + \alpha\|\varphi\|_{\mathbb{X}}^2$$

is to find the minimizer of the nonlinear functional $\|\mathcal{F}_q\varphi\|_{\mathbb{Y}}^2 + \alpha\|\varphi\|_{\mathbb{X}}^2$ which usually has multiple local minima. In Horowitz and Lee (2007) it is assumed that this minimum is known exactly which is unrealistic in practice. A convergence analysis which takes the performance of a minimization algorithm into account would typically lead to a different rate which also depends on some derivative depending on the particular minimization algorithm. The IRGNM does not have this problem since we only have to solve a linear least squares problem in every Newton step. For a fair comparison of the convergence rates, we will assume that the term $\left(\mathbb{E}\left[\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}^2\right]\right)^{\frac{2\mu}{2\mu+1}}$ dominates our rate. For sake of simplicity we assume $d_z = 1$. Hence, our rate in the case of $\mathbb{X} = L^2$ is

$$\mathbb{E}\left[\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|_{L^2}^2\right] = \mathcal{O}\left((n^{-1}h^{-2} + h^{2r})^{\frac{2\mu}{2\mu+1}}\right),$$

while the rate in Horowitz and Lee (2007) is $\mathcal{O}\left(n^{-(2\beta-1)/(2\beta+a)}\right)$ with a and β defined an their paper as α and β. Horowitz and Lee (2007) restricts the bandwidth choice to

$$h = C_h n^{-\gamma} \quad \text{with} \quad \frac{2\beta + a - 1)}{2r(2\beta + a} < \gamma < \frac{a}{2(2\beta + a)}.$$

Under their assumptions on a and β, the density estimator achieves the rate

$$n^{-1}h^{-2} + h^{2r} = \mathcal{O}\left(n^{-\frac{2\beta+a-1}{2\beta-1}}\right).$$

Note that this will only coincide with the optimal rate $n^{-\frac{2r}{2r+2}}$ in special cases. Futhermore, in their notation the source condition in our Assumption 2 holds for any $\mu < (\beta - \frac{1}{2})/a$. Hence, $2\mu/(2\mu+1) < (2\beta-1)/(2\beta+a-1)$, where μ can be chosen such that the left hand side is arbitrarily close to the right hand side. Therefore, under their assumptions our rate is arbitrarily close to

$$\left(n^{-\frac{2\beta+a-1}{2\beta-1}}\right)^{\frac{2\beta-1}{2\beta+a-1}} = n^{-\frac{2\beta-1}{2\beta+a}},$$

which is also the optimal rate in Theorem 2 and 3 in Horowitz and Lee (2007).

Our assumptions are less restrictive which leads to faster rates of convergence in many cases compared to Horowitz and Lee (2007). Firstly, we have no restrictions on h which means we can chose the optimal bandwidth $h = \mathcal{O}\left(n^{-\frac{1}{2r+2}}\right)$ and achieve

$$\mathbb{E}\left[\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|_{L^2}^2\right] = \mathcal{O}\left(n^{-\frac{2r}{2r+2}\frac{2\mu}{2\mu+1}}\right)$$

This has also the advantage that we can choose h by standard data driven bandwidth selectors like cross-validation. It is pointed out in Horowitz and Lee (2007) that a bandwidth selector for their assumptions does not yet exist.

Secondly, we have no upper bound on μ while their method requires $\mu \leq 1$. Hence, our rate can be significantly better for smooth $\varphi^\dagger$. This is not a contradiction to their optimality result, as their result only holds under their more restrictive assumptions.

4.4. Convergence rates for adaptive estimation

The parameter choice (21) in Theorem 1 and Corollary 1 depends on the unknown μ, which is unfeasible in practice. We present in this section convergence rates for a data driven choice of the stopping parameter J by Lepskiĭ's principle. This is a popular parameter choice rule in the context of statistical inverse problems, see Tsybakov (2000), Bauer and Hohage (2005), Mathé (2006), Bauer, Hohage and Munk (2009), and Hohage and Werner (2016). In the context of nonparametric IV, Lepskiĭ's principle was used for adaptive estimation in Chen and Christensen (2015). The following theorem gives convergence rates of the MISE with a Lepskiĭ type parameter choice. We lose a logarithmic factor compared to Theorem 1. The constant C_d and γ_{nl} used in the theorem are specified in the appendix in formula (31) and in Lemma 4 respectively.

Theorem 2. *Let the assumptions of Theorem 1 hold. For all sequences δ_n^{noi}, σ_n^{noi}, δ_n^{der}, and σ_n^{der} such that*

$$\delta_n^{noi} \geq \mathbb{E}(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}), \qquad (\sigma_n^{noi})^2 \geq \mathbb{V}\mathrm{ar}(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}),$$
$$\delta_n^{der} \geq \mathbb{E}(\|\widehat{T}_{n\dagger} - T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}^{1+\mu}), \qquad (\sigma_n^{der})^2 \geq \mathbb{V}\mathrm{ar}(\|\widehat{T}_{n\dagger} - T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}^{1+\mu})$$

set

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

and define the Lepskiĭ stopping parameter by

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

and the stopping parameter by

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

Then

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

This result still depends on expectation and variance of $\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}$ and $\|\widehat{T}_{n\dagger} - T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$ and usually on some smoothing parameter for the estimator $\widehat{\mathcal{F}}_n$, like the bandwidth h in the examples above. There are several strategies to derive these quantities depending on how $\widehat{\mathcal{F}}_n$ is estimated. Examples 3 and 4 in the appendix show that parameter choice for $\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}$ and $\|\widehat{T}_{n\dagger} - T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$ is equivalent to parameter choice in multivariate density estimation. Hence, the smoothing parameter or bandwidth can be chosen by adaptive density estimation, such as cross-validation. Simple variance bounds are available for most density estimators which can be used to derive σ_{noi}^2 and σ_{der}^2. For kernel density estimation such bounds are given in Corollary 2. A pilot estimate $\tilde{\varphi}$ for $\varphi^\dagger$ gives an estimator $\|\widehat{\mathcal{F}}_n(\tilde{\varphi})\|_{\mathbb{Y}}$ for $\mathbb{E}[\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}]$. The last unknown quantity $\mathbb{E}(\|\widehat{T}_{n\dagger} - T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})})$ can be derived from the bias of $\widehat{T}_{n\dagger}$ which can be estimated by a bootstrap or an oracle inequality, see Efron and Tibshirani (1994), Tsybakov (2008).

A simple alternative to the method above is to calibrate the unknown value of $\delta_n^{noi} + \ln((\sigma_n^{noi})^{-2})\sigma_n^{noi} + C_d\rho(\delta_n^{der} + \ln((\sigma_n^{der})^{-2})\sigma_n^{der})$ in $\widetilde{\Phi}_n^{noi}(j)$ with Monte Carlo simulations. We can design test examples that are similar to the data set at hand but with known true solution. Synthetic data is generated for these examples and the Lepskiĭ parameter choice for different possible values of $\delta_n^{noi} + \ln((\sigma_n^{noi})^{-2})\sigma_n^{noi} + C_d\rho(\delta_n^{der} + \ln((\sigma_n^{der})^{-2})\sigma_n^{der})$ is compared to the true solution. The value that produces the best results over different test examples and many repetitions is used to set up a surrogate for $\widetilde{\Phi}_n^{noi}(j)$ which is then used for the parameter choice on real data.

5. Numerical examples

We evaluate the small sample behavior of the estimator based on the IRGNM with simulated data. As a test problem we use a nonparametric IV regression consistent with models (6) and (8), with univariate covariate X and instrument Z. This setup allows to compare our estimator with an estimator which solves (7) with iterated Tikhonov regularization.

