Nonparametric Instrumental Variable Estimation Under Monotonicity*

Denis Chetverikov† Daniel Wilhelm‡

Abstract

The ill-posedness of the inverse problem of recovering a regression function in a nonparametric instrumental variable model leads to estimators that may suffer from a very slow, logarithmic rate of convergence. In this paper, we show that restricting the problem to models with monotone regression functions and monotone instruments significantly weakens the ill-posedness of the problem. Under these two monotonicity assumptions, we establish that the constrained estimator that imposes monotonicity possesses the same asymptotic rate of convergence as the unconstrained estimator, but the finite-sample behavior of the constrained estimator (in terms of error bounds) is much better than expected from the asymptotic rate of convergence when the regression function is not too steep. In the absence of the point-identifying assumption of completeness, we also derive non-trivial identification bounds on the regression function as implied by our two monotonicity assumptions. Finally, we provide a new adaptive test of the monotone instrument assumption and a simulation study that demonstrates significant finite-sample performance gains from imposing monotonicity.

*First version: January 2014. This version: May 5, 2015. We thank Alex Belloni, Richard Blundell, Stéphane Bonhomme, Moshe Buchinsky, Xiaohong Chen, Victor Chernozhukov, Andrew Chesher, Joachim Freyberger, Jinyong Hahn, Dennis Kristensen, Simon Lee, Zhipeng Liao, Rosa Matzkin, Ulrich Müller, Markus Reiß, Susanne Schennach, and Vladimir Spokoiny for useful comments and discussions.

†Department of Economics, University of California at Los Angeles, 315 Portola Plaza, Bunche Hall, Los Angeles, CA 90024, USA; E-Mail address: chetverikov@econ.ucla.edu.

‡Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom; E-Mail address: d.wilhelm@ucl.ac.uk. The author gratefully acknowledges financial support from the ESRC Centre for Microdata Methods and Practice at IFS (RES-589-28-0001).

1 Introduction

Despite the pervasive use of linear instrumental variable methods in empirical research, their nonparametric counterparts are far from enjoying similar popularity. Perhaps two of the main reasons for this originate from the observation that point-identification of the regression function in the nonparametric instrumental variable (NPIV) model requires completeness assumptions, which have been argued to be strong (Santos (2012)) and non-testable (Canay, Santos, and Shaikh (2013)), and from the fact that the NPIV model is ill-posed, and so the regression function estimators in this model may suffer from a very slow, logarithmic rate of convergence (e.g. Blundell, Chen, and Kristensen (2007)).

In this paper, we study the possibility of imposing shape restrictions to improve statistical properties of the NPIV estimators. We consider two monotonicity conditions. The first condition is that the regression function to be estimated is monotone, and the second is that there is a monotone relationship between the endogenous covariate and the instrument. We show that the interaction of these conditions significantly changes the structure of the NPIV model, and weakens its ill-posedness. Specifically, we demonstrate that under the second condition, a slightly modified version of the sieve measure of ill-posedness of the NPIV model defined in Blundell, Chen, and Kristensen (2007) is bounded uniformly over the dimension of the sieve space, when restricted to the set of monotone functions; see Section 2 for details. As a result, the constrained NPIV estimator that imposes monotonicity of the regression function has a fast rate of convergence in a slowly shrinking neighborhood of constant functions. In fact, the rate of convergence is similar to that of nonparametric conditional mean estimators, regardless of how severely ill-posed the unconstrained NPIV model is. In addition, we show that in the absence of completeness assumptions, that is, when the NPIV model is not point-identified, our monotonicity conditions have non-trivial identification power, and can provide partial identification of the model.

We consider the NPIV regression model for a dependent variable Y, an endogenous covariate X, and an instrumental variable (IV) W,

$$Y = g(X) + \varepsilon, \qquad \mathrm{E}[\varepsilon|W] = 0. \tag{1}$$

Our interest focuses on identification of the function g and its estimation based on n i.i.d. observations from the distribution of (Y, X, W). We assume that g is smooth and monotone, but do not impose any parametric restrictions. In addition, we assume that the relationship between the endogenous covariate X and the instrument W is also monotone in the sense that the conditional distribution of X given W corresponding to higher values of W first-order stochastically dominates the same conditional distribution corresponding

to lower values of W. We refer to this condition as monotone instrumental variable (MIV) assumption. To simplify the presentation, we assume that all variables are scalar.

When the function g is strictly increasing, it is easy to show that under appropriate conditions, as the sample size n increases, the usual unconstrained NPIV estimators, such as those derived in Blundell, Chen, and Kristensen (2007) and Horowitz (2012), will be monotone increasing with probability approaching one; see, for example, Lemma 1 below. This implies that the constrained estimators, which impose monotonicity of the function g, will coincide with the unconstrained estimators with probability approaching one, and so the convergence rate of the constrained estimators can not exceed that of the unconstrained estimators. However, our simulations in Section 6 indicate that in finite samples, imposing the monotonicity constraint on the function g yields significant performance improvements even if g is strictly increasing. Thus, the asymptotic analysis of convergence rates does not capture an important finite sample phenomenon observed in simulations. Therefore, we rely upon a non-asymptotic approach. Specifically, we derive a new non-asymptotic error bound for the constrained version of the estimator of Blundell, Chen, and Kristensen (2007) under the MIV assumption. The bound has two regimes. The first regime is active when the function g is not too steep (as the sample size n increases, this regime applies in a slowly shrinking neighborhood of constant functions), and the bound in this regime is independent of the sieve measure of ill-posedness of the NPIV model appearing in the rate of convergence of the unconstrained estimator. In fact, under some further conditions, the bound in the first regime takes the following form: with high probability,

$$\|\widehat{g}^c - g\|_{2,t} \leq C\Big(\frac{K \log n}{n} + K^{-s}\Big)$$

where $\widehat{g}^c$ is the constrained estimator, $\|\cdot\|_{2,t}$ is an appropriate L^2-norm, K is the number of series terms in the estimator $\widehat{g}^c$, s is the smoothness of the function g, and C is some constant; see Section 3 for details. Thus, the bound in the first regime is similar to that for the series estimators of conditional mean functions. The second regime is active when the function g is sufficiently steep. In this regime, the bound is similar to that for the unconstrained NPIV estimators, which is expected from the discussion of the convergence rates of the constrained and unconstrained estimators above.

We regard both monotonicity conditions as natural in many economic applications. In fact, both of these conditions often directly follow from economic theory. Consider a generic example very similar to one of the examples in Kasy (2014). Suppose that a firm produces log output Y from log labor input X. Denote by V and W the log wage and the log price of output, respectively. Suppose that the production function is $Y = g(X) + \varepsilon$, where ε summarizes determinants of output other than labor input, such as capital and total factor productivity. The profit of the firm is $\pi = e^W e^Y - e^V e^X$.

The firm chooses X optimally to maximize its profit, so that $X = r(V, W, \varepsilon)$ for some function r. If g is increasing and strictly concave, and the elasticity of output with respect to labor input is strictly smaller than one, straightforward calculations show that the function $w \mapsto r(V, w, \varepsilon)$ is decreasing for all values of V and ε, and so one can expect that W satisfies the MIV assumption (a sufficient condition is that W is independent of the pair (V, ε)). Another example is the estimation of Engel curves. In this case, the outcome variable Y is the budget share of the good, the endogenous variable X is total expenditure, and the instrument W is gross income. Our monotonicity conditions are plausible in this example because for normal goods such as food-in, the budget share is decreasing in total expenditure, and total expenditure increases with gross income. Finally, consider the estimation of (Marshallian) demand curves. The outcome variable Y is quantity of a consumed good, the endogenous variable X is the price of the good, and W could be some variable that shifts production cost of the good. For a normal good, the Slutsky inequality predicts Y to be decreasing in price X as long as income effects are not too large. Furthermore, price is increasing in production cost and, thus, increasing in the instrument W, and so our monotonicity conditions are plausible in this example as well.

In addition, both of our monotonicity assumptions are testable. For example, a test of the MIV condition can be found in Lee, Linton, and Whang (2009). In this paper, we extend their results by deriving an *adaptive* test of the MIV condition, with the value of a certain smoothness parameter being chosen automatically in a data-driven fashion. This adaptation procedure allows us to construct a test with desirable power properties when the degree of smoothness of the conditional distribution of X given W is unknown. Regarding our first monotonicity condition, to the best of our knowledge, there are no procedures in the literature that would test monotonicity of the function g in the NPIV model (1). We consider such procedures in a separate project.

The use of shape restrictions in econometrics to facilitate identification or to improve the performance of nonparametric estimators was advocated by Matzkin (1994), who argued that economic theory often provides restrictions on functions of interest, such as monotonicity, concavity, and/or Slutsky symmetry. In the context of NPIV model (1), Freyberger and Horowitz (2013) showed that, in the absence of point-identification, shape restrictions may yield informative bounds on functionals of g and developed certain inference procedures when regressors X and instruments W are discrete. Blundell, Horowitz, and Parey (2013) demonstrated via simulations that imposing Slutsky inequalities in a quantile NPIV model for gasoline demand improves finite sample properties of the NPIV estimator. Grasmair, Scherzer, and Vanhems (2013) studied the problem of demand estimation imposing various constraints implied by economic theory, such as

Slutsky inequalities, and derived the convergence rate of a constrained NPIV estimator under certain abstract projected source condition. Our results are different from theirs because we focus on non-asymptotic error bounds, with special emphasis on properties of our estimator in *the neighborhood* of the boundary, we derive a bound on the restricted sieve measure of ill-posedness under easily interpretable, low level conditions, and we find that our estimator does not suffer from ill-posedness of the problem in a slowly shrinking neighborhood of constant functions.

Other related literature. The analysis of NPIV model (1) was started by Newey and Powell (2003) who analyzed identification of the model, derived an estimator of the function g in this model, and proved its consistency. Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007), and Darolles, Fan, Florens, and Renault (2011) introduced other estimators and established (optimal) rates of convergence of these estimators. See Horowitz (2011, 2014) for recent surveys and further references. In the mildly ill-posed case, Hall and Horowitz (2005) derived the minimax risk lower bound in L^2-norm and showed that their estimator achieves this lower bound. Under different conditions, Chen and Reiß (2011) derived a similar bound for the mildly and the severely ill-posed cases and showed that the estimator by Blundell, Chen, and Kristensen (2007) achieves this bound. Chen and Christensen (2013) established minimax risk bounds in the sup-norm, again both for the mildly and the severely ill-posed cases. The optimal convergence rates in the severely ill-posed case were shown to be logarithmic, which means that the slow convergence rate of existing estimators is not a deficiency of those estimators but rather an intrinsic feature of the statistical inverse problem.

There is also large statistics literature on nonparametric estimation of monotone functions in the exogenous case (that is, when $W = X$, so that g is a conditional mean function), which can be traced back at least to Brunk (1955). Surveys of this literature and further references can be found in Yatchew (1998), Delecroix and Thomas-Agnan (2000), and Gijbels (2004). For the case in which the regression function is both smooth and monotone, many different ways of imposing monotonicity on the estimator have been studied; see, for example, Mukerjee (1988), Cheng and Lin (1981), Wright (1981), Friedman and Tibshirani (1984), Ramsay (1988), Mammen (1991), Ramsay (1998), Mammen and Thomas-Agnan (1999), Hall and Huang (2001), Mammen, Marron, Turlach, and Wand (2001), and Dette, Neumeyer, and Pilz (2006). Importantly, the standard, unconstrained nonparametric regression estimators are known to be monotone with probability approaching one when the regression function is strictly increasing under mild assumption that these estimators consistently estimate the derivative of the regression function. Therefore, such estimators have the same rate of convergence as that of corresponding

constrained estimators that impose monotonicity (Mammen (1991)). As a consequence, the gains from imposing monotonicity constraint can only be expected when the regression function is close to the boundary of the constraint. Zhang (2002) and Chatterjee, Guntuboyina, and Sen (2013) formalized this intuition by deriving risk bounds of the isotonic (monotone) regression estimators and showing that these bounds imply fast convergence rates when the regression function has flat parts. Our results are different from theirs because we focus on the endogenous case with $W \neq X$.

Notation. For a differentiable function $f : \mathbb{R} \to \mathbb{R}$, we use $Df(x)$ to denote its derivative. When a function f has several arguments, we use D with an index to denote the derivative of f with respect to corresponding argument; for example, $D_w f(w, u)$ denotes the partial derivative of f with respect to w. For random variables A and B, we denote by $f_{A,B}(a, b)$, $f_{A|B}(a, b)$, and $f_A(a)$ the joint, conditional and marginal densities of (A, B), A given B, and A, respectively. Similarly, we let $F_{A,B}(a, b)$, $F_{A|B}(a, b)$, and $F_A(a)$ refer to the corresponding cumulative distribution functions. For an operator $T : L^2[0, 1] \to L^2[0, 1]$, we let $\|T\|_2$ denote the operator norm defined as

$$\|T\|_2 = \sup_{h \in L^2[0,1]: \|h\|_2 = 1} \|Th\|_2.$$

Finally, by increasing and decreasing we mean that a function is non-decreasing and non-increasing, respectively.

Outline. The remainder of the paper is organized as follows. In the next section, we analyze ill-posedness of the model (1) under our monotonicity conditions and derive a useful bound on a certain restricted measure of ill-posedness for the model (1). Section 4 and Section 3 discuss the implications of our monotonicity assumptions for identification and estimation, respectively. In particular, we show that the rate of convergence of our estimator is always not worse than that of unrestricted estimators but may be much faster in local-to-constant asymptotics. Section 5 provides adaptive tests of our monotonicity assumptions. In Section 6, we present results of a Monte Carlo simulation study. All proofs are contained in the appendix.

2 Local Quantitative Well-Posedness under Monotonicity

In this section, we study properties of the NPIV model (1) when we impose our monotonicity conditions. In particular, we introduce a restricted measure of ill-posedness for

this model (see equation (8) below) and derive a useful upper bound on this measure (Corollary 3).

The NPIV model requires solving the equation $\mathrm{E}[Y|W] = \mathrm{E}[g(X)|W]$ for the function g. Letting $T : L^2[0,1] \to L^2[0,1]$ be the linear operator defined by $(Th)(w) := \mathrm{E}[h(X)|W = w]f_W(w)$ and denoting $m(w) := \mathrm{E}[Y|W = w]f_W(w)$, we can express this equation as

$$Tg = m. \tag{2}$$

In finite-dimensional regressions, the operator T corresponds to a finite-dimensional matrix whose singular values are typically assumed to be nonzero (rank condition). Therefore, the solution g is continuous in m, and consistent estimation of m at a fast convergence rate leads to consistent estimation of g at a fast convergence rate. In infinite-dimensional models, however, T is an operator that typically possesses infinitely many singular values that tend to zero. Therefore, small perturbations in m may lead to large perturbations in g. This discontinuity renders equation (2) ill-posed and introduces challenges in estimation of the NPIV model (1) that are not present in parametric regressions; see Horowitz (2011, 2014) for a more detailed discussion.

In this section, we show that under our monotonicity conditions, equation (2) becomes locally quantitatively well-posed at constant functions; see Definition 2 below. This property leads to the following inequality: there is a finite constant $\bar{C}$ such that for any monotone function g' and any constant function g'', with $m' = Tg'$ and $m'' = Tg''$,

$$\|g' - g''\|_{2,t} \leq \bar{C}\|m' - m''\|_2,$$

where $\|\cdot\|_{2,t}$ is a truncated L^2-norm defined below. This result will be central to our derivation of a useful bound on the restricted measure of ill-posedness, of identification bounds, and of fast convergence rates of a monotone NPIV estimator in a slowly shrinking neighborhood of constant functions.

We now introduce our assumptions. Let $0 \leq x_1 < \widetilde{x}_1 < \widetilde{x}_2 < x_2 \leq 1$ and $0 \leq w_1 < w_2 \leq 1$ be some constants. We implicitly assume that x_1, $\widetilde{x}_1$, and w_1 are close to 0 whereas x_2, $\widetilde{x}_2$, and w_2 are close to 1. Our first assumption is the Monotone Instrumental Variable (MIV) condition that requires a monotone relationship between the endogenous covariate X and the instrument W.

Assumption 1 (Monotone IV)**.** *For all $x, w', w'' \in (0,1)$,*

$$w' \leq w'' \qquad \Rightarrow \qquad F_{X|W}(x|w') \geq F_{X|W}(x|w''). \tag{3}$$

Furthermore, there exists a constant $C_F > 1$ such that

$$F_{X|W}(x|w_1) \geq C_F F_{X|W}(x|w_2), \qquad \forall x \in (0, x_2) \tag{4}$$

and

$$C_F(1 - F_{X|W}(x|w_1)) \leq 1 - F_{X|W}(x|w_2), \qquad \forall x \in (x_1, 1) \tag{5}$$

Assumption 1 is crucial for our analysis. The first part, condition (3), requires first-order stochastic dominance of the conditional distribution of the endogenous covariate X given the instrument W as we increase the value of the instrument W. This condition (3) is testable; see, for example, Lee, Linton, and Whang (2009). In Section 5 below, we extend the results of Lee, Linton, and Whang (2009) by providing an *adaptive* test of the first-order stochastic dominance condition (3).

The second and third parts of Assumption 1, conditions (4) and (5), strengthen the stochastic dominance condition (3) in the sense that the conditional distribution is required to "shift to the right" by a *strictly* positive amount at least between two values of the instrument, w_1 and w_2, so that the instrument is not redundant. Conditions (4) and (5) are rather weak as they require such a shift only in some intervals $(0, x_2)$ and $(x_1, 1)$, respectively.

Condition (3) can be equivalently stated in terms of monotonicity with respect to the instrument W of the reduced form first stage function. Indeed, by the Skorohod representation, it is always possible to construct a random variable U distributed uniformly on $[0, 1]$ such that U is independent of W, and equation $X = r(W, U)$ holds for the reduced form first stage function $r(w, u) := F^{-1}_{X|W}(u|w) := \inf\{x : F_{X|W}(x|w) \geq u\}$. Therefore, condition (3) is equivalent to the assumption that the function $w \mapsto r(w, u)$ is increasing for all $u \in [0, 1]$. Notice, however, that our condition (3) allows for general unobserved heterogeneity of dimension larger than one, for instance as in Example 2 below.

Condition (3) is related to a corresponding condition in Kasy (2014) who assumes that the (structural) first stage has the form $X = \widetilde{r}(W, \widetilde{U})$ where $\widetilde{U}$, representing (potentially multidimensional) unobserved heterogeneity, is independent of W, and the function $w \mapsto \widetilde{r}(w, \widetilde{u})$ is increasing for all values $\widetilde{u}$. Kasy employs his condition for identification of (nonseparable) triangular systems with multidimensional unobserved heterogeneity whereas we use our condition (3) to derive a useful bound on the restricted measure of ill-posedness and to obtain a fast rate of convergence of a monotone NPIV estimator of g in the (separable) model (1). Condition (3) is not related to the MIV assumption in the influential work by Manski and Pepper (2000) which requires the function $w \mapsto \mathrm{E}[\varepsilon|W = w]$ to be increasing. Instead, we maintain the mean independence condition $\mathrm{E}[\varepsilon|W] = 0$.

Assumption 2 (Density)**.** *(i) The joint distribution of the pair (X, W) is absolutely continuous with respect to the Lebesgue measure on $[0, 1]^2$ with the density $f_{X,W}(x, w)$ satisfying $\int_0^1 \int_0^1 f_{X,W}(x, w)^2 dxdw \leq C_T$ for some finite constant C_T. (ii) There exists a constant $c_f > 0$ such that $f_{X|W}(x|w) \geq c_f$ for all $x \in [x_1, x_2]$ and $w \in \{w_1, w_2\}$. (iii) There exists constants $0 < c_W \leq C_W < \infty$ such that $c_W \leq f_W(w) \leq C_W$ for all $w \in [0, 1]$.*

This is a mild regularity assumption. The first part of the assumption implies that the operator T is compact. The second and the third parts of the assumption require the conditional distribution of X given $W = w_1$ or w_2 and the marginal distribution of W to be bounded away from zero over some intervals. Recall that we have $0 \leq x_1 < x_2 \leq 1$ and $0 \leq w_1 < w_2 \leq 1$. We could simply set $[x_1, x_2] = [w_1, w_2] = [0, 1]$ in the second part of the assumption but having $0 < x_1 < x_2 < 1$ and $0 < w_1 < w_2 < 1$ is required to allow for densities such as the normal, which, even after a transformation to the interval $[0, 1]$, may not yield a conditional density $f_{X|W}(x|w)$ bounded away from zero; see Example 1 below. Therefore, we allow for the general case $0 \leq x_1 < x_2 \leq 1$ and $0 \leq w_1 < w_2 \leq 1$. The restriction $f_W(w) \leq C_W$ for all $w \in [0, 1]$ imposed in Assumption 2 is not actually required for the results in this section, but rather those of Section 3.

