



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Institut für Numerische und Angewandte Mathematik

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

F. Dunker, J.-P. Florens, T. Hohage, J. Johannes, E. Mammen

Nr. 2011-15

Preprint-Serie des
Instituts für Numerische und Angewandte Mathematik
Lotzestr. 16-18
D - 37083 Göttingen

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

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

September 27, 2011

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

[‡]Universite de Toulouse (GREMAQ and IDEI)

[§]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. Since such operator equations are almost always ill-posed in the sense that \mathcal{F}^{-1} is not continuous, the straightforward estimator $\widehat{\mathcal{F}}^{-1}(0)$ will not be consistent in general. Regularization techniques must be applied to solve (1) or its empirical version $\widehat{\mathcal{F}}(\widehat{\varphi}) = 0$. We will use a generalized version of the iteratively regularized Gauß-Newton method. In numerical analysis this is one of the most popular computational methods for solving nonlinear ill-posed operator equations. It avoids some problems of nonlinear Tikhonov regularization given by

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

where φ_0 is some initial guess of φ . The iteratively regularized Gauß-Newton method does not suffer from the problem that minima of the functional in (2) are in general not unique and it avoids computational difficulties due to the presence of local minima. Moreover, instead of a quadratic penalty, we allow for a more general penalty term $\mathcal{R} : \mathfrak{B} \rightarrow (-\infty, \infty]$ 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

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

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. The most common choice for the penalty term is $\mathcal{R}(\varphi) = \|\varphi - \varphi_0\|_{\mathcal{X}}^2$ where $\|\cdot\|_{\mathcal{X}}$ is the norm of the Hilbert space \mathcal{X} and where 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 allows for a flexible incorporation of further a-priori information. Common choices are entropy regularization, l^1 penalties and BV penalties. Loubes and Pelletier (2008) studied entropy

regularization for instrumental variable models but they gave no theoretical results for their rates of convergence of the estimators. The use of l^1 penalties enhances sparsity properties of the estimator. A BV penalty results in piecewise constant estimators.

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 1. 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 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).

For nonlinear Tikhonov regularization convergence rates 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 estimation is considered in a quantile regression model. This is one example of a statistical model where the unknown nonparametric function is given as solution of a nonlinear integral equation. We will describe this model in the next section. In Horowitz and Lee (2007) it is assumed that the singular values of the Fréchet derivative $\mathcal{F}'[\varphi^\dagger]$ decay polynomially and results are given on the rates of the estimators under these assumptions. In the paper it is pointed out 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.

A particular focus of this paper is on instrumental regression models 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 endogenous 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 the 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), Chen and Reiss (2010) and Breunig and Johannes (2009). In this model the regression function φ is defined as 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, typically in specific econometric applications, the conditional mean assumption (5b) is 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). As an example we will demonstrate that φ can be identifiable even if W is binary and Z is one-dimensional and continuously distributed. Hence, the model (4) contains strictly 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 following 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, and discuss variational forms of source conditions, which are classically defined in terms of spectral theory. In Section 4 we present our main convergence result for the iteratively regularized Gauß-Newton method with noisy operators. Section 5 reports on numerical simulations for a instrumental variable regression model with binary instruments.

2 Examples

2.1 Instrumental quantile regression

In Horowitz and Lee (2007) the following quantile regression model has been studied:

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

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

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, that is independent from an observed instrument W .

We assume that Y , Z and W have a joint density f_{YZW} with respect to the Lebesgue measure. Let $F_{YZW}(y, z, w) := \int_{-\infty}^y f_{YZW}(\tilde{y}, z, w) d\tilde{y}$, 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 given by

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

As pointed out in Horowitz and Lee (2007), the model (7) subsumes nonseparable 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 Instrumental regression with independent instruments

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 f_{YZW} denotes the joint density of (Y, Z, W) , and f_{YZ} and f_W denote the marginal densities of (Y, Z) or 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 (9) can be written as a nonlinear ill-posed operator equation (1) with the operator

$$(\mathcal{F}(\varphi))(u, w) := \left(\begin{array}{c} \int f_{YZW}(u + \varphi(z), z, w) dz - \int f_{YZ}(u + \varphi(z), z) f_W(w) dz \\ \int \varphi(z) f_Z(z) dz - \int y f_Y(y) dy \end{array} \right). \quad (10)$$

Alternatively, it may be advantageous to integrate (9a) once with respect to u . Introducing $F_{YZW}(\tilde{y}, z, w) := \int_{-\infty}^{\tilde{y}} f_{YZW}(y, z, w) dy$ and $F_{YZ}(\tilde{y}, z) := \int_{-\infty}^{\tilde{y}} f_{YZ}(y, z) dz$ yields to an alternative operator formulation of the model (4) with the operator

$$(\tilde{\mathcal{F}}(\varphi))(\tilde{u}, w) := \left(\begin{array}{c} \int F_{YZW}(\tilde{u} + \varphi(z), z, w) dz - \int F_{YZ}(\tilde{u} + \varphi(z), z) f_W(w) dz \\ \int \varphi(z) f_Z(z) dz - \int y f_Y(y) dy \end{array} \right). \quad (11)$$

