Iterative Estimation of Solutions to Noisy Nonlinear Operator Equations in Nonparametric Instrumental Regression

Fabian DUNKER[*†], Jean-Pierre FLORENS[‡]
Thorsten HOHAGE[§],Jan JOHANNES[¶],
Enno MAMMEN[‖]

June 21, 2013

Abstract: This paper discusses the solution of nonlinear integral equations with noisy integral kernels as they appear in nonparametric instrumental regression. We propose a regularized Newton-type iteration and establish convergence and convergence rate results. A particular emphasis is on instrumental regression models where the usual conditional mean assumption is replaced by a stronger independence assumption. We demonstrate for the case of a binary instrument that our approach allows the correct estimation of regression functions which are not identifiable with the standard model. This is illustrated in computed examples with simulated data.

JEL classification: C13, C14, C30, C31, C36
Keywords and phrases: Nonparametric regression, nonlinear inverse problems, iterative regularization, instrumental regression

[*]Institute of Numerical and Applied Mathematics, University of Göttingen, Lotzestr. 16–18, 37083 Göttingen, Germany
[†]Corresponding author *Email*: dunker@math.uni-goettingen.de *Tel.*: +49551394507
[‡]Toulouse School of Economics, University Toulouse I Capitole , Manufacture des Tabacs, Aile Jean-Jacques Laffont, 21 Alle de Brienne, 31000 Toulouse, France
[§]Institute of Numerical and Applied Mathematics, University of Göttingen, Lotzestr. 16–18, 37083 Göttingen, Germany
[¶]Institut de statistique, UCL, Voie du Roman Pays, 20, 1348 Louvain-la-Neuve Belgium
[‖]Department of Economics, University Mannheim, L7,3-5, 68131 Mannheim, Germany

1 Introduction

In this paper we will propose and analyze an iterative method for estimating the solution of nonlinear integral equations which appear in nonparametric instrumental regression problems. Examples will be discussed below, see eq. (4) and Section 2. Such integral equations can be written as nonlinear operator equations

$$\mathcal{F}(\varphi) = 0 \tag{1}$$

where the operator $\mathcal{F}$ is unknown, but where an estimator $\widehat{\mathcal{F}}$ of $\mathcal{F}$ is available. We will assume that $\mathcal{F} : \mathfrak{B} \subset \mathcal{X} \to \mathcal{Y}$ maps from a convex set $\mathfrak{B}$ in a Banach space $\mathcal{X}$ to a Hilbert space $\mathcal{Y}$. Typically such operator equations are ill-posed in the sense that $\mathcal{F}^{-1}$ is not continuous. In particular this is the case for integral operators with smooth kernels on a compact set. In such cases the straightforward estimator $\widehat{\mathcal{F}}^{-1}(0)$ will not be consistent since the variance of its norm $\|\widehat{\mathcal{F}}^{-1}(0)\|_{\mathcal{Y}}$ is infinite. Regularization techniques must be applied to solve (1) or its empirical version $\widehat{\mathcal{F}}(\hat{\varphi}) = 0$. The conceptually most simple method is nonlinear Tikhonov regularization, which is given by

$$\widehat{\varphi} := \operatorname*{argmin}_{\varphi} \left[\|\widehat{\mathcal{F}}(\varphi)\|_{\mathcal{Y}}^2 + \alpha \|\varphi - \varphi_0\|_{\mathcal{X}}^2 \right]. \tag{2}$$

Here φ_0 is some initial guess of φ and $\alpha > 0$ is a regularization parameter. Since for nonlinear operators the Tikhonov functional is not convex in general, it may have many local minima. For the numerical solution of (2) no algorithms with guaranteed convergence are known, and in fact the numerical minimization of (2) can be difficult, particularly for small α.

Therefore, we will use the iteratively regularized Newton method, which is one of

the most popular algorithms for the regularization of nonlinear ill-posed operator equations. Instead of a quadratic penalty $\|\varphi-\varphi_0\|_{\mathcal{X}}^2$, we allow for a more general penalty term $\mathcal{R}:\mathfrak{B}\to(-\infty,\infty]$ with domain of definition $\mathfrak{B}$. We only assume that $\mathcal{R}$ is a convex, lower semi-continuous functional that is not identically equal to ∞. With this choice an iteratively regularized Gauß-Newton method is given by the iterations

$$\widehat{\varphi}_k := \operatorname*{argmin}_{\varphi\in\mathcal{B}} \left[\|\widehat{\mathcal{F}}'[\widehat{\varphi}_{k-1}](\varphi-\widehat{\varphi}_{k-1})+\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|_{\mathcal{Y}}^2+\alpha_k\mathcal{R}(\varphi)\right]. \tag{3}$$

In each Newton step a convex optimization problem has to be solved with a sequence of regularization parameters α_k. We assume that α_k tends to 0 in a way that will be specified in Section 4. In the special case that $\mathcal{X}$ is a Hilbert space, the most common choice for the penalty term is $\mathcal{R}(\varphi)=\|\varphi-\varphi_0\|_{\mathcal{X}}^2$. Here $\|\cdot\|_{\mathcal{X}}$ is the norm of the Hilbert space $\mathcal{X}$ and φ_0 is the initial guess at which the iteration is started. This is the iteratively regularized Gauß-Newton method as suggested by Bakushinskiĭ (1992) and further analyzed by Blaschke et al. (1997) and Hohage (1997) for low order Hölder or logarithmic source conditions, respectively. We also refer to the monographs by Bakushinskiĭ and Kokurin (2004) and Kaltenbacher et al. (2008) and to further references therein.

The use of more general convex regularization terms in the general case where $\mathcal{X}$ is a Banach space allows for a flexible incorporation of further a-priori information. Common choices are entropy regularization, l^1 penalties and bounded variation (BV) penalties. Loubes and Pelletier (2008) studied entropy regularization for instrumental variable models but they gave no theoretical results for the rates of convergence of their estimators. If a basis or a frame of $\mathcal{X}$ is given, an l^1 penalty of the coefficients with respect to this basis or frame enhances sparsity properties

of the estimator with respect to this basis or frame. A bounded variation penalty is particularly appropriate for piecewise constant solutions.

Our main result gives rates of convergence for the estimator where the distance between the estimator and the solution of (1) is measured by the Bregman distance, see Theorem 2. For entropy regularization this directly implies convergence estimates measured by the L^1-norm. Our scheme allows for the incorporation of structural a-priori information of the form $\varphi \in \mathcal{C}$ where $\mathcal{C}$ is a closed convex set (e.g. a-priori information on non-negativity, monotonicity or convexity/concavity). This can be done by setting $\mathcal{R}(\varphi) := \infty$ if $\varphi \notin \mathcal{C}$.

For convex regularization terms, the analysis differs from the mathematical approaches used for studying quadratic regularization. One has to employ variational methods rather than spectral methods. Recently, a number of papers have appeared on this subject, we only mention Eggermont (1993), Burger and Osher (2004), Resmerita (2005), Hofmann et al. (2007), Scherzer et al. (2009). A first variational convergence rate analysis of Newton-type methods in a deterministic setting without errors in the operator and $\mathcal{R}$ given by Banach norms has recently been done by Kaltenbacher and Hofmann (2010). Our analysis is closest to that of the last reference. However, all the references above only treat perturbations of the right hand side of the operator equation, and hence these results are not applicable to nonparametric instrumental regression. Our treatment of nonlinear ill-posed operator equations with errors in the operator may be of independent interest and relevant for other applications.

For the special case that $\mathcal{X}$ is a Hilbert space convergence rates of the nonlinear Tikhonov regularization were discussed in Engl et al. (1989) in a deterministic setting. Rates for a model with random errors were obtained in Bissantz et al. (2004). In Horowitz and Lee (2007) nonparametric instrumental variables esti-

mation is considered in a quantile regression model. This is one example of a statistical model where the unknown nonparametric function is given as the solution of a nonlinear integral equation. We will describe this model in the next section. Series estimators for nonlinear ill-posed problems generated by conditional moment equations have been analyzed in a Banach space setting by Chen and Pouzo (2012).

In Horowitz and Lee (2007) it is assumed that the singular values of the Fréchet derivative $\mathcal{F}'[\varphi]$ decay polynomially and results are given on the rates of convergence under these assumptions. It was pointed out in Horowitz and Lee (2007) that a convergence analysis for exponentially decreasing singular values is an important open problem. We will show that singular values of integral operators with infinitely smooth kernels do in fact decrease super-algebraically and present a convergence analysis without an assumption on the rate of decay of the singular values.

Besides the analysis of the iteratively regularized Gauß-Newton method for noisy operators the second main innovation of this paper is a nonparametric instrumental regression model where the instrument W is independent from the error U:

$$Y = \varphi(Z) + U, \tag{4a}$$

$$U \perp\!\!\!\perp W, \tag{4b}$$

$$\mathbb{E}U = 0. \tag{4c}$$

Here, Y is a scalar response variable, Z is an observed random vector of endogeneous explanatory variables. It is shown in Section 2 that this model leads to a nonlinear integral equation of the form (1) with a kernel, that has to be estimated

from data.

This model slightly differs from nonparametric instrumental regression with mean independent instruments given by

$$Y = \varphi(Z) + U, \tag{5a}$$

$$\mathbb{E}[U|W] = 0. \tag{5b}$$

The latter model has been studied intensively in econometrics by a number of authors, see e.g. Florens (2003), Newey and Powell (2003), Hall and Horowitz (2005), Blundell et al. (2007), Breunig and Johannes (2009), Chen and Reiss (2010), Darolles et al. (2011), and Chen and Pouzo (2012). In this model the regression function φ is defined as the solution of a linear first kind integral equation

$$\mathcal{T}\varphi = g \tag{6}$$

where both the kernel of the integral operator $(\mathcal{T}\varphi)(w) := \mathbb{E}[\varphi(Z)|W = w]$ and the right hand side $g(w) := \mathbb{E}[Y|W = w]$ have to be estimated from the data.

Actually, in specific econometric applications, the conditional mean assumption (5b) is typically established by arguing that the stronger independence assumption (4b) holds. Therefore, it is a natural question if one can improve the accuracy of estimation of φ by using the stronger condition (4c), (4b) directly. We will give a first partial positive answer to this question: a necessary condition for identifiability in the model (5) is that the instrumental variable W must have at least as many continuously distributed components as the explanatory variable Z. This is not necessary in model (4). We will show this for the model where W is binary and Z is one-dimensional and continuously distributed. For this model we will give examples where we have strong evidence that the assumption

(4) identifies the solution, at least in some neighborhood. Hence, the model (4) contains more information on φ than the model (5). A more detailed comparison of the two models is very complex because the integral equations obtained from these two models are related only very implicitly.

The plan of this paper is as follows: in the next section we give more details on our motivating examples from instrumental variable regression. Section 3 recalls the definition of source conditions and discusses their relation to smoothness conditions. In particular, we show that for integral equations of the first kind with smooth kernels, Hölder type source conditions are too restrictive. Furthermore, we discuss variational forms of source conditions. In Section 4 we present our main convergence result for the iteratively regularized Gauß-Newton method with noisy operators. Afterwards, we discuss in Section 5 how this result applies to the regression problem (4). Section 6 reports on numerical simulations for an instrumental variable regression model with binary instruments.

