Nonparametric instrumental regression with non-convex constraints

M Grasmair[1], O Scherzer[1,2] and A Vanhems[3]

[1]Computational Science Center, University of Vienna, Austria
[2]Radon Institute of Computational and Applied Mathematics, Linz, Austria
[3]University of Toulouse, Toulouse Business School and Toulouse School of Economics, France

E-mail: `markus.grasmair@univie.ac.at`, `otmar.scherzer@univie.ac.at`, `a.vanhems@esc-toulouse.fr`

Abstract. This paper considers the nonparametric regression model with an additive error that is dependent on the explanatory variables. As is common in empirical studies in epidemiology and economics, it also supposes that valid instrumental variables are observed. A classical example in microeconomics considers the consumer demand function as a function of the price of goods and the income, both variables often considered as endogenous. In this framework, the economic theory also imposes shape restrictions on the demand function, like integrability conditions. Motivated by this illustration in microeconomics, we study an estimator of a nonparametric constrained regression function using instrumental variables by means of Tikhonov regularization. We derive rates of convergence for the regularized model both in a deterministic and stochastic setting under the assumption that the true regression function satisfies a projected source condition including, because of the non-convexity of the imposed constraints, an additional smallness condition.

1. Motivation

We consider the model

$$Y_i = g(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $(Y_i, X_i)_{i=1,\ldots,n}$ is a sample of observations of size n representing respectively the measured data and variables effecting the measurements. The function g describes the dependence of the data on the variables, and ε_i is a combination of noise (measurement errors) and *modeling errors*, often resulting from the omittance of relevant variables. The goal is the estimation of the function g. If the modeling errors ε and the variables

X are not dependent, that is, if the conditional expectation $\mathbb{E}(\varepsilon|X)$ of ε given X is zero, then it is possible to identify g by

$$g(x) := \mathbb{E}(Y|X = x). \tag{1}$$

If, however, the conditional expectation of ε given X does not vanish, then this will lead to a biased estimate, as

$$\mathbb{E}(Y|X = x) = g(x) + \mathbb{E}(\varepsilon|X = x).$$

The variables X are then called endogenous variables. This issue of endogeneity typically arises in the presence of modeling errors, in particular, if variables have been omitted from the model that simultaneously influence both X and Y. This has been illustrated in several applications, for example in epidemiology (see [11, 14, 28]) and in economics (see [31] and also the survey [2]). In the classical microeconomic setting of consumer demand, the endogeneity issue has also been raised. In this framework, the variable Y represents the observed demand of a consumer for k goods, and the explanatory variables X include the vector of prices P of the goods and the total budget $Z > 0$ of the consumer; the function $g\colon \mathbb{R}_{>0}^k \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}^k$ denotes the consumer demand. The problem of price endogeneity has been highlighted in several research articles (see for example [8, 21, 22]). In an industrial organization framework, the paper by [4] analyzes demand and supply in differentiated product markets (like the US automobile industry) and highlight the problem involved by correlation between prices and product characteristics, some of which are observed by the consumer but not by the econometrician. Similarly total expenditure endogeneity has been studied in particular for Engel Curves analysis, see for example [7].

One remedy is the usage of *instruments*, that is, different variables W, which influence both P and Z but are uncorrelated with ε (see [2] for an overview). The analysis of nonparametric instrumental regression has been conducted in several works such as [13, 17, 19, 27]. Therefore we consider the model

$$Y = g(X) + \varepsilon$$

and we assume that the random variable $X = (P, Z)$ is described by instruments W in such a way that $\mathbb{E}(\varepsilon|W) = 0$. Therefore, the equation (1) can be transformed into

$$\mathbb{E}(g(X)|W = w) = \mathbb{E}(Y|W = w). \tag{2}$$

We assume in the following that the relation between Y, X and W is described by a joint density $f_{YXW}\colon \Omega_Y \times \Omega_X \times \Omega_W \to \mathbb{R}_{>0}$, where, for simplicity, the finite measure spaces Ω_Y, Ω_X and Ω_W are assumed to be normalized. We consider L^2 spaces with respect to this joint probability density and denote for example by $L^2(\Omega_X)$ functions depending on P and Z only. In addition, we denote by f_{YW}, f_{XW}, f_W the corresponding marginal densities defined by

$$f_{YW}(y, w) = \int_{\Omega_X} f_{YXW}(y, x, w)\, dx,$$

$$f_{XW}(x, w) = \int_{\Omega_Y} f_{YXW}(y, x, w) \, dy,$$

$$f_W(w) \quad = \int_{\Omega_X} \int_{\Omega_Y} f_{YXW}(y, x, w) \, dx \, dy.$$

Now assume that the set Ω is bounded and f_{YXW} is bounded away from zero. We consider the operator $T: L^2(\Omega_X) \to L^2(\Omega_W)$ defined by

$$T\psi(w) := \mathbb{E}(\psi(X)|W = w) = \int_{\Omega_X} \psi(x) \frac{f_{XW}(x, w)}{f_W(w)} \, dx. \tag{3}$$

Then (2) can be rewritten as the Fredholm integral equation

$$Tg = h, \tag{4}$$

where

$$h(w) = \mathbb{E}(Y|W = w) = \int_{\Omega_Y} y \, \frac{f_{YW}(y, w)}{f_W(w)} \, dy.$$

In addition, classical microeconomic theory imposes some shape restrictions on the consumer demand, and the challenge is to take these constraints into account in the nonparametric estimation of the function g. More precisely, standard micro-economic theory (see [29]) states that the demand is the result of the maximization of some (unknown) utility function. That is, there exists some function $u: \mathbb{R}_{\geq 0}^k \to \mathbb{R}$ (the utility) such that

$$g(x) = \arg\max\{u(y) : y \in \mathbb{R}_{\geq 0}^k, \ \langle y, p \rangle \leq z\}, \tag{5}$$

where $x = (p, z)$. Here the utility function is assumed to be continuously differentiable, concave, and strictly monotoneously increasing. Even though the utility is unknown, the assumption of its existence (and of utility maximization) has some implications for the demand function g, called the *integrability conditions*. First, it is rather obvious that g is homogeneous of degree 0, that is, $g(tx) = g(x)$ for every $t > 0$. Moreover, the maximum in (5) is always attained at the boundary; more precisely, we have the equality

$$\langle y, g(x) \rangle = z; \tag{6}$$

this condition is usually called the *budget constraint*. Finally, defining the *Slutsky matrix*

$$S_g(x) := \nabla_p g(x) + \partial_z g(x) \cdot g(x)^T,$$

the conditions

$$S_g(x) = S_g(x)^T \qquad \text{and} \qquad S_g(x) \leq 0 \tag{7}$$

hold. That is, the Slutsky matrix is symmetric and negative semi-definite in (almost) every point $x = (p, z)$.

Therefore, the objective of this work is to recover the function g characterized by equation (4) and satisfying the constraints defined by the Slutsky matrix.

The paper is organized as follows: In Section 2, we present our model, the link with ill-posed inverse problems in the case where the transform is unknown, and the

conditions under which a regularized solution can be defined. In Section 3 we derive rates of convergence in a deterministic setting and we extend the results in Section 4 to the statistical setting.