We now give two examples of the pairs (X, W) that satisfy Assumptions 1 and 2. These examples show two possible ways in which the instrument W can shift the conditional distribution of X given W. Figure 4 displays the conditional distributions in both examples.

Example 1 (Normal density)**.** Let $(\widetilde{X}, \widetilde{W})$ be jointly normal with mean zero, variance one, and correlation $0 < \rho < 1$; that is, $\mathrm{E}[\widetilde{X}] = \mathrm{E}[\widetilde{W}] = 0$, $\mathrm{E}[\widetilde{X}^2] = \mathrm{E}[\widetilde{W}^2] = 1$, and $\mathrm{E}[\widetilde{X}\widetilde{W}] = \rho$. Let $\Phi(u)$ denote the distribution function of $N(0, 1)$ random variable. Define $X = \Phi(\widetilde{X})$ and $W = \Phi(\widetilde{W})$. Since $\widetilde{X} = \rho\widetilde{W} + (1 - \rho^2)^{1/2}U$ for some standard normal random variable U that is independent of $\widetilde{W}$, we have

$$X = \Phi(\rho\Phi^{-1}(W) + (1 - \rho^2)^{1/2}U)$$

where U is independent of W. Therefore, the pair (X, W) clearly satisfies condition (3) of our MIV Assumption 1. Lemma 8 in the appendix verifies that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied. □

Example 2 (Two-dimensional unobserved heterogeneity)**.** Let $X = U_1 + U_2W$, where U_1, U_2, W are mutually independent, $U_1, U_2 \sim U[0, 1/2]$ and $W \sim U[0, 1]$. Since U_2 is positive, it is straightforward to see that the stochastic dominance condition (3) is satisfied. Lemma 9 in the appendix shows that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied. □

We are now ready to state our first main result in this section. Define the truncated L^2-norm $\|\cdot\|_{2,t}$ by

$$\|h\|_{2,t} := \left(\int_{\widetilde{x}_1}^{\widetilde{x}_2} h(x)^2 dx\right)^{1/2}, \quad h \in L^2[0, 1].$$

Also, let $\mathcal{M}$ denote the set of all monotone functions in $L^2[0, 1]$. Finally, define $\zeta := (c_f, c_W, C_F, C_T, w_1, w_2, x_1, x_2, \widetilde{x}_1, \widetilde{x}_2)$. We have the following theorem.

Theorem 1 (Lower Bound on T)**.** *Let Assumptions 1 and 2 be satisfied. Then there exists a finite constant $\bar{C}$ depending only on ζ such that*

$$\|h\|_{2,t} \leq \bar{C}\|Th\|_2 \tag{6}$$

for any function $h \in \mathcal{M}$.

To prove this theorem, we take a function $h \in \mathcal{M}$ with $\|h\|_{2,t} = 1$ and show that $\|Th\|_2$ is bounded away from zero. A key observation that allows us to establish this bound is that, under the MIV Assumption 1, the function $w \mapsto \mathrm{E}[h(X)|W = w]$ is monotone whenever h is. Together with non-redundancy of the instrument W implied by conditions (4) and (5) of Assumption 1, this allows us to show that $\mathrm{E}[h(X)|W = w_1]$ and $\mathrm{E}[h(X)|W = w_2]$ cannot both be close to zero so that $\|\mathrm{E}[h(X)|W = \cdot]\|_2$ is bounded from below by a strictly positive constant from the values of $\mathrm{E}[h(X)|W = w]$ in the neighborhood of either w_1 or w_2. In consequence, $\|Th\|_2$ is bounded from below by the technical Assumption 2.

Theorem 1 implies that, under our MIV Assumption 1 and some regularity conditions (Assumption 2), the operator T is bounded from below on the set $\mathcal{M}$ of monotone functions in $L^2[0,1]$. There are several important consequences to this result. To start with, consider the linear equation (2). By Assumption 2(i), the operator T is compact, and so

$$\frac{\|h_k\|_2}{\|Th_k\|_2} \to \infty \text{ as } k \to \infty \text{ for some sequence } \{h_k, k \geq 1\} \subset L^2[0,1]. \tag{7}$$

Property (7) means that $\|Th\|_2$ being small does not necessarily imply that $\|h\|_2$ is small and, therefore, the inverse of the operator $T : L^2[0,1] \to L^2[0,1]$, when it exists, cannot be continuous. Therefore, (2) is ill-posed in Hadamard's sense[1], if no other conditions are imposed. Theorem 1, on the other hand, implies that, under Assumptions 1 and 2, (7) is not possible if h_k belongs to the set $\mathcal{M}$ of monotone functions in $L^2[0,1]$ for all $k \geq 1$ and we replace the L^2-norm $\|\cdot\|_2$ in the numerator of the left-hand side of (7) by the truncated L^2-norm $\|\cdot\|_{2,t}$.[2]

To understand the relationship of Theorem 1 to well-posedness, we first note that Theorem 1 implies that equation (2) is well-posed in Hadamard's sense under our conditions:

[1]Well- and ill-posedness in Hadamard's sense are defined as follows. Let $A : D \to R$ be a continuous mapping between pseudo-metric spaces (D, ρ_D) and (R, ρ_R). Then, for $d \in D$ and $r \in R$, the equation $Ad = r$ is called "well-posed" on D in Hadamard's sense (see Hadamard (1923)) if (i) A is bijective and (ii) $A^{-1} : R \to D$ is continuous, so that for each $r \in R$ there exists a unique $d = A^{-1}r \in D$ satisfying $Ad = r$, and, moreover, the solution $d = A^{-1}r$ is continous in "the data" r. Otherwise, the equation is called "ill-posed" in Hadamard's sense.

[2]In Remark 1 below, we argue that replacing the norm in the numerator is not a significant modification in the sense that most ill-posed problems, and in particular all severely ill-posed problems, imply (7) under either norm.

Corollary 1 (Well-Posedness in Hadamard's Sense). *Let Assumptions 1 and 2 be satisfied. Equip $\mathcal{M}$ with the norm $\|\cdot\|_{2,t}$ and $T(\mathcal{M})$ with the norm $\|\cdot\|_2$ where $T(\mathcal{M}) = \{Th : h \in \mathcal{M}\}$ is the image of $\mathcal{M}$ under T. Assume that $T : \mathcal{M} \to T(\mathcal{M})$ is one-to-one. Then the problem (2) is well-posed on $\mathcal{M}$ in Hadamard's sense.*

Well-posedness in Hadamard's sense is useful to establish consistency of an estimator of the solution of equation (2), but it does not yield the convergence rate of the estimator. Therefore, we are interested in strengthening this concept of well-posedness. To this end, we introduce the concept of (local) quantitative well-posedness, which is inspired by the definitions in Bejenaru and Tao (2006), who studied the Cauchy problem for the quadratic non-linear Schrödinger equation.

Definition 1 (Quantitative Well-Posedness). *Let (D, ρ_D) and (R, ρ_R) be two pseudo-metric spaces and let $A : D \to R$ be a bijective continuous mapping from D to R. We say that equation $Ad = r$ for $d \in D$ and $r \in R$ is quantitatively well-posed if there exists a finite constant $C > 0$ such that for any $d', d'' \in D$ and $r', r'' \in R$ with $Ad' = r'$ and $Ad'' = r''$, we have $\rho_D(d', d'') \leq C\rho_R(r', r'')$.*

Definition 2 (Local Quantitative Well-Posedness). *In the setting of Definition 1, we say that equation $Ad = r$ for $d \in D$ and $r \in R$ is locally quantitatively well-posed at d_0 if there exists a finite constant $C > 0$ such that for any $d \in D$ and $r \in R$ with $Ad = r$, we have $\rho_D(d, d_0) \leq C\rho_R(r, r_0)$ where $r_0 = Ad_0$.*

Well-posedness in Hadamard's sense is useful for establishing consistency of the solution to the equation $Ad = r$: if, for a sequence of estimators $\widehat{r}_n$ of r, we have $\rho_R(\widehat{r}_n, r) \to 0$ as $n \to \infty$, then $\rho_D(\widehat{d}_n, d) \to 0$ as $n \to \infty$ where $\widehat{d}_n = A^{-1}\widehat{r}_n$ is an estimator of d. The concept of quantitative well-posedness is stronger than well-posedness in Hadamard's sense because it requires a quantitative control on the modulus of continuity of the inverse map $A^{-1} : R \to D$. In particular, quantitative well-posedness guarantees that the convergence rate of $\widehat{d}_n$ to d is not slower than that of $\widehat{r}_n$ to r and in turn allows to establish not only consistency, but also a fast convergence rate. Local quantitative well-posedness is a weaker concept since it applies only to convergence to some particular value $d = d_0$.

Theorem 1 implies that, if Assumptions 1 and 2 hold, and $T : \mathcal{M} \to T(\mathcal{M})$ is one-to-one, then (2) is locally quantitatively well-posed at constant functions.

Corollary 2 (Local Quantitative Well-Posedness). *Let Assumptions 1 and 2 be satisfied. In addition, assume that $T : \mathcal{M} \to T(\mathcal{M})$ is one-to-one and equip the spaces $\mathcal{M}$ and $T(\mathcal{M})$ with the norms $\|\cdot\|_{2,t}$ and $\|\cdot\|_2$, respectively. Then (2) is locally quantitatively well-posed at any constant function.*

With the help of Corollary 2, it is possible to show that a suitable monotone NPIV estimator $\widehat{g}^c$ of the function g has a fast convergence rate when g is constant but in fact we will be able to show much more: we will show that $\widehat{g}^c$ has a fast convergence rate when g belongs to a slowly shrinking neighborhood of constant functions. To this end, we demonstrate that Theorem 1 implies an upper bound on a certain restricted measure of ill-posedness in equation (2). For $a \in \mathbb{R}$, let

$$\mathcal{H}(a) := \left\{ h \in L^2[0,1] : \inf_{0 \le x' < x'' \le 1} \frac{h(x'') - h(x')}{x'' - x'} \ge -a \right\}$$

be the space containing all functions in $L^2[0,1]$ with lower derivative bounded from below by $-a$ uniformly over the interval $[0,1]$. Note that $\mathcal{H}(a') \subset \mathcal{H}(a'')$ whenever $a' \le a''$ and that $\mathcal{H}(0) = \mathcal{M}_+$, the set of increasing functions in $L^2[0,1]$. For continuously differentiable functions, $h \in L^2[0,1]$ belongs to $\mathcal{H}(a)$ if and only if $\inf_{x \in [0,1]} Dh(x) \ge -a$. Further, define the *restricted measure of ill-posedness*:

$$\tau(a) := \sup_{\substack{h \in \mathcal{H}(a) \\ \|h\|_{2,t} = 1}} \frac{\|h\|_{2,t}}{\|Th\|_2}. \tag{8}$$

As we discussed above, under our Assumptions 1 and 2, $\tau(\infty) = \infty$ if we use the L^2-norm instead of the truncated L^2-norm in the numerator in (8). We will also show in Remark 1 below, that $\tau(\infty) = \infty$ for many ill-posed and, in particular, for all severely ill-posed problems even with the truncated L^2-norm as defined in (8). However, it follows from Theorem 1, that $\tau(0)$ is bounded from above by $\bar{C}$ and by definition, $\tau(a)$ is increasing in a; that is, $\tau(a') \le \tau(a'')$ for $a' \le a''$. It turns out that $\tau(a)$ is bounded from above even for some positive values of a:

Corollary 3 (Bound for the Restricted Measure of Ill-Posedness)**.** *Let Assumptions 1 and 2 be satisfied. Then there exist constants $c_\tau > 0$ and $0 < C_\tau < \infty$ depending only on ζ such that*

$$\tau(a) \le C_\tau \tag{9}$$

for all $a \le c_\tau$.

This is our second main result in this section. It is exactly this corollary of Theorem 1 that allows us to obtain a fast convergence rate of our monotone NPIV estimator not only when the regression function g is constant but, more generally, when g belongs to a slowly shrinking neighborhood of constant functions.

Remark 1. Under Assumptions 1 and 2, the integral operator T satisfies (7). Here we demonstrate that in many cases, and in particular in all severely ill-posed cases, (7) continues to hold if we replace the L^2-norm $\|\cdot\|_2$ by the truncated L^2-norm $\|\cdot\|_{2,t}$ in

the numerator of the left-hand side of (7), that is, there exists a sequence $\{l_k, k \geq 1\}$ in $L^2[0,1]$ such that

$$\frac{\|l_k\|_{2,t}}{\|Tl_k\|_2} \to \infty \text{ as } k \to \infty. \tag{10}$$

Indeed, under Assumptions 1 and 2, T is compact, and so the spectral theorem implies that there exists a spectral decomposition of operator T, $\{(h_j, \varphi_j), j \geq 1\}$, where $\{h_j, j \geq 1\}$ is an orthonormal basis of $L^2[0,1]$ and $\{\varphi_j, j \geq 1\}$ is a decreasing sequence of positive numbers such that $\varphi_j \to 0$ as $j \to \infty$, and $\|Th_j\|_2 = \varphi_j \|h_j\|_2 = \varphi_j$. Also, Lemma 7 in the appendix shows that if $\{h_j, j \geq 1\}$ is an orthonormal basis in $L^2[0,1]$, then for any $\alpha > 0$, $\|h_j\|_{2,t} > j^{-1/2-\alpha}$ for infinitely many j, and so there exists a subsequence $\{h_{j_k}, k \geq 1\}$ such that $\|h_{j_k}\|_{2,t} > {j_k}^{-1/2-\alpha}$. Therefore, under a weak condition that $j^{1/2+\alpha}\varphi_j \to 0$ as $j \to \infty$, using $\|h_{j_k}\|_2 = 1$ for all $k \geq 1$, we conclude that for the subsequence $l_k = h_{j_k}$,

$$\frac{\|l_k\|_{2,t}}{\|Tl_k\|_2} \geq \frac{\|h_{j_k}\|_2}{{j_k}^{1/2+\alpha}\|Th_{j_k}\|_2} = \frac{1}{{j_k}^{1/2+\alpha}\varphi_{j_k}} \to \infty \text{ as } k \to \infty$$

leading to (10). Note also that the condition that $j^{1/2+\alpha}\varphi_j \to 0$ as $j \to \infty$ necessarily holds if there exists a constant $c > 0$ such that $\varphi_j \leq e^{-cj}$ for all large j, that is, if the problem is severely ill-posed. Thus, under our Assumptions 1 and 2, the restriction in Theorem 1 that h belongs to the space $\mathcal{M}$ of *monotone* functions in $L^2[0,1]$ plays a crucial role for the result (6) to hold. On the other hand, whether the result (6) can be obtained for all $h \in \mathcal{M}$ without imposing our MIV Assumption 1 appears to be an open question. □

Remark 2. Consider Example 1 above. Because, in this example, the pair (X, W) is a transformation of the normal distribution, it is well known that the integral operator T in this example has singular values decreasing exponentially fast, that is, the spectral decomposition $\{(h_k, \varphi_k), k \geq 1\}$ of the operator T satisfies $\varphi_k = \rho^k$ for all k and some $\rho < 1$. Hence,

$$\frac{\|h_k\|_2}{\|Th_k\|_2} = \left(\frac{1}{\rho}\right)^k.$$

Since $(1/\rho)^k \to \infty$ as $k \to \infty$ exponentially fast, this example leads to a severely ill-posed problem. Moreover, by Lemma 7, for any $\alpha > 0$ and $\rho' \in (\rho, 1)$,

$$\frac{\|h_k\|_{2,t}}{\|Th_k\|_2} > \frac{1}{k^{1/2+\alpha}}\left(\frac{1}{\rho}\right)^k \geq \left(\frac{1}{\rho'}\right)^k$$

for infinitely many k. Thus, replacing the L^2 norm $\|\cdot\|_2$ by the truncated L^2 norm $\|\cdot\|_{2,t}$ preserves the severe ill-posedness of the problem. However, it follows from Theorem 1 that uniformly over all $h \in \mathcal{M}$, $\|h\|_{2,t}/\|Th\|_2 \leq \bar{C}$. Therefore, in this example, as well as in all

other severely ill-posed problems satisfying Assumptions 1 and 2, imposing monotonicity on the function $h \in L^2[0,1]$ significantly changes the properties of the ratio $\|h\|_{2,t}/\|Th\|_2$. □

Remark 3. Here we argue that our MIV Assumption 1 does not imply applicability of a control function approach to estimation of the function g. Consider Example 2 above. In this example, the relationship between X and W has a two-dimensional vector (U_1, U_2) of unobserved heterogeneity. Therefore, by Proposition 4 of Kasy (2011), there does not exist any control function $C : [0,1]^2 \to \mathbb{R}$ such that (i) C is invertible in its second argument, and (ii) X is independent of ε conditional on $V = C(X, W)$. As a consequence, our MIV Assumption 1 does not imply any of the existing control function conditions such as those in Newey, Powell, and Vella (1999) and Imbens and Newey (2009), for example. Therefore, given that it is often argued that multidimensional unobserved heterogeneity is common in economic applications (see Imbens (2007) and Kasy (2014)), we view our approach to avoiding ill-posedness as strongly complementary to the control function approach, with conditions of two approaches being neither weaker nor stronger than conditions of the other approach. □

Remark 4. Finally, let us briefly comment on the role of the truncated norm $\|\cdot\|_{2,t}$ in (6). There are two reasons why we need the truncated L^2-norm $\|\cdot\|_{2,t}$ rather than the usual L^2-norm $\|\cdot\|_2$. First, we want to allow for the normal density as in Example 1, which violates condition (ii) of Assumption 2 if we set $[x_1, x_2] = [0,1]$. Second, when $[x_1, x_2] = [\widetilde{x}_1, \widetilde{x}_2] = [w_1, w_2] = [0,1]$ and Assumptions 1 and 2 hold, we can show (see Lemma 2 in the appendix) that there exists a constant $0 < C_2 < \infty$ such that

$$\|h\|_1 \leq C_2 \|Th\|_1$$

for any increasing and continuously differentiable function $h \in L^1[0,1]$. To extend this result to $L^2[0,1]$-norms we need to introduce a positive, but arbitrarily small, amount of trimming at the boundaries, so that we have a control $\|h\|_{2,t} \leq C\|h\|_1$ for some constant C and all monotone functions $h \in \mathcal{M}$. □

3 Non-asymptotic Risk Bounds Under Monotonicity

The rate at which unconstrained NPIV estimators converge to their probability limit depends crucially on the so-called sieve measure of ill-posedness, which, unlike $\tau(a)$, does not measure ill-posedness over the space $\mathcal{H}(a)$, but rather over a finite-dimensional approximation $\mathcal{H}_n(\infty)$ to $\mathcal{H}(\infty)$. In particular, the convergence rate is slower the faster the sieve measure of ill-posedness grows with the dimensionality of the sieve space $\mathcal{H}_n(\infty)$.

The convergence rates can be as slow as logarithmic in the severely ill-posed case. Since by Corollary 3, our monotonicity assumptions imply boundedness of $\tau(a)$ for some range of finite values a, we expect the monotonicity constraints to translate into favorable performance of an estimator of g that imposes those monotonicity constraints. This intuition is confirmed by the non-asymptotic error bounds we derive in this section.