2 Examples

2.1 Instrumental quantile regression

In Horowitz and Lee (2007) and Chen and Pouzo (2012) the estimation and in Chen et al. (2011) the identification of the following quantile regression model has been studied:

$$Y = \varphi(Z) + U \tag{7a}$$

$$\mathbb{P}(U \leq 0 | W = w) = q \qquad \text{for all } w. \tag{7b}$$

Here, Y is a response variable, Z is an endogeneous explanatory variable, $q \in (0,1)$ is a fixed constant, U is an unobserved error variable and W an observable instrument. The quantile is defined conditional on W.

We assume from now on that each of the random variables Y, Z and W is a vector of continuous or discrete random variables. Further, we assume that a joint density f_{YZW} exists with respect to the Lebesgue measure, to the counting measure, or to a product of both measures. Let $G_{YZW}(y,z,w) := \int_{-\infty}^{y} f_{YZW}(\tilde{y},z,w)\,d\tilde{y}$, and let $f_W(w) := \int\int f(y,z,w)\,dy\,dz$ denote the marginal density of W. Then φ solves a nonlinear operator equation (1) with the operator

$$(\mathcal{F}(\varphi))(w) := \int G_{YZW}(\varphi(z),z,w)\,dz - q f_W(w).$$

It is pointed out in Horowitz and Lee (2007), that the model (7) subsumes non-separable quantile regression models of the form

$$Y = H(Z,V) \tag{8}$$

as studied in Chernozhukov et al. (2007), see also Chernozhukov and Hansen (2005). Here V is an unobserved, continuously distributed random variable independent of an instrument W, and the function H is strictly increasing in its second argument. Assuming w.l.o.g. that $V \sim U[0,1]$, (8) reduces to (7) with $U := Y - H(Z,q)$ and $\varphi(z) := H(z,q)$.

2.2 Nonparametric regression with independent instruments

2.2.1 Operator equations

The model (4a), (4b) leads to the nonlinear integral equation

$$\int f_{YZW}(u+\varphi(z),z,w)\,dz - \int f_{YZ}(u+\varphi(z),z) f_W(w)\,dz = 0, \quad \text{for all } u,w, \tag{9a}$$

where we assume as above that the joint density f_{YZW} of (Y,Z,W) exists. The marginal densities of (Y,Z) and W are denoted by f_{YZ} and f_W respectively. Note that if φ is a solution to (9a), then any function $\varphi + a$ with $a \in \mathbb{R}$ is another solution to (9a). The additive constant can be fixed by taking into account eq. (4c), which may be rewritten as

$$\int \varphi(z) f_Z(z)\,dz - \int y f_Y(y)\,dy = 0 \tag{9b}$$

with the marginal densities f_Y and f_Z of Y and Z. The system of equations (9a), (9b) can be written as a nonlinear ill-posed operator equation (1) with the operator

$$(\mathcal{F}(\varphi))(u,w) := \begin{pmatrix} \int \big(f_{YZW}(u+\varphi(z),z,w) - f_{YZ}(u+\varphi(z),z) f_W(w)\big)\,dz \\ \int \varphi(z) f_Z(z)\,dz - \int y f_Y(y)\,dy \end{pmatrix}. \tag{10}$$

In the next section we discuss identification properties of model (4) which involve the Gateaux derivative of the nonlinear operator $\mathcal{F}$. The derivative of $\mathcal{F}$ depends on partial derivatives of the joint density f_{YZW}. The following modification of the operator avoids this. It has a Gateaux derivative depending directly on the

joint density f_{YZW} and not on its partial derivatives.

Let us assume the existence of the joint density of (Y, Z) unconditional and conditional given W, say $f_{Y,Z}$ and $f_{Y,Z|W}$ instead of the existence of $f_{Y,Z,W}$. We integrate $f_{Y,Z}$ and $f_{Y,Z|W}$ once with respect to u and define $G_{YZ}(y,z) := \int_{-\infty}^{y} f_{YZ}(\tilde{y},z)\,d\tilde{y}$ and $G_{YZ|W}(y,z|w) := \int_{-\infty}^{y} f_{YZ|W}(\tilde{y},z|w)\,d\tilde{y}$. This yields another operator formulation of model (4) with the operator

$$(\tilde{\mathcal{F}}(\varphi))(u,w) := \begin{pmatrix} \int \left(G_{YZ|W}(u+\varphi(z),z|w) - G_{YZ}(u+\varphi(z),z)\right) dz \\ \int \varphi(z) f_Z(z)\,dz - \int y f_Y(y)\,dy \end{pmatrix}. \quad (11)$$

The Gateaux derivative of $\tilde{\mathcal{F}}$ is presented in the next section.

2.2.2 Identification

The independence assumption (4b) together with (4c) is considerably stronger than the mean independence assumption (5b). In this section we want to argue that model (4) has thereby more identifying power than model (5). For this purpose, we discuss examples where we have strong evidence that model (4) identifies the solution at least in some neighborhood, while model (5) fails. A paradigmatic case for the insufficiency of (5) is the one of a discrete instrument combined with a continuous regressor. Beyond the following discussion, further evidence for identification in this case is provided by numerical simulations in Section 6.

The following lemma gives sufficient conditions for identification by possibly ill-posed nonlinear operator equations. It appeared at first in Hanke et al. (1995).

Lemma 1. *Let φ be characterized by some operator equation $\mathcal{F}(\varphi) = h$. Assume $\mathcal{F}$ is Gateaux differentiable, the derivative $\mathcal{F}'[\varphi]$ is injective, and $\mathcal{F}$ fulfills the*

tangential cone condition

$$\|\mathcal{F}(\psi) - \mathcal{F}(\varphi) - \mathcal{F}'[\varphi](\psi - \varphi)\| \leq \eta \|\mathcal{F}(\psi) - \mathcal{F}(\varphi)\| \tag{12}$$

for all ψ in some neighborhood U_φ of φ with $\eta \in (0,1)$. Then φ is locally identified in U_φ, i.e. for all $\psi \in U_\varphi$ with $\psi \neq \varphi$ we have $\mathcal{F}(\psi) \neq h$.

It is well known that the tangential cone condition is often difficult to check for a given operator. Nevertheless, it seems to hold in many examples. We have to impose a similar condition on the estimator, see (25c). For a discussion we refer to Kaltenbacher et al. (2008), Chen et al. (2011) or Florens and Sbaï (2010).
In the last lemma the tangential cone condition is a sufficient but not necessary assumption. In the rest of the section we focus on the second assumption of the lemma. We discuss sufficient conditions for the injectivity of the derivative $\tilde{\mathcal{F}}'[\varphi]$ of the operator $\tilde{\mathcal{F}}$ defined in (11). Although this does not necessarily imply local identification it gives evidence for it. Let us set

$$(\mathcal{G}(\varphi))(u,w) := \int \big(G_{YZ|W}(u+\varphi(z),z|w) - G_{YZ}(u+\varphi(z),z)\big)\, dz \quad \text{and} \tag{13}$$

$$(\mathcal{G}'[\varphi]\phi)(u,w) := \int \phi(z)\big(f_{YZ|W}(u+\varphi(z),z|w) - f_{YZ}(u+\varphi(z),z)\big)\, dz. \tag{14}$$

Then the operator in (11) and its Gateaux derivative can be written as

$$\tilde{\mathcal{F}}(\varphi) := \begin{pmatrix} \mathcal{G}(\varphi) \\ \mathbb{E}(Y - \varphi(Z)) \end{pmatrix} \qquad \text{and} \qquad \tilde{\mathcal{F}}'[\varphi]\phi = \begin{pmatrix} \mathcal{G}'[\varphi]\phi \\ \mathbb{E}(\phi(Z)) \end{pmatrix}.$$

Injectivity of $\tilde{\mathcal{F}}'[\varphi]$ is equivalent to injectivity of $\mathcal{G}'[\varphi]$ on the linear subspace of functions ϕ with $\mathbb{E}[\phi(Z)] = 0$. We denote by f_U the marginal density of $U = Y - \varphi(Z)$. Then by employing the independence of U and W a change of

variables allows us to write

$$\left(\mathcal{G}'[\varphi]\phi\right)(u,w) = \Big(\mathbb{E}[\phi(Z)|U=u,W=w] - \mathbb{E}[\phi(Z)|U=u]\Big) f_U(u). \tag{15}$$

Alternatively, we may consider the linear operator

$$\left(\mathcal{T}\phi\right)(u,w) := \mathbb{E}[\phi(Z)|U=u,W=w] - \mathbb{E}[\phi(Z)|U=u] \tag{16}$$

mapping from a function space of mean zero functions in Z into a function space in U and W. Roughly speaking, injectivity of the operator $\mathcal{T}$ and hence local identification is possible if the dependence between the endogenous regressor Z and the error term U varies sufficiently with respect to the instrument W. We illustrate this by two examples. The first example is based on a more theoretical construction using mixtures of Gaussian distributions. The second one discusses binary instruments.

Example 1. Let U, V and W be real valued independent random variables and let ρ be a function defined on $\mathbb{R}$ and taking values in $[-1,1]$. Define the endogenous regressor

$$Z := U\,\rho(W) + V\,\sqrt{1-\rho^2(W)}.$$

If U and V are standard normally distributed, which we assume in this example, then it is easily seen that the conditional distribution of (U,Z) given W is Gaussian:

$$\begin{pmatrix} U \\ Z \end{pmatrix} \Big| W \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho(W) \\ \rho(W) & 1 \end{pmatrix}\right). \tag{17}$$

Note that in this situation U and Z are marginally standard normally distributed, both unconditional and conditional on W. In other words, U and W as well as Z and W are independent. But obviously, the random vector (U, Z) and the instrument W are dependent. Interestingly, for the commonly studied model (5) identification is guaranteed if and only if the conditional distribution of Z given W is complete (cf. Carrasco et al. (2006)). This rules out the independence of Z and W and hence this example. However, in this example the linear operator $\mathcal{T}$ defined in (16) can be injective and thus local identification might be still possible.

In order to provide sufficient conditions to ensure injectivity of $\mathcal{T}$, let us recall the eigenvalue decomposition of the conditional expectation operator for normally distributed random variables. The following development can be found in Carrasco et al. (2006) while it has been shown thoroughly in Letac (1995). Consider random variables U^* and Z^* satisfying

$$\begin{pmatrix} U^* \\ Z^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

for some $\rho \in [-1, 1]$. Obviously, U^* and Z^* are marginally identically distributed with standard normal density $f_{0,1}$. Hence, the L^2 spaces with respect to the probability measures of U^* and Z^* are the L^2 space with respect to the standard normal measure. Let us write this for short as $L^2_{U^*} = L^2_{Z^*} = L^2_{f_{0,1}}$. Note, that by an elementary symmetry argument the conditional expectation operator $S\phi := \mathbb{E}[\phi(Z^*)|U^*]$ of Z^* given U^*, mapping $L^2_{f_{0,1}}$ to itself, is self-adjoint. Hence, S permits an eigenvalue decomposition. Moreover, for $j = 0, 1, 2, \ldots$ let $f^{(j)}_{0,1}$ be the jth derivative of $f_{0,1}$ and let $H_j := (-1)^j f^{(j)}_{0,1}/f_{0,1}$ denote the jth Hermite polynomial. The Hermite polynomials form a complete orthogonal sys-

tem in $L^2_{f_{0,1}}$, see e.g. Problem IV-29 on page 117 in Letac (1995). Furthermore, $\mathbb{E}[H_j(Z^*)|U^*] = H_j(U^*)\rho^j$ holds true for all $j \in \mathbb{N}_0$, see e.g. Problem IV-30 on page 120 in Letac (1995). From these assertions we readily conclude that the eigenfunctions $\{\psi_j\}_{j=0}^{\infty}$ of S are given by the Hermite polynomials, up to some multiples. In addition, $(\rho^j)_{j\in\mathbb{N}_0}$ is the corresponding sequence of eigenvalues. Keeping in mind that the distribution of (U, Z) conditional on W given in (17) is Gaussian let us reconsider the operator $\mathcal{T}$ defined in (16). By employing that U and W are independent it is straightforward to conclude that

$$\mathbb{E}[|(\mathcal{T}\phi)(U,W)|^2] = \sum_{j=1}^{\infty} \mathbb{V}\mathrm{ar}(\rho^j(W))\mathbb{E}[|\phi(Z)\psi_j(Z)|^2]$$

for all $\phi \in L^2_Z$ with $\mathbb{E}[\phi(Z)] = 0$, where the basis $\{\psi_j\}_{j=1}^{\infty}$ are multiples of the Hermite polynomials. Consequently, the operator $\mathcal{T}$ is injective if and only if $\mathbb{V}\mathrm{ar}(\rho^j(W)) \neq 0$ for all $j \in \mathbb{N}$, (keep in mind Parseval's identity, i.e. $\mathbb{E}[f(Z)^2] = \sum_{j=1}^{\infty} \mathbb{E}[f(Z)\psi_j(Z)]^2$ for all $f \in L^2_Z$). This in turn holds if and only if the random variable $|\rho(W)|$ is not constant. Surprisingly, even in case of a binary instrument W taking only two values, say $P(W = 0) = w_0$ and $P(W = 1) = 1 - w_0$ with $0 < w_0 < 1$, the condition $|\rho(0)| \neq |\rho(1)|$ is sufficient to ensure the injectivity of the operator $\mathcal{T}$.