2.3 Binary instruments

We consider the following special case of the last subsection: the instrument W is binary and it only takes the values 0 and 1. Furthermore, the explanatory variable Z is a scalar. Then the marginal distribution f_W has the two values

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

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

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

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

$$\int w_1 f_{YZW}(u + \varphi(z), z, 0) - w_0 f_{YZW}(u + \varphi(z), z, 1) dz = 0 \quad \text{for all } u. \quad (12)$$

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

$$(\mathcal{F}(\varphi))(u) := \left(\begin{array}{c} \int w_1 f_{YZW}(u + \varphi(z), z, 0) - w_0 f_{YZW}(u + \varphi(z), z, 1) dz \\ \int \varphi(z) f_Z(z) dz - \int y f_Y(y) dy \end{array} \right). \quad (13)$$

In analogy to (11), the equation $\mathcal{F}(\varphi) = 0$ can equivalently be rewritten as $\tilde{\mathcal{F}}(\varphi) = 0$ with

$$(\tilde{\mathcal{F}}(\varphi))(u) := \left(\begin{array}{c} \int w_1 F_{YZW}(u + \varphi(z), z, 0) - w_0 F_{YZW}(u + \varphi(z), z, 1) dz \\ \int \varphi(z) f_Z(z) dz - \int y f_Y(y) dy \end{array} \right) \quad (14)$$

Remark (on identifiability): By dimensionality it is necessary for identifiability that Z is scalar-valued or discrete. In the following we discuss sufficient

conditions for the injectivity of the Fréchet derivative $\tilde{\mathcal{F}}'[\varphi^\dagger]$ at the exact solution φ^\dagger . For ill-posed problems injectivity of $\tilde{\mathcal{F}}'[\varphi^\dagger]$ does not necessarily imply local identifiability of the nonlinear problem in an open neighborhood of φ^\dagger , but we refer to Chen et al. (2011) for additional assumptions, which do guarantee local identifiability in a certain sense. See also Florens and Sbaï (2010).

Since

$$(\tilde{\mathcal{F}}'[\varphi^\dagger]\tilde{\varphi})(u) = \left(\begin{array}{c} \int [w_1 f_{YZW}(u + \varphi^\dagger(z), z, 0)\tilde{\varphi}(z) - w_0 f_{YZW}(u + \varphi^\dagger(z), z, 1)] \tilde{\varphi}(z) dz \\ \int \tilde{\varphi}(z) f_Z(z) dz \end{array} \right),$$

the first equation in $\tilde{\mathcal{F}}'[\varphi^\dagger]\tilde{\varphi} = 0$ can be rewritten as

$$\mathbb{E}[\tilde{\varphi}(Z)|U = u, W = 0] = \mathbb{E}[\tilde{\varphi}(Z)|U = u, W = 1].$$

This implies $\tilde{\varphi} = 0$ if the dependence structure between Z and U varies sufficiently with W . As an example, consider the following situation with $\rho_0 \neq \rho_1$:

$$\begin{aligned} \begin{pmatrix} Z \\ U \end{pmatrix} \Big| W = 0 &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}\right) \\ \begin{pmatrix} Z \\ U \end{pmatrix} \Big| W = 1 &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix}\right) \end{aligned}$$

Then the density of $(Z, U)|W = w$ can be written as

$$\begin{aligned} f_{ZU|W=w}(z, u) &= \frac{1}{2\pi\sqrt{1-\rho_w^2}} \exp\left(-\frac{z^2 - 2\rho_w zu + u^2}{2(1-\rho_w^2)}\right) \\ &= \frac{1}{2\pi\sqrt{1-\rho_w^2}} \exp\left(-\frac{z^2}{2(1-\rho_w^2)} + \frac{\rho_w zu}{1-\rho_w^2} - \frac{\rho_w^2 u^2}{2(1-\rho_w^2)} - \frac{u^2}{2}\right). \end{aligned}$$

Denote by H_m the Hermite polynomials and by $\psi_m(z) := (m!\sqrt{2\pi})^{-1/2} \exp(-x^2/2)H_m(z)$ the Hermite functions. In the following computation we use the substitution of variable $\tilde{z} = z/\sqrt{1-\rho_w}$, the generating function of the Hermite polynomials $\exp(\tilde{z}t - t^2/2) = \sum_{m=0}^{\infty} H_m(\tilde{z})t^m/m!$ and the L^2 -orthonormality of the Hermite functions:

$$\begin{aligned} \mathbb{E}[\tilde{\varphi}(Z)|U = u, W = w] &= \int \frac{\tilde{\varphi}(z)}{2\pi\sqrt{1-\rho_w^2}} \exp\left(-\frac{z^2}{2(1-\rho_w^2)} + \frac{\rho_w zu}{1-\rho_w^2} - \frac{\rho_w^2 u^2}{2(1-\rho_w^2)} - \frac{u^2}{2}\right) dz \\ &= \frac{1}{2\pi} \exp\left(-\frac{u^2}{2}\right) \int \tilde{\varphi}(\tilde{z}) \exp\left(-\frac{\tilde{z}^2}{2} + \frac{\rho_w \tilde{z}u}{\sqrt{1-\rho_w^2}} - \frac{\rho_w^2 u^2}{2(1-\rho_w^2)}\right) d\tilde{z} \\ &= \frac{1}{2\pi} \exp\left(-\frac{u^2}{2}\right) \int \tilde{\varphi}(\tilde{z}) \exp\left(-\frac{\tilde{z}^2}{2}\right) \left(\sum_{m=0}^{\infty} H_m(\tilde{z}) \frac{(\rho_w u)^m}{m!(\sqrt{1-\rho_w^2})^m}\right) d\tilde{z} \\ &= (2\pi)^{-3/4} \exp\left(-\frac{u^2}{2}\right) \sum_{m=0}^{\infty} \frac{(\rho_w u)^m}{\sqrt{m!(1-\rho_w^2)^m}} \int \tilde{\varphi}(\tilde{z}) \psi_m(\tilde{z}) d\tilde{z} \\ &= (2\pi)^{-3/4} \exp\left(-\frac{u^2}{2}\right) \sum_{m=0}^{\infty} \left(\frac{\rho_w}{\sqrt{1-\rho_w^2}}\right)^m \langle \tilde{\varphi}, \psi_m \rangle_{L^2} \frac{u^m}{\sqrt{m!}} \end{aligned}$$

Hence we can write $\mathbb{E}[\tilde{\varphi}(Z)|U = u, W = 0]$ and $\mathbb{E}[\tilde{\varphi}(Z)|U = u, W = 1]$ as two convergent power series. If $\mathbb{E}[\tilde{\varphi}(Z)|U = u, W = 0] = \mathbb{E}[\tilde{\varphi}(Z)|U = u, W = 1]$ the coefficients of both series must be equal by the identity theorem of power series. As $\rho_0 \neq \rho_1$, we have $(\rho_0/\sqrt{1-\rho_0^2})^m \neq (\rho_1/\sqrt{1-\rho_1^2})^m$ because $\rho \mapsto (\rho/\sqrt{1-\rho^2})^m$ is strictly monotonically increasing for all $m \geq 1$ and $\rho \in [0, 1)$. Thus $\langle \tilde{\varphi}, \psi_m \rangle_{L^2} = 0$ for all $m \geq 1$, so $\tilde{\varphi} = \text{const}$. Now the second equation in $\tilde{\mathcal{F}}'[\varphi^\dagger]\tilde{\varphi} = 0$ implies $\tilde{\varphi} = 0$, which completes the proof of injectivity of $\tilde{\mathcal{F}}'[\varphi]$ in our example.

Section 5 contains further numerical evidence of identifiability in a particular case.

We emphasize that Z is not necessarily discrete, 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.

3.1 The rate of decay of singular values of linear integral operators

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

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

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 K , then according to the Courant-Fischer characterization (see e.g. Kress (1999)) of the singular values the operator K_j with kernel $k_j(x, y) := \sum_{l=0}^{j-1} \sigma_l v_l(x) u_l(y)$ satisfies

$$\sigma_j = \|K - K_j\| = \inf\{\|K - \tilde{K}\| : \text{rank } \tilde{K} \leq j\}. \quad (15)$$

In particular, 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 $\int_{[0,1]^{d_1}} \int_{[0,1]^{d_2}} |k(x, y) - \sum_{l=0}^{j-1} \tilde{u}_l(x)\tilde{v}_l(y)|^2 dx dy \leq \tilde{\sigma}_j$, then $\sigma_j \leq \tilde{\sigma}_j$ since $\|K - K_j\| \leq \|k - k_j\|_{L^2([0,1]^{d_1+d_2})}$. 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.

3.2 Classical source conditions in terms of spectral theory

In regularization theory, smoothness of the solution φ^\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 $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{Y}$ between Hilbert spaces \mathcal{X} and \mathcal{Y} , such source conditions have the form

$$\varphi^\dagger - \varphi_0 = \Lambda(\mathcal{T}^*\mathcal{T})\psi. \quad (16)$$

Here $\psi \in \mathcal{X}$, $\Lambda : [0, \infty) \rightarrow [0, \infty)$ is a continuous, strictly monotonically increasing function with $\Lambda(0) = 0$, and φ_0 is an initial guess (typically $\varphi_0 = 0$ in the linear case). $\Lambda(\mathcal{T}^*\mathcal{T})$ denotes 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 $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ the Fréchet derivative $\mathcal{T} = \mathcal{F}'[\varphi^\dagger]$ at φ^\dagger is used. For the most common choice $\Lambda(t) = t^\mu$ for some $\mu > 0$ these conditions 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.

Let us discuss such conditions in the context of nonparametric instrumental variable models. In general, it is reasonable to assume that the kernels of the integral operators in nonparametric instrumental regression are infinitely smooth. Therefore, their singular values σ_j decay super-algebraically. As a consequence, Hölder-type source conditions are extremely restrictive smoothness conditions since the eigenvalues $\lambda_j((\mathcal{T}^*\mathcal{T})^\nu) = \sigma_j^{2\nu}$ decay super-algebraically. Hence, Hölder-type source conditions imply that the Fourier coefficients of the solution with respect to $\{u_j : j \in \mathbb{N}_0\}$ decay super-algebraically, whereas one would typically expect only polynomial decay corresponding to the finite Sobolev 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 the solution. For kernels, which are infinitely smooth, but not analytic, index functions Λ in between Hölder and logarithmic are appropriate, e.g. the family $\Lambda(t) = \exp(-\frac{1}{2}(-\ln t)^\theta)$ parameterized by $0 < \theta < 1$.