Let (Y_i, X_i, W_i), $i = 1, \ldots, n$, be an i.i.d. sample from the distribution of (Y, X, W). To define our estimator, we first introduce some notation. Let $\{p_k(x), k \geq 1\}$ and $\{q_k(w), k \geq 1\}$ be two orthonormal bases in $L^2[0,1]$. For $K = K_n \geq 1$ and $J = J_n \geq K_n$, denote

$$p(x) := (p_1(x), \ldots, p_K(x))' \text{ and } q(w) := (q_1(w), \ldots, q_J(w))'.$$

Let $\mathbf{P} := (p(X_1), \ldots, p(X_n))'$ and $\mathbf{Q} := (q(W_1), \ldots, q(W_n))'$. Similarly, stack all observations on Y in $\mathbf{Y} := (Y_1, \ldots, Y_n)'$. Throughout the paper, we assume that $\|g\|_2 \leq C_b$ where C_b is a large but finite constant known by the researcher. We define two estimators of g: the unconstrained estimator $\widehat{g}^u(x) := p(x)'\widehat{\beta}^u$ with

$$\widehat{\beta}^u := \text{argmin}_{b \in \mathbb{R}^K : \|b\| \leq C_b} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b) \tag{11}$$

which is similar to the estimator defined in Horowitz (2012) and a special case of the estimator considered in Blundell, Chen, and Kristensen (2007), and the constrained estimator $\widehat{g}^c(x) := p(x)'\widehat{\beta}^c$ with

$$\widehat{\beta}^c := \text{argmin}_{b \in \mathbb{R}^K : p(\cdot)'b \in \mathcal{H}_n(0), \|b\| \leq C_b} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b), \tag{12}$$

which imposes the constraint that the estimator is an increasing function.

To study properties of the two estimators we introduce a finite-dimensional, or sieve, counterpart of the restricted measure of ill-posedness $\tau(a)$ defined in (8). Consider the sequence of finite-dimensional spaces

$$\mathcal{H}_n(a) := \left\{ h \in \mathcal{H}(a) : \exists b_1, \ldots, b_{K_n} \in \mathbb{R} \text{ with } h = \sum_{j=1}^{K_n} b_j p_j \right\}$$

that become dense in $\mathcal{H}(a)$ as $n \to \infty$. Define the *restricted and unrestricted sieve measures of ill-posedness* $\tau_{n,t}(a)$ and τ_n as

$$\tau_{n,t}(a) := \sup_{\substack{h \in \mathcal{H}_n(a) \\ \|h\|_{2,t} = 1}} \frac{\|h\|_{2,t}}{\|Th\|_2} \text{ and } \tau_n := \sup_{h \in \mathcal{H}_n(\infty)} \frac{\|h\|_2}{\|Th\|_2}.$$

The (unrestricted) sieve measure of ill-posedness defined in Blundell, Chen, and Kristensen (2007) and also used, for example, in Horowitz (2012) is τ_n. Blundell, Chen, and Kristensen (2007) show that τ_n is related to the eigenvalues of T^*T, where T^* is the adjoint of T. If the eigenvalues converge to zero at the rate K^{-2r} as $K \to \infty$, then, under

certain conditions, τ_n diverges at a polynomial rate, that is $\tau_n = O(K_n^r)$. This case is typically called the "mildly ill-posed". On the other hand, when the eigenvalues decrease at a fast exponential rate, then $\tau_n = O(e^{cK_n})$, for some constant $c > 0$, and this case is typically called "severely ill-posed".

Our restricted sieve measure of ill-posedness, $\tau_{n,t}(a)$, is smaller than τ_n because we replace the L^2-norm in the numerator and the space $\mathcal{H}_n(\infty)$ in the definition of τ_n by the truncated L^2-norm in the numerator and the space $\mathcal{H}_n(a)$ in the definition of $\tau_{n,t}(a)$, respectively. As explained in Remark 1, replacing the L^2-norm by the truncated L^2-norm does not make a crucial difference but, as follows from Corollary 3, replacing $\mathcal{H}_n(\infty)$ by $\mathcal{H}_n(a)$ does. In particular, since $\tau(a) \leq C_\tau$ for all $a \leq c_\tau$ by Corollary 3, we also have $\tau_{n,t}(a) \leq C_\tau$ for all $a \leq c_\tau$ because $\tau_{n,t}(a) \leq \tau(a)$. Thus, for all values of a that are not too large, $\tau_{n,t}(a)$ remains bounded uniformly over all n, no matter how fast the eigenvalues of T^*T converge to zero.

We now specify conditions that we need to derive non-asymptotic error bounds for the constrained estimator $\widehat{g}^c(x)$.

Assumption 3 (Monotone regression function)**.** *The function g is monotone increasing.*

Assumption 4 (Moments)**.** *For some constant $C_B < \infty$, (i) $\mathrm{E}[\varepsilon^2|W] \leq C_B$ and (ii) $\mathrm{E}[g(X)^2|W] \leq C_B$.*

Assumption 5 (Relation between J and K)**.** *For some constant $C_J < \infty$, $J \leq CK$.*

Along with Assumption 1, Assumption 3 is our main monotonicity condition. Assumption 4 is a mild moment condition. Assumption 5 requires that the dimension of the vector $q(w)$ is not much larger than the dimension of the vector $p(x)$. Let $s > 0$ be some constant.

Assumption 6 (Approximation of g)**.** *There exist $\beta_n \in \mathbb{R}^K$ and a constant $C_g < \infty$ such that the function $g_n(x) := p(x)'\beta_n$, defined for all $x \in [0,1]$, satisfies (i) $g_n \in \mathcal{H}_n(0)$, (ii) $\|g - g_n\|_2 \leq C_g K^{-s}$, and (iii) $\|T(g - g_n)\|_2 \leq C_g \tau_n^{-1} K^{-s}$.*

The first part of this condition requires the approximating function g_n to be increasing. The second part of this condition requires a particular bound on the approximation error in the L^2-norm. De Vore (1977a,b) show that the assumption $\|g - g_n\|_2 \leq C_g K^{-s}$ holds when the approximating basis $p_1, \ldots, p_K$ consists of polynomial or spline functions and g belongs to a Hölder class with smoothness level s. Therefore, approximation by monotone functions is similar to approximation by all functions. The third part of this condition is similar to Assumption 6 in Blundell, Chen, and Kristensen (2007).

Assumption 7 (Approximation of m)**.** *There exist $\gamma_n \in \mathbb{R}^J$ and a constant $C_m < \infty$ such that the function $m_n(w) := q(w)'\gamma_n$, defined for all $w \in [0,1]$, satisfies $\|m - m_n\|_2 \leq C_m \tau_n^{-1} J^{-s}$.*

This condition is similar to Assumption 3(iii) in Horowitz (2012). Finally, define the operator $T_n : L^2[0,1] \to L^2[0,1]$ by

$$(T_n h)(w) := q(w)'\mathrm{E}[q(W)p(X)']\mathrm{E}[p(U)h(U)], \qquad w \in [0,1]$$

where $U \sim U[0,1]$.

Assumption 8 (Operator T)**.** *(i) The operator T is injective and (ii) for some constant $C_a < \infty$, $\|(T - T_n)h\|_2 \leq C_a \tau_n^{-1} K^{-s} \|h\|_2$ for all $h \in \mathcal{H}_n(\infty)$.*

This condition is similar to Assumption 5 in Horowitz (2012). Finally, let

$$\xi_{K,p} := \sup_{x \in [0,1]} \|p(x)\|, \quad \xi_{J,q} := \sup_{w \in [0,1]} \|q(w)\|, \quad \xi_n := \max(\xi_{K,p}, \xi_{J,q}).$$

We start our analysis in this section with a simple observation that under appropriate conditions, if the function g is strictly increasing, in sufficiently large samples the constrained estimator $\widehat{g}^c$ will coincide with the unconstrained estimator $\widehat{g}^u$, thus sharing with the unconstrained estimator the same rate of convergence.

Lemma 1. *Let Assumptions 1-8 be satisfied. In addition, assume that g is continuously differentiable and $Dg(x) \geq c_g$ for all $x \in [0,1]$ and some constant $c_g > 0$. If $\tau_n^2 \xi_n^2 \log n/n \to 0$, $\sup_{x \in [0,1]} \|Dp(x)\|(\tau_n (K/n)^{1/2} + K^{-s}) \to 0$, and $\sup_{x \in [0,1]} |Dg(x) - Dg_n(x)| \to 0$ as $n \to \infty$, then*

$$\mathrm{P}\Big(\widehat{g}^c(x) = \widehat{g}^u(x) \text{ for all } x \in [0,1]\Big) \to 1 \text{ as } n \to \infty. \tag{13}$$

The result in Lemma 1 is similar to that in Theorem 1 of Mammen (1991), who studied monotone estimators of conditional mean functions, and is negative in the sense that it implies that imposing the monotonicity constraint does not allow to improve the rate of convergence of the estimator if g is strictly increasing. However, the result in Lemma 1 is asymptotic and does not answer the question how large the sample size n should be for the asymptotic approximation to become appropriate. In particular, convergence in (13) may be very slow since one of the conditions of the lemma is that $\sup_{x \in [0,1]} \|Dp(x)\|(\tau_n (K/n)^{1/2} + K^{-s}) \to 0$, and convergence in this condition is typically slow since τ_n gets large as $n \to \infty$, potentially very fast. In addition, our Monte Carlo simulation study in Section 6 indicates that even if g is strictly increasing, there may be significant *finite sample* performance improvements from imposing monotonicity. Therefore, we next derive a *non-asymptotic* estimation error bound for the constrained estimator $\widehat{g}^c$.

Theorem 2 (Non-asymptotic error bound for the constrained estimator)**.** *Let Assumptions 1-8 be satisfied, and let $\delta \geq 0$ be some constant. Assume that $\xi_n^2 \log n/n \leq c$ for sufficiently small $c > 0$. Then with probability at least $1 - \alpha - n^{-1}$, we have*

$$\|\widehat{g}^c - g\|_{2,t} \leq C\Big\{\delta + \tau_{n,t}\Big(\frac{\|Dg_n\|_\infty}{\delta}\Big)\Big(\frac{K}{\alpha n} + \frac{\xi_n^2 \log n}{n}\Big)^{1/2} + K^{-s}\Big\} \tag{14}$$

and

$$\|\widehat{g}^c - g\|_{2,t} \leq C \min\Big\{\|Dg\|_\infty + \Big(\frac{K}{\alpha n} + \frac{\xi_n^2 \log n}{n}\Big)^{1/2}, \tau_n\Big(\frac{K}{\alpha n} + \frac{\xi_n^2 \log n}{n}\Big)^{1/2}\Big\} + CK^{-s}. \tag{15}$$

Here the constants $c, C < \infty$ can be chosen to depend only on the constants appearing in Assumptions 1-8.

This is the main result of this section. An important feature of the bound (14) is that it depends on the restricted sieve measure of ill-posedness, that we know to be smaller than the unrestricted sieve measure of ill-posedness, appearing in the analysis of the unconstrained estimator. Ideally, it would be of great interest to have a tight bound on the restricted sieve measure of ill-posedness $\tau_{n,t}(a)$ for all $a \geq 0$, so that it would be possible to optimize (14) over δ. Results of this form, however, are not yet available in the literature, and so the optimization is not possible. On the other hand, we know from Section 2 that $\tau_{n,t}(a) \leq \tau(a)$ and that by Corollary 3, $\tau(a)$ is uniformly bounded if a is not too large. Employing this result, we obtain the bound (15) of Theorem 2.

The bound (15) has two regimes depending on whether the following inequality

$$\|Dg\|_\infty \leq (\tau_n - 1)\Big(\frac{K}{\alpha n} + \frac{\xi_n^2 \log n}{n}\Big)^{1/2} \tag{16}$$

is satisfied. The most interesting feature of this bound is that in the first regime, when the inequality (16) is satisfied, the bound is independent of the (unrestricted) sieve measure of ill-posedness τ_n, and can be small if the function g is relatively flat regardless of whether the original NPIV model (1) is mildly or severely ill-posed (ill-posedness of the model disappears). This is the regime where the monotonicity constraint imposed on the estimator $\widehat{g}^c$ of g plays an important role.

As the sample size n gets large, the right-hand side of the inequality (16) decreases, and the bound switches to its second regime, where the dependence on the (unrestricted) sieve measure of ill-posedness τ_n appears. This is the regime where the monotonicity constraint is not used. However, since $\tau_n \to \infty$, potentially very fast, even for relatively large sample sizes n and relatively steep functions g, the bound maybe in its first regime, where the monotonicity constraint is important. The presence of the first regime and the observation that it is active in a large set of data generated processes provides a theoretical

justification of importance of imposing the monotonicity constraint on the estimators of the function g in the NPIV model (1) when the MIV Assumption 1 is satisfied.

A corollary of the existence of the first regime in the bound (15) is that the constrained estimator $\widehat{g}^c$ has a fast rate of convergence in a slowly shrinking neighborhood of constant functions regardless of the (unrestricted) sieve measure of ill-posedness τ_n:

Corollary 4 (Fast convergence rate under local-to-constant asymptotics)**.** *Consider the triangular array asymptotics where the data generating process, including the function g, is allowed to vary with n. Let Assumptions 1-8 be satisfied with the same constants for all n. In addition, assume that $\xi_n^2 \leq C_\xi K$ for some $0 < C_\xi < \infty$ and $K \log n/n \to 0$. If $\sup_{x\in[0,1]} Dg(x) = O((K \log n/n)^{1/2})$, then*

$$\|\widehat{g}^c - g\|_{2,t} = O_p((K \log n/n)^{1/2} + K^{-s}). \tag{17}$$

In particular, if $\sup_{x\in[0,1]} Dg(x) = O(n^{-s/(1+2s)}\sqrt{\log n})$ and $K = K_n = C_K n^{1/(1+2s)}$ for some $0 < C_K < \infty$, then

$$\|\widehat{g}^c - g\|_{2,t} = O_p(n^{-s/(1+2s)}\sqrt{\log n}).$$

Remark 5. Observe that the condition that $\xi_n^2 \leq C_\xi K$ holds for some $0 < C_\xi < \infty$ is satisfied if the sequences $\{p_k(x), k \geq 1\}$ and $\{q_k(w), k \geq 1\}$ consist of commonly used bases such as Fourier, spline, wavelet, or local polynomial partition series; see Belloni, Chernozhukov, Chetverikov, and Kato (2014) for details. □

The local-to-constant asymptotics considered in this corollary captures the finite sample situation in which the regression function is not too steep relative to the sample size. The convergence rate in this corollary is the standard polynomial rate of nonparametric conditional mean regression estimators up to a $(\log n)^{1/2}$ factor, regardless of whether the original NPIV problem without imposed monotonicity is mildly or severely ill-posed. One way to interpret this result is that the constrained estimator $\widehat{g}^c$ is able to recover regression functions in the shrinking neighborhood of constant functions at a fast polynomial rate. Notice that the neighborhood of functions g that satisfy $\sup_{x\in[0,1]} Dg(x) = O((K \log n/n)^{1/2})$ is shrinking at a slow rate because $K \to \infty$, in particular the rate is much slower than $n^{-1/2}$. Therefore, in finite samples, we expect the estimator to perform well for a wide range of (non-constant) regression functions g as long as the maximum slope of g is not too large relative to the sample size.

In conclusion, in general, the convergence rate of the restricted estimator is the same as the standard minimax optimal rate, which depends on the degree of ill-posedness and may, in the worst-case, be logarithmic. On the other hand, the restricted estimator converges at a very fast rate, independently of the degree of ill-posedness, in a slowly

shrinking neighborhood of constant functions. In finite samples, we expect to experience intermediate cases and the bound (15) provides information on what the performance of the estimator depends on in such intermediate cases. If the maximum slope $\|Dg\|_\infty$ is small relative to the order of the variance term $\sqrt{K \log n/n}$, then the bound in (15) is of the order $\sqrt{K \log n/n} + K^{-s}$. This order is small and independent of the measure of ill-posedness. Intuitively, the ill-posedness of the problem produces a high variability of an NPIV estimator, which can be seen in the general convergence rate from the fact that the variance term $\sqrt{K/n}$ is multiplied by the diverging measure of ill-posedness, $\tau_{n,t}(\infty)$. This high variability may manifest itself in high-frequency oscillations of the resulting NPIV estimator. Such oscillations are non-monotone so that, in finite samples, the monotonicity constraint is binding even when the population regression function is strictly monotone. In consequence, the monotonicity constraint smooths out the oscillations and thereby reduces variance. Our risk bounds are consistent with this interpretation in the sense that the large order of magnitude of the variance term $\tau_{n,t}(\infty)\sqrt{K/n}$ implies that, in a finite sample, the NPIV estimator may be indistinguishable from many functions with small maximal slope so that the exact finite sample risk of the estimator may be close to the small risk bounds for local-to-constant neighborhoods.

Remark 6. If we replace the condition $\xi_n^2 \log n/n \leq c$ in Theorem 2 by a more restrictive condition $\tau_n^2 \xi_n^2 \log n/n \leq c$, then in addition to the bounds (14) and (15), we can show that with probability at least $1 - \alpha - n^{-1}$, we have

$$\|\widehat{g}^c - g\| \leq C(\tau_n(K/(\alpha n))^{1/2} + K^{-s}).$$

This implies that the constrained estimator $\widehat{g}^c$ satisfies $\|\widehat{g}^c - g\| = O_P(\tau_n(K/n)^{1/2} + K^{-s})$, which is the standard rate of convergence established for the unconstrained estimator $\widehat{g}^u$ in Blundell, Chen, and Kristensen (2007). □

Remark 7. Implementation of the estimators $\widehat{g}^c$ and $\widehat{g}^u$ requires selecting the number of series terms $K = K_n$ and $J = J_n$. This is a difficult problem because the measure of ill-posedness $\tau_n = \tau(K_n)$, appearing in the convergence rate of both estimators, depends on $K = K_n$ and can blow up quickly as we increase K. Therefore, setting K higher than the optimal may result in a severe deterioration of the statistical properties of $\widehat{g}^u$. The problem is alleviated, however, in the case of the constrained estimator $\widehat{g}^c$ because $\widehat{g}^c$ satisfies the bound (15) of Theorem 2, which is independent of τ_n for sufficiently large K. This gives the constrained estimator $\widehat{g}^c$ some robustness against setting K too high.

Remark 8. Notice that the fast convergence rates in the local-to-constant asymptotics derived in this section are obtained under both monotonicity assumptions, Assumptions 1 and 3, but the estimator imposes only the monotonicity of the regression function, not that

of the instrument. Therefore, our proposed constrained estimator consistently estimates the regression function g even when the monotone IV assumption is violated. □

Remark 9. In the local-to-constant asymptotic framework where $\sup_{x\in[0,1]} Dg(x) = O((K \log n/n)^{1/2})$, the rate of convergence in (17) can also be obtained by simply fitting a constant. However, such an estimator, unlike our constrained estimator, is not consistent when the regression function g does not drift towards a constant. Alternatively, one consider a sequential approach to estimating g, namely one can first test whether the function g is constant, and then either fit the constant or apply the unconstrained estimator $\widehat{g}$ depending on the result of the test. However, it seems difficult to tune such a test to match the performance of the constrained estimator $\widehat{g}^c$ studied in this paper. □

Remark 10. Since the inversion of the operator T is a global inversion in the sense that the resulting estimators $\widehat{g}^c(x)$ and $\widehat{g}^u(x)$ depend not only on the shape of $g(x)$ locally at x, but on the shape of g over the whole domain, we do not expect convergence rate improvements from imposing monotonicity when the function g is partially flat. However, we leave the question about potential improvements from imposing monotonicity in this case for future research. □

Remark 11. The implementation of the constrained estimator in (12) is particularly simple when the basis vector $p(x)$ consists of polynomials or B-splines of order 2. In that case, $Dp(x)$ is linear in x and, therefore, the constraint $Dp(x)'b \geq 0$ for all $x \in [0,1]$ needs to be imposed only at the knots or endpoints of $[0,1]$, respectively. $\widehat{\beta}^c$ thus minimizes a quadratic objective function subject to a (finite-dimensional) linear inequality constraint. When the order of the polynomials or B-splines in $p(x)$ is larger than 2, then imposing the monotonicity constraint is slightly more complicated, but it can still be transformed into a finite-dimensional constraint using a representation of non-negative polynomials as a sum of squared polynomials:[3] one can represent any non-negative polynomial $f : \mathbb{R} \to \mathbb{R}$ as a sum of squares of polynomials (see the survey by Reznick (2000), for example), i.e. $f(x) = \widetilde{p}(x)'M\widetilde{p}(x)$ where $\widetilde{p}(x)$ is the vector of monomials up to some order and M a matrix of coefficients. Letting $f(x) = Dp(x)'b$, our monotonicity constraint becomes then equivalent to positive semi-definiteness of the matrix M, which depends on b. $\widehat{\beta}^c$ thus minimizes a quadratic objective function subject to a (finite-dimensional) semi-definiteness constraint.