5.1. Implementation

The test problem described in the next section 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 an equidistant grid with $100 \times 100 \times 100$ nodes. The joint density is estimated on this grid by a standard adaptive density estimator. Trying different

density estimators, we found that both method are quite robust with respect to the density estimate. They tolerate some undersmoothing of the density as long as the stopping index J and the regularization parameter α are chosen properly. In the simulations below the same density estimate is used for both the IRGNM and the iterated Tikhonov regularization which allows for a fair comparison of the methods. The initial guess for both methods is the constant function with the value $\mathbb{E}[Y]$, and the penalty functional for both methods is the squared H^1 norm.

The Fréchet derivative is implemented as in Example 1. The partial derivative of the density and the derivatives for the H^1 norm are computed by the central differencing scheme. Operators and norms are evaluated using numerical integration. We tried rectangle rule, trapezoid rule and Simpson's rule but could not find a significant difference in the output of the estimator.

The least squares problems in each step of the iterated Tikhonov regularization and in the inner iteration of the IRGNM are computed by QR decomposition. Note that only one QR decomposition is needed in every Newton step and it avoids the explicit computation of the adjoint operator. We tried different numbers of iterations m in the inner iteration of the IRGNM and the iterated Tikhonov regularization for the test example below. No significant difference in the results was observed, which indicates that μ is not large.

The regularization parameters for the example below are $\alpha_0 = 1$ and $\alpha_{n+1} = 0.9\alpha_n$. Lepskiĭ's principle is used to find the stopping parameter of the Newton iteration. For the alternative estimator using model (6) the regularization parameter α has to be chosen instead, which is done by Lepskiĭ's principle as well. The iterated Tikhonov regularization is computed for a large number of different α. Then one of these approximation is chosen by Lepskiĭ's principle. We calibrated Lepskiĭ's principle for both methods in Monte Carlos simulations as described in Section 4.4. Hence, both methods are fully data driven.

5.2. Simulations

The regressor of the test example is generated by some function g and a random variable V such that $X = g(Z) + V$ and $V \perp\!\!\!\perp Z$. In addition, an exact solution $\varphi^\dagger$ and an error term U depending on V but not on Z are chosen. Then Y is defined as $Y := \varphi^\dagger(X) + U$. With this construction both models (6) and (8) identify the true solution. The functions and probability densities that were used for the test example are

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

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

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

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

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

The densities of V and U are constructed with Gaussians in a way that the expectation of U. Figure 1 shows the exact solution (blue) compared to the solution a nonparametric regression without instrumental variables would yield asymptotically (green).

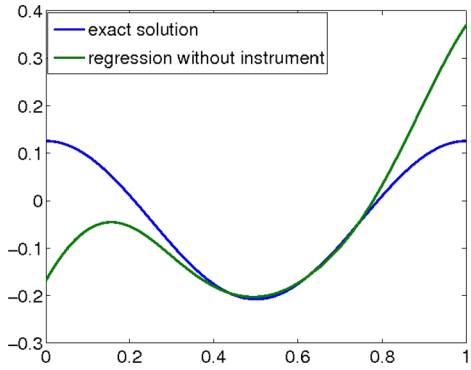

Fig 1: Necessity of the instrument: A standard nonparametric regression would asymptotically yield the green curve which is considerably different from the true curve $\varphi^\dagger$ (blue).

Both methods were tested on samples of 500 and 1000 observations. For each of the two sample sizes 1000 samples were generated and the joint density f_{YXZ} was estimated by a kernel method.

Figures 2 and 3 show histograms for the L^2 error of the reconstructions for both methods and different sample sizes. The values are normed by the initial error, so that on this scale the initial error becomes 1.

In Figure 2 we compare the errors of both methods for the sample size $n = 500$. Both methods produce acceptable results. The variance as well as the number of outliers observed for the method with independent instrument are significantly smaller than the variance or number of outliers of 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 cannot be observed for the IRGNM. In addition, the mean error of the IRGNM is smaller.

Similar histograms for sample size $n = 1000$ are displayed in Figures 3. Both methods perform well. The advantages of the IRGNM with less outliers, smaller variance and smaller mean error can be observed again. The following table provides the mean and some quantiles of the errors normed by the initial error.

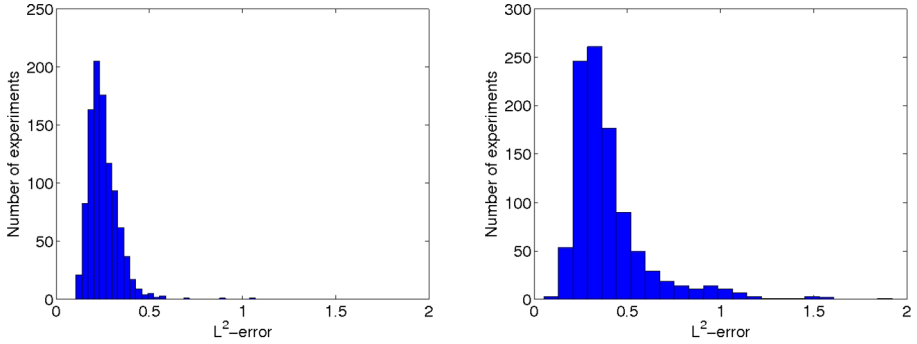

Fig 2: L^2 error for the sample size $n = 500$. Left panel: IRGNM with the assumption $U \perp\!\!\!\perp Z$. Right panel: iterated Tikhonov regularization with the assumption $\mathbb{E}[U|Z] = 0$

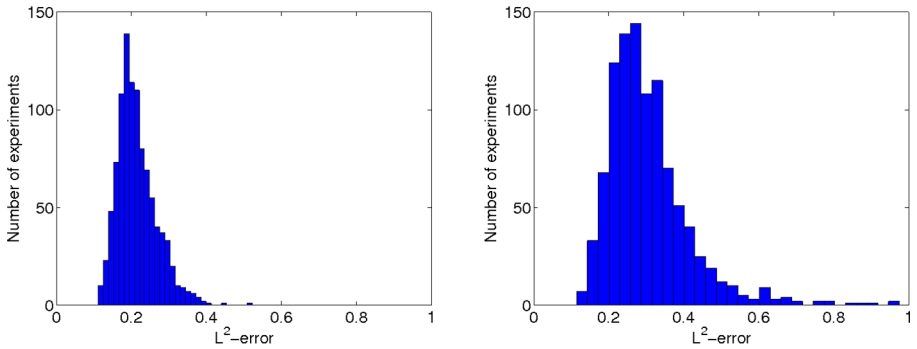

Fig 3: L^2 error for the sample size $n = 1000$. Left panel: IRGNM with the assumption $U \perp\!\!\!\perp Z$. Right panel: iterated Tikhonov regularization with the assumption $\mathbb{E}[U|Z] = 0$.

sample size	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
$n = 500$	$\mathbb{E}[U\|W] = 0$	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
$n = 1000$	$\mathbb{E}[U\|W] = 0$	0.3067	0.2339	0.2846	0.3482	0.4325

We close this section with examples of median reconstructions for both sample sizes displayed in Figure 4. They illustrate the advantage of the regression model with independent instruments solved with the IRGNM.

These results suggest that both methods give consistent estimators for the nonparametric instrumental regression with clear advantages for the regression model with independent instruments (8) and the IRGNM.

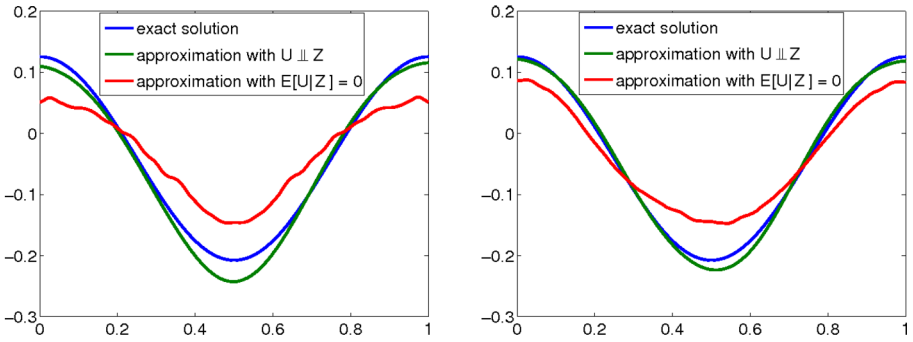

Fig 4: Examples for reconstructions with sample size $n = 500$ (left) and $n = 1000$ (right). The blue line shows the exact solution, the red curve the reconstruction with the conditional mean assumption and the green curve the reconstruction with independent instrument.