Example 2. We now give another example for injectivity of $\mathcal{T}$ in the case of a binary instrument W. We assume that the conditional copula function of U and Z, given $W = w$ does not depend on w. This assumption has been made by Imbens and Newey (2009) in case of a continuous instrument. Under this assumption it holds that (U, V) is independent of W where $V = F_{Z|W}(Z|W)$ for the conditional distribution function $F_{Z|W}(z|w)$ of Z given $W = w$. Note that in

the case of a binary instrument injectivity of $\mathcal{T}$ is equivalent to the injectivity of the map

$$\phi \mapsto \mathbb{E}[\phi(Z)|U, W = 1] - \mathbb{E}[\phi(Z)|U, W = 0]$$

on the space of all functions ϕ with $\mathbb{E}[\phi(Z)] = 0$. We use that

$$\begin{aligned} 0 &= \mathbb{E}[\phi(Z)|U, W = 1] - \mathbb{E}[\phi(Z)|U, W = 0] \\ &= \mathbb{E}[\phi(F^{-1}_{V|W}(V|1))|U, W = 1] - \mathbb{E}[\phi(F^{-1}_{V|W}(V|0))|U, W = 0] \\ &= \mathbb{E}[\phi(F^{-1}_{V|W}(V|1))|U] - \mathbb{E}[\phi(F^{-1}_{V|W}(V|0))|U] \\ &= \mathbb{E}[\phi(F^{-1}_{V|W}(V|1)) - \phi(F^{-1}_{V|W}(V|0))|U], \end{aligned}$$

because of independence of (U, V) and W. If the family of conditional densities of V given U is complete this equation implies that $\phi(F^{-1}_{V|W}(v|1)) = \phi(F^{-1}_{V|W}(v|0))$ almost surely. The latter equation can be used to get that under some additional assumptions on $F_{Z|W}$ the function ϕ is almost surely constant, see the arguments used in Torgovitsky (2012) and D'Haultfœuille and Février (2011). Because of $\mathbb{E}[\phi(Z)] = 0$ we get that $\phi(z) = 0$ a.s. Thus $\mathcal{T}$ is invertible. Note that our discussion differs from the results in Imbens and Newey (2009), Torgovitsky (2012) and D'Haultfœuille and Février (2011). We make the assumption on the conditional copula function only for the underlying distribution and argue that - under additional conditions - local identifiability holds for a neighborhood of distributions for which this assumption may not apply whereas in the latter papers the conditional copula assumption is used as a model assumption for all distributions of the statistical model. This heuristic discussion can be generalized to more general instruments with discrete and/or continuous components.

2.2.3 Binary instruments

As mentioned above, a particular interest of this paper is in the special case of a binary instrument W and continuously distributed Z. This is an example where model (4) is superior to (5) and it serves as a test scenario for numerical simulations in Section 6. We therefore give a short discussion how much the operator (10) degenerates in this case. Assume W only takes the values 0 and 1. Then the marginal density f_W (w.r.t. the counting measure) has the two values

$$f_W(0) = w_0 \qquad \text{and} \qquad f_W(1) = w_1 = 1 - w_0\,.$$

Equation (9a) is equivalent to the system of equations

$$\begin{aligned} \textstyle\int f_{YZW}(u+\varphi(z),z,0)dz &= w_0 \textstyle\int f_{YZ}(u+\varphi(z),z)\,dz \\ \textstyle\int f_{YZW}(u+\varphi(z),z,1)dz &= w_1 \textstyle\int f_{YZ}(u+\varphi(z),z)\,dz \end{aligned} \qquad \text{for all } u\,.$$

It follows from the identity $f_{YZ}(y,z) = f_{YZW}(y,z,0) + f_{YZW}(y,z,1)$ that these two equations are linearly dependent and can be rewritten as

$$\int w_1 f_{YZW}(u+\varphi(z),z,0) - w_0 f_{YZW}(u+\varphi(z),z,1)\,dz = 0 \quad \text{for all } u\,. \tag{18}$$

So φ is a root of the nonlinear ill-posed operator

$$(\mathcal{F}(\varphi))(u) := \begin{pmatrix} \int w_1 f_{YZW}(u+\varphi(z),z,0) - w_0 f_{YZW}(u+\varphi(z),z,1)\,dz \\ \int \varphi(z) f_Z(z)\,dz - \int y f_Y(y)\,dy \end{pmatrix}. \tag{19}$$

This operator formulation has been used for the numerical simulations in Section 6. as a result of our discussion Z does not have to be discrete for identifiability as it is the case when the conditional mean assumption (5b) is used instead of

the independence assumption (4c).

3 Smoothness in terms of source conditions

In this section we collect some material on source conditions that will be needed in the next section to state our main result. We are primarily interested in source conditions in Banach spaces. However, we start with a motivation for L^2 spaces and present in a first step a definition of source conditions in the special case of Hilbert spaces. For the sake of simplicity we discuss the relevance of source conditions for nonparametric instrumental regression problems in this special case. Afterward, we introduce source conditions for the general case of Banach spaces.

3.1 Source conditions in Hilbert spaces

Let us recall the relationship between the smoothness of a kernel k of a compact linear integral operator $\mathcal{T} : L^2([0,1]^{d_1}) \to L^2([0,1]^{d_2})$,

$$(\mathcal{T}\varphi)(x) := \int_{[0,1]^{d_1}} k(x,y)\varphi(y)\,dy, \qquad x \in [0,1]^{d_2}$$

and the decay of its singular values σ_j. If $\{(u_j, v_j, \sigma_j) : j \in \mathbb{N}_0\}$ is a singular system of $\mathcal{T}$, then according to the Courant-Fischer characterization (see e.g. Kress (1999)) of the singular values the operator $\mathcal{T}_j$ with kernel $k_j(x,y) := \sum_{l=0}^{j-1} \sigma_l v_l(x) u_l(y)$ satisfies

$$\sigma_j = \|\mathcal{T} - \mathcal{T}_j\| = \inf\{\|\mathcal{T} - \tilde{\mathcal{T}}\| : \operatorname{rank}\tilde{\mathcal{T}} \le j\}.$$

The norm used in the last equation is the usual operator norm

$$\|\mathcal{T}\| := \inf_{1=\|g\|_{L^2([0,1]^{d_1})}} \|\mathcal{T}g\|_{L^2([0,1]^{d_2})}.$$

If there exist functions $\tilde{u}_l \in L^2([0,1]^{d_1})$, $\tilde{v}_l \in L^2([0,1]^{d_2})$, and numbers $\tilde{\sigma}_l$ for all $l \in \mathbb{N}_0$ such that $\int_{[0,1]^{d_1}} \int_{[0,1]^{d_2}} |k(x,y) - \sum_{l=0}^{j-1} \tilde{u}_l(x)\tilde{v}_l(y)|^2 \, dx \, dy \leq \tilde{\sigma}_l$, then $\sigma_j \leq \tilde{\sigma}_j$ since $\|\mathcal{T} - \mathcal{T}_j\| \leq \|k - k_j\|_{L^2([0,1]^{d_1+d_2})}$. It follows from standard results in approximation theory (see e.g. Prössdorf and Silbermann (1991)) that for smooth bounded domains the singular values σ_j decay at least polynomially if k belongs to a Sobolev space, super-algebraically if $k \in C^\infty([0,1]^{d_1+d_2})$, and at least exponentially if k is analytic.

In regularization theory, smoothness of the solution $\varphi^\dagger$ to an inverse problem is usually formulated in terms of source conditions, which describe smoothness relative to the smoothing properties of the operator. For a linear operator $\mathcal{T} : \mathcal{X} \to \mathcal{Y}$ between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, such source conditions have the form

$$\varphi^\dagger - \varphi_0 = \Lambda(\mathcal{T}^*\mathcal{T})\psi \,. \tag{20}$$

Here $\psi \in \mathcal{X}$, φ_0 is an initial guess (typically $\varphi_0 = 0$ in the linear case), $\mathcal{T}^*$ is the adjoint operator of $\mathcal{T}$ with respect to the scalar product of the Hilbert space, and $\Lambda : [0,\infty) \to [0,\infty)$ is a continuous, strictly monotonically increasing function with $\Lambda(0) = 0$. $\Lambda(\mathcal{T}^*\mathcal{T})$ is defined by using the spectral calculus. So with the notations above $\Lambda(\mathcal{T}^*\mathcal{T})\psi = \sum_{l=0}^{\infty} \Lambda(\sigma_l^2)\langle\psi, u_l\rangle u_l$. For a nonlinear operator between Hilbert spaces $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ the Gateaux derivative $\mathcal{T} = \mathcal{F}'[\varphi^\dagger]$ at $\varphi^\dagger$ is used.

If we choose a fixed Λ the source condition is the more restrictive the faster the singular values decay. This means for integral operators that it is the more

restrictive the smoother the kernel. For the most common choice $\Lambda(t) = t^\mu$ for some $\mu > 0$ these condition are called *Hölder-type source conditions*. We refer to the monographs Engl et al. (1996); Bakushinskiĭ and Kokurin (2004); Kaltenbacher et al. (2008) for further information.

3.2 Impact on nonparametric instrumental regression

Let us discuss source conditions in the context of nonparametric instrumental variable models. The kernel of the integral operator in (14) is composed of probability densities. For the derivatives of the alternative operators (10) and (11) it is composed of partial derivatives of densities. Many typical probability density functions are analytic, e.g. the density of the normal distribution. Hence, in applications it will frequently occur that the kernel of the operator in the source condition is infinitely smooth or even analytic.