3.3 Variational source conditions

In our analysis we will make use of variational methods in regularization theory, which have been explored recently in a number of papers as an alternative to spectral methods. Note that if \mathcal{X} is a Banach space, which we assume in the following, 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. We will prove convergence results in terms

of the Bregman distance in \mathcal{X} with respect to the convex functional \mathcal{R} . If $\varphi^\dagger \in \partial\mathcal{R}(\varphi^\dagger)$ (i.e. $\varphi^\dagger_* = \mathcal{R}'[\varphi^\dagger]$ if \mathcal{R} is differentiable at φ^\dagger), then the Bregman distance with respect to \mathcal{R} and φ^\dagger_* is defined as

$$\Delta(\varphi, \varphi^\dagger) := \mathcal{R}(\varphi) - \mathcal{R}(\varphi^\dagger) - \langle \varphi^\dagger_*, \varphi - \varphi^\dagger \rangle. \quad (17)$$

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, if \mathcal{X} is a Hilbert space and $\mathcal{R}(\varphi) = \|\varphi - \varphi_0\|_{\mathcal{X}}^2$ for some $\varphi_0 \in \mathcal{X}$, then

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

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 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

$$\langle \varphi^\dagger_*, \varphi^\dagger - \varphi \rangle \leq \beta \Delta(\varphi, \varphi^\dagger)^{1/2} \Lambda \left(\frac{\|\mathcal{F}'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2}{\Delta(\varphi, \varphi^\dagger)} \right) \quad \text{for all } \varphi \in \mathfrak{B}. \quad (18)$$

If \mathcal{X} is a Hilbert space, $\mathcal{R}(\varphi) = \|\varphi - \varphi_0\|^2$, and a classical source condition (16) is satisfied with a function Λ such that $(\Lambda^2)^{-1}$ is convex, then it follows from Jensen's inequality that

$$\langle \varphi^\dagger_*, \varphi^\dagger - \varphi \rangle = \langle \psi, \Lambda(\mathcal{T}^* \mathcal{T}) \psi \rangle \leq \|\psi\| \|\varphi - \varphi^\dagger\| \Lambda \left(\frac{\|\mathcal{T}(\varphi - \varphi^\dagger)\|^2}{\|\varphi - \varphi^\dagger\|^2} \right)$$

for all $\varphi \in \mathfrak{B}$, i.e. classical source conditions imply variational source conditions. Note that if \mathfrak{B} is chosen such that φ^\dagger is on the boundary of \mathfrak{B} , then possibly Λ can be chosen smaller than for the case where φ^\dagger is in the interior of \mathfrak{B} . In this context it is important that no absolute values appear on the left hand side of (18) as opposed to the formulation in Kaltenbacher and Hofmann (2010). Therefore, our analysis captures the fact that convex constraints may lead to improved rates of convergence.

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 $\mathcal{F} : \mathfrak{B} \rightarrow \mathcal{Y}$:

$$\mathcal{F}(\varphi^\dagger) = 0. \quad (19)$$

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

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

which maps to some (possibly finite-dimensional and/or data dependent) Hilbert space $\widehat{\mathcal{Y}}_n$. \mathcal{F} and all $\widehat{\mathcal{F}}_n$ 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 $\sup_{\{\tilde{\varphi} \in \mathfrak{B} : \Delta(\tilde{\varphi}, \varphi) \neq 0\}} \|\mathcal{F}'[\varphi](\tilde{\varphi} - \varphi)\|^2 / \Delta(\tilde{\varphi}, \varphi) < \infty$ and $\mathcal{F}'[\varphi](\tilde{\varphi} - \varphi) \neq 0$ 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 1. *Let (18) 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:*

$$\|\widehat{\mathcal{F}}_n(\varphi^\dagger)\| = O_p(\delta_n), \quad (20a)$$

$$\left(\sup_{\varphi \in \mathfrak{B}} \frac{\|\mathcal{F}'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2 - \|\widehat{\mathcal{F}}'_n[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2}{\Delta(\varphi, \varphi^\dagger)} \right)^{1/2} = O_p(\gamma_n), \quad (20b)$$

$$P\{\|\widehat{\mathcal{F}}_n(\varphi_1) - \widehat{\mathcal{F}}_n(\varphi_2) - \widehat{\mathcal{F}}'_n[\varphi_2](\varphi_1 - \varphi_2)\| > \eta \|\widehat{\mathcal{F}}_n(\varphi_1) - \widehat{\mathcal{F}}_n(\varphi_2)\| \text{ for some } \varphi_1, \varphi_2 \in \mathfrak{B}\} \xrightarrow{(20c)} 0.$$

Here η is sufficiently small. Suppose that the convex minimization problems (3) are uniquely solvable 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

$$\alpha_{K_n+1} \leq \max(\Theta^{-1}(\delta_n), \gamma_n^2), \quad \text{where } \Theta(t) := \sqrt{t}\Lambda(t). \quad (21)$$

Then

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

Remarks:

1. Sufficient conditions for uniqueness of solutions to the minimization problem (3) are strict convexity of \mathcal{R} or injectivity of $\widehat{\mathcal{F}}'_n[\widehat{\varphi}_{k-1}]$.
2. Sufficient conditions for existence 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 $\varphi_* \in \mathfrak{B}$. Since the Tikhonov functional is convex and lower semi-continuous, it is also weakly lower semi-continuous, and hence φ_* is a minimizer.