Appendix A: Proofs

A.1. Nonlinearity restriction

We prove in this section that for the IV regression examples Assumption 5 is an alternative to Assumption 4.

Lemma 2. *For the operator equations* (10) *and* (4) *Assumption 5 implies Assumption 4.*

Proof. Let $\widehat{k}_n(\xi_1(x), x, z)$ denote the estimate for k_{ind} or k_q. Since the second derivatives with respect to y are bounded we have

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

If $\widehat{k}_n$ is a strongly consistent estimator, for any constant $c > 0$

$$\sup_{y,x,z} \left|\frac{\partial^2}{\partial y^2}\widehat{k}_n(y, x, z)\right| \leq \sup_{y,x,z} \left|\frac{\partial^2}{\partial y^2}k(y, x, z)\right| + c \qquad \text{almost surely for large n.}$$

Hence, Assumption 4 holds with

$$L = \sqrt{\mu(\operatorname{supp}(Z))} \sup_{y,x,z} \left|\frac{\partial^2}{\partial y^2}k(y, x, z)\right| + c$$

for any $c > 0$. □

A.2. Concentration inequalities

This section proves Lemma 1 using McDiamid's extension of Hoeffding's inequality.

Theorem 3 (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 $1 \le i \le n$*

$$\sup_{\substack{(w_1,\dots,w_n),(w_1',\dots,w_n') \\ \in \operatorname{supp}(W_1,\dots,W_n)}} |f(w_1,\dots,w_n) - f(w_1,\dots,w_{i-1},w_i',w_{i+1},\dots,w_n)| \le c_i. \quad (22)$$

Then for $\tau \ge 0$

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

Now we can prove Lemma 1.

Proof. (of Lemma 1) Denote the kernels by

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

The proof has 4 parts in which we prove for each of the operators (10) and (4) the inequalities (19) and (20).

part 1 (19) for operator (10)

We show (22) with $W = (Y, X, Z)$ and

$$\begin{aligned} &f\big((y_1,x_1,z_1),\dots,(y_n,x_n,z_n)\big) := \|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)(u,z)\|_{L^2(U,Z)} \\ &= \Big\| \int n^{-1}\sum_{k=1}^n K_{Y,h}(\varphi^\dagger(x) - u - y_k)K_{X,h}(x - x_k)K_{Z,h}(z - z_k) \\ &\qquad - n^{-2}\sum_{k=1}^n\sum_{l=1}^n K_{Y,h}(\varphi^\dagger(x) - u - y_k)K_{X,h}(x - x_k)K_{Z,h}(z - z_l)dx\Big\|_{L^2(U,Z)}. \end{aligned}$$

Iterated application of the triangular inequality and dropping the terms that cancel in the sums yields

$$\begin{aligned} &|f(w_1,\dots,w_n) - f(w_1,\dots,w_{i-1},w_i',w_{i+1},\dots,w_n)| \\ &\le n^{-1}\Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i)K_{X,h}(x - x_i)K_{Z,h}(z - z_i) \\ &\qquad\qquad - K_{Y,h}(\varphi^\dagger(x) - u - y_i')K_{X,h}(x - x_i')K_{Z,h}(z - z_i')dx\Big\|_{L^2(U,Z)} \\ &+ n^{-2}\Big\| \int \sum_{k\ne i} K_{Y,h}(\varphi^\dagger(x) - u - y_k)K_{X,h}(x - x_k) \\ &\qquad\qquad \big[K_{Z,h}(z - z_i') - K_{Z,h}(z - z_i)\big]dx\Big\|_{L^2(U,Z)} \end{aligned}$$

$$
\begin{aligned}
&+ n^{-2}\Big\| \int \sum_{l\neq i} \big[K_{Y,h}(\varphi^\dagger(x) - u - y_i')K_{X,h}(x - x_i') \\
&\qquad\qquad - K_{Y,h}(\varphi^\dagger(x) - u - y_i)K_{X,h}(x - x_i)\big] K_{Z,h}(z - z_l)dx\Big\|_{L^2(U,Z)} \\
&+ n^{-2}\Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i')K_{X,h}(x - x_i')K_{Z,h}(z - z_i') \\
&\qquad\qquad - K_{Y,h}(\varphi^\dagger(x) - u - y_i)K_{X,h}(x - x_i)K_{Z,h}(z - z_i)dx\Big\|_{L^2(U,Z)}
\end{aligned}
$$

We use the triangular inequality again and substitute x_i, y_i, z_i for x_i', y_i', z_i' and for x_k, y_k, z_l.

$$
\begin{aligned}
&\le 2n^{-1}\Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i)K_{X,h}(x - x_i)K_{Z,h}(z - z_i)dx\Big\|_{L^2(U,Z)} \\
&\quad + 2n^{-2}\Big\| \int \sum_{k\neq i} K_{Y,h}(\varphi^\dagger(x) - u - y_k)K_{X,h}(x - x_k)K_{Z,h}(z - z_i)dx\Big\|_{L^2(U,Z)} \\
&\quad + 2n^{-2}\Big\| \int \sum_{l\neq i} K_{Y,h}(\varphi^\dagger(x) - u - y_i)K_{X,h}(x - x_i)K_{Z,h}(z - z_l)dx\Big\|_{L^2(U,Z)} \\
&\quad + 2n^{-2}\Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i')K_{X,h}(x - x_i')K_{Z,h}(z - z_i')dx\Big\|_{L^2(U,Z)} \\
&\le \left(\frac{2}{n} + \frac{4(n-1)}{n^2} + \frac{2}{n^2}\right)\Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i')K_{X,h}(x - x_i') \\
&\qquad\qquad K_{Z,h}(z - z_i')dx\Big\|_{L^2(U,Z)} \\
&< 6n^{-1}\Big\| \int K_{Y,h}(\varphi^\dagger(x) - u - y_i')K_{X,h}(x - x_i')K_{Z,h}(z - z_i')dx\Big\|_{L^2(U,Z)} \\
&= 6n^{-1}\left(\int\Big(\int K_{Y,h}(\varphi^\dagger(hx + x_i) - u - y_i)K_X(x)K_{Z,h}(z - z_i)dx\Big)^2 d(u,z)\right)^{1/2} \\
&\le 6n^{-1}\left(\int K_X^2(x)\int K_{Y,h}^2(\varphi^\dagger(hx + x_i) - u - y_i)K_{Z,h}^2(z - z_i)d(u,z)\,dx\right)^{1/2} \\
&= 6n^{-1}\left(\int K_X^2(x)\|K_{Y,h}(u)K_{Z,h}(z)\|_{L^2(U,Z)}^2 dx\right)^{1/2} \\
&= 6n^{-1}\|K_X\|_{L^2}\|K_{Y,h}(u)K_{Z,h}(z)\|_{L^2(U,Z)} \\
&= 6n^{-1}h^{-(d_Z+1)/2}\|K_X\|_{L^2}\|K_Y(u)K_Z(z)\|_{L^2(U,Z)}.
\end{aligned}
$$