2. Constrained Inversion of T

Let now T be the operator defined in (3) (operating on vector valued functions). Then, in order to recover g, we have to solve the equation

$$Tg = h,$$

where h denotes the right hand side of (4) subject to the constraints that g is homogeneous of degree 0 and satisfies the budget constraint (6) and the Slutsky condition (7) almost everywhere in $\Omega := \Omega_X = \Omega_P \times \Omega_Z$. In the following we will always assume that the set Ω is bounded, open, connected and has a Lipschitz boundary.

Apart from the constraints, there are three problems: First, the operator T is defined by the density f_{XW}, which is not known exactly but can only by estimated up to a certain error δ. Consequently, we will only have an approximation T^δ of T available. Second, the right hand side h is only known up to some error γ, as it may be prone to measurement errors (in a deterministic setting) or is the realization of a random variable (in a stochastic setting), and, again, it depends on the density f_{YW}. In addition, the assumption $\mathbb{E}(\varepsilon|w) = 0$ need not hold exactly. Finally, the operator T (and also its approximation T^δ) is not boundedly invertible in $L^2(\Omega; \mathbb{R}^k)$. Thus a direct solution of the operator equation

$$T^\delta g = h^\gamma$$

does not make sense, as its solution $g^{\delta,\gamma}$ (if it exists) need not be close to the true solution $g^\dagger$, even if the errors δ and γ are small. In addition, there is no reason why the exact solution of the perturbed operator equation (if it exists) should satisfy the required constraints, in particular, as the constraints are non-linear and describe a non-convex set.

In order to find a solution nevertheless, it is necessary to consider some kind of regularized solution. In the following, we consider the application of (constrained) Tikhonov regularization, where we use the (weighted) first order Sobolev norm as regularization functional. That is, denoting for $\mu \geq 0$ by

$$\|g\|_\mu^2 := \mu\|g\|_{L^2}^2 + \|\nabla g\|_{L^2}^2 \tag{8}$$

the weighted Sobolev norm, one minimizes, for some regularization parameter $\alpha > 0$ depending on δ and γ, the functional

$$\mathcal{T}_\alpha(g; T^\delta, h^\gamma) := \|T^\delta g - h^\gamma\|_{L^2}^2 + \alpha\|g\|_\mu^2$$

subject to the constraints of positivity, 0-homogeneity, the Slutsky condition, and the budget constraint. For the sake of simplicity, we will omit in the following the subscripts in the L^2-norms and we will assume that Ω is compactly contained in $\mathbb{R}_{>0}^k \times \mathbb{R}_{>0}$.

We use in the following the abbreviation

$$\mathcal{X} := \{g \in H^1(\Omega; \mathbb{R}^k) : g \geq 0 \text{ is 0-homogeneous, } \langle p, g(x) \rangle = z \text{ and } S_g = S_g^T \leq 0 \text{ a.e.}\}.$$

Then one can define

$$g_\alpha^{\delta, \gamma} := \arg \min\{\|T^\delta g - h^\gamma\|^2 + \alpha \|g\|_\mu^2 : g \in \mathcal{X}\},$$

provided the Tikhonov functional attains its minimum in $\mathcal{X}$. In the following, we will show that this is indeed the case. The proof is based on the direct method in the calculus of variations. As a first important result, we prove that the set $\mathcal{X}$ is weakly closed in $H^1(\Omega; \mathbb{R}^k)$, which is not an obvious assertion, as $\mathcal{X}$ is non-convex, and the weak closedness of a subset of a Hilbert space is usually strongly tied to its convexity.

Lemma 2.1. The set $\mathcal{X}$ is weakly sequentially closed in $H^1(\Omega; \mathbb{R}^k)$.

Proof. Obviously the set of non-negative 0-homogeneous functions satisfying the budget constraint $\langle p, g(x) \rangle = z$ is convex and closed in $H^1(\Omega; \mathbb{R}^k)$, implying that it is also weakly closed.

Next we show that the mapping $S \colon H^1(\Omega; \mathbb{R}^k) \to L^1(\Omega; \mathbb{R}^{k \times k})$,

$$g \mapsto S(g) = \nabla_p g + \partial_z g \cdot g^T$$

is weak–weak continuous. To that end assume that the sequence $(g_n)_{n \in \mathbb{N}}$ weakly converges to $g \in H^1(\Omega; \mathbb{R}^k)$. Then ∇g_n weakly converges to ∇g in $L^2(\Omega; \mathbb{R}^{k \times (k+1)})$ (which in particular implies that the sequence is bounded) and the Rellich–Kondrachov compactness theorem (see [1, Thm. 6.2]) implies that the functions g_n converge strongly to g with respect to the L^2 topology. Thus, if $1 \leq i, j \leq k$ and $u \in L^2(\Omega; \mathbb{R})$, we have

$$|\langle \partial_z g_n^{(i)} g_n^{(j)} - \partial_z g^{(i)} g^{(j)}, u \rangle| \leq |\langle \partial_z g_n^{(i)} (g_n^{(j)} - g^{(j)}), u \rangle| + |\langle (\partial_z g_n^{(i)} - \partial_z g^{(i)}) g^{(j)}, u \rangle|$$
$$\leq \|g_n^{(j)} - g^{(j)}\| \|u\| \|\partial_z g_n^{(i)}\| + |\langle \partial_z g_n^{(i)} - \partial_z g^{(i)}, g^{(j)} u \rangle| \to 0.$$

Consequently the product $\partial_z g_n \cdot g_n^T$ converges to $\partial_z g \cdot g^T$ with respect to the weak topology on $L^1(\Omega; \mathbb{R}^{k \times k})$.

Now note that the set Sym_k^- of all symmetric and negative semi-definite $(k \times k)$-matrices is a closed and convex cone in $\mathbb{R}^{k \times k}$. Consequently also the set of all summable functions on Ω with values in Sym_k^- is a closed and convex cone in $L^1(\Omega; \mathbb{R}^{k \times k})$ and therefore, in particular, also weakly closed. Therefore the weak-weak continuity of the mapping S implies that the set of functions $g \in H^1(\Omega; \mathbb{R}^k)$ satisfying the Slutsky condition $S(g) = S(g)^T \leq 0$ is weakly closed.

This shows that the set $\mathcal{X}$ is the intersection of the (weakly closed) set of 0-homogeneous, non-negative functions satisfying the budget constraint with a weakly closed set, which proves that $\mathcal{X}$ itself is weakly closed in $H^1(\Omega; \mathbb{R}^k)$. □

For the usage of the direct method in the calculus of variations, we still have to prove the coercivity of the regularization functional. In the case $\mu > 0$, the coercivity is obvious, as the regularization term is equivalent to the H^1-norm; in the case $\mu = 0$, however, the equivalence only holds, if the operator T does not annihilate constant

functions (see [3, 30] for a related result on total variation regularization). In the next result, we provide a detailed proof of this assertion by explicitly computing constants defining this equivalence of norms. In particular, the results show that these constants depend continuously on the operator T, which will be required in the proof of the convergence result, where we also treat the case of operator errors.