3. Note that if \mathcal{X} is a Hilbert space and $\widehat{\mathcal{F}}_n$ Fréchet differentiable, 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 $\widehat{\mathcal{F}}_n$

$$\|\widehat{\mathcal{F}}_n(x) - \widehat{\mathcal{F}}_n(y) - \widehat{\mathcal{F}}'_n[y](x - y)\| \leq \eta \|\widehat{\mathcal{F}}_n(x) - \widehat{\mathcal{F}}_n(y)\|, \quad (22)$$

used in (20c) is known as the tangential cone condition. This condition is commonly used in the analysis of regularization methods for nonlinear ill-posed problems Kaltenbacher et al. (2008). The right hand side of (22) may be replaced by $\|F'[y](x - y)\|$ (see (33) below), and in this form it corresponds to Assumption 2 in Chen et al. (2011).

Corollary 2. *Let the assumptions of Theorem 1 hold true.*

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

$$\Delta(\widehat{\varphi}_K, \varphi^\dagger)^{1/2} = O_p(\max(\delta^{2\mu/(2\mu+1)}, \gamma^{2\mu})). \quad (23)$$

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*

$$\Delta(\widehat{\varphi}_K, \varphi^\dagger)^{1/2} \leq O_p((-\ln \max(\delta, \gamma))^{-p}) \quad (24)$$

for all δ, γ sufficiently small.

5 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 (13). 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

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

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

$$\begin{aligned} f_{YZW}(y, z, 0) &= f_{ZW}(z, 0) \frac{1}{0.09\sqrt{2}} \exp\left(-\frac{1}{2} \left(\frac{y - \varphi^\dagger(z) - \mu_0(z)}{0.09}\right)^2\right), \\ f_{YZW}(y, z, 1) &= f_{ZW}(z, 1) \frac{1}{0.09\sqrt{2}} \exp\left(-\frac{1}{2} \left(\frac{y - \varphi^\dagger(z) - \mu_1(z)}{0.09}\right)^2\right). \end{aligned} \quad (25)$$

Now f_{ZW} has to be determined such that W and U are independent, which is equivalent to (12). Let us set $f_{ZW}(z, 1) := 0.625f_{ZW}(1.25z - 0.125, 0)$ for this purpose. With a substitution of variables we compute

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

This shows that (12) 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.

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

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: $\tilde{\varphi}(z) = w_0 \mu_0(z) + w_1 \mu_1(z) + \varphi^\dagger$

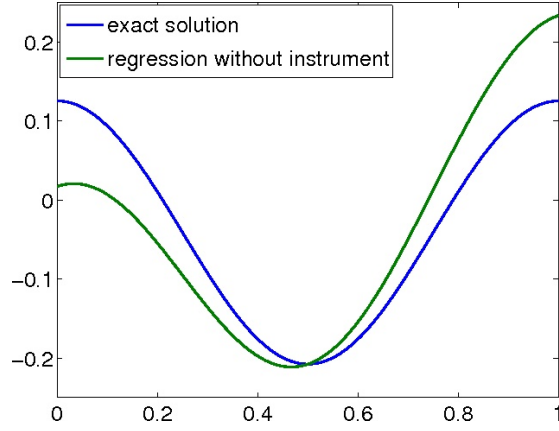


Figure 1: Necessity of the instrument: A standard regression would asymptotically yield the green curve $\tilde{\varphi}$ which is far away from the true curve φ^\dagger in blue.

To approximately solve the integral operator equation (13) 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). The singular values of $\mathcal{F}'[\varphi^\dagger]$ are shown in Figure 3. They exhibit an exponential decay, so according to Corollary 2 we can only expect slow rates of convergence.

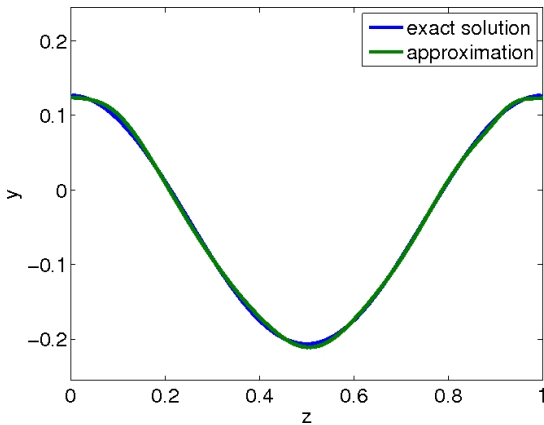


Figure 2: Result of the iterative inversion using the exact density f_{YZW} .