For polynomials defined not over whole $\mathbb{R}$ but only over a compact sub-interval of $\mathbb{R}$, one can use the same reasoning as above together with a result attributed to M. Fekete (see Powers and Reznick (2000), for example): for any polynomial $f(x)$ with $f(x) \geq 0$ for $x \in [-1,1]$, there are polynomials $f_1(x)$ and $f_2(x)$, non-negative over whole $\mathbb{R}$, such that $f(x) = f_1(x) + (1-x^2)f_2(x)$. Letting again $f(x) = Dp(x)'b$, one can therefore impose

[3] We thank A. Belloni for pointing out this possibility.

our monotonicity constraint by imposing the positive semi-definiteness of the coefficients in the sums-of-squares representation of $f_1(x)$ and $f_2(x)$. □

Remark 12. In practice, the constraint $\|b\| \leq C_b$, or similar constraints in terms of Sobolev norms, which also impose bounds on derivatives of g, are often not enforced when computing an NPIV estimator. Horowitz (2012) and Horowitz and Lee (2012), for example, observe that imposing the constraint does not seem to have an effect in their simulations. On the other hand, especially when one includes many series terms in the computation of the estimator, Blundell, Chen, and Kristensen (2007) argue that penalizing the norm of g and of its derivatives may stabilize the estimator by reducing its variance. In this sense, penalizing the derivative of g may have a similar effect as imposing monotonicity. However, there are at least two important differences between penalization and imposing monotonicity. First, penalization is motivated by an assumption of the form $\|g\|_s < B$, where $\|\cdot\|_s$ is some Sobolev norm and B some finite constant. Because of the strict inequality, Horowitz (2012) argues that asymptotically the constraint is not binding and satisfied by the unconstrained estimator. The monotonicity constraint, on the other hand, is a weak inequality and may be binding in the limit. The second, perhaps more important difference between the two methods, is that penalization requires the choice of a tuning parameter that governs the strength of penalization. Economic theory does not provide any guidance on how to choose this parameter whereas imposing monotonicity does not require such choices and can often be motivated directly from economic theory arguments. □

4 Identification Bounds under Monotonicity

Point-identification of the function g in model (1) requires the linear operator T to be invertible. Completeness of the conditional distribution of X given W is known to be a sufficient condition for identification (Newey and Powell (2003)), but completeness has been argued to be a strong requirement (Santos (2012)) that cannot be tested (Canay, Santos, and Shaikh (2013)). In this section, we therefore explore the identification power of our monotonicity conditions, which appear natural in many economic applications, in the absence of completeness. Specifically, we derive informative bounds on the identified set of functions g satisfying (1). This means that, under our two monotonicity assumptions, the identified set is a proper subset of all monotone functions $g \in L^2[0,1]$.

Define the sign function

$$\text{sign}(w) := 1\{w > 0\} - 1\{w < 0\}, \qquad w \in \mathbb{R}.$$

We first show that if the function g is monotone (increasing or decreasing), the sign of its slope is identified under our conditions:

Theorem 3 (Identification of the sign of the slope)**.** *Let Assumptions 1 and 2 be satisfied. In addition, assume that the function g is monotone (increasing or decreasing) and continuously differentiable. Then $sign(Dg(x))$ is identified.*

This theorem is very useful in the sense that, under the regularity conditions of Assumption 2, monotone instruments and a monotone regression function suffice to identify the sign of the regression function's slope, even though the regression function itself is, in general, not point-identified. In many empirical applications it is natural to assume a monotone relationship between outcome variable Y and endogenous covariate X, given by the function g, but the main question of interest concerns not the exact shape of g itself, but whether the effect of X on Y, given by the slope of g, is positive, zero, or negative; see, for example, the discussion in Abrevaya, Hausman, and Khan (2010)). By Theorem 3, this question can be answered in large samples under our conditions.

Remark 13. In fact, Theorem 3 yields a surprisingly simple way to test the sign of the slope of the function g. Indeed, the proof of Theorem 3 reveals that g is increasing, constant, or decreasing if the function $w \mapsto \mathrm{E}[Y|W = w]$ is increasing, constant, or decreasing, respectively. By Chebyshev's association inequality (Lemma 6 in the appendix), the latter assertions are equivalent to the coefficient β in the linear regression model

$$Y = \alpha + \beta W + U, \quad E[UW] = 0 \tag{18}$$

being positive, zero, or negative since $sign(\beta) = sign(cov(W, Y))$ and

$$\begin{aligned} cov(W, Y) &= E[WY] - E[W]E[Y] \\ &= E[WE[Y|W]] - E[W]E[Y|W] = cov(W, E[Y|W]) \end{aligned}$$

by the law of iterated expectations. Therefore, under our conditions, hypotheses about the sign of the slope of the function g can be tested by testing corresponding hypotheses about the sign of the slope coefficient β in the linear regression model (18). □

It turns out that our two monotonicity assumptions possess identifying power beyond the slope of the regression function.

Definition 3 (Identified set)**.** *Let g satisfy (1). We say that two functions $g', g'' \in L^2[0, 1]$ are observationally equivalent if $E[g'(X) - g''(X)|W] = 0$. The identified set Θ is defined as the set of all functions $g' \in \mathcal{M}$ that are observationally equivalent to g.*

The following Theorem provides necessary conditions for observational equivalence.

Theorem 4 (Identification bounds)**.** *Let Assumptions 1 and 2 be satisfied, and let* $g', g'' \in L^2[0,1]$*. Further, let* $\bar{C} := C_1/c_p$ *where* $C_1 := (\widetilde{x}_2 - \widetilde{x}_1)^{1/2} \,/ \min\{\widetilde{x}_1 - x_1, x_2 - \widetilde{x}_2\}$ *and* $c_p := \min\{1-w_2, w_1\} \min\{C_F - 1, 2\} c_w c_f / 4$*. If there exists a function* $h \in L^2[0,1]$ *such that* $g' - g'' + h \in \mathcal{M}$ *and* $\|h\|_{2,t} + \bar{C}\|T\|_2\|h\|_2 < \|g' - g''\|_{2,t}$*, then* g' *and* g'' *are not observationally equivalent.*

Theorem 4 suggests the construction of bounds on the regression function as $\Theta' := \mathcal{M} \backslash \Delta$ with

$$\Delta := \Big\{ g' \in \mathcal{M} : \text{ there exists } h \in L^2[0,1] \text{ such that } g' - g + h \in \mathcal{M} \text{ and } \|h\|_{2,t} + \bar{C}\|T\|_2\|h\|_2 < \|g' - g\|_{2,t} \Big\}. \quad (19)$$

Then, under Assumptions 1, 2, and 3, the identified set Θ is contained in Θ'. Interestingly, Δ is not the empty set which means that our Assumptions 1, 2, and 3 possess identifying power leading to nontrivial bounds on g. Notice that the constant $\bar{C}$ depends only on the observable quantities c_w, c_f, and C_F from Assumptions 1–2, and on the known constants $\widetilde{x}_1$, $\widetilde{x}_2$, x_1, x_2, w_1, and w_2. Therefore, the set Θ' could, in principle, be estimated, but we leave estimation and inference on this set to future research.

It is possible to provide more insight into which functions are in Δ and thus not in Θ'. First, all functions in Θ' have to intersect, otherwise they are not observationally equivalent. Second, for a given $g' \in \mathcal{M}$ and $h \in L^2[0,1]$ such that $g' - g + h$ is monotone, the inequality in condition (19) is satisfied if $\|h\|_2$ is not too large relative to $\|g' - g\|_{2,t}$. In the extreme case, consider setting $h = 0$ to see that Θ' does not contain elements g' such that $g' - g$ is monotone. More generally, Θ' does not contain elements g' whose difference with g is too close to a monotone function. Therefore, functions g' that are much steeper than g are excluded from Θ'. Finally, since by Theorem 3 the sign of g is identified, the set Θ' can only contain increasing or decreasing functions, but not both.

5 Testing the Monotonicity Assumptions

In this section, we propose tests for our two monotonicity assumptions, the stochastic dominance condition (3) in our monotone IV Assumption 1 and the monotonicity of the regression function $g(x)$. Consider first testing (3), i.e. the null hypothesis

$$H_0 : F_{X|W}(x|w') \geq F_{X|W}(x|w'') \text{ for all } x, w', w'' \in (0,1) \text{ with } w' \leq w''$$

against the alternative,

$$H_a : F_{X|W}(x|w') < F_{X|W}(x|w'') \text{ for some } x, w', w'' \in (0,1) \text{ with } w' \leq w''$$

based on an i.i.d. sample (X_i, W_i), $i = 1, \ldots, n$, from the distribution of (X, W).

The null hypothesis, H_0, is equivalent to stochastic monotonicity of the conditional distribution function $F_{X|W}(x|w)$. Although there exist several good tests of H_0 in the literature (see Lee, Linton, and Whang (2009), Delgado and Escanciano (2012) and Lee, Song, and Whang (2014), for example), it is unknown how to construct such a test that obtains the optimal rate of consistency simultaneously over a reasonably large range of smoothness levels of $F_{X|W}(x|w)$. We solve this problem and develop an adaptive test that tunes to the smoothness level of $F_{X|W}(x|w)$, and has the optimal rate of consistency against the distributions in H_a with this smoothness level. Adaptiveness of our test is theoretically attractive but also important in practice: it delivers a data-driven choice of the smoothing parameter h_n (bandwidth value) of the test whereas nonadaptive tests are usually based on the assumption that $h_n \to 0$ with some rate in a range of prespecified rates, leaving the problem of selecting an appropriate value of h_n in a given data set to the researcher. We develop the critical value for the test that takes into account the data dependence induced by the data-driven choice of the smoothing parameter, leads to a test that controls size, and is asymptotically non-conservative.

Our test is based on the ideas in Chetverikov (2012) who in turn builds on the methods for adaptive specification testing in Horowitz and Spokoiny (2001) and on the theoretical results of high dimensional distributional approximations in Chernozhukov, Chetverikov, and Kato (2013c) (CCK). Note that $F_{X|W}(x|w) = \mathrm{E}[1\{X \leq x\}|W = w]$, so that for fixed $x \in (0, 1)$, the hypothesis that $F_{X|W}(x|w') \geq F_{X|W}(x, w'')$ for all $0 < w' \leq w'' \leq 1$ is equivalent to the hypothesis that the regression function $w \mapsto \mathrm{E}[1\{X \leq x\}|W = w]$ is decreasing. An adaptive test of this hypothesis was developed in Chetverikov (2012). In our case, H_0 requires that the regression function $w \mapsto \mathrm{E}[1\{X \leq x\}|W = w]$ is decreasing not only for a particular value $x \in (0, 1)$ but for all $x \in (0, 1)$, an extension of the results in Chetverikov (2012) that the remainder of this section develops.

Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function satisfying the following conditions:

Assumption 9 (Kernel)**.** *The kernel function $K : \mathbb{R} \to \mathbb{R}$ is such that (i) $K(w) > 0$ for all $w \in (-1, 1)$, (ii) $K(w) = 0$ for all $w \notin (-1, 1)$, (iii) K is continuous, and (iv) $\int_{-\infty}^{\infty} K(w)dw = 1$.*

We assume that the kernel function $K(w)$ has bounded support, is continuous, and is strictly positive on the support. The later condition excludes higher-order kernels. For a bandwidth value $h > 0$, define $K_h(w) := h^{-1}K(w/h)$.

Suppose H_0 is satisfied. Then, by the law of iterated expectations,

$$\mathrm{E}\left[(1\{X_i \leq x\} - 1\{X_j \leq x\})\mathrm{sign}(W_i - W_j)K_h(W_i - w)K_h(W_j - w)\right] \leq 0 \qquad (20)$$

for all $x, w \in (0,1)$ and $i, j = 1, \ldots, n$. Denoting

$$K_{ij,h}(w) := \text{sign}(W_i - W_j) K_h(W_i - w) K_h(W_j - w),$$

taking the sum of the left-hand side in (20) over $i, j = 1, \ldots, n$, and rearranging give

$$\mathrm{E}\left[\sum_{i=1}^{n} 1\{X_i \le x\} \sum_{j=1}^{n} (K_{ij,h}(w) - K_{ji,h}(w))\right] \le 0,$$

or, equivalently,

$$\mathrm{E}\left[\sum_{i=1}^{n} k_{i,h}(w) 1\{X_i \le x\}\right] \le 0, \tag{21}$$

where

$$k_{i,h}(w) := \sum_{j=1}^{n} (K_{ij,h}(w) - K_{ji,h}(w)).$$

To define the test statistic T, let $\mathcal{H}_n$ be a collection of bandwidth values satisfying the following conditions:

Assumption 10 (Bandwidth values)**.** *The collection of bandwidth values is* $\mathcal{H}_n := \{h \in \mathbb{R} : h = u^l/2, l = 0, 1, 2, \ldots, h \ge h_{\min}\}$ *for some* $u \in (0,1)$ *where* $h_{\min} := h_{\min,n}$ *is such that* $1/(nh_{\min}) \le C_h n^{-c_h}$ *for some constants* $c_h, C_h > 0$.

The collection of bandwidth values $\mathcal{H}_n$ is a geometric progression with the coefficient $u \in (0,1)$, the largest value $1/2$, and the smallest value converging to zero not too fast. As the sample size n increases, the collection of bandwidth values $\mathcal{H}_n$ expands.

Let $\mathcal{W}_n := \{W_1, \ldots, W_n\}$, and $\mathcal{X}_n := \{\epsilon + l(1-2\epsilon)/n : l = 0, 1, \ldots, n\}$ for some small $\epsilon > 0$. We define our test statistic by

$$T := \max_{(x,w,h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} \frac{\sum_{i=1}^{n} k_{i,h}(w) 1\{X_i \le x\}}{\left(\sum_{i=1}^{n} k_{i,h}(w)^2\right)^{1/2}}. \tag{22}$$

The statistic T is most closely related to that in Lee, Linton, and Whang (2009). The main difference is that we take the maximum with respect to the set of bandwidth values $h \in \mathcal{H}_n$ to achieve adaptiveness of the test.

We now discuss the construction of a critical value for the test. Suppose that we would like to have a test of level (approximately) α. As succinctly demonstrated by Lee, Linton, and Whang (2009), the derivation of the asymptotic distribution of T is complicated even when $\mathcal{H}_n$ is a singleton. Moreover, when $\mathcal{H}_n$ is not a singleton, it is generally unknown whether T converges to some nondegenerate asymptotic distribution after an appropriate normalization. We avoid these complications by employing the nonasymptotic approach developed in CCK and using a multiplier bootstrap critical value for the test.

Let $e_1, \ldots, e_n$ be an i.i.d. sequence of $N(0,1)$ random variables that are independent of the data. Also, let $\widehat{F}_{X|W}(x|w)$ be an estimator of $F_{X|W}(x|w)$ satisfying the following conditions:

Assumption 11 (Estimator of $F_{X|W}(x|w)$)**.** *The estimator $\widehat{F}_{X|W}(x|w)$ of $F_{X|W}(x|w)$ is such that (i)*

$$\mathrm{P}\left(\mathrm{P}\left(\max_{(x,w)\in\mathcal{X}_n\times\mathcal{W}_n}|\widehat{F}_{X|W}(x|w)-F_{X|W}(x|w)|>C_F n^{-c_F}|\{\mathcal{W}_n\}\right)>C_F n^{-c_F}\right)\leq C_F n^{-c_F}$$

for some constants $c_F, C_F > 0$, and (ii) $|\widehat{F}_{X|W}(x|w)| \leq C_F$ for all $(x,w) \in \mathcal{X}_n \times \mathcal{W}_n$.

This is a mild assumption implying uniform consistency of an estimator $\widehat{F}_{X|W}(x|w)$ of $F_{X|W}(x|w)$ over $(x,w) \in \mathcal{X}_n \times \mathcal{W}_n$. For completeness, the Supplemental Material provides sufficient conditions for Assumption 11 when $\widehat{F}_{X|W}(x|w)$ is a series estimator.

Define a bootstrap test statistic by

$$T^b := \max_{(x,w,h)\in\mathcal{X}_n\times\mathcal{W}_n\times\mathcal{H}_n}\frac{\sum_{i=1}^n e_i\left(k_{i,h}(w)(1\{X_i\leq x\}-\widehat{F}_{X|W}(x|W_i))\right)}{\left(\sum_{i=1}^n k_{i,h}(w)^2\right)^{1/2}}.$$

Then we define the critical value $c(\alpha)$ for the test as

$$c(\alpha) := (1-\alpha) \text{ conditional quantile of } T^b \text{ given the data.}$$

In the terminology of the moment inequalities literature, $c(\alpha)$ can be considered a "one-step" or "plug-in" critical value. Following Chetverikov (2012), we could also consider two-step or even multi-step (stepdown) critical values. For brevity of the paper, however, we do not consider these options here.

We reject H_0 if and only if $T > c(\alpha)$. To prove validity of this test, we assume that the conditional distribution function $F_{X|W}(x|w)$ satisfies the following condition:

Assumption 12 (Conditional Distribution Function $F_{X|W}(x|w)$)**.** *The conditional distribution function $F_{X|W}(x|w)$ is such that $c_\epsilon \leq F_{X|W}(\epsilon|w) \leq F_{X|W}(1-\epsilon|w) \leq C_\epsilon$ for all $w \in (0,1)$ and some constants $0 < c_\epsilon < C_\epsilon < 1$.*

The first theorem in this section shows that our test controls size asymptotically and is not conservative:

Theorem 5 (Polynomial Size Control)**.** *Let Assumptions 2, 9, 10, and 11 be satisfied. If H_0 holds, then*

$$\mathrm{P}\left(T > c(\alpha)\right) \leq \alpha + Cn^{-c}. \tag{23}$$

If the functions $w \mapsto F_{X|W}(x|w)$ are constant for all $x \in (0,1)$, then

$$|\mathrm{P}(T > c(\alpha)) - \alpha| \leq Cn^{-c}. \tag{24}$$

In both, (23) and (24), the constants c and C depend only on $c_W, C_W, c_h, C_h, c_F, C_F, c_\epsilon, C_\epsilon$, and the kernel K.

Remark 14 (Weak Condition on the Bandwidth Values)**.** Our theorem requires

$$\frac{1}{nh} \leq C_h n^{-c_h} \tag{25}$$

for all $h \in \mathcal{H}_n$, which is considerably weaker than the analogous condition in Lee, Linton, and Whang (2009) who require $1/(nh^3) \to 0$, up-to logs. This is achieved by using a conditional test and by applying the results of CCK. As follows from the proof of the theorem, the multiplier bootstrap distribution approximates the conditional distribution of the test statistic given $\mathcal{W}_n = \{W_1, \ldots, W_n\}$. Conditional on $\mathcal{W}_n$, the denominator in the definition of T is fixed, and does not require any approximation. Instead, we could try to approximate the denominator of T by its probability limit. This is done in Ghosal, Sen, and Vaart (2000) using the theory of Hoeffding projections but they require the condition $1/nh^2 \to 0$. Our weak condition (25) also crucially relies on the fact that we use the results of CCK. Indeed, it has already been demonstrated (see Chernozhukov, Chetverikov, and Kato (2013a,b), and Belloni, Chernozhukov, Chetverikov, and Kato (2014)) that, in typical nonparametric problems, the techniques of CCK often lead to weak conditions on the bandwidth value or the number of series terms. Our theorem is another instance of this fact. □

Remark 15 (Polynomial Size Control)**.** Note that, by (23) and (24), the probability of rejecting H_0 when H_0 is satisfied can exceed the nominal level α only by a term that is polynomially small in n. We refer to this phenomenon as a *polynomial size control*. As explained in Lee, Linton, and Whang (2009), when $\mathcal{H}_n$ is a singleton, convergence of T to the limit distribution is logarithmically slow. Therefore, Lee, Linton, and Whang (2009) used higher-order corrections derived in Piterbarg (1996) to obtain polynomial size control. Here we show that the multiplier bootstrap also gives higher-order corrections and leads to polynomial size control. This feature of our theorem is also inherited from the results of CCK. □

Remark 16 (Uniformity)**.** The constants c and C in (23) and (24) depend on the data generating process only via constants (and the kernel) appearing in Assumptions 2, 9, 10, and 11. Therefore, inequalities (23) and (24) hold uniformly over all data generating processes satisfying Assumptions 2, 9, 10, and 11 with the same constants. The issue of

uniformity has been studied intensively in the recent econometric literature and several techniques have been developed to prove uniformity (for instance, Mikusheva (2007) and Andrews and Guggenberger (2009)). Here we obtain uniformity of the result essentially for free since the distributional approximation theorems of CCK, which we employ, are nonasymptotic, and do not rely on convergence arguments. □

The final result of this section concerns the ability of our test to detect models in the alternative H_a. Let $\epsilon > 0$ be the constant appearing in the definition of T via the set $\mathcal{X}_n$.