Let us have a closer look at these cases. The singular values of the operator in (14) will decay super-algebraically or even exponentially. As a consequence, Hölder-type source conditions are extremely restrictive smoothness conditions, since the eigenvalues $\lambda_j((\mathcal{T}^*\mathcal{T})^\nu) = \sigma_j^{2\nu}$ will decay super-algebraically or exponentially, too. Hence, Hölder-type source conditions imply that the Fourier coefficients with respect to $\{u_j : j \in \mathbb{N}_0\}$ of the difference between initial guess and regression function $\varphi^\dagger - \varphi_0$ decay super-algebraically or exponentially. For standard Fourier coefficients this entails that $\varphi^\dagger - \varphi_0$ has to be infinitely smooth or even analytic. Hence, the initial guess must be very good and already capture some features of the unknown function $\varphi^\dagger$. In applications, one would typically expect only polynomial decay of the Fourier coefficients of $\varphi^\dagger - \varphi_0$ which corresponds to finite Sobolev smoothness instead of infinite smoothness. Therefore, it is desirable to consider also functions Λ which decay to 0 more slowly than $t \mapsto t^\nu$. For

exponentially decaying singular values the logarithmic functions

$$\Lambda(t) = (-\ln t)^{-p}$$

with a parameter $p > 0$ are a natural choice corresponding to a polynomial decay of the Fourier coefficients of $\varphi^\dagger - \varphi_0$. (Here we always assume that the operator is scaled so that $\|\mathcal{T}^*\mathcal{T}\| \leq \exp(-1)$ or alternatively we use a dilated version of the above function Λ.) The importance of logarithmic source conditions for nonparametric instrumental regression is also pointed out in Blundell et al. (2007) and Horowitz and Lee (2007).

3.3 Variational source conditions for Banach spaces

In our analysis we will not restrict ourselves to Hilbert spaces, but study the more general situation where $\mathcal{X}$ is a Banach space, which we assume in the following. Note that in this case the operator $\mathcal{T}^*\mathcal{T}$ maps from $\mathcal{X}$ to the dual space $\mathcal{X}'$, so even integer powers of $\mathcal{T}^*\mathcal{T}$ are not well-defined. Therefore, spectral source conditions as introduced above must be generalized. For this purpose we use variational methods which have been explored in regularization theory recently in a number of papers. We will prove convergence results with these methods in terms of the Bregman distance in $\mathcal{X}$ with respect to the convex func-

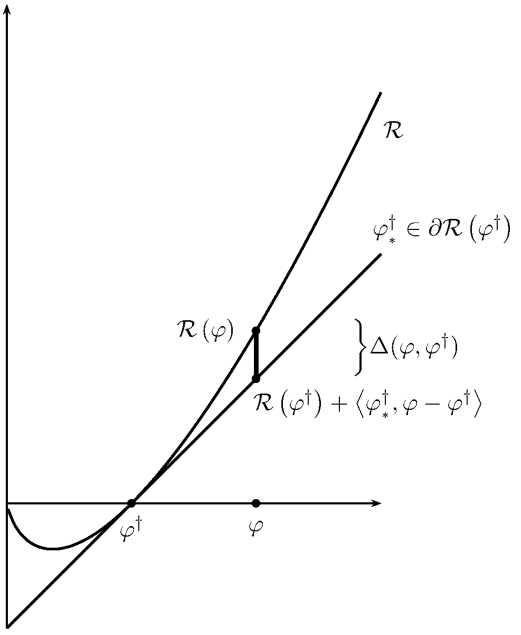

Figure 1: Bregman distance

tional $\mathcal{R}$.

Let $\varphi_*^\dagger \in \partial\mathcal{R}(\varphi^\dagger)$ be a fixed element of the subdifferential of $\mathcal{R}$ at $\varphi^\dagger$ (i.e. $\varphi_*^\dagger = \mathcal{R}'[\varphi^\dagger]$, if $\mathcal{R}$ is differentiable at $\varphi^\dagger$). Then the Bregman distance with respect to $\mathcal{R}$ and $\varphi_*^\dagger$ is defined as

$$\Delta(\varphi, \varphi^\dagger) := \mathcal{R}(\varphi) - \mathcal{R}(\varphi^\dagger) - \left\langle \varphi_*^\dagger, \varphi - \varphi^\dagger \right\rangle. \tag{21}$$

Here $\langle \cdot, \cdot \rangle$ denotes the classical dual pairing $\langle \mathcal{X}', \mathcal{X} \rangle$, i.e. $\left\langle \varphi_*^\dagger, \varphi - \varphi^\dagger \right\rangle$ is the evaluation of the functional $\varphi_*^\dagger$ at $\varphi - \varphi^\dagger$. Hence, the Bregman distance measures how much the linearization of $\mathcal{R}$ at $\varphi^\dagger$ and $\mathcal{R}$ differ at the point φ. This is illustrated in Figure 1. For strictly convex $\mathcal{R}$ we have $\Delta(\varphi, \varphi^\dagger) = 0$ if and only if $\varphi = \varphi^\dagger$. The Bregman distance Δ is nonnegative and convex in the first argument, but it does not define a metric since it is neither symmetric nor does it satisfy the triangle inequality in general. However, Bregman distances provide a generalization of the simpler case, where $\mathcal{X}$ is a Hilbert space and $\mathcal{R}(\varphi) = \|\varphi - \varphi_0\|_{\mathcal{X}}^2$ for some $\varphi_0 \in \mathcal{X}$. Because, in this situation

$$\Delta(\varphi, \varphi^\dagger) = \|\varphi - \varphi^\dagger\|_{\mathcal{X}}^2.$$

Although, Bregman distances are in general not metrics they have meaningful interpretations in some Banach space settings. If $\mathcal{X} = L^1(D)$ and $\mathcal{R}(\varphi) = \int_D \varphi(x) \ln(\varphi(x))\, dx$ (entropy regularization), then $\Delta(\varphi, \varphi^\dagger)$ can be bounded from below by $\|\varphi - \varphi^\dagger\|_{L^1}^2$ (see e.g. Resmerita (2005)), i.e. the error bounds formulated in the next theorem can be interpreted as bounds with respect to the squared L^1 norm. Our framework also allows the incorporation of convex constraints by setting $\mathcal{R}(\varphi) := \infty$ if φ does not belong to some convex set $\mathcal{C}$. Obviously, this does not change Δ in $\mathcal{C}$.

Following Kaltenbacher and Hofmann (2010) we formulate the source condition as a variational inequality

$$\left\langle \varphi_*^\dagger, \varphi^\dagger - \varphi \right\rangle \leq \beta \Delta(\varphi, \varphi^\dagger)^{1/2} \Lambda \left(\frac{\|\mathcal{F}'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2}{\Delta(\varphi, \varphi^\dagger)} \right) \qquad \text{for all } \varphi \in \mathfrak{B}. \quad (22)$$

Here Λ is the same kind of function as above. The constant $\beta > 0$ corresponds to the norm $\|\psi\|$ in the spectral source condition (20). It can influence the convergence as a multiplicative constant. However, it does not influence the rate of convergence as we will see in (27).

The variational source condition is a generalization of the Hilbert space case. It is shown in Kaltenbacher and Hofmann (2010) that if $\mathcal{X}$ is a Hilbert space, $\mathcal{R}(\varphi) = \|\varphi - \varphi_0\|^2$ and $(\Lambda^2)^{-1}$ is convex, the classical source condition (20) implies the variational one (22).

Let us close this section with a technical remark. Note that if $\mathfrak{B}$ is chosen so that $\varphi^\dagger$ is on the boundary of $\mathfrak{B}$, then possibly Λ can be chosen smaller than in the case where $\varphi^\dagger$ is in the interior of $\mathfrak{B}$. Theorem 2 yields that this may lead to faster rates of convergence. Hence, a convex constraint on the regression function can improve estimation. To capture this fact it is important that, opposed to the formulation in Kaltenbacher and Hofmann (2010), no absolute values appear on the left hand side of (22). A typical example where $\varphi^\dagger$ is on the boundary of $\mathfrak{B}$ is the assumption that $\varphi^\dagger$ is a positive function.

4 Convergence results

Let $\mathcal{X}$ be a Banach space, $\mathcal{Y}$ a Hilbert space, $\mathfrak{B} \subset \mathcal{X}$ convex and $\varphi^\dagger \in \mathfrak{B}$ a root of the operator $\mathcal{F} : \mathfrak{B} \to \mathcal{Y}$:

$$\mathcal{F}(\varphi^\dagger) = 0. \quad (23)$$

Assume that $\mathcal{F}$ is approximated by a series of estimators

$$\widehat{\mathcal{F}}_n : \mathfrak{B} \to \widehat{\mathcal{Y}}_n$$

which maps to some (possibly finite-dimensional and/or data dependent) Hilbert space $\widehat{\mathcal{Y}}_n$. For all n larger than some $N \in \mathbb{N}$ the operators $\widehat{\mathcal{F}}_n$ and the operator $\mathcal{F}$ are assumed to be Gateaux differentiable on $\mathfrak{B}$ with linear derivatives $\mathcal{F}'[\varphi]$ and $\widehat{\mathcal{F}}_n'[\varphi]$, which are "bounded with respect to Δ" in the sense that

$$\sup_{\{\tilde{\varphi}\in\mathfrak{B}:\Delta(\tilde{\varphi},\varphi)\neq 0\}} \|\mathcal{F}'[\varphi](\tilde{\varphi}-\varphi)\|^2/\Delta(\tilde{\varphi},\varphi) < \infty \quad \text{and} \quad \mathcal{F}'[\varphi](\tilde{\varphi}-\varphi) \neq 0 \tag{24}$$

whenever $\Delta(\tilde{\varphi},\varphi) \neq 0$ and analogously for all $\widehat{\mathcal{F}}_n$. Now we can state the main theorem of this paper, which is proved in Appendix A:

Theorem 2. *Let* (22) *hold true with a concave* Λ *for which* $t \mapsto \sqrt{t}/\Lambda(t)$ *is monotonically increasing. Assume that the sequence* $\widehat{\mathcal{F}}_n$ *has the following convergence properties:*

$$\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\| = O_p(\delta_n), \tag{25a}$$

$$\left(\left|\sup_{\varphi\in\mathfrak{B}} \frac{\|\mathcal{F}'[\varphi^\dagger](\varphi-\varphi^\dagger)\|^2 - \|\widehat{\mathcal{F}}_n'[\varphi^\dagger](\varphi-\varphi^\dagger)\|^2}{\Delta(\varphi,\varphi^\dagger)}\right|\right)^{1/2} = O_p(\gamma_n), \tag{25b}$$

$$P\{\|\widehat{\mathcal{F}}_n(\varphi_1) - \widehat{\mathcal{F}}_n(\varphi_2) - \widehat{\mathcal{F}}_n'[\varphi_2](\varphi_1-\varphi_2)\| > \eta\|\widehat{\mathcal{F}}_n(\varphi_1) - \widehat{\mathcal{F}}_n(\varphi_2)\| \text{ for some } \varphi_1,\varphi_2 \in \mathfrak{B}\} \to 0. \tag{25c}$$

Here η *must be sufficiently small, such that* $(1-\eta)/(\eta+\eta^2) > 4q^{3/2}+1$. *Suppose that the convex minimization problems* (3) *have a solution for every* $\widehat{\mathcal{F}}_n$ *(see Remark 1 for sufficient conditions), i.e. the method is well defined. Further assume that* $\alpha_0 > \max(\Theta^{-1}(\delta_n), \gamma_n^2)$ *and that* $\alpha_k \leq q\alpha_{k+1}$ *for all* k *with a constant* $q > 1$.