Lemma 2.2. Assume that $T: L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ is a bounded linear operator. If $\mu = 0$, assume in addition that $Tc \neq 0$ for every non-zero constant function $c: \Omega \to \mathbb{R}^k$. Define for $g \in H^1(\Omega; \mathbb{R}^k)$

$$\|g\|_T^2 := \|g\|_\mu^2 + \|Tg\|_{L^2}^2. \tag{9}$$

Then $\|\cdot\|_T$ is a norm on $H^1(\Omega; \mathbb{R}^k)$ that is equivalent to the standard H^1-norm. More precisely, we have the following estimates: For every $\mu \geq 0$,

$$\|g\|_T \leq \|T\| \|g\|_{H^1}; \tag{10}$$

if $\mu > 0$ in (9), then

$$\|g\|_{H^1} \leq \frac{1}{\min\{\sqrt{\mu}, 1\}} \|g\|_T, \tag{11}$$

and if $\mu = 0$ in (9), then there exists a constant $A > 0$ only depending on the set Ω such that

$$\|g\|_{H^1} \leq A(\|T\| D(T)^{-1} + D(T)^{-1} + 1) \|g\|_T, \tag{12}$$

where

$$D(T) := \inf\{\|Tc\| : c: \Omega \to \mathbb{R}^k \text{ is constant with } |c| = 1\}.$$

Proof. Inequality (10) follows from

$$\|g\|_T \leq \|Tg\|_{L^2} \leq \|T\| \|g\|_{L^2} \leq \|T\| \|g\|_{H^1},$$

and (11) is trivial.

Now assume that $\mu = 0$. Then the assertion $Tc \neq 0$ for every non-zero constant function $c: \Omega \to \mathbb{R}^k$ implies that $0 < D(T) < +\infty$. Define now the projection $P: L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$, $g \mapsto \frac{1}{|\Omega|} \int_\Omega g$. Then

$$\begin{aligned}
\|g\|^2 &= \|g - Pg\|^2 + \|Pg\|^2 \\
&\leq \|g - Pg\|^2 + D(T)^{-2} \|TPg\|^2 \\
&\leq \|g - Pg\|^2 + 2D(T)^{-2}(\|Tg\|^2 + \|T(g - Pg)\|^2) \\
&\leq (1 + 2D(T)^{-2} \|T\|^2) \|g - Pg\|^2 + 2D(T)^{-2} \|Tg\|^2.
\end{aligned}$$

From the Poincaré Inequality (see e.g. [32, Thm. 4.8.1]) it follows that there exists $C > 0$ such that $\|g - Pg\| \leq C \|\nabla g\|$. Thus

$$\begin{aligned}
\|g\|_{H^1}^2 &= \|g\|^2 + \|\nabla g\|^2 \\
&\leq (C^2(1 + 2D(T)^{-2} \|T\|^2) + 1) \|\nabla g\|^2 + 2D(T)^{-2} \|Tg\|^2.
\end{aligned}$$

Setting $A = 2(C + 1)$ we obtain (12). $\qquad\square$

Lemma 2.3. Assume that $T^\delta: L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ is bounded linear, $h^\gamma \in L^2(\Omega; \mathbb{R}^k)$, $\alpha > 0$, and $\mu \geq 0$. If $\mu = 0$, assume in addition that $T^\delta c \neq 0$ for every non-zero constant function $c: \Omega \to \mathbb{R}^k$. Then the regularization functional $\mathcal{T}_\alpha: L^2(\Omega; \mathbb{R}^k) \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$.

$$\mathcal{T}_\alpha(g; T^\delta, h^\gamma) := \begin{cases} \|T^\delta g - h^\gamma\|^2 + \alpha \|g\|_\mu^2 & \text{if } g \in \mathcal{X}, \\ +\infty & \text{else,} \end{cases}$$

attains its minimum.

Proof. The weak closedness of the set $\mathcal{X}$ and the weak lower semi-continuity of the mapping $g \mapsto \frac{1}{2}\|T^\delta g - h^\gamma\|^2 + \frac{\alpha}{2}\|g\|_\mu^2$ on the space $H^1(\Omega)$ imply that also the mapping $\mathcal{T}_\alpha(\cdot; T^\delta, h^\gamma)$ is weakly lower semi-continuous. Moreover, Lemma 2.2 implies that $\mathcal{T}_\alpha(\cdot; T^\delta, h^\gamma)$ is weakly coercive. Applying the direct method in the calculus of variations, we obtain the existence of a minimizer. $\square$

Note that the previous result does not say anything about the uniqueness of the minimizer. Because of the non-convexity of the set $\mathcal{X}$, it is probable that the Tikhonov functional has multiple local minima, but also possible that it has several global minima.

The following result is very similar to the convergence result in [25]. The main difference is that we also consider the homogeneous Sobolev semi-norm as a regularization term, which is not coercive by itself. The coercivity (or rather the equi-coercivity of the functionals $\mathcal{T}_\alpha(\cdot; T^\delta, h^\gamma)/\alpha$) is only obtained by means of Lemma 2.2.

Proposition 2.4. *Assume that $T: L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ is bounded linear satisfying $Tc \neq 0$ for every non-zero constant function $c: \Omega \to \mathbb{R}^k$ and that the operator equation $Tg = h$ has a solution in $\mathcal{X}$. Let $\delta_j \to 0$, $\gamma_j \to 0$ and assume that $T^{\delta_j}: L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ are bounded linear operators satisfying $\|T^{\delta_j} - T\| \leq \delta_j$ and that the functions $h^{\gamma_j} \in L^2(\Omega; \mathbb{R}^k)$ satisfy $\|h^{\gamma_j} - h\| \leq \gamma_j$. Let $\mu \geq 0$ be fixed; if $\mu = 0$, assume in addition that $T^{\delta_j} c \neq 0$ for every non-zero constant function $c: \Omega \to \mathbb{R}^k$.*

Assume that $\alpha_j > 0$ is chosen such that $\alpha_j \to 0$ and $(\delta_j + \gamma_j)^2/\alpha_j \to 0$. Then every sequence $(g_j)_{j \in \mathbb{N}} \subset \mathcal{X}$ satisfying

$$g_j \in \arg\min\{\mathcal{T}_{\alpha_j}(g; T^{\delta_j}, h^{\gamma_j}) : g \in \mathcal{X}\}$$

has a subsequence $g_{j(i)}$ converging with respect to the H^1-norm to some

$$g^\dagger \in \arg\min\{\|g\|_\mu^2 : Tg = h, \ g \in \mathcal{X}\}.$$

Proof. Let $\tilde{g}$ be any solution of $Tg = h$ in $\mathcal{X}$. Then

$$\begin{aligned} \|T^{\delta_j} g_j - h^{\gamma_j}\|^2 + \alpha \|g_j\|_\mu^2 &\leq \|T^{\delta_j} \tilde{g} - h^{\gamma_j}\|^2 + \alpha \|\tilde{g}\|_\mu^2 \\ &\leq (\|T^{\delta_j} - T\| \|\tilde{g}\| + \|h - h^{\gamma_j}\|)^2 + \alpha \|\tilde{g}\|_\mu^2 \\ &\leq (\delta_j \|\tilde{g}\| + \gamma_j)^2 + \alpha \|\tilde{g}\|_\mu^2. \end{aligned}$$