Theorem 6 (Consistency)**.** *Let Assumptions 2, 9, 10, and 11 be satisfied and assume that $F_{X|W}(x|w)$ is continuously differentiable. If H_a holds with $D_w F_{X|W}(x|w) > 0$ for some $x \in (\epsilon, 1-\epsilon)$ and $w \in (0,1)$, then*

$$\mathrm{P}\left(T > c(\alpha)\right) \to 1 \text{ as } n \to \infty. \tag{26}$$

This theorem shows that our test is consistent against any model in H_a (with smooth $F_{X|W}(x|w)$) whose deviation from H_0 is not on the boundary, so that the deviation $D_w F_{X|W}(x|w) > 0$ occurs for $x \in (\epsilon, 1-\epsilon)$. It is also possible to extend our results and to show that Theorems 5 and 6 hold with $\epsilon = 0$ at the expense of additional technicalities.

We conclude this section by proposing a test of our second monotonicity assumption, i.e. testing the monotonicity of the regression function,

$$H_0 : g(x') \leq g(x'') \text{ for all } x', x'' \in (0,1) \text{ with } x' \leq x''$$

against

$$H_a : g(x') > g(x'') \text{ for some } x', x'' \in (0,1) \text{ with } x' \leq x''.$$

The discussion in Remark 13 revealed that, Assumptions 1 and 2, monotonicity of $g(x)$ implies monotonicity of $w \mapsto \mathrm{E}[Y|W=w]$. Therefore, under Assumptions 1 and 2, we can test H_0 by testing monotonicity of the conditional expectation $w \mapsto \mathrm{E}[Y|W=w]$ using existing tests such as Lee, Linton, and Whang (2009), Delgado and Escanciano (2012), Chetverikov (2012), and Lee, Song, and Whang (2014), among others. This procedure does not require solving the model for $g(x)$ and therefore avoids the ill-posedness of the problem.

6 Simulations

In this section, we study the finite sample behavior of our restricted estimator that imposes monotonicity and compare its performance to that of the unrestricted estimator. We

consider the NPIV model $Y = g(X)+\varepsilon$, $E[\varepsilon|W] = 0$, for two different regression functions, one that is strictly increasing and a weakly increasing one that is constant over part of its domain:

Model 1: $g(x) = \kappa \sin(\pi x - \pi/2)$

Model 2: $g(x) = 10\kappa\left[-(x-0.25)^2\mathbf{1}\{x \in [0, 0.25]\} + (x-0.75)^2\mathbf{1}\{x \in [0.75, 1]\}\right]$

where $\varepsilon = \kappa\sigma_\varepsilon\bar{\varepsilon}$ and $\bar{\varepsilon} = \eta\epsilon + \sqrt{1-\eta^2}\nu$. The regressors and instruments are generated by $X = \Phi(\xi)$ and $W = \Phi(\zeta)$, respectively, where Φ is the standard normal cdf and $\xi = \rho\zeta + \sqrt{1-\rho^2}\epsilon$. The errors are generated by $(\nu, \zeta, \epsilon) \sim N(0, I)$.

We vary the parameter κ in $\{1, 0.5, 0.1\}$ to study how the restricted and unrestricted estimators' performance compares depending on the maximum slope of the regression function. η governs the dependence of X on the regression error ε and ρ the strength of the first stage. All results are based on $1,000$ MC samples and the normalized B-spline basis for $p(x)$ and $q(w)$ of degree 3 and 4, respectively.

Tables 1–4 report the Monte Carlo approximations to the squared bias, variance, and mean squared error ("MSE") of the two estimators, each averaged over a grid on the interval $[0, 1]$. We also show the ratio of the restricted estimator's MSE divided by the unrestricted estimator's MSE. k_X and k_W denote, respectively, the number of knots used for the basis $p(x)$ and $q(w)$. The first two tables vary the number of knots, and the latter two the dependence parameters ρ and η. Different sample sizes and different values for ρ, η, and σ_ε yield qualitatively similar results. Figures 2 and 3 show the two estimators for a particular combination of the simulation parameters. The dashed lines represent confidence bands, computed as two times the (pointwise) empirical standard deviation of the estimators across simulation samples. Both, the restricted and the unrestricted, estimators are computed by ignoring the bound $\|b\| \leq C_b$ in their respective definitions. Horowitz and Lee (2012) and Horowitz (2012) also ignore the constraint $\|b\| \leq C_b$ and state that it does not affect the qualitative results of their simulation experiment.

The MSE of the restricted estimator (and, interestingly, also of the unrestricted estimator) decreases as the regression function becomes flatter. This observation is consistent with the risk bound in Theorem 2 depending positively on the maximum slope of g.

Because of the joint normality of (X, W), the simulation design is severely ill-posed and we expect high variability of both estimators. In all simulation scenarios, we do in fact observe a very large variance relative to bias. However, the magnitude of the variance differs significantly across the two estimators: in all scenarios, even in the design with a strictly increasing regression function, imposing the monotonicity constraint significantly reduces the variance of the NPIV estimator. The MSE of the restricted estimator is therefore much smaller than that of the unrestricted estimator, from about a factor of

two smaller when g is strictly increasing and the noise level is low ($\sigma_\varepsilon = 0.1$), to around 20 times smaller when g contains a flat part and the noise level is high ($\sigma_\varepsilon = 0.7$). Generally, the gains in MSE from imposing monotonicity are larger the higher the noise level σ_ε in the regression equation and the higher the first-stage correlation ρ.[4]

7 Gasoline Demand in the United States

In this section, we revisit the problem of estimating demand functions for gasoline in the United States. Because of the dramatic changes in the oil price over the last few decades, understanding the elasticity of gasoline demand is fundamental to evaluating tax policies. Consider the following partially linear specification of the demand function:

$$Y = g(X, Z_1) + \gamma' Z_2 + \varepsilon, \qquad E[\varepsilon | W, Z_1, Z_2] = 0,$$

where Y denotes annual log-gasoline consumption of a household, X log-price of gasoline (average local price), Z_1 log-household income, Z_2 are control variables (such as population density, urbanization, and demographics), and W distance to major oil platform. We allow for price X to be endogenous, but assume that (Z_1, Z_2) is exogenous. W serves as an instrument for price by capturing transport cost and, therefore, shifting the cost of gasoline production. We use the same sample of size $4,812$ from the 2001 National Household Travel Survey and the same control variables Z_2 as Blundell, Horowitz, and Parey (2012). More details can be found in their paper.

Moving away from constant price and income elasticities is likely very important as individuals' responses to price changes vary greatly with price and income level. Since economic theory does not provide guidance on the functional form of g, finding an appropriate parametrization is difficult. Hausman and Newey (1995) and Blundell, Horowitz, and Parey (2012), for example, demonstrate the importance of employing flexible estimators of g that do not suffer from misspecification bias due to arbitrary restrictions in the model. Blundell, Horowitz, and Parey (2013) argue that prices at the local market level vary for several reasons and that they may reflect preferences of the consumers in the local market. Therefore, one would expect prices X to depend on unobserved factors in ε that determine consumption, rendering price an endogenous variable. Furthermore, the theory of the consumer requires downward-sloping compensated demand curves. Assuming a positive income derivative[5] $\partial g / \partial z_1$, the Slutsky condition implies that the uncompensated

[4] Since Tables 1 and 2 report results for the lower level of ρ, and Tables 3 and 4 results for the lower noise level σ_ε, we consider the selection of results as, if at all, favoring the unrestricted estimator.

[5] Blundell, Horowitz, and Parey (2012) estimate this income derivative and do, in fact, find it to be positive over the price range of interest.

(Marshallian) demand curves are also downward-sloping, i.e. $g(\cdot, z_1)$ should be monotone for any z_1, as long as income effects do not completely offset price effects. Finally, we expect the cost shifter W to monotonically increase cost of producing gasoline and thus satisfy our monotone IV condition. In conclusion, our restricted NPIV estimator appears to be an attractive estimator of demand functions in this setting.

We consider three benchmark estimators. First, we compute the unrestricted nonparametric ("unrestr. NP") series estimator of the regression of Y on X and Z_1, treating price as exogenous. As in Blundell, Horowitz, and Parey (2012), we accommodate the high-dimensional vector of additional, exogenous covariates Z_2 by (i) estimating γ by Robinson (1988)'s procedure, (ii) then removing these covariates from the outcome, and (iii) estimating g by regressing the adjusted outcomes on X and Z_1. The second benchmark estimator ("restr. NP") repeats the same steps (i)–(iii) except that it imposes monotonicity (in price) of g in steps (i) and (iii). The third benchmark estimator is the unrestricted NPIV estimator ("unrestr. NPIV") that accounts for the covariates Z_2 in similar fashion as the first, unrestricted nonparametric estimator, except that (i) and (iii) employ NPIV estimators that impose additive separability and linearity in Z_2.

The fourth estimator we consider is the restricted NPIV estimator ("restr. NPIV") that we compare to the three benchmark estimators. We allow for the presence of the covariates Z_2 in the same fashion as the unrestricted NPIV estimator except that, in steps (i) and (iii), we impose monotonicity in price.

We report results for the following choice of bases. All estimators employ a quadratic B-spline basis with 3 knots for price X and a cubic B-spline with 10 knots for the instrument W. Denote these two bases by $\mathbf{P}$ and $\mathbf{Q}$, using the same notation as in Section 3. In step (i), the NPIV estimators include the additional exogenous covariates (Z_1, Z_2) in the respective bases for X and W, so they use the estimator defined in Section 3 except that the bases $\mathbf{P}$ and $\mathbf{Q}$ are replaced by $\widetilde{\mathbf{P}} := [\mathbf{P}, \mathbf{P} \times \mathbf{Z}_1, \mathbf{Z}_2]$ and $\widetilde{\mathbf{Q}} := [\mathbf{Q}, \mathbf{Q} \times (\mathbf{Z}_1, \mathbf{Z}_2)]$, respectively, where $\mathbf{Z}_k := (Z_{k,1}, \ldots, Z_{k,n})'$, $k = 1, 2$, stacks the observations $i = 1, \ldots, n$ and $\mathbf{P} \times \mathbf{Z}_1$ denotes the tensor product of the columns of the two matrices. Since, in the basis $\widetilde{\mathbf{P}}$, we include interactions of $\mathbf{P}$ with $\mathbf{Z}_1$, but not with $\mathbf{Z}_2$, the resulting estimator allows for a nonlinear, nonseparable dependence of Y on X and Z_1, but imposes additive separability in Z_2. The conditional expectation of Y given W, Z_1, and Z_2 does not have to be additively separable in Z_2, so that, in the basis $\widetilde{\mathbf{Q}}$, we include interactions of $\mathbf{Q}$ with both $\mathbf{Z}_1$ and $\mathbf{Z}_2$.[6]

We estimated the demand functions for many different combinations of the order of B-spline for W, the number of knots in both bases, and even with various penalization terms

[6]Notice that $\mathbf{P}$ and $\mathbf{Q}$ include constant terms so it is not necessary to separately include $\mathbf{Z}_k$ in addition to its interactions with $\mathbf{P}$ and $\mathbf{Q}$, respectively.

(as discussed in Remark 12). While the shape of the unconstrained NPIV estimate varied slightly across these different choices of tuning parameters (mostly near the boundary of the support of X), the constrained NPIV estimator did not exhibit any visible changes at all.

Figure 4 shows a nonparametric kernel estimate of the conditional distribution of the price X given the instrument W. Overall the graph indicates an increasing relationship between the two variables as required by our stochastic dominance condition (3). We formally test this monotone IV assumption by applying our new test proposed in Section 5. We find a test statistic value of 0.139 and 95%-critical value of 1.720.[7] Therefore, we fail to reject the monotone IV assumption.

Figure 5 shows the estimates of the demand function at three income levels, at the lower quartile ($\$42,500$), the median ($\$57,500$), and the upper quartile ($\$72,500$). The area shaded in grey represents the 90% uniform confidence bands around the unrestricted NPIV estimator as proposed in Horowitz and Lee (2012).[8] The black lines correspond to the estimators assuming exogeneity of price and the red lines to the NPIV estimators that allow for endogeneity of price. The dashed black line shows the kernel estimate of Blundell, Horowitz, and Parey (2012) and the solid black line the corresponding series estimator that imposes monotonicity. The dashed and solid red lines similarly depict the unrestricted and restricted NPIV estimators, respectively.

All estimates show an overall decreasing pattern of the demand curves, but the two unconstrained estimators are both increasing over some parts of the price domain. We view these implausible increasing parts as finite sample phenomena that arise because the unconstrained nonparametric estimators are too imprecise. The wide confidence bands of the unconstrained NPIV estimator are consistent with this view. Hausman and Newey (1995) and Horowitz and Lee (2012) find similar anomalies in their nonparametric estimates, assuming exogenous prices. Unlike the unconstrained estimates, our constrained NPIV estimates are downward-sloping everywhere and smoother. They lie within the 90% uniform confidence bands of the unconstrained estimator so that the monotonicity constraint appears compatible with the data.

The two restricted estimates are very similar, indicating that endogeneity of prices may not be important in this problem, but they are both significantly flatter than the unrestricted estimates across all three income groups, which implies that households appear to be less sensitive to price changes that the unconstrained estimates suggest. The small

[7] The critical value is computed from $1,000$ bootstrap samples, using the bandwidth set $\mathcal{H}_n = \{2, 1, 0.5, 0.25, 0.125, 0.0625\}$, and a kernel estimator for $\widehat{F}_{X|W}$ with bandwidth 0.3 which produces the estimate in Figure 4.

[8] Critical values are computed from $1,000$ bootstrap samples and the bands are computed on a grid of 100 equally-spaced points in the support of the data for X.

maximum slope of the constrained NPIV estimator also suggests that the risk bound in Theorem 2 may be small and therefore we expect the constrained NPIV estimate to be precise for this data set.

8 Conclusions

In this paper, we provide a theoretical explanation for the dramatic gains in finite sample performance that are possible when imposing monotonicity in the NPIV estimation procedure. In particular, we show that monotone instruments together with a monotone regression function lead to a so-called locally quantitatively well-posed problem. This feature of the restricted problem significantly reduces the statistical difficulty in nonparametric estimation of the regression function. We show that the restricted NPIV estimator may possess finite-sample risk much lower than the unrestricted estimator, especially when the regression function is not too steep and the unrestricted estimator exhibits high variability. In fact, the constrained estimator's risk may be comparable to that of standard conditional mean estimators.

A Proofs for Section 2

For any $h \in L^1[0,1]$, let $\|h\|_1 := \int_0^1 |h(x)|dx$, $\|h\|_{1,t} := \int_{x_1}^{x_2} |h(x)|dx$ and define the operator norm by $\|T\|_2 := \sup_{h \in L^2[0,1]: \|h\|_2 > 0} \|Th\|_2 / \|h\|_2$. Note that $\|T\|_2 \leq \int_0^1 \int_0^1 f_{X,W}^2(x,w)dxdw$, and so under Assumption 2, $\|T\|_2 \leq C_T$.

Proof of Theorem 1. We first show that for any $h \in \mathcal{M}$,

$$\|h\|_{2,t} \leq C_1 \|h\|_{1,t} \tag{27}$$

for $C_1 := (\widetilde{x}_2 - \widetilde{x}_1)^{1/2} / \min\{\widetilde{x}_1 - x_1, x_2 - \widetilde{x}_2\}$. Indeed, by monotonicity of h,

$$\begin{aligned}
\|h\|_{2,t} &= \left(\int_{\widetilde{x}_1}^{\widetilde{x}_2} h(x)^2 dx \right)^{1/2} \leq \sqrt{\widetilde{x}_2 - \widetilde{x}_1} \max\{|h(\widetilde{x}_1)|, |h(\widetilde{x}_2)|\} \\
&\leq \sqrt{\widetilde{x}_2 - \widetilde{x}_1} \, \frac{\int_{x_1}^{x_2} |h(x)|dx}{\min\{\widetilde{x}_1 - x_1, x_2 - \widetilde{x}_2\}}
\end{aligned}$$

so that (27) follows. Therefore, for any increasing continuously differentiable $h \in \mathcal{M}$,

$$\|h\|_{2,t} \leq C_1 \|h\|_{1,t} \leq C_1 C_2 \|Th\|_1 \leq C_1 C_2 \|Th\|_2,$$

where the first inequality follows from (27), the second from Lemma 2 below (which is the main step in the proof of the theorem), and the third by Jensen's inequality. Hence, conclusion (6) of Theorem 1 holds for increasing continuously differentiable $h \in \mathcal{M}$ with $\bar{C} := C_1 C_2$ and C_2 as defined in Lemma 2.

Next, for any increasing function $h \in \mathcal{M}$, it follows from Lemma 10 that one can find a sequence of increasing continuously differentiable functions $h_k \in \mathcal{M}$, $k \geq 1$, such that $\|h_k - h\|_2 \to 0$ as $k \to \infty$. Therefore, by the triangle inequality,

$$\begin{aligned}
\|h\|_{2,t} &\leq \|h_k\|_{2,t} + \|h_k - h\|_{2,t} \leq \bar{C}\|Th_k\|_2 + \|h_k - h\|_{2,t} \\
&\leq \bar{C}\|Th\|_2 + \bar{C}\|T(h_k - h)\|_2 + \|h_k - h\|_{2,t} \\
&\leq \bar{C}\|Th\|_2 + \bar{C}\|T\|_2\|(h_k - h)\|_2 + \|h_k - h\|_{2,t} \\
&\leq \bar{C}\|Th\|_2 + (\bar{C}\|T\|_2 + 1)\|(h_k - h)\|_2 \\
&\leq \bar{C}\|Th\|_2 + (\bar{C}C_T + 1)\|h_k - h\|_2
\end{aligned}$$

where the third line follows from the Cauchy-Schwarz inequality, the fourth from $\|h_k - h\|_{2,t} \leq \|h_k - h\|_2$, and the fifth from Assumption 2(i). Taking the limit as $k \to \infty$ of both the left-hand and the right-hand sides of this chain of inequalities yields conclusion (6) of Theorem 1 for all increasing $h \in \mathcal{M}$.

Finally, since for any decreasing $h \in \mathcal{M}$, we have that $-h \in \mathcal{M}$ is increasing, $\|-h\|_{2,t} = \|h\|_{2,t}$ and $\|Th\|_2 = \|T(-h)\|_2$, conclusion (6) of Theorem 1 also holds for all decreasing $h \in \mathcal{M}$, and thus for all $h \in \mathcal{M}$. This completes the proof of the theorem. Q.E.D.

Lemma 2. *Let Assumptions 1 and 2 hold. Then for any increasing continuously differentiable $h \in L^1[0,1]$,*

$$\|h\|_{1,t} \leq C_2 \|Th\|_1$$

where $C_2 := 1/c_p$ and $c_p := c_w c_f / 2 \min\{1 - w_2, w_1\} \min\{(C_F - 1)/2, 1\}$.

Proof. Take any increasing continuously differentiable function $h \in L^1[0,1]$ such that $\|h\|_{1,t} = 1$. Define $M(w) := E[h(X)|W = w]$ for all $w \in [0,1]$ and note that

$$\|Th\|_1 = \int_0^1 |M(w) f_W(w)| dw \geq c_W \int_0^1 |M(w)| dw$$

where the inequality follows from Assumption 2(iii). Therefore, the asserted claim follows if we can show that $\int_0^1 |M(w)| dw$ is bounded away from zero by a constant that depends only on ζ.

First, note that $M(w)$ is increasing. This is because, by integration by parts,

$$M(w) = \int_0^1 h(x) f_{X|W}(x|w) dx = h(1) - \int_0^1 Dh(x) F_{X|W}(x|w) dx,$$

so that condition (3) of Assumption 1 and $Dh(x) \geq 0$ for all x imply that the function $M(w)$ is increasing.