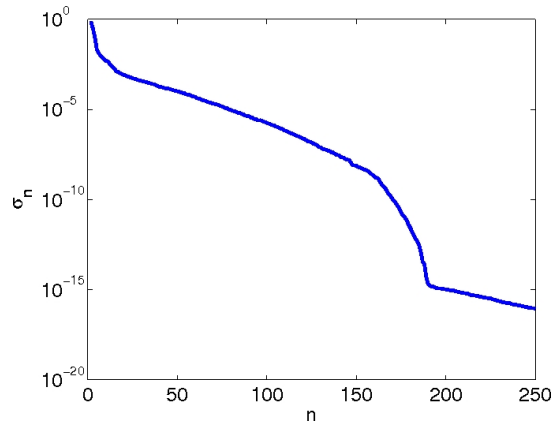


Figure 3: Singular values of $\mathcal{F}'[\varphi^\dagger]$

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 es-

timated non-parametrically by the kernel density estimator developed by Botev et al. (2010). Afterwards again (13) 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. 4–6 show the L^2 -errors of the approximate solution normed by the error of the initial guess (i.e. the error of the 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. 7–9 show median reconstructions for each sample size. The results demonstrate that our method computes an asymptotically correct estimator of the regression function φ^\dagger with an endogeneous explanatory variable Z using only a binary instrument W .