Consequently,

$$\begin{aligned} \|g_j\|_{T^{\delta_j}}^2 &\leq 2\|Tg_j - h^{\gamma_j}\|^2 + 2\|h^{\gamma_j}\|^2 + \|g_j\|_\mu^2 \\ &\leq (\delta_j \|\tilde{g}\| + \gamma_j)^2 + \alpha \|\tilde{g}\|_\mu^2 + 2\|h^{\gamma_j}\|^2 + \frac{(\delta_j \|\tilde{g}\|^2 + \gamma_j)^2}{\alpha} + \|\tilde{g}\|_\mu^2. \end{aligned}$$

From Lemma 2.2, it follows that $\|g_j\|_{T^{\delta_j}} \geq C(T^{\delta_j})\|g_j\|_{H^1}$ for some constants $C(\cdot) > 0$ depending continuously on the operator T^{δ_j}. Therefore, the assumption $(\delta_j + \gamma_j)^2/\alpha \to 0$ implies that the sequence $(g_j)_{j \in \mathbb{N}}$ is bounded. The proof of the subsequential convergence is now along the lines of [26, Thm. 3.26]. $\qquad\square$

3. Convergence Rates

Lemma 3.1. Assume that $T\colon L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ is bounded linear and that the equation $Tg = h$ has a solution in $\mathcal{X}$. Let

$$g^\dagger \in \arg\min\{\|g\|_\mu^2 : Tg = h, \ g \in \mathcal{X}\}.$$

Let moreover $T^\delta\colon L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ satisfy $\|T^\delta - T\|_{L^2} \leq \delta$, and let $h^\gamma \in L^2(\Omega; \mathbb{R}^k)$ satisfy $\|h^\gamma - h\| \leq \gamma$. If $\mu = 0$, assume in addition that $\delta < \|T\|$ and $Tc \neq 0$ for every non-zero constant function $c\colon \Omega \to \mathbb{R}^k$. Assume that there exists a set $L \subset \mathcal{X}$ such that, for some $\beta > 0$, $C \geq 0$ and every $g \in L$, we have

$$\beta\|g - g^\dagger\|_\mu^2 \leq \|g\|_\mu^2 - \|g^\dagger\|_\mu^2 + C\|T(g - g^\dagger)\|. \tag{13}$$

Let moreover

$$g_\alpha^{\delta,\gamma} \in \arg\min\{\mathcal{T}_\alpha(g; T^\delta, h^\gamma) : g \in \mathcal{X}\}.$$

Define for $\mu > 0$

$$D_\mu(\alpha, \delta, \gamma) := \frac{\delta\|g^\dagger\| + \gamma}{\sqrt{\mu\alpha}} + \frac{\|g^\dagger\|_\mu}{\sqrt{\mu}},$$

and let

$$D_0(\alpha, \delta, \gamma) := A\frac{\|T\| + D(T) + 1}{D(T) - \delta}\left[\|h\| + \gamma + \sqrt{2}\frac{\delta\|g^\dagger\| + \gamma + \sqrt{\alpha}\|\nabla g^\dagger\|}{\min\{\sqrt{\alpha}, 1\}}\right]$$

with $A > 0$ and $D(T) > 0$ as in Lemma 2.2.

Then the estimates

$$\beta\|g_\alpha^{\delta,\gamma} - g^\dagger\|_\mu^2 \leq \frac{(\gamma + \delta\|g^\dagger\|)^2}{\alpha} + C(\gamma + \delta D_\mu(\alpha, \delta, \gamma)) + \frac{C^2\alpha}{4}$$

and

$$\|T(g_\alpha^{\delta,\gamma} - g^\dagger)\|^2 \leq 2(\gamma + \delta\|g^\dagger\|)^2 + 2\alpha C(\gamma + \delta D_\mu(\alpha, \delta, \gamma)) + C^2\alpha^2$$

hold whenever $g_\alpha^{\delta,\gamma} \in L$.

Proof. The inequality (13) and the optimality of $g_\alpha^{\delta,\gamma}$ imply that

$$\begin{aligned}
\beta\|g_\alpha^{\delta,\gamma} - g^\dagger\|_\mu^2 &\leq \|g_\alpha^{\delta,\gamma}\|_\mu^2 - \|g^\dagger\|_\mu^2 + C\|T(g_\alpha^{\delta,\gamma} - g^\dagger)\| \\
&\leq \frac{1}{\alpha}\left(\|T^\delta g^\dagger - h^\gamma\|^2 - \|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\|^2\right) + C\|T(g_\alpha^{\delta,\gamma} - g^\dagger)\| \\
&\leq \frac{(\delta\|g^\dagger\|_{L^2} + \gamma)^2}{\alpha} + C(\delta\|g_\alpha^{\delta,\gamma}\|_{L^2} + \gamma) \\
&\quad + C\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\| - \frac{\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\|^2}{\alpha}.
\end{aligned}$$

Estimating

$$C\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\| - \frac{\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\|^2}{\alpha} \leq \sup_{t\geq 0}\left[Ct - \frac{t^2}{\alpha}\right] = \frac{C^2\alpha}{4},$$

we obtain the inequality

$$\beta\|g_\alpha^{\delta,\gamma} - g^\dagger\|_\mu^2 \leq \frac{(\gamma + \delta\|g^\dagger\|_{L^2})^2}{\alpha} + C(\gamma + \delta\|g_\alpha^{\delta,\gamma}\|) + \frac{C^2\alpha}{4}. \tag{14}$$

Moreover, using the estimate

$$C\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\| - \frac{\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\|^2}{\alpha}$$
$$\leq \sup_{t\geq 0}\left[Ct - \frac{t^2}{2\alpha}\right] - \frac{\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\|^2}{2\alpha} = \frac{C^2\alpha}{2} - \frac{\|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\|^2}{2\alpha},$$

we obtain

$$\|T(g_\alpha^{\delta,\gamma} - g^\dagger)\|^2 \leq 2(\gamma + \delta\|g^\dagger\|_{L^2})^2 + 2\alpha C(\gamma + \delta\|g_\alpha^{\delta,\gamma}\|) + C^2\alpha^2. \tag{15}$$

Assume first that $\mu > 0$. Then the definition of $\|\cdot\|_\mu$ and the optimality of g_α^δ imply the estimate

$$\|g_\alpha^{\delta,\gamma}\| \leq \frac{1}{\sqrt{\mu}}\|g_\alpha^{\delta,\gamma}\|_\mu \leq \frac{1}{\sqrt{\mu}}\left(\frac{(\delta\|g^\dagger\| + \gamma)^2}{\alpha} + \|g^\dagger\|_\mu^2\right)^{1/2} \leq \frac{1}{\sqrt{\mu}}\left(\frac{\delta\|g^\dagger\| + \gamma}{\sqrt{\alpha}} + \|g^\dagger\|_\mu\right),$$

which proves the assertion for μ strictly positive.