Consider the case in which $h(x) \geq 0$ for all $x \in [0,1]$. Then $M(w) \geq 0$ for all $w \in [0,1]$. Therefore,

$$\begin{aligned}
\int_0^1 |M(w)| dw &\geq \int_{w_2}^1 |M(w)| dw \geq (1 - w_2) M(w_2) = (1 - w_2) \int_0^1 h(x) f_{X|W}(x|w_2) dx \\
&\geq (1 - w_2) \int_{x_1}^{x_2} h(x) f_{X|W}(x|w_2) dx \geq (1 - w_2) c_f \int_{x_1}^{x_2} h(x) dx \\
&= (1 - w_2) c_f \|h\|_{1,t} = (1 - w_2) c_f > 0
\end{aligned}$$

by Assumption 2(ii). Similarly,

$$\int_0^1 |M(w)| dw \geq w_1 c_f > 0$$

when $h(x) \leq 0$ for all $x \in [0,1]$. Therefore, it remains to consider the case in which there exists $x^* \in (0,1)$ such that $h(x) \leq 0$ for $x \leq x^*$ and $h(x) \geq 0$ for $x > x^*$. Since $h(x)$ is continuous, $h(x^*) = 0$, and so integration by parts yields

$$\begin{aligned}
M(w) &= \int_0^{x^*} h(x) f_{X|W}(x|w) dx + \int_{x^*}^1 h(x) f_{X|W}(x|w) dx \\
&= -\int_0^{x^*} Dh(x) F_{X|W}(x|w) dx + \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w)) dx. \qquad (28)
\end{aligned}$$

For $k = 1, 2$, let $A_k := \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_k))$ and $B_k := \int_0^{x^*} Dh(x)F_{X|W}(x|w_k)dx$, so that $M(w_k) = A_k - B_k$.

Consider the following three cases separately, depending on where x^* lies relative to x_1 and x_2.

Case I ($x_1 < x^* < x_2$): First, we have

$$\begin{aligned}
A_1 + B_2 &= \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx + \int_0^{x^*} Dh(x)F_{X|W}(x|w_2)dx \\
&= \int_{x^*}^1 h(x)f_{X|W}(x|w_1)dx - \int_0^{x^*} h(x)f_{X|W}(x|w_2)dx \\
&\geq \int_{x^*}^{x_2} h(x)f_{X|W}(x|w_1)dx - \int_{x_1}^{x^*} h(x)f_{X|W}(x|w_2)dx \\
&\geq c_1 \int_{x^*}^{x_2} h(x)dx + c_f \int_{x_1}^{x^*} |h(x)|dx = c_f \int_{x_1}^{x_2} |h(x)|dx \\
&= c_f \|h\|_{1,t} = c_f > 0
\end{aligned} \tag{29}$$

where the fourth line follows from Assumption 2(ii). Second, by (3) and (4) of Assumption 1,

$$\begin{aligned}
M(w_1) &= \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx - \int_0^{x^*} Dh(x)F_{X|W}(x|w_1)dx \\
&\leq \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_2))dx - C_F \int_0^{x^*} Dh(x)F_{X|W}(x|w_2)dx \\
&= A_2 - C_F B_2
\end{aligned}$$

so that, together with $M(w_2) = A_2 - B_2$, we obtain

$$M(w_2) - M(w_1) \geq (C_F - 1)B_2. \tag{30}$$

Similarly, by (3) and (5) of Assumption 1,

$$\begin{aligned}
M(w_2) &= \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_2))dx - \int_0^{x^*} Dh(x)F_{X|W}(x|w_2)dx \\
&\geq C_F \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx - \int_0^{x^*} Dh(x)F_{X|W}(x|w_1)dx \\
&= C_F A_1 - B_1
\end{aligned}$$

so that, together with $M(w_1) = A_1 - B_1$, we obtain

$$M(w_2) - M(w_1) \geq (C_F - 1)A_1. \tag{31}$$

In conclusion, equations (29), (30), and (31) yield

$$M(w_2) - M(w_1) \geq (C_F - 1)(A_1 + B_2)/2 \geq (C_F - 1)c_f/2 > 0. \tag{32}$$

Consider the case $M(w_1) \geq 0$ and $M(w_2) \geq 0$. Then $M(w_2) \geq M(w_2) - M(w_1)$ and thus

$$\int_0^1 |M(w)|dw \geq \int_{w_2}^1 |M(w)|dw \geq (1 - w_2)M(w_2) \geq (1 - w_2)(C_F - 1)c_f/2 > 0. \tag{33}$$

Similarly,

$$\int_0^1 |M(w)|dw \geq w_1(C_F - 1)c_f/2 > 0 \tag{34}$$

when $M(w_1) \leq 0$ and $M(w_2) \leq 0$.

Finally, consider the case $M(w_1) \leq 0$ and $M(w_2) \geq 0$. If $M(w_2) \geq |M(w_1)|$, then $M(w_2) \geq (M(w_2) - M(w_1))/2$ and the same argument as in (33) shows that

$$\int_0^1 |M(w)|dw \geq (1 - w_2)(C_F - 1)c_f/4.$$

If $|M(w_1)| \geq M(w_2)$, then $|M(w_1)| \geq (M(w_2) - M(w_1))/2$ and we obtain

$$\int_0^1 |M(w)|dw \geq \int_0^{w_1} |M(w)|dw \geq w_1(C_F - 1)c_f/4 > 0.$$

This completes the proof of Case I.

Case II ($x_2 \leq x^*$): Suppose $M(w_1) \geq -c_f/2$. As in Case I, we have $M(w_2) \geq C_F A_1 - B_1$. Together with $M(w_1) = A_1 - B_1$, this inequality yields

$$\begin{aligned}
M(w_2) - M(w_1) &= M(w_2) - C_F M(w_1) + C_F M(w_1) - M(w_1) \\
&\geq (C_F - 1)B_1 + (C_F - 1)M(w_1) \\
&= (C_F - 1)\left(\int_0^{x^*} Dh(x)F_{X|W}(x|w_1)dx + M(w_1)\right) \\
&= (C_F - 1)\left(\int_0^{x^*} |h(x)|f_{X|W}(x|w_1)dx + M(w_1)\right) \\
&\geq (C_F - 1)\left(\int_{x_1}^{x_2} |h(x)|f_{X|W}(x|w_1)dx - \frac{c_f}{2}\right) \\
&\geq (C_F - 1)\left(c_f \int_{x_1}^{x_2} |h(x)|dx - \frac{c_f}{2}\right) = \frac{(C_F - 1)c_f}{2} > 0
\end{aligned}$$

With this inequality we proceed as in Case I to show that $\int_0^1 |M(w)|dw$ is bounded from below by a positive constant that depends only on ζ. On the other hand, when $M(w_1) \leq -c_f/2$ we bound $\int_0^1 |M(w)|dw$ as in (34), and the proof of Case II is complete.

Case III ($x^* \leq x_1$): Similarly as in Case II, suppose first that $M(w_2) \leq c_f/2$. As in Case I we have $M(w_1) \leq A_2 - C_F B_2$ so that together with $M(w_2) = A_2 - B_2$,

$$\begin{aligned} M(w_2) - M(w_1) &= M(w_2) - C_F M(w_2) + C_F M(w_2) - M(w_1) \\ &\geq (1 - C_F) M(w_2) + (C_F - 1) A_2 \\ &= (C_F - 1) \left(\int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_2))dx - M(w_2) \right) \\ &= (C_F - 1) \left(\int_{x^*}^1 h(x) f_{X|W}(x|w_2)dx - M(w_2) \right) \\ &\geq (C_F - 1) \left(\int_{x_1}^{x_2} h(x) f_{X|W}(x|w_2)dx - M(w_2) \right) \\ &\geq (C_F - 1) \left(c_f \int_{x_1}^{x_2} h(x)dx - \frac{c_f}{2} \right) = \frac{(C_F - 1)c_f}{2} > 0 \end{aligned}$$

and we proceed as in Case I to bound $\int_0^1 |M(w)|dw$ from below by a positive constant that depends only on ζ. On the other hand, when $M(w_2) > c_f/2$, we bound $\int_0^1 |M(w)|dw$ as in (33), and the proof of Case III is complete. The lemma is proven. Q.E.D.

Proof of Corollary 1. Let h_k be a sequence in $T(\mathcal{M})$ and $h \in T(\mathcal{M})$ such that $\|h_k - h\|_2 \to 0$ as $k \to \infty$. Define $g_k := T^{-1}h_k$ and $g := T^{-1}h$. We want to show that $\|g_k - g\|_{2,t} \to 0$ as $k \to \infty$. To this end, we have

$$\sup_{x \in [\widetilde{x}_1, \widetilde{x}_2]} |g_k(x)| \leq C\|g_k\|_{2,t} \leq C\bar{C}\|Tg_k\|_2 = C\bar{C}\|h_k\|_2 \to C\bar{C}\|h\|_2$$

as $k \to \infty$, where the first inequality follows for some $C > 0$ depending only on $x_1, x_2, \widetilde{x}_1, \widetilde{x}_2$ by an argument similar to that used in the proof of Theorem 1 since g_k is monotone, and the second inequality follows from the statement of Theorem 1. Therefore, there exists some k_0 such that for all $k \geq k_0$, $\sup_{x \in [\widetilde{x}_1, \widetilde{x}_2]} |g_k(x)| \leq C\bar{C}(\|h\|_2 + 1) < \infty$. This means that for all $k \geq k_0$, the functions g_k belong to the space of monotone functions in $L^2[0,1]$ that are uniformly bounded by $CC_2(\|h\|_2 + 1) < \infty$ over the interval $[\widetilde{x}_1, \widetilde{x}_2]$. Since this space is compact under the norm $\|\cdot\|_{2,t}$ (see, for example, discussion on p. 18 in van de Geer (2000)), it follows from Lemma 5 that $\|g_k - g\|_{2,t} \to 0$ as $k \to \infty$ as desired. Q.E.D.

Proof of Corollary 2. Observe that the operator T has a bounded inverse on $T(\mathcal{M})$: for any function $m \in L^2[0,1]$ such that $m = Th$ for some $h \in \mathcal{M}$,

$$\|T^{-1}m\|_{2,t} = \|h\|_{2,t} \leq \bar{C}\|Th\|_2 = \bar{C}\|m\|_2 \tag{35}$$

by Theorem 1. Now, let g be a constant function, and let $m = Tg$. Then for any function m' such that $m' = Tg'$ for some monotone $g' \in \mathcal{M}$, $g' - g = T^{-1}(m' - m)$ and $g' - g \in \mathcal{M}$. Therefore,

$$\|g' - g\|_{2,t} \leq \bar{C}\|m' - m\|_2$$

by (35). The asserted claim follows. Q.E.D.

Proof of Corollary 3. Note that since $\tau(a') \leq \tau(a'')$ whenever $a' \leq a''$, the claim for $a \leq 0$, follows from $\tau(a) \leq \tau(0) \leq \bar{C}$, where the second inequality holds by Theorem 1. Therefore, assume that $a > 0$. Fix any $\alpha \in (0,1)$. Take any function $h \in \mathcal{H}(a)$ such that $\|h\|_{2,t} = 1$. Set $h'(x) = ax$ for all $x \in [0,1]$. Note that the function $x \mapsto h(x) + ax$ is increasing and so belongs to the class $\mathcal{M}$. Also, $\|h'\|_{2,t} \leq \|h'\|_2 \leq a/\sqrt{3}$. Thus, the bound (36) in Lemma 3 below applies whenever $(1 + \bar{C}\|T\|_2)a/\sqrt{3} \leq \alpha$. Therefore, for all a satisfying the inequality

$$a \leq \frac{\sqrt{3}\alpha}{1 + \bar{C}\|T\|_2},$$

we have $\tau(a) \leq \bar{C}/(1-\alpha)$. This completes the proof of the corollary. Q.E.D.

Lemma 3. *Let Assumptions 1 and 2 be satisfied. Consider any function $h \in L^2[0,1]$. If there exist $h' \in L^2[0,1]$ and $\alpha \in (0,1)$ such that $h + h' \in \mathcal{M}$ and $\|h'\|_{2,t} + \bar{C}\|T\|_2\|h'\|_2 \leq (<)\alpha\|h\|_{2,t}$, then*

$$\|h\|_{2,t} \leq (<)\frac{\bar{C}}{1-\alpha}\|Th\|_2 \tag{36}$$

for the constant $\bar{C}$ defined in Theorem 1.

Proof. Define

$$\widetilde{h}(x) := \frac{h(x) + h'(x)}{\|h\|_{2,t} - \|h'\|_{2,t}}, \quad x \in [0,1].$$

By assumption, $\|h'\|_{2,t} < \|h\|_{2,t}$, and so the triangle inequality yields

$$\|\widetilde{h}\|_{2,t} \geq \frac{\|h\|_{2,t} - \|h'\|_{2,t}}{\|h\|_{2,t} - \|h'\|_{2,t}} = 1.$$

Therefore, since $\widetilde{h} \in \mathcal{M}$, Theorem 1 gives

$$\|T\widetilde{h}\|_2 \geq \|\widetilde{h}\|_{2,t}/\bar{C} \geq 1/\bar{C}.$$

Hence, applying the triangle inequality once again yields

$$\begin{aligned}\|Th\|_2 &\geq (\|h\|_{2,t} - \|h'\|_{2,t})\|T\widetilde{h}\|_2 - \|Th'\|_2 \geq (\|h\|_{2,t} - \|h'\|_{2,t})\|T\widetilde{h}\|_2 - \|T\|_2\|h'\|_2 \\ &\geq \frac{\|h\|_{2,t} - \|h'\|_{2,t}}{\bar{C}} - \|T\|_2\|h'\|_2 = \frac{\|h\|_{2,t}}{\bar{C}}\left(1 - \frac{\|h'\|_{2,t} + \bar{C}\|T\|_2\|h'\|_2}{\|h\|_{2,t}}\right)\end{aligned}$$

Since the expression in the last parentheses is bounded from below (weakly or strictly) by $1 - \alpha$ by assumption, we obtain the inequality

$$\|Th\|_2 \geq (>)\frac{1-\alpha}{\bar{C}}\|h\|_{2,t},$$

which is equivalent to (36). Q.E.D.

B Proofs for Section 3

In this section, we use C to denote a strictly positive constant, which value may change from place to place.

Proof of Lemma 1. Observe that if $D\widehat{g}^u(x) \geq 0$ for all $x \in [0,1]$, then $\widehat{g}^c$ coincides with $\widehat{g}^u$, so that to prove (13), it suffices to prove that

$$P\Big(D\widehat{g}^u(x) \geq 0 \text{ for all } x \in [0,1]\Big) \to 1 \text{ as } n \to \infty. \tag{37}$$

In turn, (37) follows if

$$\sup_{x\in[0,1]} |D\widehat{g}^u(x) - Dg(x)| = o_P(1) \tag{38}$$

since $Dg(x) \geq c_g$ for all $x \in [0,1]$ and some $c_g > 0$.

To prove (38), define a function $\widehat{m} \in L^2[0,1]$ by

$$\widehat{m}(w) = q(w)'E_n[q(W_i)Y_i], \qquad w \in [0,1], \tag{39}$$

and an operator $\widehat{T} : L^2[0,1] \to L^2[0,1]$ by

$$(\widehat{T}h)(w) = q(w)'E_n[q(W_i)p(X_i)']\mathrm{E}[p(U)h(U)], \quad w \in [0,1],\ h \in L^2[0,1].$$

Throughout the proof, we assume that the following events hold:

$$\|E_n[q(W_i)p(X_i)'] - E[q(W)p(X)']\| \leq C(\xi_n^2 \log n/n)^{1/2}, \tag{40}$$
$$\|E_n[q(W_i)q(W_i)'] - E[q(W)q(W)']\| \leq C(\xi_n^2 \log n/n)^{1/2}, \tag{41}$$
$$\|\widehat{m} - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1}J^{-s}) \tag{42}$$

for some sufficiently large constant $0 < C < \infty$. It follows from Lemmas 4 and 11 that all three events hold jointly with probability at least $1 - \alpha - n^{-1}$.

Next, we derive a bound on $\|\widehat{g}^u - g_n\|_2$. By the definition of τ_n,

$$\begin{aligned}\|\widehat{g}^u - g_n\|_2 &\leq \tau_n\|T(\widehat{g}^u - g_n)\|_2 \\ &\leq \tau_n\|T(\widehat{g}^u - g)\|_2 + \tau_n\|T(g - g_n)\|_2 \leq \tau_n\|T(\widehat{g}^u - g)\|_2 + C_gK^{-s}\end{aligned}$$

where the second inequality follows from the triangle inequality, and the third inequality from Assumption 6(iii). Next, since $m = Tg$,

$$\|T(\widehat{g}^u - g)\|_2 \leq \|(T - T_n)\widehat{g}^u\|_2 + \|(T_n - \widehat{T})\widehat{g}^u\|_2 + \|\widehat{T}\widehat{g}^u - \widehat{m}\|_2 + \|\widehat{m} - m\|_2$$

by the triangle inequality. The bound on $\|\widehat{m} - m\|_2$ is given in (42). Also, since $\|\widehat{g}^u\|_2 \leq C_b$ by construction,

$$\|(T - T_n)\widehat{g}^u\|_2 \leq C_bC_a\tau_n^{-1}K^{-s}$$

by Assumption 8(ii). In addition, by the triangle inequality,

$$\begin{aligned}\|(T_n - \widehat{T})\widehat{g}^u\|_2 &\leq \|(T_n - \widehat{T})(\widehat{g}^u - g_n)\|_2 + \|(T_n - \widehat{T})g_n\|_2 \\ &\leq \|T_n - \widehat{T}\|_2\|\widehat{g}^u - g_n\|_2 + \|(T_n - \widehat{T})g_n\|_2.\end{aligned}$$

Moreover,

$$\|T_n - \widehat{T}\|_2 = \|E_n[q(W_i)p(X_i)'] - E[q(W)p(X)']\| \leq C(\xi_n^2 \log n/n)^{1/2}$$

by (40), and as in the proof of Lemma 4,

$$\|(T_n - \widehat{T})g_n\|_2 = \|E_n[q(W_i)g_n(X_i)] - E[q(W)g_n(X)]\| \leq C(J/(\alpha n))^{1/2}$$

with probability at least $1 - \alpha$.

Further, by Assumption 2(iii), all eigenvalues of $E[q(W)q(W)']$ are bounded from below by c_w and from above by C_w, and so it follows from (41) that for large n, all eigenvalues of $Q_n := E_n[q(W_i)q(W_i)']$ are bounded below from zero and from above. Therefore,

$$\begin{aligned}\|\widehat{T}\widehat{g}^u - \widehat{m}\|_2 &= \|\mathbb{E}_n[q(W_i)(p(X_i)'\widehat{\beta}^u - Y_i)]\| \\ &\leq C\|E_n[(Y_i - p(X_i)'\widehat{\beta}^u)q(W_i)']Q_n^{-1}E_n[q(W_i)(Y_i - p(X_i)'\widehat{\beta}^u)]\|^{1/2} \\ &\leq C\|E_n[(Y_i - p(X_i)'\beta_n)q(W_i)']Q_n^{-1}E_n[q(W_i)(Y_i - p(X_i)'\beta_n)]\|^{1/2} \\ &\leq C\|\mathbb{E}_n[q(W_i)(p(X_i)'\beta_n - Y_i)]\|\end{aligned}$$

by optimality of $\widehat{\beta}^u$. Moreover,

$$\begin{aligned}\|\mathbb{E}_n[q(W_i)(p(X_i)'\beta_n - Y_i)]\| \leq \|(\widehat{T} - T_n)g_n\|_2 + \|(T_n - T)g_n\|_2 \\ + \|T(g_n - g)\|_2 + \|m - \widehat{m}\|_2\end{aligned}$$

by the triangle inequality. The terms $\|(\widehat{T} - T_n)g_n\|_2$ and $\|m - \widehat{m}\|_2$ have been bounded above. Also, by Assumptions 8(ii) and 6(iii),

$$\|(T_n - T)g_n\|_2 \leq C\tau_n^{-1}K^{-s}, \qquad \|T(g - g_n)\|_2 \leq C_g\tau_n^{-1}K^{-s}.$$

Combining the inequalities above gives

$$\|\widehat{g}^u - g_n\|_2 \leq C\Big(\tau_n(J/(\alpha n))^{1/2} + K^{-s} + \tau_n(\xi_n^2 \log n/n)^{1/2}\|\widehat{g} - g_n\|_2\Big) \tag{43}$$

with probability at least $1 - 2\alpha - n^{-c}$. Since $\tau_n^2\xi_n^2 \log n/n \to 0$, it follows that with the same probability,

$$\|\widehat{\beta}^u - \beta_n\| = \|\widehat{g}^u - g_n\|_2 \leq C\Big(\tau_n(J/(\alpha n))^{1/2} + K^{-s}\Big).$$

Conclude that with the same probability, by the triangle inequality,

$$\begin{aligned}|D\widehat{g}^u(x) - Dg(x)| &\leq |D\widehat{g}^u(x) - Dg_n(x)| + |Dg_n(x) - Dg(x)| \\ &\leq C \sup_{x\in[0,1]} \|Dp(x)\|(\tau_n(K/(\alpha n))^{1/2} + K^{-s}) + o(1)\end{aligned}$$

uniformly over $x \in [0,1]$ since $J \leq C_J K$ by Assumption 5. Since by the conditions of the lemma, $\sup_{x\in[0,1]} \|Dp(x)\|(\tau_n(K/n)^{1/2} + K^{-s}) \to 0$, (38) follows by taking $\alpha = \alpha_n \to 0$ slowly enough. This completes the proof of the lemma. Q.E.D.