Let the iteration be stopped at the smallest index $K_n \in \mathbb{N}_0$ for which

$$\alpha_{K_n+1} \leq \max(\Theta^{-1}(\delta_n), \gamma_n^2)\,, \qquad \textit{where } \Theta(t) := \sqrt{t}\Lambda(t). \tag{26}$$

Then

$$\Delta(\widehat{\varphi}_{K_n}, \varphi^\dagger) = O_p\left(\Lambda^2\left(\max(\Theta^{-1}(\delta_n), \gamma_n^2)\right)\right) \tag{27}$$

Remarks:

1. Sufficient conditions for the existence of solutions to the minimization problems (3) are reflexivity of $\mathcal{X}$, weak closedness of $\mathfrak{B}$, and the boundedness of the sets $\{\varphi \in \mathfrak{B} : \mathcal{R}(\varphi) \leq R\}$ in $\mathcal{X}$ for any $R \in \mathbb{R}$. This is a standard argument: If (φ_n) is a minimizing sequence, it must be bounded due to our last condition. Since $\mathcal{X}$ is reflexive, there exists a weakly convergent subsequence, and by weak closedness of $\mathfrak{B}$ a weak limit point $\varphi_* \in \mathfrak{B}$. The Tikhonov functional is convex and lower semi-continuous due to the convexity and lower semi continuity of $\mathcal{R}$. Hence, it is also weakly lower semi-continuous and φ_* is a minimizer.

2. Moreover, we can guarantee uniqueness of the solutions to the minimization problems (3) under mild assumptions. Sufficient conditions are strict convexity of $\mathcal{R}$ or injectivity of $\widehat{\mathcal{F}}_n'[\widehat{\varphi}_{k-1}]$.

3. Note that if $\mathcal{X}$ is a Hilbert space, $\widehat{\mathcal{F}}_n$ Fréchet differentiable, and $\mathcal{R}$ quadratic penalty, then $\|\mathcal{F}'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2 - \|\widehat{\mathcal{F}}_n'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2 \leq \|\mathcal{F}'[\varphi^\dagger]^*\mathcal{F}'[\varphi^\dagger] - \widehat{\mathcal{F}}_n'[\varphi^\dagger]^*\widehat{\mathcal{F}}_n'[\varphi^\dagger]\|\ \|\varphi - \varphi^\dagger\|^2$, so $\gamma_n \leq \|\mathcal{F}'[\varphi^\dagger]^*\mathcal{F}'[\varphi^\dagger] - \widehat{\mathcal{F}}_n'[\varphi^\dagger]^*\widehat{\mathcal{F}}_n'[\varphi^\dagger]\|^{1/2}$.

4. The bound on the Taylor remainder of $\widehat{\mathcal{F}}_n$

$$\|\widehat{\mathcal{F}}_n(x) - \widehat{\mathcal{F}}_n(y) - \widehat{\mathcal{F}}_n'[y](x-y)\| \leq \eta \|\widehat{\mathcal{F}}_n(x) - \widehat{\mathcal{F}}_n(y)\|, \tag{28}$$

used in (25c) is known as the tangential cone condition. This condition is commonly used in the analysis of regularization methods for nonlinear ill-posed problems, see Kaltenbacher et al. (2008). The right hand side of (28) may be replaced by $\|F'[y](x-y)\|$ (see (40) below), and in this form it corresponds to Assumption 2 in Chen et al. (2011).

Corollary 3. *Let the assumptions of Theorem 2 hold true.*

1. *If $\Lambda(t) = t^\mu$ for some $\mu \in (0, 1/2]$ (Hölder-type source conditions), then*

$$\Delta(\widehat{\varphi}_K, \varphi^\dagger)^{1/2} = O_p\left(\max(\delta_n^{2\mu/(2\mu+1)}, \gamma_n^{2\mu})\right) . \tag{29}$$

2. *If $\mathcal{F}$ is scaled such that $\|\mathcal{F}'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2 / \Delta(\varphi, \varphi^\dagger) \leq \frac{1}{2}$ and $\Lambda(t) = (-\ln t)^{-p}$ for some $p > 0$ (logarithmic source conditions), then*

$$\Delta(\widehat{\varphi}_K, \varphi^\dagger)^{1/2} = O_p\left((-\ln \max(\delta_n, \gamma_n))^{-p}\right) \tag{30}$$

for all δ_n, γ_n sufficiently small.

Let us discuss some properties of the method. First of all it is a local method like any Newton method. Convergence is only guaranteed if the initial guess φ_0 is sufficiently close to the true solution $\varphi^\dagger$. How close it has to be depends on the special problem, i.e. the operator $\mathcal{F}$. This property appears in the assumptions (22) and (25c) in Theorem 2.

Unlike for nonlinear Tikhonov regularization our theoretical results do not require

the strong assumption that we can always find the minimum of a functional with an arbitrary number of local minima. In turn we have to assume (25c), which is usually hard to check. Although, rigorous proofs for (25c) are often missing, it seems to hold in many cases at least in a neighborhood of $\varphi^\dagger$.

An important advantage for the numerical implementation is that a lot of efficient algorithms converging always towards the true solution are known for convex minimization problems. The error of these minimization algorithms plays a minor role compared to the regularization error for the applications of Section 2. We refer to Langer and Hohage (2007) for a detailed discussion of the interplay of these errors in other applications.

5 Examples revisited

The assumptions of Theorem 2 and Corollary 3 are rather abstract and need some explanations concerning the application to the nonparametric regression with independent instrument (4). They are applicable in a similar way to the nonparametric quantile regression (7). In (10) the operator $\mathcal{F}$ for the regression with independent instrument is an integral operator with a kernel composed of f_{YZW} and its marginals. Hence, an estimator $\widehat{f}_{YZW}$ yields an estimator of the kernel and thereby of $\widehat{\mathcal{F}}$.

Condition (24) that all $\widehat{\mathcal{F}}_n'[\varphi]$ must be bounded with respect to the Bregman distance is fulfilled if the derivatives of $\mathcal{F}$ are bounded according to (24), the estimation of f_{YZW} is strongly consistent and n is large enough. Strong consistency is established for many density estimators. The boundedness of $\mathcal{F}$ with respect to the Bregman distance is reasonable. It holds if the partial derivative of the joint density $\frac{\partial}{\partial y} f_{YZW}$ is bounded for the operator (10) or if $f_{YZ|W}$ is bounded for

the operator (11) and the Bregman distance is bounded from below by the power of a norm. As mentioned in Section 3.3, the latter is for example the case for quadratic and maximum entropy penalty.

It can be argued with strong consistency as well that the probabilistic tangential cone condition (25c) holds if the exact operator $\mathcal{F}$ fulfills the tangential cone condition (12). But, it is known that a verification, whether or not the tangential cone condition is true, is often difficult for a given operator.

In analogy to (11) the operator

$$\tilde{\mathcal{F}}(\varphi)(u,w) := \begin{pmatrix} \mathbb{P}\big(Y-\varphi(Z)\leq u\big) - \mathbb{P}\big(Y-\varphi(Z)\leq u|W=w\big) \\ \mathbb{E}[\varphi(Z)-Y] \end{pmatrix}$$

can be considered for model (4). With this formulation of the operator the conditions (24) and (12) are more explicitly assumptions on the primitives of the model.

In the rates for Hölder source conditions (29) in Corollary 3, δ has a smaller exponent than γ. However δ does not necessarily dominate the convergence. In the nonparametric instrumental regression δ corresponds to the estimation of a density, while γ is determined by the estimation of a partial derivative of that density. Hence, γ decays usually slower than δ. Which of the terms $\delta^{2\mu/(2\mu+1)}$ or $\gamma^{2\mu}$ dominates the convergence depends on the properties of the special problem, namely the number of instruments and covariates as well as on the smoothness of the density and the initial error $\varphi^\dagger - \varphi_0$.

The situation becomes clearer in the case of logarithmic source conditions. If the kernel of the operator is analytic, but the initial error in the regression function is not smooth or has only finite Hölder smoothness, merely a logarithmic rate of convergence can be expected. As discussed in Section 3.2 this situation can

occur in many applications. Even for estimating an analytic density a nonparametric density estimator will attain only a polynomial rate in n. Due to (30) our estimator for φ will end up asymptotically with the logarithmic rate $(-\ln(n))^{-p}$. Furthermore, the reconstructions are heavily influenced by the choice of the penalty functional $\mathcal{R}$. Therefore, $\mathcal{R}$ must be chosen carefully in applications. If shape constraints on $\varphi^\dagger$ are known, they can be enforced by the choice of $\mathcal{R}$, and this can significantly improve rates of convergence (see (22) and (27) and the discussion at the end of Section 3). E.g., a-priori known nonnegativity can be incorporated by an entropy functional or a hard constraint ($R(\varphi) := \infty$ if not $\varphi \geq 0$ a.e.). Concerning smoothness it is generally better to choose a penalty functional $\mathcal{R}$ involving higher order derivatives as far as rates are concerned (see (Engl et al., 1996, §8.5)). If jumps in $\varphi^\dagger$ are of interest, a total variation penalty is a good choice.

6 Numerical simulations

In this section we present some numerical simulations for nonparametric instrumental regression with independent binary instrument and real-valued continuous explanatory and dependent variables. This leads to the nonlinear operator equation (19). Our simulations show that the solution computed by the method (3) approximates the exact solution. As mentioned above, due to dimensionality, the regression function cannot be identified with a binary instrument if the standard regression model (5) is used.

In our simulations we choose Y as real valued, Z with values in $[0, 1]$ and W with

values in $\{0,1\}$. We assume the regression function is

$$\varphi^\dagger(z) = \frac{1}{6}\sin(2\pi(z+0.25)) + 0.41\,, \qquad z \in [0,1]\,.$$

Moreover, we take $w_0 = P(W=0) = 2/3$ and $w_1 = P(W=1) = 1/3$. To make Z endogenous, let us choose the error term as $(U|Z=z, W=w) \sim \mathcal{N}(\mu_w(z), 0.09^2)$ with $\mu_0(z) := 0.2z - 0.1$ and $\mu_1(z) := 0.25z - 0.125$. The functions $\mu_0(z)$ and $\mu_1(z)$ describe precisely the correlation between the explanatory variable and the error term, which should be removed using the information contained in the instrumental variable. Although U varies with Z and W the condition $W \perp\!\!\!\perp U$ can be assured by a proper choice of $f_{Z,W}(z,w)$. We write the joint density as

$$\begin{aligned} f_{YZW}(y,z,0) &= f_{ZW}(z,0)\frac{1}{0.09\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\varphi^\dagger(z)-\mu_0(z)}{0.09}\right)^2\right), \\ f_{YZW}(y,z,1) &= f_{ZW}(z,1)\frac{1}{0.09\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\varphi^\dagger(z)-\mu_1(z)}{0.09}\right)^2\right). \end{aligned} \tag{31}$$

Now f_{ZW} has to be determined such that W and U are independent, which is equivalent to (18). Let us show that setting $f_{ZW}(z,1) := 0.625 f_{ZW}(1.25z - 0.125, 0)$ achieves this. With a substitution of variables we compute

$$\begin{aligned} &\int w_1 f_{YZW}(u+\varphi^\dagger(z), z, 0)\,dz \\ &= \int \frac{1}{3} f_{ZW}(z,0)\frac{1}{0.09\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{u-0.2z+0.1)}{0.09}\right)^2\right) dz \\ &= \int \frac{1.25}{3} f_{ZW}(1.25v-0.125,0)\frac{1}{0.09\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{u-0.25v+0.125)}{0.09}\right)^2\right) dv \\ &= \int w_0 f_{YZW}(u+\varphi^\dagger(v), v, 1)\,dv. \end{aligned}$$

This shows that (18) holds with our definition of $f_{ZW}(z,1)$ what ever $f_{ZW}(z,0)$

looks like. Here we take it to be normally distributed with variance 0.3^2 and expectation $1/2$ truncated to the interval $[0,1]$, i.e.