We used the standard substitution arguments for kernel methods to get from $K_{X,h}$ to K_X and from $K_{Y,h}K_{Z,h}$ to K_YK_Z. Together with Theorem 3 this proves

$$
\begin{aligned}
&\mathbb{P}\Big\{\Big|\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2} - \mathbb{E}\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}\Big| \ge \sqrt{\tau}\Big\} \\
&\qquad\qquad \le 2\exp\left(\frac{-\tau n h^{d_Z+1}}{18\|K_X\|_{L^2}^2\|K_YK_Z\|_{L_2}^2}\right). \qquad (23)
\end{aligned}
$$

Hence, there exists a constant $c_2 > 0$ such that

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

part 2 (19) for operator (4)

A similar argument applies to the quantile regression operator in (4). We set $\bar{K}_{Y,h}(y) = \int_{-\infty}^{y} K_Y(\tilde{y})d\tilde{y}$ and $\bar{C}_Y := |\sup_y \bar{K}_{Y,h}(y) - \inf_t \bar{K}_{Y,h}(y)|$. Note that $\bar{C}_Y = |\sup_y \bar{K}_{Y,1}(y) - \inf_y \bar{K}_{Y,1}(y)|$ does not depend on h. Theorem 3 is now applied with

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

The estimation

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

proves together with Theorem 3

$$\mathbb{P}\left\{\left|\|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2} - \mathbb{E}\|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2}\right| \geq \sqrt{\tau}\right\} \leq 2\exp\left(\frac{-2\tau nh^{d_Z}}{\bar{C}_Y^2(1 + 2q)^2\|K_Z\|_{L_2}^2}\right). \quad (24)$$

Hence, there exists a constant $c_2 > 0$ such that (20) holds for $\widehat{\mathcal{F}}_q$.

part 3 (20) for operator (10)

We follow the same strategy and adopt the notation above. Theorem 3 is applied with $W = (Y, X, Z)$ with

$$\begin{aligned} &f\big((y_1,x_1,z_1),\ldots,(y_n,x_n,z_n)\big) := \|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} \\ &= \Bigg\|n^{-1}\sum_{i=1}^{n} K_{Y,h}'(\varphi^\dagger(x) - u - y_i)K_{X,h}(x - x_i)K_{Z,h}(z - z_i) \\ &\qquad - n^{-2}\sum_{k=1}^{n}\sum_{l=1}^{n} K_{Y,h}'(\varphi^\dagger(x) - u - y_k)K_{X,h}(x - x_k)K_{Z,h}(z - z_l) \end{aligned}$$

$$- \frac{\partial}{\partial y} f_{YXZ}(\varphi^\dagger(x) - u, x, z) + \frac{\partial}{\partial y} f_{YX}(\varphi^\dagger(x) - u, x) f_Z(z) \Big\|_{L_2}^{1+\mu}$$

where $K'_{Y,h}$ is the derivative of $K_{Y,h}$. With the same steps as in part 1 with repeated application of the triangular inequality and substitution we get

$$\begin{aligned}
&|f(w_1, \dots, w_n) - f(w_1, \dots, w_{i-1}, w'_i, w_{i+1}, \dots, w_n)| \\
&\le \big\| n^{-1} K'_{Y,h}(\varphi^\dagger(x) - u - y_i) K_{X,h}(x - x_i) K_{Z,h}(z - z_i) \\
&\quad - n^{-1} K'_{Y,h}(\varphi^\dagger(x) - u - y'_i) K_{X,h}(x - x'_i) K_{Z,h}(z - z'_i) \\
&\quad - n^{-2} \sum_{k=1}^n \sum_{l=1}^n K'_{Y,h}(\varphi^\dagger(x) - u - y_k) K_{X,h}(x - x_k) K_{Z,h}(z - z_l) \\
&\quad - n^{-2} \sum_{k=1}^n \sum_{l=1}^n K'_{Y,h}(\varphi^\dagger(x) - u - y'_k) K_{X,h}(x - x'_k) K_{Z,h}(z - z'_l) \big\|_{L_2}^{1+\mu} \\
&\le 6^{1+\mu} n^{-1-\mu} \big\| K'_{Y,h}(\varphi^\dagger(x) - u - y_i) K_{X,h}(x - x_i) K_{Z,h}(z - z_i) \big\|_{L_2}^{1+\mu} \\
&= 6^{1+\mu} n^{-1-\mu} \|K'_{Y,h}\|_{L_2}^{1+\mu} \|K_{X,h}\|_{L_2}^{1+\mu} \|K_{Z,h}\|_{L_2}^{1+\mu} \\
&= 6^{1+\mu} n^{-1-\mu} h^{-\frac{(1+\mu)(d_X+d_Z+3)}{2}} \|K'_Y\|_{L_2}^{1+\mu} \|K_X\|_{L_2}^{1+\mu} \|K_Z\|_{L_2}^{1+\mu}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{P}\Big\{ \Big| \|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} - \mathbb{E}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|_{HS}^{1+\mu} \right) \Big| \ge \sqrt{\tau} \Big\} \\
&\qquad \le 2 \exp\left(\frac{-2\tau n^{1+2\mu} h^{(1+\mu)(d_X+d_Z+3)}}{36^{1+\mu} \|K_X\|_{L^2}^{2+2\mu} \|K_Y\|_{L^2}^{2+2\mu} \|K_Z\|_{L_2}^{2+2\mu}} \right).
\end{aligned} \tag{25}$$

Thus, there exist a constant c_4 such that

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

part 4 (20) for operator (4)

A similar argument holds for the instrumental quantile regression problem. The kernel of the Fréchet derivative of the operator $\mathcal{F}_q$ in (4) at $\varphi^\dagger$ is simply $f_{YXZ}(\varphi^\dagger(x), x, z)$. So Theorem 3 is applied to

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

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

$$|f(w_1, \dots, w_n) - f(w_1, \dots, w_{i-1}, w'_i, w_{i+1}, \dots, w_n)|$$

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

This implies

$$\begin{aligned}
&\mathbb{P}\Big\{\Big|\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}-\mathbb{E}\Big(\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\Big)\Big|\geq\sqrt{\tau}\Big\}\\
&\qquad\qquad\leq 2\exp\left(\frac{-2\tau n^{1+2\mu}h^{(1+\mu)(2+d_X+d_Z)}}{C_Y^2\|K_X\|_{L_2}^{2+2\mu}\|K_Z\|_{L_2}^{2+2\mu}}\right).
\end{aligned}\tag{26}$$

Hence, there exist a constants c_4 such that

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

□

Corollary 2. *Under the assumptions of Lemma 1 we have for the operator* $\mathcal{F}_{ind}$ *in* (10)

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}\right)&\leq\frac{18\|K_X\|_{L^2}^2\|K_YK_Z\|_{L_2}^2}{nh^{d_Z+1}}=\mathcal{O}(n^{-1}h^{-d_Z-1})\\
\mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\right)&\leq\frac{36^{1+\mu}\|K_X\|_{L^2}^{2+2\mu}\|K_Y\|_{L^2}^{2+2\mu}\|K_Z\|_{L_2}^{2+2\mu}}{2n^{1+2\mu}h^{(1+\mu)(d_X+d_Z+3)}},
\end{aligned}$$

and for the operator $\mathcal{F}_q$ *in* (4)

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_q(\varphi^\dagger)\|_{L_2}\right)&\leq\frac{\bar{C}_Y^2(1+2q)^2\|K_Z\|_{L_2}^2}{2nh^{d_Z}}=\mathcal{O}(n^{-1}h^{-d_Z})\\
\mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|_{HS}^{1+\mu}\right)&\leq\frac{C_Y^2\|K_X\|_{L_2}^{2+2\mu}\|K_Z\|_{L_2}^{2+2\mu}}{2n^{1+2\mu}h^{(1+\mu)(2+d_X+d_Z)}}
\end{aligned}$$

Proof. It follows from (23) that $\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}$ is a subgaussian random variable. The moment condition for subgaussian variables implies that

$$\mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_{ind}(\varphi^\dagger)\|_{L_2}\right)=\frac{18\|K_X\|_{L^2}^2\|K_YK_Z\|_{L_2}^2}{nh^{d_Z+1}}\ .$$

The other three statements follow in the same way from (25), (24), and (26) respectively. □

A.3. Error analysis

In this section we prepare the proofs of the convergence rate theorems by decompose the error $e_{j+1}:=\widehat{\varphi}_{j+1}-\varphi^\dagger$ of our method (14) into different components and derive estimates for each component.