Proof of Theorem 2. Applying the same arguments as those in the proof of Lemma 1 with $\widehat{g}^c$ replacing $\widehat{g}^u$, $\alpha/2$ replacing α, and using the bound

$$\|(T_n - \widehat{T})\widehat{g}^c\|_2 \leq \|T_n - \widehat{T}\|_2 \|\widehat{g}^c\|_2 \leq C_b \|T_n - \widehat{T}\|_2$$

instead of the bound for $\|(T_n - \widehat{T})\widehat{g}^u\|_2$ used in the proof of Lemma 1, it follows that with probability at least $1 - \alpha - n^{-1}$,

$$\|T(\widehat{g}^c - g_n)\|_2 \leq C\Big((K/(\alpha n))^{1/2} + (\xi_n^2 \log n/n)^{1/2} + \tau_n^{-1} K^{-s}\Big). \tag{44}$$

Further,

$$\|\widehat{g}^c - g_n\|_{2,t} \leq \delta + \tau_{n,t}\Big(\frac{\|Dg_n\|_\infty}{\delta}\Big)\|T(\widehat{g}^c - g_n)\|_2$$

since $\widehat{g}^c$ is increasing (indeed, if $\|\widehat{g}^c - g\|_{2,t} \leq \delta$, the bound is trivial; otherwise, apply the definition of $\tau_{n,t}$ to the function $(\widehat{g}^c - g_n)/\|\widehat{g}^c - g_n\|_{2,t}$ and use the inequality $\tau_{n,t}(\|Dg_n\|_\infty/\|\widehat{g}^c - g_n\|_{2,t}) \leq \tau_{n,t}(\|Dg_n\|_\infty/\delta)$). Finally, by the triangle inequality,

$$\|\widehat{g}^c - g\|_{2,t} \leq \|\widehat{g}^c - g_n\|_{2,t} + \|g_n - g\|_{2,t} \leq \|\widehat{g}^c - g_n\|_{2,t} + C_g K^{-s}.$$

Combining these inequalities gives the asserted claim (14).

To prove (15), observe that combining (44) and Assumption 6(iii) and applying the triangle inequality yield

$$\|T(\widehat{g}^c - g)\|_2 \leq C\Big((K/(\alpha n))^{1/2} + (\xi_n^2 \log n/n)^{1/2} + \tau_n^{-1} K^{-s}\Big),$$

which, by the same argument as that used to prove 14, gives

$$\|\widehat{g}^c - g\|_{2,t} \leq C\Big\{\delta + \tau\Big(\frac{\|Dg_n\|_\infty}{\delta}\Big)\Big(\frac{K}{\alpha n} + \frac{\xi_n^2 \log n}{n}\Big)^{1/2} + K^{-s}\Big\}. \tag{45}$$

The asserted claim (15) now follows by applying (14) with $\delta = 0$ and (45) with $\delta = \|Dg_n\|/c_\tau$ and using Corollary 3. This completes the proof of the theorem. Q.E.D.

Lemma 4. *Under conditions of Theorem 2, $\|\widehat{m} - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1} J^{-s})$ with probability at least $1 - \alpha$ where $\widehat{m}$ is defined in (39).*

Proof. Using the triangle inequality and an elementary inequality $(a+b)^2 \leq 2a^2+2b^2$ for all $a, b \geq 0$,

$$\|E_n[q(W_i)Y_i] - E[q(W)g(X)]\|^2 \leq 2\|E_n[q(W_i)\varepsilon_i]\|^2 + 2\|E_n[q(W_i)g(X_i)] - E[q(W)g(X)]\|^2.$$

To bound the first term on the right-hand side of this inequality, we have

$$E\left[\|E_n[q(W_i)\varepsilon_i]\|^2\right] = n^{-1}E[\|q(W)\varepsilon\|^2] \leq (C_B/n)E[\|q(W)\|^2] \leq CJ/n$$

where the first and the second inequalities follow from Assumptions 4 and 2, respectively. Similarly,

$$\begin{aligned} E\left[\|E_n[q(W_i)g(X_i)] - E[q(W)g(X)]\|^2\right] &\leq n^{-1}E[\|q(W)g(X)\|^2] \\ &\leq (C_B/n)E[\|q(W)\|^2] \leq CJ/n \end{aligned}$$

by Assumption 4. Therefore, denoting $\bar{m}_n(w) := q(w)'E[q(W)g(X)]$ for all $w \in [0,1]$, we obtain

$$E[\|\widehat{m} - \bar{m}_n\|_2^2] \leq CJ/n,$$

and so by Markov's inequality, $\|\widehat{m} - \bar{m}_n\|_2 \leq C(J/(\alpha n))^{1/2}$ with probability at least $1-\alpha$. Further, using $\gamma_n \in \mathbb{R}^J$ from Assumption 7, so that $m_n(w) = q(w)'\gamma_n$ for all $w \in [0,1]$, and denoting $r_n(w) := m(w) - m_n(w)$ for all $w \in [0,1]$, we obtain

$$\begin{aligned} \bar{m}_n(w) &= q(w)' \int_0^1 \int_0^1 q(t)g(x)f_{X,W}(x,t)dxdt \\ &= q(w)' \int_0^1 q(t)m(t)dt = q(w)' \int_0^1 q(t)(q(t)'\gamma_n + r_n(t))dt \\ &= q(w)'\gamma_n + q(w)' \int_0^1 q(t)r_n(t)dt = m(w) - r_n(w) + q(w)' \int_0^1 q(t)r_n(t)dt. \end{aligned}$$

Hence, by the triangle inequality,

$$\|\bar{m}_n - m\|_2 \leq \|r_n\|_2 + \left\| \int_0^1 q(t)r_n(t)dt \right\| \leq 2\|r_n\|_2 \leq 2C_m\tau_n^{-1}J^{-s}$$

by Bessel's inequality and Assumption 7. Applying the triangle inequality one more time, we obtain

$$\|\widehat{m} - m\|_2 \leq \|\widehat{m} - \bar{m}_n\| + \|\bar{m}_n - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1}J^{-s})$$

with probability at least $1-\alpha$. This completes the proof of the lemma. Q.E.D.

C Proofs for Section 4

Proof of Theorem 3. From the proof of Lemma 2 we know that g being increasing, constant, or decreasing implies that $M(w) := E[Y|W = w]$ is increasing, constant, or decreasing, respectively. Therefore, the sign of $Dg(x)$ is equal to the sign of $DM(w)$, which is identified from the observed distribution of (Y, W). Q.E.D.

Proof of Theorem 4. Suppose g' and g'' are observationally equivalent. Then $\|T(g' - g'')\|_2 = 0$. On the other hand, since $\|h\|_{2,t} + \bar{C}\|T\|_2\|h\|_2 < \|g' - g''\|_{2,t}$, there exists $\alpha \in (0, 1)$ such that $\|h\|_{2,t} + \bar{C}\|T\|_2\|h\|_2 < \alpha\|g' - g''\|_{2,t}$. Therefore, by Lemma 3, $\|T(g' - g'')\|_2 > \|g' - g''\|_{2,t}(1 - \alpha)/\bar{C} \geq 0$ which is a contradiction. This completes the proof of the theorem. Q.E.D.

D Proofs for Section 5

Proof of Theorem 5. In this proof, c and C are understood as sufficiently small and large constants, respectively, whose values may change at each appearance but can be chosen to depend only on $c_W, C_W, c_h, C_H, c_F, C_F, c_\epsilon, C_\epsilon$, and the kernel K.

To prove the asserted claims, we apply Corollary 3.1, Case (E.3), from CCK conditional on $\mathcal{W}_n = \{W_1, \ldots, W_n\}$. Under H_0,

$$T \leq \max_{(x,w,h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} \frac{\sum_{i=1}^n k_{i,h}(w)(1\{X_i \leq x\} - F_{X|W}(x|W_i))}{(\sum_{i=1}^n k_{i,h}(w)^2)^{1/2}} =: T_0 \tag{46}$$

with equality if the functions $w \mapsto F_{X|W}(x|w)$ are constant for all $x \in (0, 1)$. Using the notation of CCK,

$$T_0 = \max_{1 \leq j \leq p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij}$$

where $p = |\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n|$, the number of elements in the set $\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n$, $x_{ij} = z_{ij}\varepsilon_{ij}$ with z_{ij} having the form $\sqrt{n}k_{i,h}(w)/(\sum_{i=1}^n k_{i,h}(w)^2)^{1/2}$, and ε_{ij} having the form $1\{X_i \leq x\} - F_{X|W}(x|W_i)$ for some $(x, w, h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n$. The dimension p satisfies $\log p \leq C \log n$. Also, $n^{-1}\sum_{i=1}^n z_{ij}^2 = 1$. Further, since $0 \leq 1\{X_i \leq x\} \leq 1$, we have $|\varepsilon_{ij}| \leq 1$, and so $\mathrm{E}[\exp(|\varepsilon_{ij}|/2)|\mathcal{W}_n] \leq 2$. In addition, $\mathrm{E}[\varepsilon_{ij}^2|\mathcal{W}_n] \geq c_\epsilon(1 - C_\epsilon) > 0$ by Assumption 12. Thus, T_0 satisfies the conditions of Case (E.3) in CCK with a sequence of constants B_n as long as $|z_{ij}| \leq B_n$ for all $j = 1, \ldots, p$. In turn, Proposition B.2 in Chetverikov (2012) shows that under Assumptions 2, 9, and 10, with probability at least $1 - Cn^{-c}$, $z_{ij} \leq C/\sqrt{h_{\min}} =: B_n$ uniformly over all $j = 1, \ldots, p$ (Proposition B.2 in Chetverikov (2012) is stated with "w.p.a.1" replacing "$1 - Cn^{-c}$"; however, inspecting the proof of Proposition B.2 (and supporting Lemma H.1) shows that the result applies with

"$1 - Cn^{-c}$" instead of "w.p.a.1"). Let $\mathcal{B}_{1,n}$ denote the event that $|z_{ij}| \leq C/\sqrt{h_{\min}} = B_n$ for all $j = 1, \ldots, p$. As we just established, $\mathrm{P}(\mathcal{B}_{1,n}) \geq 1 - Cn^{-c}$. Since $(\log n)^7/(nh_{\min}) \leq C_h n^{-c_h}$ by Assumption 10, we have that $B_n^2(\log n)^7/n \leq Cn^{-c}$, and so condition (i) of Corollary 3.1 in CCK is satisfied on the event $\mathcal{B}_{1,n}$.

Let $\mathcal{B}_{2,n}$ denote the event that

$$\mathrm{P}\left(\max_{(x,w)\in\mathcal{X}_n\times\mathcal{W}_n} |\widehat{F}_{X|W}(x|w) - F_{X|W}(x|w)| > C_F n^{-c_F} | \{\mathcal{W}_n\}\right) \leq C_F n^{-c_F}.$$

By Assumption 11, $\mathrm{P}(\mathcal{B}_{2,n}) \geq 1 - C_F n^{-c_F}$. We will apply Corollary 3.1 from CCK conditional on $\mathcal{W}_n$ on the event $\mathcal{B}_{1,n} \cap \mathcal{B}_{2,n}$. For this, we need to show that on the event $\mathcal{B}_{2,n}$, $\zeta_{1,n}\sqrt{\log n} + \zeta_{2,n} \leq Cn^{-c}$ where $\zeta_{1,n}$ and $\zeta_{2,n}$ are positive sequences such that

$$\mathrm{P}\left(\mathrm{P}_e(|T^b - T_0^b| > \zeta_{1,n}) > \zeta_{2,n} | \mathcal{W}_n\right) < \zeta_{2,n} \tag{47}$$

where

$$T_0^b := \max_{(x,w,h)\in\mathcal{X}_n\times\mathcal{W}_n\times\mathcal{H}_n} \frac{\sum_{i=1}^n e_i \left(k_{i,h}(w)(1\{X_i \leq x\} - F_{X|W}(x|W_i))\right)}{\left(\sum_{i=1}^n k_{i,h}(w)^2\right)^{1/2}}$$

and where $\mathrm{P}_e(\cdot)$ denotes the probability distribution with respect to the distribution of $e_1, \ldots, e_n$ and keeping everything else fixed. To find such sequences $\zeta_{1,n}$ and $\zeta_{2,n}$, note that $\zeta_{1,n}\sqrt{\log n} + \zeta_{2,n} \leq Cn^{-c}$ follows from $\zeta_{1,n} + \zeta_{2,n} \leq Cn^{-c}$ (with different constants $c, C > 0$), so that it suffices to verify the latter condition. Also,

$$|T^b - T_0^b| \leq \max_{(x,w,h)\in\mathcal{X}_n\times\mathcal{W}_n\times\mathcal{H}_n} \left| \frac{\sum_{i=1}^n e_i k_{i,h}(w)(\widehat{F}_{X|W}(x|W_i) - F_{X|W}(x|W_i))}{\left(\sum_{i=1}^n k_{i,h}(w)^2\right)^{1/2}} \right|.$$

For fixed $W_1, \ldots, W_n$ and $X_1, \ldots, X_n$, the random variables under the modulus on the right-hand side of this inequality are normal with zero mean and variance bounded from above by $\max_{(x,w)\in\mathcal{X}_n\times\mathcal{W}_n} |\widehat{F}_{X|W}(x|w) - F_{X|W}(x|w)|^2$. Therefore,

$$\mathrm{P}_e\left(|T^b - T_0^b| > C\sqrt{\log n} \max_{(x,w)\in\mathcal{X}_n\times\mathcal{W}_n} \left|\widehat{F}_{X|W}(x|w) - F_{X|W}(x|w)\right|\right) \leq Cn^{-c}.$$

Hence, on the event that

$$\max_{(x,w)\in\mathcal{X}_n\times\mathcal{W}_n} \left|\widehat{F}_{X|W}(x|w) - F_{X|W}(x|w)\right| \leq C_F n^{-c_F},$$

whose conditional probability given $\mathcal{W}_n$ on $\mathcal{B}_{2,n}$ is at least $1 - C_F n^{-c_F}$ by the definition of $\mathcal{B}_{2,n}$,

$$\mathrm{P}_e\left(|T^b - T_0^b| > Cn^{-c}\right) \leq Cn^{-c}$$

implying that (47) holds for some $\zeta_{1,n}$ and $\zeta_{2,n}$ satisfying $\zeta_{1,n} + \zeta_{2,n} \leq Cn^{-c}$.

Thus, applying Corollary 3.1, Case (E.3), from CCK conditional on $\{W_1, \ldots, W_n\}$ on the event $\mathcal{B}_{1,n} \cap \mathcal{B}_{2,n}$ gives

$$\alpha - Cn^{-c} \leq \mathrm{P}(T_0 > c(\alpha)|\mathcal{W}_n) \leq \alpha + Cn^{-c}.$$

Since $\mathrm{P}(\mathcal{B}_{1,n} \cap \mathcal{B}_{2,n}) \geq 1 - Cn^{-c}$, integrating this inequality over the distribution of $\mathcal{W}_n = \{W_1, \ldots, W_n\}$ gives (24). Combining this inequality with (46) gives (23). This completes the proof of the theorem. Q.E.D.

Proof of Theorem 6. Conditional on the data, the random variables

$$T^b(x, w, h) := \frac{\sum_{i=1}^n e_i \left(k_{i,h}(w)(1\{X_i \leq x\} - \widehat{F}_{X|W}(x|W_i)) \right)}{(\sum_{i=1}^n k_{i,h}(w)^2)^{1/2}}$$

for $(x, w, h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n$ are normal with zero mean and variances bounded from above by

$$\begin{aligned} &\frac{\sum_{i=1}^n \left(k_{i,h}(w)(1\{X_i \leq x\} - \widehat{F}_{X|W}(x|W_i)) \right)^2}{\sum_{i=1}^n k_{i,h}(w)^2} \\ &\quad \leq \max_{(x,w,h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} \max_{1 \leq i \leq n} \left(1\{X_i \leq x\} - \widehat{F}_{X|W}(x|W_i) \right)^2 \leq (1 + C_h)^2 \end{aligned}$$

by Assumption 11. Therefore, $c(\alpha) \leq C(\log n)^{1/2}$ for some constant $C > 0$ since $c(\alpha)$ is the $(1 - \alpha)$ conditional quantile of T^b given the data, $T^b = \max_{(x,w,h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} T^b(x, w, h)$, and $p := |\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n|$, the number of elements of the set $\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n$, satisfies $\log p \leq C \log n$ (with a possibly different constant $C > 0$). Thus, the growth rate of the critical value $c(\alpha)$ satisfies the same upper bound $(\log n)^{1/2}$ as if we were testing monotonicity of one particular regression function $w \mapsto \mathrm{E}[1\{X \leq x_0\}|W = w]$ with $\mathcal{X}_n$ replaced by x_0 for some $x_0 \in (0, 1)$ in the definition of T and T^b. Hence, the asserted claim follows from the same arguments as those given in the proof of Theorem 4.2 in Chetverikov (2012). This completes the proof of the theorem. Q.E.D.

E Technical tools

In this section, we provide a set of technical results that are used to prove the statements from the main text.

Lemma 5 (Tikhonov)**.** *Let (D, ρ_D) and (R, ρ_R) be two pseudo-metric spaces and assume that D is compact. Further, suppose there is a one-to-one continuous operator $A : D \to R$. Then the inverse operator A^{-1} exists and is continuous over the range $A(D)$ of A.*

Remark 17. This Tikhonov's lemma is essentially well known but it is typically presented for metric spaces whereas we require pseudo-metric spaces. Following Dudley (2002), we define a pseudo-metric ρ_D on the space D as a function $\rho_D : D \times D \to \mathbb{R}$ that satisfies for any $d_1, d_2, d_3 \in D$, (i) $\rho_D(d_1, d_2) \geq 0$, (ii) $\rho_D(d_1, d_1) = 0$, (iii) $\rho_D(d_1, d_2) = \rho_D(d_2, d_1)$, and (iv) $\rho_D(d_1, d_3) \leq \rho_D(d_1, d_2) + \rho_D(d_2, d_3)$. Importantly for our application of the above lemma, the pseudo-metric ρ_D allows for the case that $\rho_D(d_1, d_2) = 0$, but $d_1 \neq d_2$ and thus $A(d_1) \neq A(d_2)$. □

Proof. Since A is one-to-one on D, the inverse operator $A^{-1} : A(D) \to A$ exists. To prove its continuity, take any $r \in A(D)$ and any sequence r_k in $A(D)$ such that $r_k \to r$ as $k \to \infty$. Let $d_k := A^{-1} r_k$ for all k and $d := A^{-1} r$. We want to show that $d_k \to d$ as $k \to \infty$. Suppose the contrary. Then there exist $\varepsilon > 0$ and a subsequence d_{k_l} of d_k, $k_l \to \infty$ as $l \to \infty$, such that $\rho_D(d_{k_l}, d) \geq \varepsilon$ for all l. Also, by compactness of D, there exists a further subsequence $d_{k_{l_m}}$ of d_{k_l}, $l_m \to \infty$ as $m \to \infty$, that converges to some element $\widetilde{d} \in D$ as $m \to \infty$. Clearly, $\rho_D(\widetilde{d}, d) \geq \varepsilon$, and so $\widetilde{d} \neq d$. On the other hand, by continuity of A, we also have $r_{k_{l_m}} = A(d_{k_{l_m}}) \to A(\widetilde{d})$ as $m \to \infty$. However, since $r_{k_{l_m}} \to r$ as $m \to \infty$, $A(\widetilde{d}) = r$ and thus $\widetilde{d} = d$, a contradiction. Q.E.D.