Now assume that $\mu = 0$. Then Lemma 2.2 implies that, using the same notation as in the Lemma,

$$\|g_\alpha^{\delta,\gamma}\| \leq A(\|T^\delta\|D(T^\delta)^{-1} + D(T^\delta)^{-1} + 1)\|g_\alpha^{\delta,\gamma}\|_{T^\delta}.$$

Moreover, for $\|T\| > \delta$, we have

$$D(T^\delta) = \inf\{\|T^\delta c\| : c\colon \Omega \to \mathbb{R}^k \text{ is constant with } |c| = 1\}$$
$$\geq \inf\{\|Tc\| : c\colon \Omega \to \mathbb{R}^k \text{ is constant with } |c| = 1\} - \delta$$
$$= D(T) - \delta.$$

Therefore,

$$\|g_\alpha^{\delta,\gamma}\| \leq A((\|T\| + \delta)(D(T) - \delta)^{-1} + (D(T) - \delta)^{-1} + 1)\|g_\alpha^{\delta,\gamma}\|_{T^\delta}$$
$$= A\frac{\|T\| + D(T) + 1}{D(T) - \delta}\|g_\alpha^{\delta,\gamma}\|_{T^\delta}.$$

Now the optimality of $g_\alpha^{\delta,\gamma}$ implies that

$$\|g_\alpha^{\delta,\gamma}\|_{T^\delta} \leq \|T^\delta g_\alpha^{\delta,\gamma} - h^\gamma\| + \|\nabla g_\alpha^{\delta,\gamma}\| + \|h^\gamma\|$$
$$\leq \|h\| + \gamma + \frac{(2\mathcal{T}_\alpha(g_\alpha^{\delta,\gamma}; T^\delta, h^\gamma))^{1/2}}{\min\{\sqrt{\alpha}, 1\}}$$
$$\leq \|h\| + \gamma + \frac{(2\mathcal{T}_\alpha(g^\dagger; T^\delta, h^\gamma))^{1/2}}{\min\{\sqrt{\alpha}, 1\}}$$
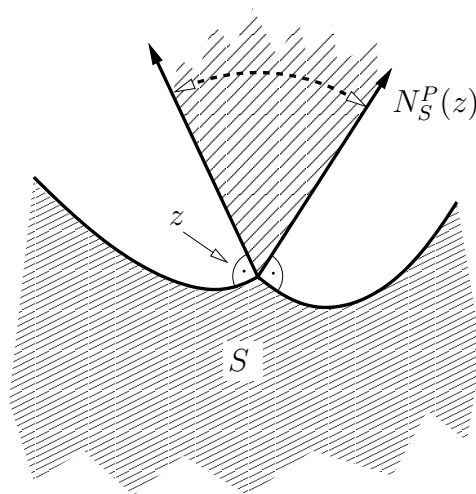
Figure 1. Proximal normal cone to the non-convex set S at the point $z \in \partial S$.

$$\leq \|h\| + \gamma + \sqrt{2}\frac{\delta\|g^\dagger\| + \gamma + \sqrt{\alpha}\|\nabla g^\dagger\|}{\min\{\sqrt{\alpha}, 1\}}.$$

Together, these estimates show that

$$\|g_\alpha^{\delta,\gamma}\| \leq A\frac{\|T\| + D(T) + 1}{D(T) - \delta}\Big[\|h\| + \gamma + \sqrt{2}\frac{\delta\|g^\dagger\| + \gamma + \sqrt{\alpha}\|\nabla g^\dagger\|}{\min\{\sqrt{\alpha}, 1\}}\Big].$$

Inserting this inequality in (14) and (15) proves the assertion for $\mu = 0$. $\square$

In the next result, we will present concrete conditions that imply the inequality (13). These conditions are a generalization of *projected source conditions*, which are a classical concept in the theory of inverse problems with convex contraints (see [10, 15, 24]), to a non-convex setting. Recently, the relation between projected source conditions and variational inequalities of the type (13) has also been studied in [16], though still in a convex setting. In order to generalize this concept to non-convex constraints, we recall the notion of a proximal normal cone to a subset of a Hilbert space (see [12]).

Definition 3.2. Let Y be a Hilbert space and let $S \subset Y$ be non-empty. We define for $y \in Y$ the set $\text{proj}_S(y) \subset S$ as the set of all points $z \in S$ for which the distance to y is minimal. Moreover we define for $z \in S$ the *proximal normal cone* $N_S^P(z)$ to S at z as

$$N_S^P(z) := \{\zeta = t(y - z) \in Y : t \geq 0, \ z \in \text{proj}_S(y)\}.$$

See also Figure 1.

For the following result, see [12, Prop. 1.5].

Proposition 3.3. *A vector ζ belongs to $N_S^P(z)$, if and only if there exists $\tau \geq 0$ (possibly depending on ζ and z) such that*

$$\langle \zeta, y - z \rangle \leq \tau\|y - z\|^2 \tag{16}$$

for all $y \in S$.

In the following we will denote, for given $z \in S$ and $\zeta \in N_S^P(z)$, by $\tau(\zeta, z)$ the smallest $\tau \geq 0$ for which (16) holds. Then the function τ is positively homogeneous with respect to its first variable, that is, $\tau(t\zeta, z) = t\tau(\zeta, z)$ whenever $\zeta \in N_S^P(z)$ and $t > 0$ (note that the fact that $N_S^P(z)$ is a cone implies that $t\zeta \in N_S^P(z)$).

Theorem 3.4. *Assume that $g^\dagger \in \mathcal{X}$ satisfies $Tg^\dagger = h$. In addition, assume that $\partial_\nu g^\dagger = 0$ on $\partial\Omega$. Denote moreover by $T^*\colon L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ the adjoint of T and let $N_\mathcal{X}^P(g^\dagger) \subset L^2(\Omega; \mathbb{R}^k)$ be the proximal normal cone to the set $\mathcal{X}$ at the point $g^\dagger$. Assume that there exist $\omega \in L^2(\Omega; \mathbb{R}^k)$ and $\zeta \in N_\mathcal{X}^P(g^\dagger)$ such that*

$$2(\mu g^\dagger - \Delta g^\dagger) = T^*\omega + \zeta.$$

- *If $\mu > 0$ and $\tau(\zeta, g^\dagger) < \mu$, then (13) holds for every $g \in \mathcal{X}$ with $C = \|\omega\|$ and $\beta = 1 - \tau(\zeta, g^\dagger)/\mu$.*
- *If $\mu = 0$, assume in addition that $Tc \neq 0$ for every non-zero constant function $c\colon \Omega \to \mathbb{R}^k$ and that*

$$E := A^2(\|T\|D(T)^{-1} + D(T)^{-1} + 1)^2 \tau(\zeta, g^\dagger)$$

with A and $D(T)$ as in Lemma 2.2 satisfies $E < 1$. Then for every $s > 0$ the inequality (13) holds with $\beta = 1 - E$ and $C = \|\omega\| + sE$ whenever $g \in \mathcal{X}$ satisfies $\|T(g - g^\dagger)\| < s$.