$$f_{ZW}(z,0) := a \exp\left(-\frac{1}{2}\left(\frac{z-1/2}{0.3}\right)^2\right), \qquad z \in [0,1]$$

with some scaling factor a chosen such that $\int_0^1 f_{ZW}(z,0)dz = 2/3$. By this construction, the error term also meets the condition $\mathbb{E}U = 0$ of the regression model (4): To see this, note that $f_{ZW}(\cdot,0)$ and $f_{ZW}(\cdot,1)$ are even, while μ_0 and μ_1 are odd functions with respect to the point 0.5. Hence,

$$\begin{aligned}\mathbb{E}U &= \int w_0 f_{Z,W}(z,0)\mathbb{E}(U|Z=z,W=0) + w_1 f_{Z,W}(z,1)\mathbb{E}(U|Z=z,W=1)\,dz \\ &= \int w_0 f_{Z,W}(z,0)\mu_0(z) + w_1 f_{Z,W}(z,1)\mu_1(z)\,dz = 0.\end{aligned}$$

This construction allows an easy formulation of how the solution of a nonparametric regression without instrumental variable and without noise would look like: $\widetilde{\varphi}(z) = w_0\mu_0(z) + w_1\mu_1(z) + \varphi^\dagger$

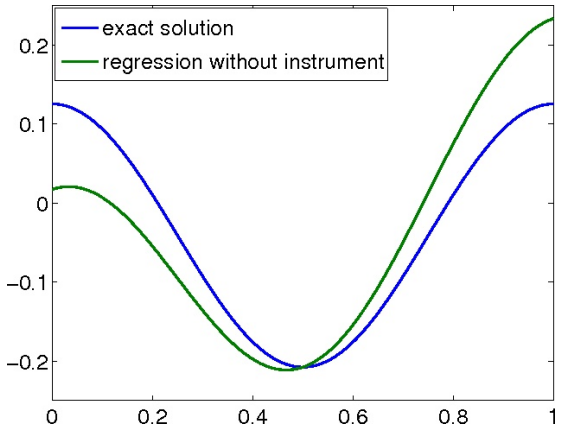

Figure 2: Necessity of the instrument: A standard regression would asymptotically yield a curve which is far away from the true solution $\varphi^\dagger$.

To approximately solve the integral operator equation (19) by the method (3) we discretized the domain $[0,1] \times [0,1] \times \{0,1\}$ by $256 \times 256 \times 2$ points and chose

the regularization parameters by $\alpha_0 = 1$ and $\alpha_{n+1} = 0.9\alpha_n$. The iteration was stopped using Lepskiĭ's principle as in Bauer et al. (2009). The initial guess was chosen as the constant function $E[Y]$. For a first test we used the exact density f_{YZW}, which actually has to be estimated from the data, of course. The L^2-error was reduced from 0.1294 to 0.0028. The remaining error is due to discretization noise. This suggests that the example is identifiable and can be solved by the method (3). Compared to the error for densities estimated from simulated data below, the observed discretization error is very small. Hence, the discretization is fine enough and the discretization error is insignificant for our simulations. The singular values of $\mathcal{F}'[\varphi^\dagger]$ are shown in Figure 4. They exhibit an exponential decay, so according to Corollary 3 we can only expect slow rates of convergence.

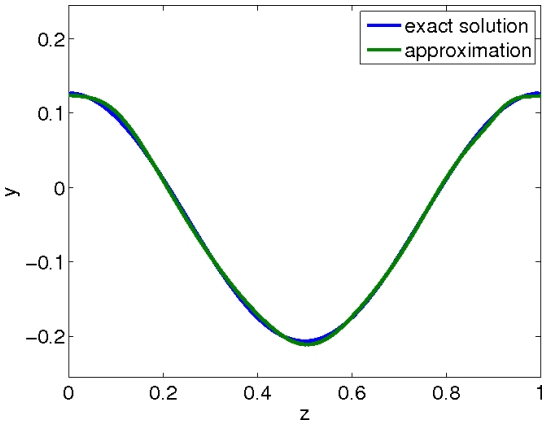

Figure 3: Result of the iterative inversion using the exact density f_{YZW}.

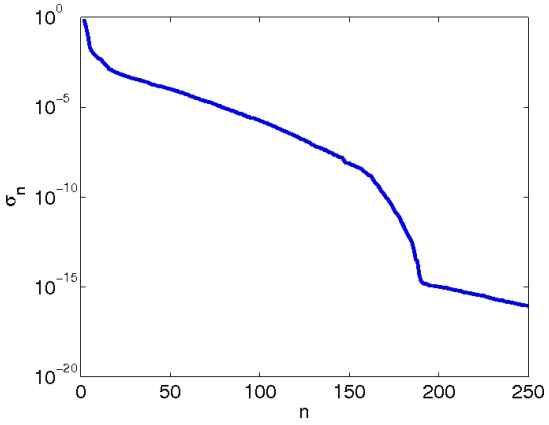

Figure 4: Singular values of $\mathcal{F}'[\varphi^\dagger]$

In further tests the algorithm was evaluated for finite samples of (Y, Z, W) with 10^3, 10^4 and 10^5 points. Given such a sample, the joint density f_{YZW} was estimated non-parametrically by the kernel density estimator developed by Botev et al. (2010). Afterwards again (19) was solved, but the exact density was replaced by the estimated one. We made 1000 samples for each tested sample size. The following table and histograms in Fig. 5–7 show the L^2-errors of the approx-

imate solution normed by the error of the initial guess (i.e. the error of the initial guess is 1). It can be seen that small samples produce unwanted outliers, but the method becomes reliable when the sample size is large enough. Fig. 8–10 show median reconstructions for each sample size. The results demonstrate that our method computes an asymptotically correct estimator of the regression function $\varphi^\dagger$ with an endogeneous explanatory variable Z using only a binary instrument W.

the exact solution is 0 and the error of the initial guess is 1. It can be seen that small samples produce unwanted outliers, but that the method becomes reliable, when the sample is large enough.

sample size N	mean	quantiles $p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 0.9$
10^3	0.6159	0.4057	0.5751	0.7921	0.9575
10^4	0.3694	0.2496	0.3524	0.4574	0.5729
10^5	0.3264	0.2592	0.3278	0.3882	0.4610

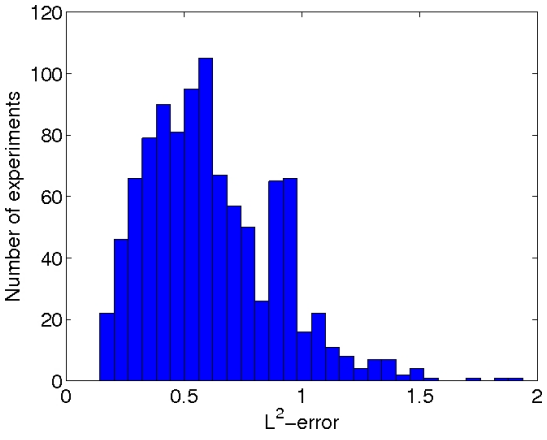

Figure 5: L^2 error for sample size $N = 10^3$

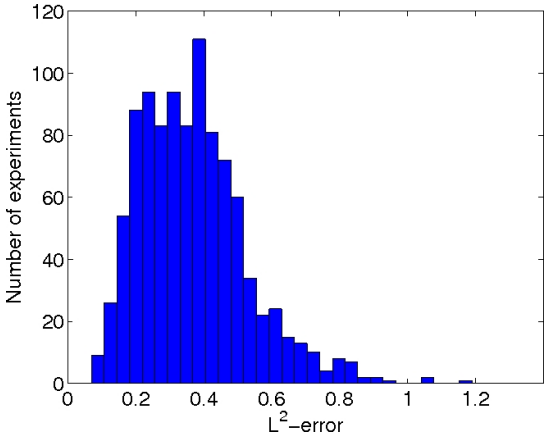

Figure 6: L^2 error for sample size $N = 10^4$

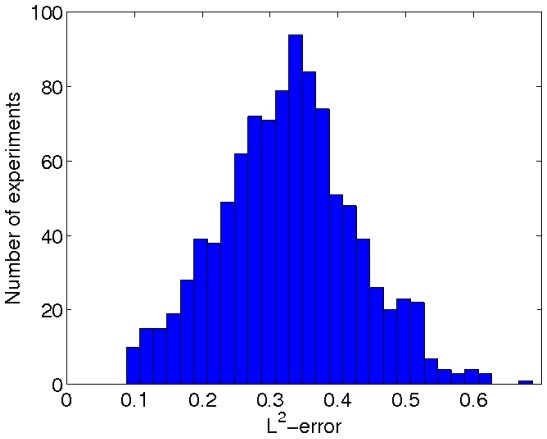

Figure 7: L^2 error for sample size $N = 10^5$

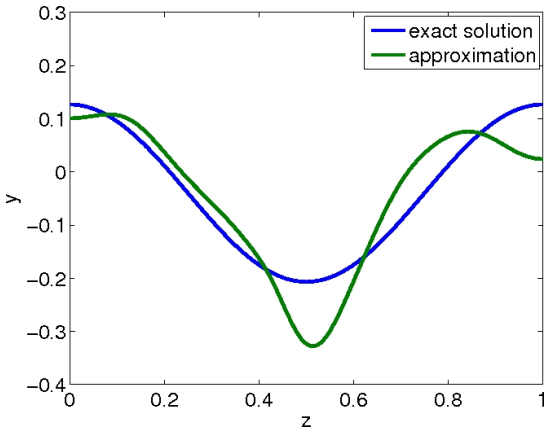

Figure 8: Median reconstruction, $N = 10^3$

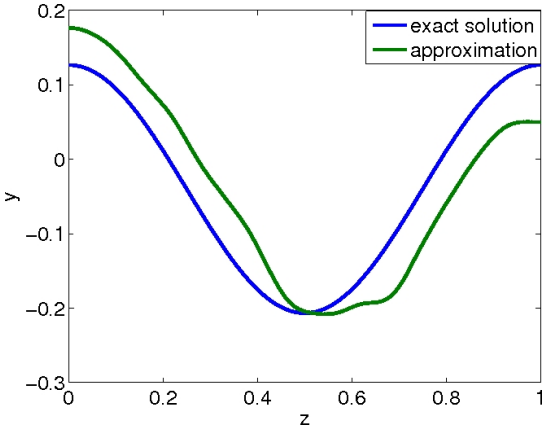

Figure 9: Median reconstruction, $N = 10^4$

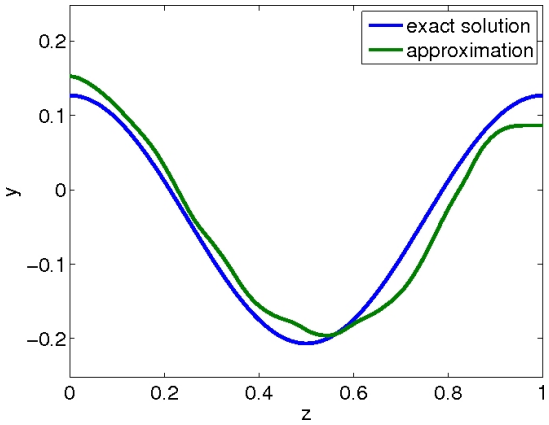

Figure 10: Median reconstruction, $N = 10^5$

A Proofs

Proof of Lemma 1. The tangential cone condition implies for all $\psi \in U_\varphi$

$$(1-\eta)\|\mathcal{F}(\psi)-\mathcal{F}(\varphi)\| \leq \|\mathcal{F}'[\varphi](\psi-\varphi)\| \leq (1+\eta)\|\mathcal{F}(\psi)-\mathcal{F}(\varphi)\|$$

Thereby, $0 = \|\mathcal{F}'[\varphi](\psi-\varphi)\|$ if and only if $0 = \|\mathcal{F}(\psi)-\mathcal{F}(\varphi)\|$. Hence, the injectivity of $\mathcal{F}'$ implies that $\mathcal{F}(\psi) \neq \mathcal{F}(\varphi)$ for all $\psi \in U_\varphi$ with $\psi \neq \varphi$. This

proves Lemma 1.