A.3.1. Error decomposition

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

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

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

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

In the decomposition we use a function r_α defined as $r_\alpha(\lambda) := 1 - \lambda g_\alpha(\lambda)$. With g_α as in (13) we have: $r_\alpha(\lambda) = \left(\frac{\alpha}{\lambda+\alpha}\right)^m$.

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

A similar related decomposition without e_{j+1}^{der} was proposed in Bakushinskiĭ (1992) for the case of known operators. In the rest of the section we will derive bounds on each error component.

A.3.2. Approximation error

Assumption 3 implies the existence of a constant C_Λ such that

$$\sup_{0 < x \le \|\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger}\|} r_\alpha(x) \Lambda(x) \le C_\Lambda \Lambda(\alpha) \qquad \text{for all } \alpha \ge 0.$$

Hence, with $\rho = \|\omega\|$ the approximation error is bounded by

$$\|e_{j+1}^{app}\|_{\mathbb{X}} \le C_\Lambda \Lambda(\alpha_j) \rho. \tag{27}$$

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

$$\begin{aligned} &\|e_{j+1}^{app}\|_{\mathbb{X}} \le \|e_j^{app}\|_{\mathbb{X}} \le \gamma_{app} \|e_{j+1}^{app}\|_{\mathbb{X}} && \text{for } j \ge 1 \\ &\text{and } \|e_0^{app}\|_{\mathbb{X}} \le \gamma_{app} \|e_1^{app}\|_{\mathbb{X}} && \text{since } \alpha_0 \ge \frac{\|\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger}\|_{\mathcal{L}(\mathbb{X},\mathbb{X})}}{1 - q_\alpha}. \end{aligned} \tag{28}$$

Note that the bound on the approximation error behave like a bias term. It tends to 0 with increasing j because α_j is decreasing while Λ is strictly increasing and $\Lambda(0) = 0$.

A.3.3. 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 for any linear bounded operator $T : \mathbb{X} \to \mathbb{Y}$ and $\psi \in \mathbb{Y}$

$$\begin{aligned}
\|g_\alpha(TT^*)\|_{\mathcal{L}(\mathbb{Y},\mathbb{Y})} &\le \|g_\alpha\|_\infty = \sup_{x\ge 0}\left(\frac{(x+\alpha)^m - \alpha^m}{x(x+\alpha)^m}\right) \le \frac{m}{\alpha}\ , \\
\|g_\alpha(TT^*)TT^*\|_{\mathcal{L}(\mathbb{Y},\mathbb{Y})} &\le \sup_{x\ge 0}|g_\alpha(x)x| = \sup_{x\ge 0}\left(\frac{(x+\alpha)^m - \alpha^m}{(x+\alpha)^m}\right) = 1\ , \text{ and} \\
\|g_\alpha(T^*T)T^*\psi\|_{\mathbb{X}}^2 &= \langle g_\alpha(TT^*)\psi,\, g_\alpha(TT^*)TT^*\psi\rangle_{\mathbb{Y}} \\
&\le \sup_{x\ge 0}|xg_\alpha(x)|\|g_\alpha\|_\infty\|\psi\|_{\mathbb{Y}}^2 \le \frac{m}{\alpha}\|\psi\|_{\mathbb{Y}}^2.
\end{aligned} \tag{29}$$

where $\|\cdot\|_\infty$ denotes the sup norm. Hence,

$$\begin{aligned}
&\|e_{j+1}^{noi}\|_{\mathbb{X}} = \|g_{\alpha_j}(\widehat{T}_{n,j}^*\widehat{T}_{n,j})\widehat{T}_{n,j}^*\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{X}} \le \sqrt{\frac{m}{\alpha_j}}\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}} \quad \text{and} \\
&\mathbb{E}\left(\|e_{j+1}^{noi}\|_{\mathbb{X}}^2\right) \le \frac{m}{\alpha_j}\mathbb{E}\left(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}^2\right).
\end{aligned} \tag{30}$$

Note that 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 $\varphi^\dagger$. It behaves like a variance term, i.e. it increases with a decreasing regularization parameter α_j. In addition to the bound (30) a concentration inequality is needed to bound the MISE of the Newton method. This is part of Assumption 6. The following example illustrates the asymptotic behavior of the bound (30) for the nonparametric IV operators (10) and (4).

Example 3. We work under the assumptions of Lemma 1. We need different smoothness condition for the operators $\mathcal{F}_{ind}$ and $\mathcal{F}_q$. Assume for $\mathcal{F}_{ind}$ that all derivatives of degree r of the density f_{YXZ} exist and are bounded. For the operator in (4) 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 also gives straight forward estimators $\widehat{f}_Y$, $\widehat{f}_{YZ}$, $\widehat{F}_{YXZ}$, $\widehat{f}_{XZ}$, and consequently for $\widehat{k}_{ind}$ and $\widehat{\mathcal{F}}_{ind}$, and for $\widehat{k}_q$ and $\widehat{\mathcal{F}}_q$. The convergence rates of k_{ind} and k_q will be dominated by $\widehat{f}_{YXZ}$ and $\widehat{F}_{YXZ}$ respectively.

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

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

It follows from Corollary 2 that

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

$$= \mathcal{O}(n^{-\frac{2r}{2r+d_Z+1}}) \quad \text{when } h \sim n^{-\frac{1}{2r+d_Z+1}} \qquad \text{and}$$

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

With the bound in (30) the rate of the MISE of the propagated noise error is

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

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

A.3.4. Derivative noise error

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 $\|e_{j+1}^{der}\|_{\mathbb{X}} \leq \rho\|\Lambda(T_\dagger^* T_\dagger) - \Lambda(\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger})\|_{\mathcal{L}(\mathbb{X},\mathbb{X})}$. Here the norm on the right hand side of the inequality is the usual operator norm. A way to simplify the term $\|\Lambda(T_\dagger^* T_\dagger) - \Lambda(\widehat{T}_{n\dagger}^* \widehat{T}_{n\dagger})\|_{\mathcal{L}(\mathbb{X},\mathbb{X})}$ is provided by the following lemma.

Lemma 3 (Egger (2005) Lemma 3.2.)**.** *For two linear bounded operators between Hilbert spaces A and B and $\mu > \frac{1}{2}$ there exists a constant c_μ such that with the corresponding operator norms*

$$\|(A^*A)^\mu - (B^*B)^\mu\| \leq c_\mu \|A - B\| \left|\|A\| - \|B\|\right|^\mu .$$

Hence, with some constant C_d and a norm $\|\cdot\|_D$ which is either the operator norm or some norm dominating the operator norm

$$\|e_{j+1}^{der}\|_{\mathbb{X}} \leq C_d \rho \|\widehat{T}_{n\dagger} - T_\dagger\|_D^{1+\mu} . \tag{31}$$

The bound in (31) is independent of the regularization parameter α and of the number of Newton steps j. It depends only on the noise in the Fréchet derivative of $\mathcal{F}$ at $\varphi^\dagger$. In addition to this bound we have to assume a concentration inequality for the right hand side of (31) which is part of Assumption 6.

The following example interprets $\|e_{j+1}^{der}\|_{L_2}$ for the IV regression operators (10) and (4) in the setup of the previous example.