Proof. First note that

$$\begin{aligned}
\langle 2\mu g^\dagger - 2\Delta g^\dagger - \zeta, g^\dagger - g \rangle &= \langle T^*\omega, g^\dagger - g \rangle \\
&= \langle \omega, T(g^\dagger - g) \rangle \\
&\leq \|\omega\|\|T(g^\dagger - g)\|.
\end{aligned}$$

Now the assumption $\zeta \in N_\mathcal{X}^P(g^\dagger)$ implies that

$$\langle \zeta, g^\dagger - g \rangle \leq \tau(\zeta, g^\dagger)\|g^\dagger - g\|^2$$

for all $g \in \mathcal{X}$. In addition, Stoke's theorem and the assumption $\partial_\nu g^\dagger = 0$ on $\partial\Omega$ imply that

$$2\langle \mu g^\dagger - \Delta g^\dagger, g^\dagger - g \rangle = 2\mu\langle g^\dagger, g^\dagger - g \rangle + 2\langle \nabla g^\dagger, \nabla(g^\dagger - g) \rangle$$
$$= \|g^\dagger - g\|_\mu^2 + \|g^\dagger\|_\mu^2 - \|g\|_\mu^2.$$

Thus we obtain the estimate

$$\|\omega\|\|T(g - g^\dagger)\| \geq \|g^\dagger\|_\mu^2 - \|g\|_\mu^2 + \|g^\dagger - g\|_\mu^2 - \tau(\zeta, g^\dagger)\|g^\dagger - g\|^2. \quad (17)$$

In the case $\mu > 0$, it follows that

$$(1 - \tau(\zeta, g^\dagger)/\mu)\|g^\dagger - g\|_\mu^2 \leq \|g^\dagger\|_\mu^2 - \|g\|_\mu^2 + \|\omega\|\|T(g - g^\dagger)\|,$$

which proves the first part of the assertion.

On the other hand, if $\mu = 0$, then (17) and Lemma 2.2 imply that

$$(1 - E)\|\nabla(g^\dagger - g)\|^2 \leq \|\nabla g\|^2 - \|\nabla g^\dagger\|^2 + \|\omega\|\|T(g - g^\dagger)\| + E\|T(g - g^\dagger)\|^2.$$

Thus (13) holds for $\|T(g - g^\dagger)\| \leq s$. $\qquad\square$

Corollary 3.5. *Assume that the assumptions of Theorem 3.4 are satisfied. Then we have, with the notation of Lemma 3.1, the estimates*

$$(1 - \tau(\zeta, g^\dagger)/\mu)\|g_\alpha^{\delta,\gamma} - g^\dagger\|_\mu^2 \leq \frac{(\gamma + \delta\|g^\dagger\|)^2}{\alpha} + \|\omega\|(\gamma + \delta D_\mu(\alpha, \delta, \gamma)) + \frac{\|\omega\|^2\alpha}{4}$$

in the case $\mu > 0$, and

$$(1 - E)\|\nabla(g_\alpha^{\delta,\gamma} - g^\dagger)\|^2$$
$$\leq \frac{(\gamma + \delta\|g^\dagger\|)^2}{\alpha} + (\|\omega\| + sE)(\gamma + \delta D_0(\alpha, \delta, \gamma)) + \frac{(\|\omega\| + sE)^2\alpha}{4}$$

in the case $\mu = 0$. In particular, we have in both cases with a parameter choice $\alpha \asymp \max\{\delta, \gamma\}$ a convergence rate

$$\|g_\alpha^{\delta,\gamma} - g^\dagger\|_\mu^2 = O(\max\{\delta, \gamma\}).$$

Remark 3.6. Consider for the moment the setting where the constraint set $\mathcal{X}$ is closed and convex. Then the convexity of $\mathcal{X}$ implies that $\tau(\zeta, g^\dagger) = 0$ whenever $\zeta \in N_{\mathcal{X}}^P(g^\dagger)$; in other words, the proximal normal cone $N_{\mathcal{X}}^P(g^\dagger)$ coincides with the (usual) normal cone $N_{\mathcal{X}}(g^\dagger) = \{\zeta : \langle \zeta, \tilde{g} - g \rangle \leq 0 \text{ for all } \tilde{g} \in \mathcal{X}\}$. Thus in the condition $T^*\omega + \zeta = 2(\mu g^\dagger - \Delta g^\dagger)$ for some $\zeta \in N_{\mathcal{X}}^P(g^\dagger)$ no smallness condition is required for ζ, and therefore this condition reduces to the classical projected source condition found in [10, 24].

Remark 3.7. The conditions and results of Theorem 3.4 and Corollary 3.5 can also be translated into the context of convex analysis with subgradients and Bregman distances (see [9, 20, 26]). Recall that the subdifferential $\partial\mathcal{R}(g^\dagger) \subset X$ of a convex mapping $\mathcal{R} \colon X \to [0, +\infty]$ at $g^\dagger$ consists of all elements $\xi \in X$ satisfying $\mathcal{R}(g) \geq \mathcal{R}(g^\dagger) + \langle \xi, g - g^\dagger \rangle$ for all $g \in X$. Moreover, the Bregman distance $\mathcal{D}^\xi(\cdot; g^\dagger)$ is defined as

$$\mathcal{D}^\xi(g; g^\dagger) := \mathcal{R}(g) - \mathcal{R}(g^\dagger) - \langle \xi, g - g^\dagger \rangle.$$

If $\mathcal{R}(g) := \|g\|_\mu^2$ (setting $\mathcal{R}(g) = +\infty$ if $g \notin H^1(\Omega; \mathbb{R}^k)$), we obtain that the subdifferential is non-empty if and only if $\partial_\nu g^\dagger = 0$ on $\partial\Omega$. Moreover, in this case its unique element is the function $2(\mu g^\dagger - \Delta g^\dagger)$. Finally, it is easy to see that the Bregman distance between with g and $g^\dagger$ with respect to $\|\cdot\|_\mu^2$ is precisely $\|g - g^\dagger\|_\mu^2$.

In this setting, Corollary 3.5 with $\mu > 0$ reads as follows: If there exist $\xi \in \partial\mathcal{R}(g^\dagger)$ and $\zeta \in N_{\mathcal{X}}^P(g^\dagger)$ with $\tau(\zeta, g^\dagger) < \mu$, then

$$\mathcal{D}^\xi(g_\alpha^{\delta,\gamma}) = O(\max\{\delta, \gamma\}).$$

Note moreover that in [18] a theory based on abstract convex analysis has been developed in order to derive convergence rates for non-convex regularization terms. Again, the results of Corollary 3.5 can be seen as special cases of the results in [18, Section 4] by realizing that the function $2(\mu g^\dagger - \Delta g^\dagger) - \zeta$ is a generalized subgradient of the mapping

$$\mathcal{R}(g) = \begin{cases} \|g\|_\mu^2 & \text{if } g \in \mathcal{X}, \\ +\infty & \text{else.} \end{cases}$$

4. Extension to the stochastic setting

In this section, we allow the approximation errors $\|T^\delta - T\|$ and $\|h^\gamma - h\|$ to be stochastic and depend on the sample size n. More precisely, T^δ is a nonparametric estimator of the operator T depending on the random sample $(Y_i, X_i, W_i)_{i=1,\dots,n}$ and we will denote it by $\hat{T}$. Similarly, h^γ is a nonparametric estimator of the function h depending on the random sample $(Y_i, X_i, W_i)_{i=1,\dots,n}$ and we will denote it by $\hat{h}$. Finally, the approximated regularized solution $g_\alpha^{\delta,\gamma}$ will be denoted by $\hat{g}_\alpha$.