Before we come to the proof of Theorem 2, let us first formulate a result with deterministic error in the operator. We assume that $\mathcal{F}$ is approximated by some deterministic operator

$$\widehat{\mathcal{F}} : \mathfrak{B} \to \widehat{\mathcal{Y}}.$$

Let both $\mathcal{F}$ and $\widehat{\mathcal{F}}$ be Gateaux differentiable on $\mathfrak{B}$ with derivatives $\mathcal{F}'[\varphi]$ and $\widehat{\mathcal{F}}'[\varphi]$, which are "bounded with respect to Δ" in the sense that $\sup_{\{\tilde{\varphi}\in\mathfrak{B}:\Delta(\tilde{\varphi},\varphi)\neq 0\}} \|\mathcal{F}'[\varphi](\tilde{\varphi}-\varphi)\|^2/\Delta(\tilde{\varphi},\varphi) < \infty$ and $\mathcal{F}'[\varphi](\tilde{\varphi}-\varphi) \neq 0$ whenever $\Delta(\tilde{\varphi},\varphi) \neq 0$ and analogously for $\widehat{\mathcal{F}}$. The error of the approximation is described by:

$$\delta := \|\widehat{\mathcal{F}}(\varphi^\dagger)\|, \tag{32a}$$

$$\gamma := \left(\left| \sup_{\{\varphi\in\mathfrak{B}:\Delta(\varphi,\varphi^\dagger)\neq 0\}} \frac{\|\mathcal{F}'[\varphi^\dagger](\varphi-\varphi^\dagger)\|^2 - \|\widehat{\mathcal{F}}'[\varphi^\dagger](\varphi-\varphi^\dagger)\|^2}{\Delta(\varphi,\varphi^\dagger)} \right| \right)^{1/2}. \tag{32b}$$

Moreover, we assume that the tangential cone condition

$$\|\widehat{\mathcal{F}}(x) - \widehat{\mathcal{F}}(y) - \widehat{\mathcal{F}}'[y](x-y)\| \leq \eta \|\widehat{\mathcal{F}}(x) - \widehat{\mathcal{F}}(y)\|, \tag{33}$$

holds for all x, y in some neighborhood of $\varphi^\dagger$.

Lemma 4. *Assume that* (22), (32) *and* (33) *hold true with η sufficiently small, such that*

$$\frac{1-\eta}{\eta+\eta^2} > 4q^{3/2} + 1. \tag{34}$$

Further assume that solutions of the convex minimization problems (3) *exist and*

that the iteration is stopped at the smallest index $K \in \mathbb{N}_0$ for which

$$\alpha_{K+1} \leq \max(\Theta^{-1}(\delta), \gamma^2)\,, \qquad \textit{where } \Theta(t) := \sqrt{t}\Lambda(t). \tag{35}$$

In addition it should hold that $\alpha_0 > \max(\Theta^{-1}(\delta), \gamma^2)$ and $\alpha_k \leq q\alpha_{k+1}$ for all k with a constant $q > 1$. Moreover, let Λ be concave and assume that $t \mapsto \sqrt{t}/\Lambda(t)$ is monotonically increasing.

Then there exists a constant $C > 0$ independent of the $\widehat{\mathcal{F}}$ such that

$$\Delta(\widehat{\varphi}_K, \varphi^\dagger) \;\; \leq \;\; C\left(\Lambda\left(\max(\Theta^{-1}(\delta), \gamma^2)\right)\right)^2. \tag{36}$$

Proof. Let us introduce the following notation:

$$\begin{aligned} &\mathcal{T} := \mathcal{F}'[\varphi^\dagger], \qquad \widehat{\mathcal{T}} := \widehat{\mathcal{F}}'[\varphi^\dagger], \qquad \widehat{\mathcal{T}}_{k-1} := \widehat{\mathcal{F}}'[\widehat{\varphi}_{k-1}], \\ &\Delta_k := \Delta(\widehat{\varphi}_k, \varphi^\dagger), \qquad e_k := \widehat{\varphi}_k - \varphi^\dagger. \end{aligned}$$

From the optimality condition (3) with $\varphi = \varphi^\dagger$ we find that

$$\begin{aligned} &\|\widehat{\mathcal{T}}_{k-1}(\widehat{\varphi}_k - \widehat{\varphi}_{k-1}) + \widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2 + \alpha_k \mathcal{R}(\widehat{\varphi}_k) \\ &\qquad \leq \|\widehat{\mathcal{T}}_{k-1}(\varphi^\dagger - \widehat{\varphi}_{k-1}) + \widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2 + \alpha_k \mathcal{R}(\varphi^\dagger). \end{aligned} \tag{37}$$

From the definition (21) of the Bregman distance and the source condition (22) we obtain

$$\mathcal{R}(\varphi^\dagger) - \mathcal{R}(\widehat{\varphi}_k) = \left\langle \varphi_*^\dagger, \varphi^\dagger - \widehat{\varphi}_k \right\rangle - \Delta_k \;\leq\; \beta \Delta_k^{1/2} \Lambda\left(\frac{\|\mathcal{T} e_k\|^2}{\Delta_k}\right) - \Delta_k. \tag{38}$$

Plugging this into (37) yields

$$\begin{aligned}&\|\widehat{\mathcal{T}}_{k-1}(\widehat{\varphi}_k-\widehat{\varphi}_{k-1})+\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2+\alpha_k\Delta_k\\&\quad\le\|\widehat{\mathcal{T}}_{k-1}(\varphi^\dagger-\widehat{\varphi}_{k-1})+\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2+\beta\alpha_k\Delta_k^{1/2}\Lambda\left(\frac{\|\mathcal{T}e_k\|^2}{\Delta_k}\right).\end{aligned}\tag{39}$$

Note that the tangential cone condition (33) implies

$$(1+\eta)^{-1}\|\widehat{\mathcal{T}}e_k\|\le\|\widehat{\mathcal{F}}(\widehat{\varphi}_k)-\widehat{\mathcal{F}}(\varphi^\dagger)\|\le(1-\eta)^{-1}\|\widehat{\mathcal{T}}e_k\|.\tag{40}$$

To estimate the first term on the left hand side of (39) we use (33) and (40) to get that

$$\begin{aligned}&\|\widehat{\mathcal{F}}(\widehat{\varphi}_k)\|-\|\widehat{\mathcal{T}}_{k-1}(\widehat{\varphi}_k-\widehat{\varphi}_{k-1})+\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|\\&\quad\le\ \|\widehat{\mathcal{T}}_{k-1}(\widehat{\varphi}_k-\widehat{\varphi}_{k-1})+\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})-\widehat{\mathcal{F}}(\widehat{\varphi}_k)\|\\&\quad\le\ \eta\|\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})-\widehat{\mathcal{F}}(\widehat{\varphi}_k)\|\\&\quad\le\ \eta\|\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})-\widehat{\mathcal{F}}(\varphi^\dagger)\|+\eta\|\widehat{\mathcal{F}}(\widehat{\varphi}_k)-\widehat{\mathcal{F}}(\varphi^\dagger)\|\\&\quad\le\ \frac{\eta}{1-\eta}(\|\widehat{\mathcal{T}}e_k\|+\|\widehat{\mathcal{T}}e_{k-1}\|).\end{aligned}$$

Together with $\|\|\widehat{\mathcal{F}}(\widehat{\varphi}_k)\|\ge\|\widehat{\mathcal{F}}(\widehat{\varphi}_k)-\widehat{\mathcal{F}}(\varphi^\dagger)\|-\delta\ge(1+\eta)^{-1}\|\widehat{\mathcal{T}}e_k\|-\delta$ this yields

$$\|\widehat{\mathcal{T}}_{k-1}(\widehat{\varphi}_k-\widehat{\varphi}_{k-1})+\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|\ge\frac{1-2\eta-\eta^2}{1-\eta^2}\|\widehat{\mathcal{T}}e_k\|-\frac{\eta}{1-\eta}\|\widehat{\mathcal{T}}e_{k-1}\|-\delta.$$

For the right hand side of (39) we get from (32) and another application of (33) that

$$\|\widehat{\mathcal{T}}_{k-1}(\varphi^\dagger-\widehat{\varphi}_{k-1})+\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|\le\eta\|\widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})-\widehat{\mathcal{F}}(\varphi^\dagger)\|+\delta\le\frac{\eta}{1-\eta}\|\widehat{\mathcal{T}}e_{k-1}\|+\delta.$$

Plugging the last two inequalities into (39) and using the simple inequalities $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$ and $(a+b)^2 \leq 2a^2 + 2b^2$ we obtain that

$$\underbrace{\frac{1}{2}\left(\frac{1-2\eta-\eta^2}{1-\eta^2}\right)^2}_{=:C_\eta} \left\|\widehat{\mathcal{T}}e_k\right\|^2 + \alpha_k\Delta_k \\ \leq \underbrace{\frac{4\eta^2}{(1-\eta)^2}}_{=:c_\eta} \left\|\widehat{\mathcal{T}}e_{k-1}\right\|^2 + 4\delta^2 + \beta\alpha_k\Delta_k^{1/2}\Lambda\left(\frac{\|\mathcal{T}e_k\|^2}{\Delta_k}\right).$$

Using (32b) and the monotonicity of Λ we find that $\Lambda\left(\frac{\|\mathcal{T}e_k\|^2}{\Delta_k}\right) \leq \Lambda\left(\frac{\|\widehat{\mathcal{T}}e_k\|^2}{\Delta_k} + \gamma^2\right)$. Together with the stopping rule (35) this implies

$$C_\eta\|\widehat{\mathcal{T}}e_k\|^2 + \alpha_k\Delta_k \leq c_\eta\|\widehat{\mathcal{T}}e_{k-1}\|^2 + 4\Theta(\alpha_k)^2 + \beta\alpha_k\Delta_k^{1/2}\Lambda\left(\frac{\|\widehat{\mathcal{T}}e_k\|^2}{\Delta_k} + \alpha_k\right). \quad (41)$$

We will show the following error bounds

$$\|\widehat{\mathcal{T}}e_k\|^2 \leq C_1\Theta(\alpha_k)^2, \quad (42a)$$

$$\Delta(\widehat{\varphi}_k, \varphi^\dagger) \leq C_2\Lambda(\alpha_k)^2 \quad (42b)$$

with

$$C_1 := \max\left(\frac{\|\widehat{\mathcal{T}}e_0\|^2}{\Theta(\alpha_0)^2}, \frac{8}{C_\eta - 2q^3c_\eta}, \frac{16\beta^2}{C_\eta+1}, \frac{16\beta^2}{C_\eta^2}\right),$$
$$C_2 := \max\left(\frac{\Delta(\varphi_0,\varphi^\dagger)}{\Lambda(\alpha_0)^2}, 2C_1c_\eta q^3 + 8, 16\beta^2, \frac{16\beta^2}{C_\eta}\right).$$

We will prove these claims by induction in $k \leq K$. For $k = 0$ this is arranged by the definitions of C_1 and C_2. For the induction step we distinguish two cases:

Case 1: $c_\eta\|\widehat{\mathcal{T}}e_{k-1}\|^2 + 4\Theta(\alpha_k)^2 \geq \beta\alpha_k\Delta_k^{1/2}\Lambda\left(\frac{\|\widehat{\mathcal{T}}e_k\|^2}{\Delta_k} + \alpha_k\right)$.