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

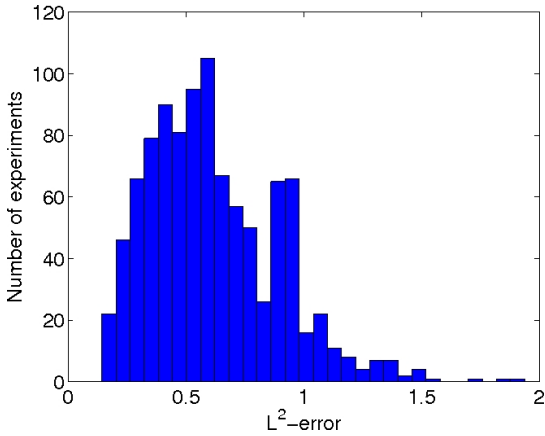


Figure 4: L^2 error for sample size $N = 10^3$

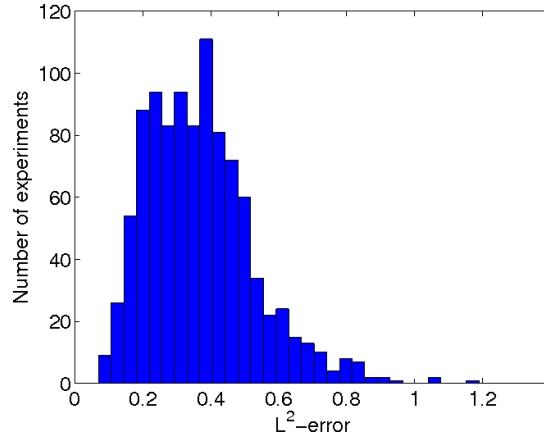


Figure 5: L^2 error for sample size $N = 10^4$

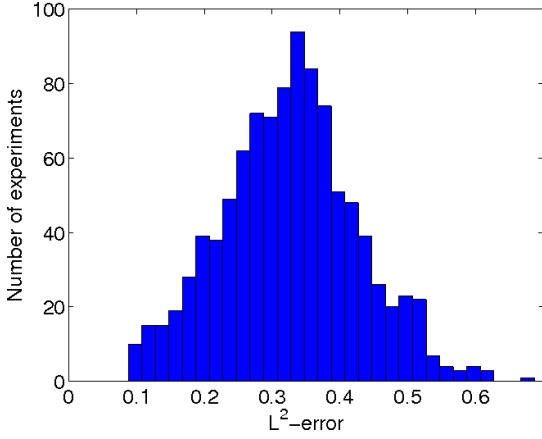


Figure 6: L^2 error for sample size $N = 10^5$

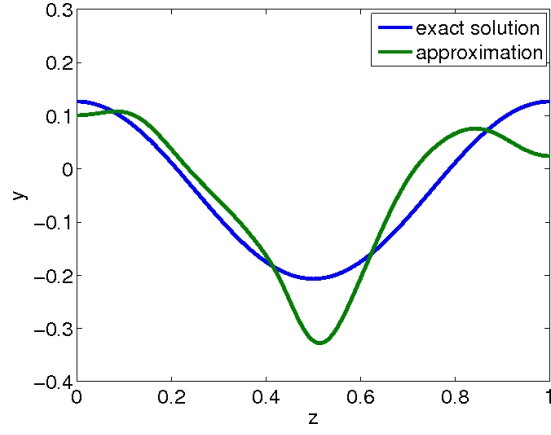


Figure 7: Median reconstruction, $N = 10^3$

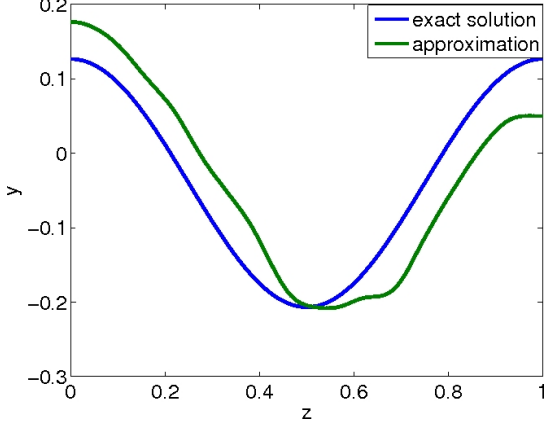


Figure 8: Median reconstruction, $N = 10^4$

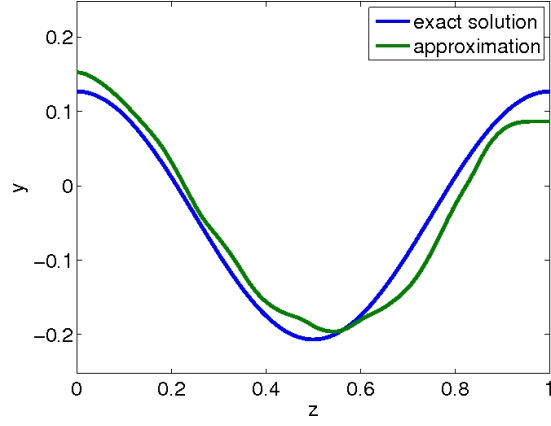


Figure 9: Median reconstruction, $N = 10^5$

A Proof of the main theorem

Before we come to the proof of Theorem 1, let us first formulate a result with deterministic error in the operator. We assume that \mathcal{F} is approximated by some deterministic operator

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

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:

$$\delta := \|\widehat{\mathcal{F}}(\varphi^\dagger)\|, \quad (26a)$$

$$\gamma := \left(\left| \sup_{\{\varphi \in \mathfrak{B}: \Delta(\varphi, \varphi^\dagger) \neq 0\}} \frac{\|\mathcal{F}'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2 - \|\widehat{\mathcal{F}}'[\varphi^\dagger](\varphi - \varphi^\dagger)\|^2}{\Delta(\varphi, \varphi^\dagger)} \right| \right)^{1/2}. \quad (26b)$$

Moreover, we assume that the tangential cone condition

$$\|\widehat{\mathcal{F}}(x) - \widehat{\mathcal{F}}(y) - \widehat{\mathcal{F}}'[y](x - y)\| \leq \eta \|\widehat{\mathcal{F}}(x) - \widehat{\mathcal{F}}(y)\|, \quad (27)$$

holds for all x, y in some neighborhood of \mathcal{B} .

Lemma 3. *Assume that (18), (26) and (27) hold true with η sufficiently small, and that the convex minimization problems (3) are uniquely solvable. Further assume that the iteration is stopped at the smallest index $K \in \mathbb{N}_0$ for which*

$$\alpha_{K+1} \leq \max(\Theta^{-1}(\delta), \gamma^2), \quad \text{where } \Theta(t) := \sqrt{t}\Lambda(t) \quad (28)$$

and that $\alpha_0 > \max(\Theta^{-1}(\delta_1), \delta_2^2)$. In addition it should hold that $\alpha_k \leq q\alpha_{k+1}$ 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

$$\Delta(\widehat{\varphi}_K, \varphi^\dagger) \leq C (\Lambda(\max(\Theta^{-1}(\delta), \gamma^2)))^2. \quad (29)$$

Proof. Let us introduce the following notation:

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

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

$$\begin{aligned} &\|\widehat{\mathcal{T}}_{k-1}(\widehat{\varphi}_k - \widehat{\varphi}_{k-1}) + \widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2 + \alpha_k \mathcal{R}(\widehat{\varphi}_k) \\ &\leq \|\widehat{\mathcal{T}}_{k-1}(\varphi^\dagger - \widehat{\varphi}_{k-1}) + \widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2 + \alpha_k \mathcal{R}(\varphi^\dagger). \end{aligned} \quad (30)$$

From the definition (17) of the Bregman distance and the source condition (18) we obtain

$$\mathcal{R}(\varphi^\dagger) - \mathcal{R}(\widehat{\varphi}_k) = \langle \varphi^\dagger, \varphi^\dagger - \widehat{\varphi}_k \rangle - \Delta_k \leq \beta \Delta_k^{1/2} \Lambda \left(\frac{\|\mathcal{T}e_k\|^2}{\Delta_k} \right) - \Delta_k. \quad (31)$$

Plugging this into (30) yields

$$\begin{aligned} &\|\widehat{\mathcal{T}}_{k-1}(\widehat{\varphi}_k - \widehat{\varphi}_{k-1}) + \widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2 + \alpha_k \Delta_k \\ &\leq \|\widehat{\mathcal{T}}_{k-1}(\varphi^\dagger - \widehat{\varphi}_{k-1}) + \widehat{\mathcal{F}}(\widehat{\varphi}_{k-1})\|^2 + \beta \alpha_k \Delta_k^{1/2} \Lambda \left(\frac{\|\mathcal{T}e_k\|^2}{\Delta_k} \right). \end{aligned} \quad (32)$$

Note that the tangential cone condition (27) implies

$$(1 - \eta)\|\widehat{\mathcal{T}}e_k\| \leq \|\widehat{\mathcal{F}}(\widehat{\varphi}_k) - \widehat{\mathcal{F}}(\varphi^\dagger)\| \leq (1 + \eta)\|\widehat{\mathcal{T}}e_k\|. \quad (33)$$

To estimate the first term on the left hand side of (32) we use (27) and (33) to get that

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

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

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

For the right hand side of (32) we get from (26) and another application of (27) that

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

Plugging the last two inequalities into (32) 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