In the following, we will derive convergence rates in probability for $\hat{g}_\alpha$. To that end, recall that a sequence of random variables Q_n, $n \in \mathbb{N}$, in a normed space is bounded in probability, if for every $\epsilon > 0$ there exists $C > 0$ and $n_0 \in \mathbb{N}$ such that

$$\mathbb{P}(\|Q_n\| > C) < \epsilon \qquad \text{for all } n \geq n_0.$$

In this case, we say that

$$Q_n = O_P(1).$$

Similarly, if c_n, $n \in \mathbb{N}$, is any real sequence, we write

$$Q_n = O_P(c_n) \qquad \text{if} \qquad \frac{Q_n}{c_n} = O_P(1).$$

Note that an alternative to convergence rates in probability is the derivation of convergence rates in expectation, which has been carried out for Tikhonov regularization and generalizations in [5, 6]. In this paper, however, we will restrict ourselves to rates in probability in order to be able to exploit the results in [13] on unconstrained instrumental regression.

Following [13], we introduce the kernel approach with generalized kernel functions of order l for estimating $\hat{T}$ and $\hat{h}$. Note that the kernel is considered in generalized form only to overcome edge effects. Let $\sigma \equiv \sigma_n \to 0$ denote a bandwidth and $K_\sigma(\cdot, \cdot)$ denote a univariate generalized kernel function with the properties $K_\sigma(u, t) = 0$ if $u > t$ or $u < t - 1$; for all $t \in [0, 1]$,

$$\sigma^{-(j+1)} \int_{t-1}^{t} u^j K_\sigma(u, t) du = \begin{cases} 1 & \text{if } j = 0, \\ 0 & \text{if } 1 \leq j \leq l - 1. \end{cases}$$

We call $K_\sigma(\cdot, \cdot)$ a univariate generalized kernel function of order l (see [23]). A special class of multivariate generalized kernel functions of order l is given by that of products of univariate generalized kernel functions of order l. Let $K_{X,\sigma}$ and $K_{W,\sigma}$ denote two generalized multivariate kernel functions of dimension $k + 1$ and $K_{Y,\sigma}$ a kernel function of dimension 1. First we estimate the density functions f_{YW}, f_{XW} and f_W. Note that, for simplicity of notation, we use the same bandwidth to estimate the three densities

$$\hat{f}_{YW}(y, w) = \frac{1}{n\sigma^{k+2}} \sum_{i=1}^{n} K_{Y,\sigma}(y - Y_i, y) K_{W,\sigma}(w - W_i, w),$$

$$\hat{f}_{XW}(x, w) = \frac{1}{n\sigma^{2k+2}} \sum_{i=1}^{n} K_{X,\sigma}(x - X_i, x) K_{W,\sigma}(w - W_i, w),$$

$$\hat{f}_W(w) \quad = \frac{1}{n\sigma^{k+1}} \sum_{i=1}^n K_{W,\sigma}(w - W_i, w).$$

Then the estimators of T and h are

$$\hat{T}\psi(w) = \int \psi(x)\frac{\hat{f}_{XW}(x, w)}{\hat{f}_W(w)}dx,$$

$$\hat{h}(w) \quad = \int y\frac{\hat{f}_{YW}(y, w)}{\hat{f}_W(w)}dy.$$

In order to derive a rate of convergence for $\hat{g}_\alpha$, we require

Assumption 4.1. We assume that the following conditions are satisfied:

(i) The data (Y_i, X_i, W_i), $i = 1, \ldots, n$, define an *i.i.d.* sample of (Y, X, W).

(ii) The probability density function f_{YXW} is l times continuously differentiable in the interior of $\Omega_Y \times \Omega \times \Omega_W$ and bounded away from zero on $\Omega_Y \times \Omega \times \Omega_W$.

(iii) The conditional expectation $\mathbb{E}(\epsilon^2|W = w)$ is uniformly bounded on Ω_W.

(iv) Both multivariate kernels $K_{X,\sigma}$ and $K_{W,\sigma}$ are product kernels generated from the univariate generalized kernel function K_σ with the following properties:

    (a) The kernel function K_σ is a generalized kernel function of order l.

    (b) For each $t \in [0, 1]$, the function $K_\sigma(\sigma\cdot, t)$ is supported on a set of the form $[(t-1)/\sigma, t/\sigma] \cap \mathcal{K}$ where $\mathcal{K}$ is a compact interval not depending on t and $\sup_{\sigma>0, t\in[0,1], u\in\mathcal{K}} |K_\sigma(\sigma u, t)| < \infty$.

(v) The bandwidth parameter satisfies $\sigma \to 0$ and $(n\sigma^{2k+2})^{-1}\log(n) \to 0$.

Proposition 4.2. *Suppose Assumption 4.1 holds. Let $\rho = \min\{l, k+1\} \geq 2$ and $\mu \geq 0$. Let*

$$g^\dagger \in \arg\min\{\|g\|_\mu^2 : Tg = h, \ g \in \mathcal{X}\}.$$

and

$$\hat{g}_\alpha \in \arg\min\{\mathcal{T}_\alpha(g; \hat{T}, \hat{h}) : g \in \mathcal{X}\}.$$

Assume that $\partial_\nu g^\dagger = 0$ on $\partial\Omega$. Denote moreover by $T^*: L^2(\Omega; \mathbb{R}^k) \to L^2(\Omega; \mathbb{R}^k)$ the adjoint of T and let $N_{\mathcal{X}}^P(g^\dagger) \subset L^2(\Omega; \mathbb{R}^{k+1})$ be the proximal normal cone to the set $\mathcal{X}$ at the point $g^\dagger$.

(i) *Let $\mu > 0$. Assume that there exist $\omega \in L^2(\Omega; \mathbb{R}^k)$ and $\zeta \in N_{\mathcal{X}}^P(g^\dagger)$ with $\tau(\zeta, g^\dagger) < \mu$ such that*

$$2(\mu g^\dagger - \Delta g^\dagger) = T^*\omega + \zeta.$$

Then the estimate

$$\|\hat{g}_\alpha - g^\dagger\|_\mu^2 = O_P\left((1 + \sqrt{\alpha})\frac{\frac{1}{n\sigma^{2k+2}} + \sigma^{2\rho}}{\alpha} + \frac{1}{n\sigma^{k+1}} + \sigma^\rho + \alpha\right)$$

holds.

(ii) *Let $\mu = 0$ and assume that $\alpha \to 0$, $\left(\frac{1}{n\sigma^{2k+2}} + \sigma^{2\rho}\right)/\alpha \to 0$ as $n \to \infty$. Assume moreover that there exist $\omega \in L^2(\Omega; \mathbb{R}^k)$ and $\zeta \in N_{\mathcal{X}}^P(g^\dagger)$ such that*

$$2(\mu g^\dagger - \Delta g^\dagger) = T^*\omega + \zeta$$

and

$$A^2(\|T\|D(T)^{-1} + D(T)^{-1} + 1)^2 \tau(\zeta, g^\dagger) < 1,$$

where A and $D(T)$ are as in Lemma 2.2. Then the estimate

$$\|\nabla\hat{g}_\alpha - \nabla g^\dagger\|^2 = O_P\left(\frac{\frac{1}{n\sigma^{2k+2}} + \sigma^{2\rho}}{\alpha} + \frac{1}{n\sigma^{k+1}} + \sigma^\rho + \alpha\right)$$

holds.