Now by using the induction hypothesis (42a) equation (41) simplifies to

$$C_\eta \|\widehat{\mathcal{T}} e_k\|^2 + \alpha_k \Delta_k \leq 2c_\eta C_1 \Theta(\alpha_{k-1})^2 + 8\Theta(\alpha_k)^2.$$

We have $\Theta(\alpha_{k-1}) = (\alpha_{k-1})^{1/2}\Lambda(\alpha_{k-1}) \leq (q\alpha_k)^{1/2}\Lambda(q\alpha_k)$ as Λ is monotonically increasing. While Λ is concave and $\Lambda(0) = 0$ the definition of concavity implies $t\Lambda(x) \leq \Lambda(tx)$ for $0 \geq t \geq 1$. Now taking $x = q\alpha_k$ and $t = q^{-1}$ gives $\Lambda(q\alpha_k) \leq q\Lambda(\alpha_k)$ and therefore

$$\Theta(\alpha_{k-1}) \leq q^{3/2}\Theta(\alpha_k).$$

Putting the last two equations together results into the bound

$$C_\eta \|\widehat{\mathcal{T}} e_k\|^2 + \alpha_k \Delta_k \leq (2c_\eta C_1 q^3 + 8)\Theta(\alpha_k)^2 = (2c_\eta C_1 q^3 + 8)\alpha_k \Lambda(\alpha_k)^2.$$

Firstly, this implies by omitting the second term on the left hand side that

$$\|\widehat{\mathcal{T}} e_k\|^2 \leq \frac{2c_\eta C_1 q^3 + 8}{C_\eta}\Theta(\alpha_k)^2 \quad \text{and hence} \quad C_1 \geq \frac{2c_\eta C_1 q^3 + 8}{C_\eta}.$$

Hence, it is necessary that $C_\eta > 2q^3 c_\eta$, which is equivalent to the inequality (34) assumed in the Lemma. Then (42a) is true with $C_1 \geq \dfrac{8}{C_\eta - 2q^3 c_\eta}$.
Secondly, omitting the first term of the left hand side shows $\Delta_k \leq (2c_\eta C_1 q^3 + 8)\Lambda(\alpha_k)^2$, so we have (42b) with $C_2 \geq 2c_\eta C_1 q^3 + 8$.

Case 2: $\beta\alpha_k \Delta_k^{1/2} \Lambda\left(\frac{\|\widehat{\mathcal{T}} e_k\|^2}{\Delta_k} + \alpha_k\right) \geq c_\eta \|\widehat{\mathcal{T}} e_{k-1}\|^2 + 4\Theta(\alpha_k)^2$.
In this case (41) simplifies to

$$C_\eta \|\widehat{\mathcal{T}} e_k\|^2 + \alpha_k \Delta_k \leq 2\beta\alpha_k \Delta_k^{1/2} \left(\Lambda\left(\frac{\|\widehat{\mathcal{T}} e_k\|^2}{\Delta_k} + \alpha_k\right)\right).$$

Using again $\Lambda(0) = 0$ and the concavity we get $\Lambda(x) \geq \frac{x}{(a+b)}\Lambda(a+b)$ for all $0 \leq x \leq a+b$. Taking now $x = a$ and $x = b$ respectively implies $\Lambda(a) + \Lambda(b) \geq \Lambda(a+b)$. Thus we have

$$C_\eta \|\widehat{\mathcal{T}} e_k\|^2 + \alpha_k \Delta_k \leq 2\beta \alpha_k \Delta_k^{1/2} \left(\Lambda\left(\frac{\|\widehat{\mathcal{T}} e_k\|^2}{\Delta_k} \right) + \Lambda(\alpha_k) \right). \tag{43}$$

It is again convenient to study two cases:

Case 2.1: $\|\widehat{\mathcal{T}} e_k\|^2 \leq \alpha_k \Delta_k$.

Now the monotonicity of Λ entails

$$C_\eta \|\widehat{\mathcal{T}} e_k\|^2 + \alpha_k \Delta_k \leq 4\beta \alpha_k \Delta_k^{1/2} \Lambda\left(\alpha_k\right).$$

This shows that $\Delta_k^{1/2} \leq 4\beta \Lambda\left(\alpha_k\right)$ and thereby (42b) with $C_2 \geq 16\beta^2$. Plugging this into the right hand side of the last inequality and using the case assumption for the left hand side we get

$$(1 + C_\eta) \|\widehat{\mathcal{T}} e_k\|^2 \leq 16\beta^2 \alpha_k \Lambda\left(\alpha_k\right)^2 = 16\beta^2 \Theta\left(\alpha_k\right)^2.$$

Hence (42a) holds with $C_1 \geq \dfrac{16\beta^2}{1 + C_\eta}$.

Case 2.2: $\alpha_k \Delta_k \leq \|\widehat{\mathcal{T}} e_k\|^2$.

Dividing formula (43) by $\|\widehat{\mathcal{T}} e_k\|$ results in

$$C_\eta \|\widehat{\mathcal{T}} e_k\| + \frac{\alpha_k \Delta_k}{\|\widehat{\mathcal{T}} e_k\|} \leq 2\beta \alpha_k \left(\frac{\Delta_k}{\|\widehat{\mathcal{T}} e_k\|^2} \right)^{1/2} \left(\Lambda\left(\frac{\|\widehat{\mathcal{T}} e_k\|^2}{\Delta_k} \right) + \Lambda(\alpha_k) \right).$$

Since the functions $t^{-1/2}\Lambda(t)$ and $t^{-1/2}$ are monotonically decreasing, we obtain

$$C_\eta \|\widehat{\mathcal{T}} e_k\| + \frac{\alpha_k \Delta_k}{\|\widehat{\mathcal{T}} e_k\|} \leq 4\beta \alpha_k^{1/2} \Lambda(\alpha_k).$$

This shows that $C_\eta \|\widehat{\mathcal{T}} e_k\| \leq 4\beta\Theta(\alpha_k)$, so (42a) is true with $C_1 \geq \dfrac{16\beta^2}{C_\eta^2}$. Plugging this into the left hand side of the last equation gives

$$\frac{\alpha_k \Delta_k C_\eta}{4\beta \alpha_k^{1/2} \Lambda(\alpha_k)} \leq 4\beta \alpha_k^{1/2} \Lambda(\alpha_k).$$

Now we see that $\Delta_k \leq 16\beta^2 \Lambda(\alpha_k)^2 / C_\eta$ and therefore that (42b) is valid with $C_2 \geq 16\beta^2/C_\eta$. This completes the proof.

Now Theorem 2 follows easily:

Proof of Theorem 2. The constant C in the last lemma is independent of δ and γ. So if δ and γ converge to 0 in probability and if the probability that the tangential cone condition is not fulfilled goes to 0, this implies convergence in probability of $\Delta(\widehat{\varphi}_K, \varphi^\dagger)$. That is the assertion of Theorem 2.

References

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

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

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

Bissantz, N., Hohage, T., and Munk, A. 2004. Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems*, 20:1773–1791.

Blaschke, B., Neubauer, A., and Scherzer, O. 1997. On convergence rates for the iteratively regularized Gauss-Newton method. *IMA Journal of Numerical Analysis*, 17:421–436.

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

Botev, Z. I., Grotowski, J. F., and Kroese, D. P. 2010. Kernel density estimation via diffusion. *Annals of Statistics*, 38(5):2916–2957.

Breunig, C. and Johannes, J. 2009. On rate optimal local estimation in nonparametric instrumental regression. *arXiv:0902.2103v1.*

Burger, M. and Osher, S. 2004. Convergence rates of convex variational regularization. *Inverse Problems*, 20(5):1411–1421.

Carrasco, M., Florens, J.-P., and Renault, E. 2006. Linear inverse problems in structural econometrics: Estimation based on spectral decomposition and regularization. In *Handbook of Econometrics*, volume 6. North Holland.

Chen, X., Chernozhukov, V., Lee, S., and Newey, W. K. 2011. Local identification of nonparametric and semiparametric models. *Cowles Foundation Discussion Paper No. 1795.*

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

Chen, X. and Reiss, M. 2010. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27:497–521.

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

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

Darolles, S., Fan, Y., Florens, J.-P., and Renault, E. 2011. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

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

Eggermont, P. P. B. 1993. Maximum entropy regularization for fredholm integral equations of the first kind. *SIAM J. Math. Anal.*, 24:1557–1576.

Engl, H. W., Hanke, M., and Neubauer, A. 1996. *Regularization of Inverse Problems*. Kluwer Academic Publisher, Dordrecht, Boston, London.

Engl, H. W., Kunisch, K., and Neubauer, A. 1989. Convergence rates for Tikhonov regularization of nonlinear ill-posed problems. *Inverse Problems*, 5:523–540.

Florens, J.-P. 2003. Inverse problems and structural economics: The example of instrumental variables. In Dewatripont, M., Hansen, L. P., and Turnovsky, S.,

editors, *Advances in Economics and Econometrics: Theory and Applications*, pages 284–311. Cambridge Univ. Press.

Florens, J.-P. and Sbaï, E. 2010. Local identification in empirical games of incomplete information. *Econometric Theory*, 26:1638–1662.

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

Hanke, M., Neubauer, A., and Scherzer, O. 1995. A convergence analysis of the Landweber iteration for nonlinear ill-posed problems. *Numer. Math.*, 72(1):21–37.

Hofmann, B., Kaltenbacher, B., Pöschl, C., and Scherzer, O. 2007. A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Problems*, 23(3):987–1010.

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

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

Imbens, G. W. and Newey, W. K. 2009. Identification and estimation of triangular simulataneous equations without monotonicity. *Econometrica*, 77:1481–1512.

Kaltenbacher, B. and Hofmann, B. 2010. Convergence rates for the iteratively regularized Gauss-Newton method in Banach spaces. *Inverse Problems*, 26(3):035007, 21.

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.

Kress, R. 1999. *Linear Integral Equations*. Springer Verlag, Berlin, Heidelberg, New York, 2nd edition.

Langer, S. and Hohage, T. 2007. Convergence analysis of an inexact iteratively regularized Gauss-Newton method under general source conditions. *J. Inverse Ill-Posed Probl.*, 15(3):311–327.

Letac, G. 1995. *Exercises and Solutions Manual for Integration and Probability by Paul Malliavin*. Springer. Translated by Kay, L.

Loubes, J.-M. and Pelletier, B. 2008. Maximum entropy solution to ill-posed inverse problems with approximately known operator. *Journal of Mathematical Analysis and Applications*, 344(1):260–273.

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

Prössdorf, S. and Silbermann, B. 1991. *Numerical Analysis for Integral and Related Operator Equations*. Birkhäuser, Basel.

Resmerita, E. 2005. Regularization of ill-posed problems in Banach spaces: convergence rates. *Inverse Problems*, 21(4):1303–1314.

Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., and Lenzen, F. 2009. *Variational methods in imaging*, volume 167 of *Applied Mathematical Sciences*. Springer, New York.

Torgovitsky, A. 2012. Identification of nonseparable models with general instruments. *Preprint*.