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

Using (26b) 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 (28) this implies

$$C_\eta \|\widehat{\mathcal{T}}e_k\|^2 + \alpha_k \Delta_k \leq c_\eta \|\widehat{\mathcal{T}}e_{k-1}\|^2 + 4\Theta(\alpha_k)^2 + \beta\alpha_k \Delta_k^{1/2} \Lambda \left(\frac{\|\widehat{\mathcal{T}}e_k\|^2}{\Delta_k} + \alpha_k \right). \quad (34)$$

We will show the following error bounds

$$\|\widehat{\mathcal{T}}e_k\|^2 \leq C_1 \Theta(\alpha_k)^2, \quad (35a)$$

$$\Delta(\widehat{\varphi}_k, \varphi^\dagger) \leq C_2 \Lambda(\alpha_k)^2 \quad (35b)$$

with

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

$$C_2 := \max \left(\frac{\Delta(\varphi_0, \varphi^\dagger)}{\Lambda(\alpha_0)^2}, 2C_1 c_\eta q^3 + 8, 16\beta^2 \right).$$

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:

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

Now by using the induction hypothesis (35a) equation (34) 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 $0 \leq t \leq 1$. 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

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

Hence it is necessary that $C_\eta > \frac{2q^3 c_\eta}{8}$, which means that η must be small enough.

Then (35a) is true with $C_1 \geq \frac{8}{C_\eta - 2q^3 c_\eta}$.

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 (35b) with $C_2 \geq 2c_\eta C_1 q^3 + 8$.

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

In this case (34) 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

$$C_\eta \|\widehat{\mathcal{T}}e_k\|^2 + \alpha_k\Delta_k \leq 2\beta\alpha_k\Delta_k^{1/2}\left(\Lambda\left(\frac{\|\widehat{\mathcal{T}}e_k\|^2}{\Delta_k}\right) + \Lambda(\alpha_k)\right). \quad (36)$$

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

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

This shows that $\Delta_k^{1/2} \leq 4\beta \Lambda(\alpha_k)$ and thereby (35b) 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

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

Hence (35a) holds with $C_1 \geq \frac{16\beta^2}{1 + C_\eta}$.

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

Dividing formula (36) by $\|\widehat{\mathcal{T}}e_k\|$ results in

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

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 (35a) is true with $C_1 > \frac{16\beta^2}{C_\eta^2}$. Plugging this into the right hand side of the last equation gives

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

Now we see that $\Delta_k \leq 16\beta^2 \Lambda(\alpha_k)^2$ and therefore that (35b) is valid with $C_1 = 16\beta^2$. This completes the proof.

Now Theorem 1 follows easily:

Proof of Theorem 1. 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 1.

Acknowledgement

Fabian Dunker, Thorsten Hohage and Enno Mammen acknowledge support by DFG through the research group FOR 916.

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.
- 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 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.
- 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.
- 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.
- 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.
- 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 nonparametric 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.

Institut für Numerische und Angewandte Mathematik
Universität Göttingen
Lotzestr. 16-18
D - 37083 Göttingen

Telefon: 0551/394512

Telefax: 0551/393944

Email: trapp@math.uni-goettingen.de URL: <http://www.num.math.uni-goettingen.de>

Verzeichnis der erschienenen Preprints 2011:

2011-1	M. Braack, G. Lube, L. Röhe	Divergence preserving interpolation on anisotropic quadrilateral meshes
2011-2	M.-C. Körner, H. Martini, A. Schöbel	Minsum hyperspheres in normed spaces
2011-3	R. Bauer, A. Schöbel	Rules of Thumb – Practical Online-Strategies for Delay Management
2011-4	S. Cicerone, G. Di Stefano, M. Schachtebeck, A. Schöbel	Multi-Stage Recovery Robustness for Optimization Problems: a new Concept for Planning under Disturbances
2011-5	E. Carrizosa, M. Goerigk, M. Körner, A. Schöbel	Recovery to feasibility in robust optimization
2011-6	M. Goerigk, M. Knoth, M. Müller-Hannemann, A. Schöbel, M. Schmidt	The Price of Robustness in Timetable Information
2011-7	L. Nannen, T. Hohage, A. Schädle, J. Schöberl	High order Curl-conforming Hardy space infinite elements for exterior Maxwell problems
2011-8	D. Mirzaei, R. Schaback, M. Dehghan	On Generalized Moving Least Squares and Diffuse Derivatives
2011-9	M. Pazouki, R. Schaback	Bases of Kernel-Based Spaces
2011-10	D. Mirzaei, R. Schaback	Direct Meshless Local Petrov-Galerkin (DMLPG) Method: A Generalized MLS Approximation
2011-11	T. Hohage, F. Werner	Iteratively regularized Newton methods with general data misfit functionals and applications to Poisson data
2011-12	D. Rosca, G. Plonka	Uniform spherical grids via equal area projection from the cube to the sphere
2011-13	S. Hein, W. Koch, L. Nannen	Trapped modes and Fano resonances in two-dimensional acoustical duct-cavity systems
2011-14	T. Peter, D. Rosca, G. Plonka-Hoch	Representation of sparse Legendre expansions
2011-15	F. Dunker, J.-P. Florens, T. Hohage, J. Johannes, E. Mammen	Iterative Estimation of Solutions to Noisy Nonlinear Operator Equations in Nonparametric Instrumental Regression