In particular, if

$$\alpha \asymp n^{-\frac{\rho}{2(k+\rho+1)}} \qquad \text{and} \qquad \sigma \asymp n^{-\frac{1}{2(k+\rho+1)}},$$

then we obtain in both cases the rate

$$\|\hat{g}_\alpha - g^\dagger\|_\mu^2 = O_P(n^{-\frac{\rho}{2(k+\rho+1)}}).$$

Proof. Note first that the assumption that the density f_{YXW} is bounded away from zero implies that the operator T is bounded and satisfies $Tc \neq 0$ for every constant function c. Moreover, in [13] the convergence rate result

$$\|\hat{T} - T\|^2 = O_P\left(\frac{1}{n\sigma^{2k+2}} + \sigma^{2\rho}\right),$$

$$\|\hat{h} - h\|^2 = O_P\left(\frac{1}{n\sigma^{2k+2}} + \sigma^{2\rho}\right)$$

has been derived under Assumption 4.1. Together with the results of Lemma 3.1, Theorem 3.4 and Corollary 3.5, this immediately proves the assertion in the case $\mu > 0$.

In the case $\mu = 0$, note that the assumption on the behaviour of α and Proposition 2.4 imply that the regularized solutions $\hat{g}_\alpha$ converge in probability to $g^\dagger$. Moreover, the convergence in probability of $\hat{T}$ to T implies that $1/(D(T) - \|\hat{T} - T\|) = O_P(1)$, and therefore, as $\sigma \to 0$ and $1/(n\sigma^{2k+2}) \to 0$, we obtain in the notation of Lemma 3.1 the estimate

$$D_0(\alpha, \|\hat{T} - T\|, \|\hat{h} - h\|) = O_P(1).$$

Then the result follows again immediately from Lemma 3.1, Theorem 3.4 and Corollary 3.5. □

5. Conclusion

In this paper, we have studied the problem of nonparametric regression in the presence of endogenous variables and additional non-convex shape constraints. The main motivation is the estimation of the consumer demand function, which, according to standard microeconomic theory, satisfies certain (non-linear) integrability conditions. We have used instruments in order to tackle the issue of endogeneity, which, in the case where the coupling between the instruments and the explanatory variables is weak (that is, only given by a density), leads to the solution of an ill-posed operator equation.

We propose to solve the resulting inverse problem by (constrained) Tikhonov regularization using a weighted Sobolev norm as a regularization term. Because of the weak closedness of the constrained set in the Sobolev space, the regularization method is convergent. In addition, we have derived convergence rates under the additional assumption that the true solution $g^\dagger$ satisfies a certain variational inequality, which is shown to hold if $g^\dagger$ satisfies a projected source condition. In contrast to the usual convex case, however, this condition is coupled with a smallness condition. The convergence rates are derived in both a deterministic and a stochastic setting. In the latter situation we have the additional problem that the correspondence between the instruments and the explanatory variables, and thus the operator itself, is not known exactly but has to be estimated in a first step. Here we propose to use a kernel estimator, which allows us to obtain rates in probability for the operator error in dependence of the number of measurements.

References

[1] R. A. Adams. *Sobolev Spaces.* Academic Press, New York, 1975.

[2] J.D. Angrist and A.B. Krueger. Instrumental variables and the search for identifications: from supply and demand to natural experiments. *J. Econ. Persp.*, 15(4):69–85, 2001.

[3] G. Aubert and L. Vese. A variational method in image recovery. *SIAM J. Numer. Anal.*, 34(5):1948–1979, 1997.

[4] S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica* 63:841-890, 1995.

[5] N. Bissantz, T. Hohage, and A. Munk. Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Probl.*, 20(6):1773–1789, 2004.

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

[7] R. Blundell, X. Chen, and D. Kristensen. Nonparametric IV estimation of shape-invariant Engel curves *Econometrica*, 75:1613–1669, 2007.

[8] B. Brown, and M. Walker. The random utility hypothesis and inference in demand systems. *Econometrica*, 57:81529, 1989.

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

[10] G. Chavent and K. Kunisch. Convergence of Tikhonov regularization for constrained ill-posed inverse problems. *Inverse Probl.*, 10:63–76, 1994.

[11] L. Chen, G.D. Smith, R. Harbord, and S.J. Lewis. Alcohol Intake and Blood Pressure: A

Systematic Review Implementing a Mendelian Randomization Approach. *PLoS Med*, 5(3), 2008.

[12] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth analysis and control theory*, volume 178 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

[13] S. Darolles, Y. Fan, J.P. Florens, and E. Renault. Nonparametric Instrumental Regression. *Econometrica*, 79:1541–1565, 2011.

[14] V. Didelez, S. Meng, and N.A. Sheehan. Assumptions of IV Methods for Observational Epidemiology. *Statistical Science*, 25:22–40, 2010.

[15] B. Eicke. Iteration methods for convexly constrained ill-posed problems in Hilbert space. *Numer. Funct. Anal. Optim.*, 13(5-6):413–429, 1992.

[16] J. Flemming and B. Hofmann. Convergence rates in constrained Tikhonov regularization: equivalence of projected source conditions and variational inequalities. *Inverse Probl.*, 27(8):085001, 11, 2011.

[17] J.P. Florens. *Inverse Problems and Structural Econometrics: The Example of Instrumental Variables*, in Advances in Economics and Econometrics: Theory and Applications, 284–311. Cambridge: Cambridge University Press, 2003.

[18] M. Grasmair. Generalized Bregman distances and convergence rates for non-convex regularization methods. *Inverse Probl.*, 26(11):115014, 2010.

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

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

[21] A. Lewbel. Demand systems with and without errors. *American Economic Review* 91:61118, 2001.

[22] R. Matzkin. *Heterogeneous choice.* In N. W. Blundell, R. and T. Persson (Eds.), Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometrics Society, Volume 43:111–121. Cambridge: Cambridge University Press, 2007.

[23] H.-G. Müller. Smooth Optimum Kernel Estimators Near Endpoints. *Biometrika*, 78:521-530, 1991.

[24] A. Neubauer. Finite-dimensional approximation of constrained Tikhonov-regularized solutions of ill-posed linear operator equations. *Math. Comp.*, 48(178):565–583, 1987.

[25] A. Neubauer and O. Scherzer. Finite-dimensional approximation of Tikhonov regularized solutions of nonlinear ill-posed problems. *Numer. Funct. Anal. Optim.*, 11(1-2):85–99, 1990.

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

[27] A. Vanhems. Nonparametric estimation of exact consumer surplus with endogeneity in price. *Econometrics Journal*, 13(3):80-98, 2010.

[28] S. Vansteelandt, J. Bowden, M. Babanezhad, and E. Goetghebeur. On Instrumental Variables Estimation of Causal Odds Ratios. *Statistical Science*, 26:403–422, 2011.

[29] H. R. Varian. *Microeconomic Analysis*. W. W. Norton & Company, New York, third edition, 1992.

[30] L. Vese. A study in the BV space of a denoising-deblurring variational problem. *Appl. Math. Optim.*, 44(2):131–161, 2001.

[31] J. Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western College Pub, fourth edition, 2008.

[32] W. P. Ziemer. *Weakly Differentiable Functions. Sobolev Spaces and Functions of Bounded Variation*, volume 120 of *Graduate Texts in Mathematics*. Springer Verlag, Berlin etc., 1989.