Example 4. We adopt the assumptions and constructions of $\widehat{\mathcal{F}}_{ind}$, $\widehat{\mathcal{F}}_q$, $\widehat{k}_{ind}$ and $\widehat{k}_q$ from Example 3. When Assumption 1 holds, the Fréchet derivatives have the form

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

$$\widehat{\mathcal{F}}'_q[\varphi]\psi(z) = \int \frac{\partial}{\partial y}\widehat{k}_q(\varphi(x),\, x,\, z)\psi(z)\, dz.$$

The Hilbert-Schmidt norm bounds the operator norm from above and is the L^2 norm of the integral kernels $\frac{\partial}{\partial y}\widehat{k}_{ind}(\varphi(x) + u, x, z)$ and $\frac{\partial}{\partial y}\widehat{k}_q(\varphi(x), x, z)$. We

will denote the Hilbert-Schmidt norm by $\|\cdot\|_{HS}$. We introduce the notation $\kappa(u,x,z) := \frac{\partial}{\partial y}k_{ind}(u,x,z)$ and $\widehat{\kappa}_{n,h}(u,x,z) := \frac{\partial}{\partial y}\widehat{k}_{ind}(u,x,z)$ 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.

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

Jensen's inequality was 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

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

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

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

A similar computation can be carried out for the quantile regression problem with operator (4).

A.3.5. Nonlinearity error

A restriction on the nonlinearity of $\widehat{\mathcal{F}}_n$ is necessary to control $\|e_{j+1}^{nl}\|$.A suitable constraint is the Lipschitz condition (16) in Assumption 4. It allows to bound the Taylor reminder of the first term in the nonlinearity error by

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

For the norm of the second term in e_{j+1}^{nl} 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 $\|[r_\alpha(A^*A) - r_\alpha(B^*B)](B^*B)^\mu\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} \leq C_\mu\|A - B\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}$. This yields in our case

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

Putting both estimates together and use (29) gives

$$\|e_{j+1}^{nl}\|_{\mathbb{X}} \leq \frac{L\sqrt{m}}{2\sqrt{\alpha_j}}\|e_j\|_{\mathbb{X}}^2 + C_\mu\rho L\|e_j\|_{\mathbb{X}}. \tag{32}$$

The next Lemma computes an appropriate stopping parameter $J_{\max}$ such that $\|e_j^{nl}\|$ is dominated by the other error components for all $j \leq J_{\max}$.

Lemma 4. *Let Assumptions 1, 2, 4, 3 hold true with a sufficiently small ρ in Assumption 2. 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 $\|e_j^{noi} + e_j^{der}\|_{\mathbb{X}} \leq \Phi(j)$ for all $j \geq 0$. Define*

$$\begin{aligned}J_{\max} &:= \max\left\{j \in \mathbb{N} : \frac{\Phi(j)}{\sqrt{\alpha_j}} \leq C_{stop}\right\} \\ \text{with } 0 < C_{stop} &\leq \min\left\{\frac{1}{8L\sqrt{m}}, \frac{R}{4\sqrt{\alpha_0}}\right\}.\end{aligned} \tag{33}$$

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

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

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

Proof. We generalize the proof strategy of Lemma 2.2 in Bauer, Hohage and Munk (2009) to our setting. The proposition follows by induction in 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 (28)

$$\|e_{j-1}\| \leq (1 + \gamma_{nl})\left(\|e_{j-1}^{app}\| + \Phi(j-1)\right)$$

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

Combining this with inequality (32) and using $(a+b)^2 \leq 2a^2+2b^2$ yields

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

If $\rho \leq \gamma_{nl}/(2C_\mu(1+\gamma_{nl})\gamma_{app})$, 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 $\rho \leq \gamma_{nl}/(2C_\Lambda\alpha_0^{\mu-1/2}L\sqrt{m}(1+\gamma_{nl})^2\gamma_{app}^2)$. It follows from (27) that

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

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

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

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

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

The base case $j=1$ of the induction follows in exactly the same way, as long as α_0 is large enough which is guaranteed by Assumption 3 and (28).

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

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

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

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

This shows together with the first part of the proof that

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

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

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.

A.4. Convergence rates with a priori parameter choice

This section presents the proof to Theorem 1. We generalize the proof strategy in Bauer, Hohage and Munk (2009) to our setting. We start with a lemma about deterministic errors, i.e. $0 = \mathbb{V}\mathrm{ar}(\|\widehat{\mathcal{F}}(\varphi^\dagger)\|_{\mathbb{Y}}) = \mathbb{V}\mathrm{ar}(\|\widehat{T}_{n\dagger}\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})})$. The crucial point is to show that the maximal stopping parameter $J_{\max}$ from in Lemma 4 is larger or equal to a suitable stopping parameter.

Lemma 5. *Suppose that the Assumptions 1, 2, 3, and 4 are fulfilled. Assume that $B_{2R}(\varphi_0) \subset \mathrm{dom}(\mathcal{F})$, and that ρ is small enough as in Lemma 4. Let $\widetilde{\delta}_n^{noi}$ and $\widetilde{\delta}_n^{der}$ be a sequence such that $\widetilde{\delta}_n^{noi} \geq \|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}$ and $\widetilde{\delta}_n^{der} \geq \|\widehat{T}_{n\dagger} - T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}^{1+\mu}$. Set*

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

Then there exists a constant C such that

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

Proof. Notice that J also minimizes

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

because $C_d\rho\widetilde{\delta}_n^{der}$ does not depend on j. Set $\Phi(j) := \sqrt{m/\alpha_j}\widetilde{\delta}_n^{noi} + C_d\rho\widetilde{\delta}_n^{der}$. If $\widetilde{J} \leq J_{\max}$, the theorem is proven by Lemma 4 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

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

□

This lemma implies convergence in probability of the estimator with the same rate. We can now proof Theorem 1.

Proof. (of Theorem 1) We introduce the notation

$$\delta_n^{noi} = \mathbb{E}\big[\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}\big], \qquad (\sigma_n^{noi})^2 = \mathbb{V}\mathrm{ar}\left(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}\right),$$
$$\delta_n^{der} = \mathbb{E}\big[\|\widehat{T}_{n\dagger} - T_\dagger\|^{1+\mu}_{\mathcal{L}(\mathbb{X},\mathbb{Y})}\big], \qquad (\sigma_n^{der})^2 = \mathbb{V}\mathrm{ar}\left(\|\widehat{T}_{n\dagger} - T_\dagger\|^{1+\mu}_{\mathcal{L}(\mathbb{X},\mathbb{Y})}\right).$$

Similar to the last proof J is also a minimizer of

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

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 as $A_1 \subset A_2 \subset \ldots \subset A_{k_{\max}}$ by

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

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 (17) and (18), and with

$$\Phi_n^{noi}(\tau, j) := \sqrt{\frac{m}{\alpha_j}}\delta_n^{noi} + C_d\rho\delta_n^{der} + \sqrt{\tau(j)}\left(\sqrt{\frac{m}{\alpha_j}}\sigma_n^{noi} + C_d\rho\sigma_n^{der}\right). \quad (36)$$

Set $\tau_k(j) := k + c_2^{-1}\ln(\kappa)(J-j)$ with some $\kappa > 1$ small enough such that

$$\tau(j+1)q_\alpha \ge \tau(j) \quad (37)$$

with q_α as in (12) for all j. Consequently, $\Phi_n^{noi}(\tau_k, j)$ is monotonically increasing in j as required for the application of Lemma 4. Notice that $k_{\max}$ is chosen in a way such that

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

Lemma 4 and Lemma 7 below show that $\|e_j^{noi} + e_j^{der}\| \le \Phi_n^{noi}(\tau_k, j)$ 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.

In order to prepare the final step of the proof, we bound the probability of $A_k \backslash A_{k-1}$ and the probability of the event complementary to A_k. The following computation uses (16), (17), (18).

$$P(A_k\backslash A_{k-1}) = P\left\{\Phi_n^{noi}(\tau_{k-1}, j) < \|e_j^{noi} + e_j^{der}\| \le \Phi_n^{noi}(\tau_k, j) \text{ for all } j = 1, \ldots, J\right\}$$
$$\le P\left\{\Phi_n^{noi}(\tau_{k-1}, j) < \|e_j^{noi} + e_j^{der}\| \text{ for all } j = 1, \ldots, J\right\}$$

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

In every event A_k we have $J = J^*$. The assumptions of Lemma 4 are fulfilled in A_k. This implies the following error bound

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

By the construction of the algorithm (11) the worst case error is $\|\widehat{\varphi}_{J^*} - \varphi^\dagger\| \leq 3R$. This will serve as an error bound in the event $\mathcal{C}A_{k_{\max}}$. Putting everything together yields

$$\mathbb{E}(\|\widehat{\varphi}_{J^*} - \varphi^\dagger\|^2) \leq P(A_1)C_1 + \sum_{k=2}^{k_{\max}} P(A_k\backslash A_{k-1})C_k + P(\mathcal{C}A_{k_{\max}})9R^2$$
$$\leq 10\left(\|e_J^{app}\|^2 + \frac{m}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2 C_d^2\rho^2\right)$$
$$+ 10P(A_1)\left(\frac{m}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2 C_d^2\rho^2\right)$$
$$+ \sum_{k=2}^{k_{\max}} P(A_k\backslash A_{k-1})\left(10k\frac{m}{\alpha_J}(\sigma_n^{noi})^2 + 10k(\sigma_n^{der})^2 C_d^2\rho^2\right) + P(\mathcal{C}A_{k_{\max}})9R^2$$
$$\leq 10\left(\|e_J^{app}\|^2 + \frac{m}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2 C_d^2\rho^2\right) + P(\mathcal{C}A_{k_{\max}})9R^2$$
$$+ 10\left(\frac{m}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2 C_d^2\rho^2\right)\left(2 + \sum_{k=3}^{k_{\max}} kP(A_k\backslash A_{k-1})\right)$$
$$\leq 10\left(\|e_J^{app}\|^2 + \frac{m}{\alpha_J}(\delta_n^{noi})^2 + (\delta_n^{der})^2 C_d^2\rho^2\right) + \left(\frac{c_1 e^{-c_2 k_{\max}} + c_3 e^{-c_4 k_{\max}}}{1-\kappa^{-1}}\right)9R^2$$
$$+ 10\left(\frac{m}{\alpha_J}(\sigma_n^{noi})^2 + (\sigma_n^{der})^2 C_d^2\rho^2\right)\left(2 + \sum_{k=2}^{k_{\max}-1}(k+1)\frac{c_1 e^{-c_2 k} + c_3 e^{-c_4 k}}{1-\kappa^{-1}}\right)$$

$$
\begin{aligned}
&\leq 10\left(\|e_J^{app}\|^2+\frac{m}{\alpha_J}(\delta_n^{noi})^2+(\delta_n^{der})^2C_d^2\rho^2\right)+(c'\max\{(\sigma_n^{noi})^2,(\sigma_n^{der})^2\})9R^2\\
&\quad+10\left(\frac{m}{\alpha_J}(\sigma_n^{noi})^2+(\sigma_n^{der})^2C_d^2\rho^2\right)\left(2+\sum_{k=2}^{\infty}(k+1)\frac{c_1e^{-c_2k}+c_3e^{-c_4k}}{1-\kappa^{-1}}\right)\\
&\leq 10\left(\|e_J^{app}\|^2+\frac{m}{\alpha_J}(\delta_n^{noi})^2+(\delta_n^{der})^2C_d^2\rho^2\right)+(c'\max\{(\sigma_n^{noi})^2,(\sigma_n^{der})^2\})9R^2\\
&\quad+10c''\left(\frac{m}{\alpha_J}(\sigma_n^{noi})^2+(\sigma_n^{der})^2C_d^2\rho^2\right)\\
&\leq C\left(\|e_J^{app}\|^2+\frac{m}{\alpha_J}\left[(\delta_n^{noi})^2+(\sigma_n^{noi})^2\right]+C_d^2\rho^2\left[(\delta_n^{der})^2+(\sigma_n^{der})^2\right]\right)\\
&=C\left(\|e_J^{app}\|^2+\frac{m}{\alpha_J}\mathbb{E}\left[\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}^2\right]+C_d^2\rho^2\mathbb{E}\left[\|\widehat{T}_{n\dagger}-T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}^{2+2\mu}\right]\right)\\
&=\mathcal{O}\left(\left(\mathbb{E}(\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\|_{\mathbb{Y}}^2)\right)^{\frac{2\mu}{2\mu+1}}+\mathbb{E}\left(\|\widehat{T}_{n\dagger}-T_\dagger\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})}^{2+2\mu}\right)\right)\ .
\end{aligned}
$$

We used $P(A_1)+\sum_{k=2}^{k_{\max}}P(A_k\backslash A_{k-1})+P(\mathcal{C}A_{k_{\max}})=1$ 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 1 above.

Lemma 6. *Let the assumptions of Theorem 1 hold and define:*

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

The following two bounds hold for the stopping index J in Theorem 1:

$$
(1-\underline{\Gamma}_{noi}^{-1})\tilde{\Phi}(J)\leq(\gamma_{app}-1)\|e_J^{app}\|, \tag{38}
$$

$$
J\geq\sup\left\{k\in\mathbb{N}\middle|\|e_1^{app}\|\gamma_{app}^{1-k}>\inf_{l\in\mathbb{N}}\left(C_\Lambda\rho\sqrt{\alpha_l}+\tilde{\Phi}(1)\overline{\Gamma}_{noi}^{l-1}\right)\right\}. \tag{39}
$$

Proof. Note that (28) implies

$$
1<\underline{\Gamma}_{noi}\leq\frac{\tilde{\Phi}(j+1)}{\tilde{\Phi}(j)}\leq\overline{\Gamma}_{noi},\quad\text{for all } j\in\mathbb{N}. \tag{40}
$$

We start with inequality (38). 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 (28) and (40) 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 (38).

In order to prove (39) 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 (27), (28) and (40) 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 proves the lemma. □

Lemma 7. *Let the assumptions of Theorem 1 hold true. Define $J_{\max}$ as in Lemma 4 with $\tau_k(j) := k + \frac{\ln(\kappa)}{c_2}(J-j)$*

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

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 $k = 1, \dots, k_{\max}$ it holds that $J \leq J_{\max}$.

Proof. Since $\tau_k(j)$ fulfills inequality (37) 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,

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

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 (39)

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

decays with a polynomial rate in σ_n^{noi} and σ_n^{der}. Therefore, there exists a constant b for which $J \geq -b\max\{\ln(\sigma_n^{noi}),\ln(\sigma_n^{der})\}$, 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:

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

with $\xi_n = \max\left\{\sqrt{\ln((\sigma_n^{noi})^{-2})},\sqrt{\ln((\sigma_n^{der})^{-2})}\right\}\frac{b}{2}(\mu-\frac{1}{2})$. Together with the first estimate this proves the lemma. □

Finally, we can proof Corallary 1.

Proof. (of Corollary 1)

Combining the results of Theorem 1 and Examples 3 and 4 we get the rate

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

A.5. Convergence rates for adaptive estimation

Proof. (of Theorem 2) When $\widetilde{\Phi}_n^{noi}$ is used in the definition of $J_{\max}$ in (33), it follows that $J_{\max}=\mathcal{O}\big[\ln((\sigma_n^{noi})^{-1})+\ln((\sigma_n^{der})^{-1})\big]$. 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

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

In Lemma 7 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 $\mathbb{N}$

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

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

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

with two constants C' and C''. We used in the third row the fact that $J_{\max} = \mathcal{O}\big[\ln((\sigma_n^{noi})^{-1}) + \ln((\sigma_n^{der})^{-1})\big]$ 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 MISE

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

The rate in the theorem follows from the last line and the bound (27) on $\|e_j^{app}\|$. □

Acknowledgments

The author would like to thank Thorsten Hohage and Johannes Schmidt-Hieber for interesting and fruitful discussions on this topic. He also would like to thank two anonymous referees for valuable comments that improved the paper.

References

Andrews, D. W. K. (2017). Examples of L2-complete and boundedly-complete distributions. *Journal of Econometrics* **199** 213 - 220. MR3681027

Babii, A. (2020). Honest confidence sets in nonparametric IV regression and other ill-posed models. *Econometric Theory* **36** 658–706. MR4125475

Babii, A. and Florens, J.-P. (2020). Is completeness necessary? Estimation in nonidentified linear models. *arXiv:1709.03473*.

Bakushinskiĭ, A. B. (1992). On a convergence problem of the iterative-regularized Gauss-Newton method. *Zhurnal Vychislitel' noĭ Matematiki i Matematicheskoĭ Fiziki* **32** 1503–1509. MR1185952

Bakushinskiĭ, A. B. and Kokurin, M. (2004). *Iterative Methods for Approximate Solution of Inverse Problems*. Springer, Dordrecht. MR2133802

Bauer, F. and Hohage, T. (2005). A Lepskij-type stopping rule for regularized Newton methods. *Inverse Problems* **21** 1975–1991. MR2183662

Bauer, F., Hohage, T. and Munk, A. (2009). Iteratively Regularized Gauss–Newton Method for Nonlinear Inverse Problems with Random Noise. *SIAM Journal on Numerical Analysis* **47** 1827-1846. MR2505875

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** 1749–1797. MR3268395

Blundell, R., Chen, X. and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75** 1613–1669. MR2351452

Breunig, C. (2015). Goodness-of-fit tests based on series estimators in nonparametric instrumental regression. *Journal of Econometrics* **184** 328–346. MR3291006

Breunig, C. and Johannes, J. (2016). Adaptive estimation of functionals in nonparametric instrumental regression. *Econometric Theory* **32** 612–654. MR3506434

Carrasco, M., Florens, J.-P. and Renault, E. (2007). Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization. In *Handbook of Econometrics*, (J. J. Heckman and E. E. Leamer, eds.). *Handbook of Econometrics* **6** 5633-5751. Elsevier.

Chen, X. and Christensen, T. M. (2015). Optimal Sup-norm Rates, Adaptivity and Inference in Nonparametric Instrumental Variables Estimation. *ArXiv:1508.03365v1*.

Chen, X. and Christensen, T. M. (2018). Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression. *Quantitative Economics* **9** 39–84. MR3789729

Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* **80** 277–321. MR2920758

Chen, X. and Reiss, M. (2011). On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory* **27** 497–521. MR2806258

Chen, X., Chernozhukov, V., Lee, S. and Newey, W. K. (2014). Local Identification of Nonparametric and Semiparametric Models. *Econometrica*

82 785–809. MR3191719
Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica* **73** 245–261. MR2115636
Chernozhukov, V., Imbens, G. W. and Newey, W. K. (2007). Instrumental variable estimation of nonseparable models. *Journal of Econometrics* **139** 4–14. MR2380756
Darolles, S., Fan, Y., Florens, J.-P. and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541–1565. MR2883763
D'Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory* **27** 460–471. MR2806256
D'Haultfoeuille, X. and Février, P. (2015). Identification of Nonseparable Triangular Models With Discrete Instruments. *Econometrica* **83** 1199-1210. MR3357488
Dunker, F., Hoderlein, S. and Kaido, H. (2014). Nonparametric identification of endogenous and heterogeneous aggregate demand models: complements, bundles and the market level. *cemmap Working Papers* **CWP23/14**.
Dunker, F., Florens, J.-P., Hohage, T., Johannes, J. and Mammen, E. (2014). Iterative estimation of solutions to noisy nonlinear operator equations in nonparametric instrumental regression. *Journal Econometrics* **178** 444–455. MR3132443
Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press. MR1270903
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 *Advances in Economics and Econometrics: Theory and Applications* (M. Dewatripont, L. P. Hansen and S. J. Turnovsky, eds.) 284–311. Cambridge Univ. Press. MR0908740
Florens, J.-P., Johannes, J. and Van Bellegem, S. (2011). Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory* **27** 472–496. MR2806257
Florens, J.-P. and Sbaï, E. (2010). Local identification in empirical games of incomplete information. *Econometric Theory* **26** 1638–1662. MR2738012
Gagliardini, P. and Scaillet, O. (2012a). Nonparametric Instrumental Variable Estimation of Structural Quantile Effects. *Econometrica* **80** 1533-1562. MR2977430
Gagliardini, P. and Scaillet, O. (2012b). Tikhonov regularization for nonparametric instrumental variable estimators. *Journal of Econometrics* **167** 61-75. MR2885439
Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.* **33** 2904–2929. MR2253107
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. MR1474369
Hohage, T. and Werner, F. (2016). Inverse problems with Poisson data: sta-

tistical regularization theory, applications and algorithms. *Inverse Problems* **32** 093001. MR3543331

Horowitz, J. L. (2011). Applied Nonparametric Instrumental Variables Estimation. *Econometrica* **79** 347–394. MR2809374

Horowitz, J. L. (2014a). Adaptive nonparametric instrumental variables estimation: Empirical choice of the regularization parameter. *Journal of Econometrics* **180** 158 - 173. MR3197791

Horowitz, J. L. (2014b). Ill-Posed Inverse Problems in Economics. *Annual Review of Economics* **6** 21-51.

Horowitz, J. L. and Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* **75** 1191–1208. MR2333498

Johannes, J., Van Bellegem, S. and Vanhems, A. (2011). Convergence rates for ill-posed inverse problems with an unknown operator. *Econometric Theory* **27** 522–545. MR2806259

Kaido, H. and Wüthrich, K. (2021). Decentralization estimators for instrumental variable quantile regression models. *Quantitative Economics* **12** 443-475. MR4325591

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. MR2459012

Kato, K. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics* **40** 3108 – 3136. MR3097971

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

Linton, O. and Mammen, E. (2005). Estimating Semiparametric ARCH (∞) Models by Kernel Smoothing Methods. *Econometrica* **73** 771–836. MR2135143

Little, G. and Reade, J. B. (1984). Eigenvalues of analytic kernels. *SIAM Journal on Mathematical Analysis* **15** 133–136. MR0728688

Loh, I. (2019). Nonparametric Identification and Estimation with Independent, Discrete Instruments. *arXiv:1906.05231*.

Mair, B. A. (1994). Tikhonov regularization for finitely and infinitely smoothing operators. *SIAM Journal on Mathematical Analysis* **25** 135–147. MR1257145

Mathé, P. (2006). The Lepskii principle revisited. *Inverse Problems* **22** L11–L15. MR2235633

Mathé, P. and Hofmann, B. (2008). How general are general source conditions? *Inverse Problems* **24** 015009, 5. MR2384768

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989). London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge Univ. Press, Cambridge. MR1036755

Morozov, V. A. (1968). The principle of disparity in solving operator equations by the method of regularization. *Žurnal Vyčislitel′noĭ Matematiki i Matematičeskoĭ Fiziki* **8** 295–309. MR0243738

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation

of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257
Reade, J. B. (1984). Eigenvalues of smooth kernels. *Math. Proc. Cambridge Philos. Soc.* **95** 135–140. MR0727087
Torgovitsky, A. (2015). Identification of Nonseparable Models Using Instruments With Small Support. *Econometrica* **83** 1185-1197. MR3357487
Tsybakov, A. (2000). On the best rate of adaptive estimation in some inverse problems. *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique* **330** 835–840. MR1769957
Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer. MR2724359