Lemma 6. *Let W be a random variable with the density function bounded below from zero on its support $[0, 1]$, and let $M : [0, 1] \to \mathbb{R}$ be a monotone function. If M is constant, then $cov(W, M(W)) = 0$. If M is increasing in the sense that there exist $0 < w_1 < w_2 < 1$ such that $M(w_1) < M(w_2)$, then $cov(W, M(W)) > 0$.*

Proof. The first claim is trivial. The second claim follows by introducing an independent copy W' of the random variable W, and rearranging the inequality

$$E[(M(W) - M(W'))(W - W')] > 0,$$

which holds for increasing M since $(M(W) - M(W'))(W - W') \geq 0$ almost surely and $(M(W) - M(W'))(W - W') > 0$ with strictly positive probability. This completes the proof of the lemma. Q.E.D.

Lemma 7. *For any orthonormal basis $\{h_j, j \geq 1\}$ in $L^2[0, 1]$, any $0 \leq x_1 < x_2 \leq 1$, and any $\alpha > 0$,*

$$\|h_j\|_{2,t} = \left(\int_{x_1}^{x_2} h_j^2(x) dx \right)^{1/2} > j^{-1/2-\alpha}$$

for infinitely many j.

Proof. Fix $M \in \mathbb{N}$ and consider any partition $x_1 = t_0 < t_1 < \cdots < t_M = x_2$. Further, fix $m = 1, \ldots, M$ and consider the function

$$h(x) = \begin{cases} \frac{1}{\sqrt{t_m - t_{m-1}}} & x \in (t_{m-1}, t_m], \\ 0, & x \notin (t_{m-1}, t_m]. \end{cases}$$

Note that $\|h\|_2 = 1$, so that

$$h = \sum_{j=1}^{\infty} \beta_j h_j \text{ in } L^2[0,1], \quad \beta_j := \frac{\int_{t_{m-1}}^{t_m} h_j(x)dx}{(t_m - t_{m-1})^{1/2}}, \quad \text{and } \sum_{j=1}^{\infty} \beta_j^2 = 1.$$

Therefore, by the Cauchy-Schwartz inequality,

$$1 = \sum_{j=1}^{\infty} \beta_j^2 = \frac{1}{t_m - t_{m-1}} \sum_{j=1}^{\infty} \left(\int_{t_{m-1}}^{t_m} h_j(x)dx \right)^2 \leq \sum_{j=1}^{\infty} \int_{t_{m-1}}^{t_m} (h_j(x))^2 dx.$$

Hence, $\sum_{j=1}^{\infty} \|h_j\|_{2,t}^2 \geq M$. Since M is arbitrary, we obtain $\sum_{j=1}^{\infty} \|h_j\|_{2,t}^2 = \infty$, and so for any J, there exists $j > J$ such that $\|h_j\|_{2,t} > j^{-1/2-\alpha}$. Otherwise, we would have $\sum_{j=1}^{\infty} \|h_j\|_{2,t}^2 < \infty$. This completes the proof of the lemma. Q.E.D.

Lemma 8. *Let (X, W) be a pair of random variables defined as in Example 1. Then Assumptions 1 and 2 of Section 2 are satisfied if $0 < x_1 < x_2 < 1$ and $0 < w_1 < w_2 < 1$.*

Proof. As noted in Example 1, we have

$$X = \Phi(\rho \Phi^{-1}(W) + (1 - \rho^2)^{1/2} U)$$

where $\Phi(x)$ is the distribution function of a $N(0,1)$ random variable and U is a $N(0,1)$ random variable that is independent of W. Therefore, the conditional distribution function of X given W is

$$F_{X|W}(x|w) := \Phi\left(\frac{\Phi^{-1}(x) - \rho \Phi^{-1}(w)}{\sqrt{1 - \rho^2}} \right).$$

Since the function $w \mapsto F_{X|W}(x|w)$ is decreasing for all $x \in (0,1)$, condition (3) of Assumption 1 follows. Further, to prove condition (4) of Assumption 1, it suffices to show that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} \leq c_F \tag{48}$$

for some constant $c_F < 0$, all $x \in (0, x_2)$, and all $w \in (w_1, w_2)$ because, for every $x \in (0, x_2)$ and $w \in (w_1, w_2)$, there exists $\bar{w} \in (w_1, w_2)$ such that

$$\log\left(\frac{F_{X|W}(x|w_1)}{F_{X|W}(x|w_2)} \right) = \log F_{X|W}(x|w_1) - \log F_{X|W}(x|w_2) = -(w_2 - w_1) \frac{\partial \log F_{X|W}(x|\bar{w})}{\partial w}.$$

Therefore, $\partial \log F_{X|W}(x|w)/\partial w \leq c_F < 0$ for all $x \in (0, x_2)$ and $w \in (w_1, w_2)$ implies

$$\frac{F_{X|W}(x|w_1)}{F_{X|W}(x|w_2)} \geq e^{-c_F(w_2 - w_1)} > 1$$

for all $x \in (0, x_2)$. To show (48), observe that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} = -\frac{\rho}{\sqrt{1-\rho^2}} \frac{\phi(y)}{\Phi(y)} \frac{1}{\phi(\Phi^{-1}(w))} \leq -\frac{\sqrt{2\pi}\rho}{\sqrt{1-\rho^2}} \frac{\phi(y)}{\Phi(y)} \tag{49}$$

where $y := (\Phi^{-1}(x) - \rho\Phi^{-1}(w))/(1-\rho^2)^{1/2}$. Thus, (48) holds for some $c_F < 0$ and all $x \in (0, x_2)$ and $w \in (w_1, w_2)$ such that $\Phi^{-1}(x) \geq \rho\Phi^{-1}(w)$ since $x_2 < 1$ and $0 < w_1 < w_2 < 1$. On the other hand, when $\Phi^{-1}(x) < \rho\Phi^{-1}(w)$, so that $y < 0$, it follows from Proposition 2.5 in Dudley (2014) that $\phi(y)/\Phi(y) \geq (2/\pi)^{1/2}$, and so (49) implies that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} \leq -\frac{2\rho}{\sqrt{1-\rho^2}}$$

in this case. Hence, condition (4) of Assumption 1 is satisfied. Similar argument also shows that condition (5) of Assumption 1 is satisfied as well.

We next consider Assumption 2. Since W is distributed uniformly on $[0,1]$ (remember that $\widetilde{W} \sim N(0,1)$ and $W = \Phi(\widetilde{W})$), condition (iii) of Assumption 2 is satisfied. Further, differentiating $x \mapsto F_{X|W}(x|w)$ gives

$$f_{X|W}(x|w) := \frac{1}{\sqrt{1-\rho^2}} \phi\left(\frac{\Phi^{-1}(x) - \rho\Phi^{-1}(w)}{\sqrt{1-\rho^2}}\right) \frac{1}{\phi(\Phi^{-1}(x))}. \tag{50}$$

Since $0 < x_1 < x_2 < 1$ and $0 < w_1 < w_2 < 1$, condition (ii) of Assumption 2 is satisfied as well. Finally, to prove condition (i) of Assumption 2, note that since $f_W(w) = 1$ for all $w \in [0,1]$, (50) combined with the change of variables formula with $x = \Phi(\widetilde{x})$ and $w = \Phi(\widetilde{w})$ give

$$\begin{aligned}
(1-\rho^2) \int_0^1 \int_0^1 f_{X,W}^2(x,w) dx dw &= (1-\rho^2) \int_0^1 \int_0^1 f_{X|W}^2(x|w) dx dw \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \phi^2\left(\frac{\widetilde{x} - \rho\widetilde{w}}{\sqrt{1-\rho^2}}\right) \frac{\phi(\widetilde{w})}{\phi(\widetilde{x})} d\widetilde{x} d\widetilde{w} \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left[\left(\frac{1}{2} - \frac{1}{1-\rho^2}\right)\widetilde{x}^2 + \frac{2\rho}{1-\rho^2}\widetilde{x}\widetilde{w} - \left(\frac{\rho^2}{1-\rho^2} + \frac{1}{2}\right)\widetilde{w}^2\right] d\widetilde{x} d\widetilde{w} \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\left[-\frac{1+\rho^2}{2(1-\rho^2)}\left(\widetilde{x}^2 - \frac{4\rho}{1+\rho^2}\widetilde{x}\widetilde{w} + \widetilde{w}^2\right)\right] d\widetilde{x} d\widetilde{w}.
\end{aligned}$$

Since $4\rho/(1+\rho^2) < 2$, the integral in the last line is finite implying that condition (i) of Assumption 2 is satisfied. This completes the proof of the lemma. Q.E.D.

Lemma 9. *Let $X = U_1 + U_2 W$ where U_1, U_2, W are mutually independent, $U_1, U_2 \sim U[0, 1/2]$ and $W \sim U[0,1]$. Then Assumptions 1 and 2 of Section 2 are satisfied if $0 < w_1 < w_2 < 1$, $0 < x_1 < x_2 < 1$, and $w_1 > w_2 - \sqrt{w_2/2}$.*

Proof. Since $X|W=w$ is a convolution of the random variables U_1 and $U_2 w$,

$$\begin{aligned} f_{X|W}(x|w) &= \int_0^{1/2} f_{U_1}(x-u_2 w) f_{U_2}(u_2) du_2 \\ &= 4\int_0^{1/2} 1\left\{0 \le x - u_2 w \le \frac{1}{2}\right\} du_2 \\ &= 4\int_0^{1/2} 1\left\{\frac{x}{w} - \frac{1}{2w} \le u_2 \le \frac{x}{w}\right\} du_2 \\ &= \begin{cases} \frac{4x}{w}, & 0 \le x < \frac{w}{2} \\ 2, & \frac{w}{2} \le x < \frac{1}{2} \\ \frac{2(1+w)}{w} - \frac{4x}{w}, & \frac{1}{2} \le x < \frac{1+w}{2} \\ 0, & \frac{1+w}{2} \le x \le 1 \end{cases} \end{aligned}$$

and, thus,

$$F_{X|W}(x|w) = \begin{cases} \frac{2x^2}{w}, & 0 \le x < \frac{w}{2} \\ 2x - \frac{w}{2}, & \frac{w}{2} \le x < \frac{1}{2} \\ 1 - \frac{2}{w}\left(x - \frac{1+w}{2}\right)^2, & \frac{1}{2} \le x < \frac{1+w}{2} \\ 1, & \frac{1+w}{2} \le x \le 1 \end{cases}.$$

It is easy to check that $\partial F_{X|W}(x|w)/\partial w \le 0$ for all $x, w \in [0,1]$ so that condition (3) of Assumption 1 is satisfied. To check conditions (4) and (5), we proceed as in Lemma 8 and show $\partial \log F_{X|W}(x|w)/\partial w < 0$ uniformly for all $x \in [\underline{x}_2, \overline{x}_1]$ and $w \in (\widetilde{w}_1, \widetilde{w}_2)$. First, notice that, as required by Assumption 2(iv), $[\underline{x}_k, \overline{x}_k] = [0, (1+\widetilde{w}_k)/2]$, $k = 1, 2$. For $0 \le x < w/2$ and $w \in (\widetilde{w}_1, \widetilde{w}_2)$,

$$\frac{\partial F_{X|W}(x|w)}{\partial w} = \frac{-2x^2/w^2}{2x^2/w} = -\frac{1}{w} < -\frac{1}{\widetilde{w}_1} < 0,$$

and, for $w/2 \le x < 1/2$ and $w \in (\widetilde{w}_1, \widetilde{w}_2)$,

$$\frac{\partial F_{X|W}(x|w)}{\partial w} = \frac{-1/2}{2x - w/2} < \frac{-1/2}{w - w/2} < -\frac{1}{\widetilde{w}_1} < 0.$$

Therefore, (4) holds uniformly over $x \in (\underline{x}_2, 1/2)$ and (5) uniformly over $x \in (x_1, 1/2)$. Now, consider $1/2 \le x < (1+\widetilde{w}_1)/2$ and $w \in (\widetilde{w}_1, \widetilde{w}_2)$. Notice that, on this interval, $\partial(F_{X|W}(x|\widetilde{w}_1)/F_{X|W}(x|\widetilde{w}_2))/\partial x \le 0$ so that

$$\frac{F_{X|W}(x|\widetilde{w}_1)}{F_{X|W}(x|\widetilde{w}_2)} = \frac{1 - \frac{2}{\widetilde{w}_1}\left(x - \frac{1+\widetilde{w}_1}{2}\right)^2}{1 - \frac{2}{\widetilde{w}_2}\left(x - \frac{1+\widetilde{w}_2}{2}\right)^2} \ge \frac{1}{1 - \frac{2}{\widetilde{w}_2}\left(\frac{1+\widetilde{w}_1}{2} - \frac{1+\widetilde{w}_2}{2}\right)^2} = \frac{\widetilde{w}_2}{\widetilde{w}_2 - 2(\widetilde{w}_1 - \widetilde{w}_2)^2} > 1,$$

where the last inequality uses $\widetilde{w}_1 > \widetilde{w}_2 - \sqrt{\widetilde{w}_2/2}$, and thus (4) holds also uniformly over $1/2 \le x < x_2$. Similarly,

$$\frac{1 - F_{X|W}(x|\widetilde{w}_2)}{1 - F_{X|W}(x|\widetilde{w}_1)} = \frac{\frac{2}{\widetilde{w}_2}\left(x - \frac{1+\widetilde{w}_2}{2}\right)^2}{\frac{2}{\widetilde{w}_1}\left(x - \frac{1+\widetilde{w}_1}{2}\right)^2} \ge \frac{\frac{2}{\widetilde{w}_2}\left(\frac{\widetilde{w}_2}{2}\right)^2}{\frac{2}{\widetilde{w}_1}\left(\frac{\widetilde{w}_1}{2}\right)^2} = \frac{\widetilde{w}_2}{\widetilde{w}_1} > 1$$

so that (5) also holds uniformly over $1/2 \leq x < \overline{x}_1$. Assumption 2(i) trivially holds. Parts (ii) and (iii) of Assumption 2 hold for any $0 < \widetilde{x}_1 < \widetilde{x}_2 \leq \overline{x}_1 \leq 1$ and $0 \leq w_1 < \widetilde{w}_1 < \widetilde{w}_2 < w_2 \leq 1$ with $[\underline{x}_k, \overline{x}_k] = [0, (1 + \widetilde{w}_k)/2]$, $k = 1, 2$. Q.E.D.

Lemma 10. *For any increasing function $h \in L^2[0,1]$, one can find a sequence of increasing continuously differentiable functions $h_k \in L^2[0,1]$, $k \geq 1$, such that $\|h_k - h\|_2 \to 0$ as $k \to \infty$.*

Proof. Fix some increasing $h \in L^2[0,1]$. For $a > 0$, consider the truncated function:

$$\widetilde{h}_a(x) := h(x)1\{|h(x)| \leq a\} + a1\{h(x) > a\} - a1\{h(x) < -a\}$$

for all $x \in [0,1]$. Then $\|\widetilde{h}_a - h\|_2 \to 0$ as $a \to \infty$ by Lebesgue's dominated convergence theorem. Hence, by scaling and shifting h if necessary, we can assume without loss of generality that $h(0) = 0$ and $h(1) = 1$.

To approximate h, set $h(x) = 0$ for all $x \in \mathbb{R}\backslash[0,1]$ and for $\sigma > 0$, consider the function

$$h_\sigma(x) := \frac{1}{\sigma}\int_0^1 h(y)\phi\left(\frac{y-x}{\sigma}\right)dy = \frac{1}{\sigma}\int_{-\infty}^{\infty} h(y)\phi\left(\frac{y-x}{\sigma}\right)dy$$

for $y \in \mathbb{R}$ where ϕ is the density function of a $N(0,1)$ random variable. Theorem 6.3.14 in Stroock (1999) shows that

$$\begin{aligned}\|h_\sigma - h\|_2 &= \left(\int_0^1 (h_\sigma(x) - h(x))^2 dx\right)^{1/2} \\ &\leq \left(\int_{-\infty}^{\infty} (h_\sigma(x) - h(x))^2 dx\right)^{1/2} \to 0\end{aligned}$$

as $\sigma \to 0$. The function h_σ is continuously differentiable but it is not necessarily increasing, and so we need to further approximate it by an increasing continuously differentiable function. However, integration by parts yields for all $x \in [0,1]$,

$$\begin{aligned}Dh_\sigma(x) &= -\frac{1}{\sigma^2}\int_0^1 h(y)D\phi\left(\frac{y-x}{\sigma}\right)dy \\ &= -\frac{1}{\sigma}\left(h(1)\phi\left(\frac{1-x}{\sigma}\right) - h(0)\phi\left(\frac{-x}{\sigma}\right) - \int_0^1 \phi\left(\frac{y-x}{\sigma}\right)dh(y)\right) \\ &\geq -\frac{1}{\sigma}\phi\left(\frac{1-x}{\sigma}\right)\end{aligned}$$

since $h(0) = 0$, $h(1) = 1$, and $\int_0^1 \phi((y-x)\sigma)dh(y) \geq 0$ by h being increasing. Therefore, the function

$$h_{\sigma,\bar{x}}(x) = \begin{cases} h_\sigma(x) + (x/\sigma)\phi((1-\bar{x})/\sigma), & \text{for } x \in [0, \overline{x}] \\ h_\sigma(\bar{x}) + (\bar{x}/\sigma)\phi((1-\bar{x})/\sigma), & \text{for } x \in (\overline{x}, 1] \end{cases}$$

defined for all $x \in [0,1]$ and some $\bar{x} \in (0,1)$ is increasing and continuously differentiable for all $x \in (0,1)\backslash\bar{x}$, where it has a kink. Also, setting $\bar{x} = \bar{x}_\sigma = 1 - \sqrt{\sigma}$ and observing that $0 \leq h_\sigma(x) \leq 1$ for all $x \in [0,1]$, we obtain

$$\|h_{\sigma,\bar{x}_\sigma} - h_\sigma\|_2 \leq \frac{1}{\sigma}\phi\left(\frac{1}{\sqrt{\sigma}}\right)\left(\int_0^{1-\sqrt{\sigma}} dx\right)^{1/2} + \left(1 + \frac{1}{\sigma}\phi\left(\frac{1}{\sqrt{\sigma}}\right)\right)\left(\int_{1-\sqrt{\sigma}}^1 dx\right)^{1/2} \to 0$$

as $\sigma \to 0$ because $\sigma^{-1}\phi(\sigma^{-1/2}) \to 0$. Smoothing the kink of $h_{\sigma,\bar{x}_\sigma}$ and using the triangle inequality, we obtain the asserted claim. This completes the proof of the lemma. Q.E.D.

Lemma 11. *Let $(p_1', q_1')', \ldots, (p_n', q_n')'$ be a sequence of i.i.d. random vectors where p_i's are vectors in $\mathbb{R}^K$ and q_i's are vectors in $\mathbb{R}^J$. Assume that $\|p_1\| \leq \xi_n$, $\|q_1\| \leq \xi_n$, $\|E[p_1 p_1']\| \leq C_p$, and $\|E[q_1 q_1']\| \leq C_q$ where $\xi_n \geq 1$. Then for all $t \geq 0$,*

$$P\left(\|E_n[p_i q_i'] - E[p_1 q_1']\| \geq t\right) \leq \exp\left(\log(K+J) - \frac{At^2}{\xi_n^2(1+t)}\right)$$

where $A > 0$ is a constant depending only on C_p and C_q.

Remark 18. A closely related result can be found in Belloni, Chernozhukov, Chetverikov, and Kato (2014) who were the first to use it in the econometrics literature. The current version of the result is more useful for our purposes.

Proof. The proof follows from Corollary 6.2.1 in Tropp (2012). Below we perform some auxiliary calculations. For any $a \in \mathbb{R}^K$ and $b \in \mathbb{R}^J$,

$$\begin{aligned} a'E[p_1 q_1']b &= E[(a'p_1)(b'q_1)] \\ &\leq \left(E[(a'p_1)^2]E[(b'q_1)^2]\right)^{1/2} \leq \|a\|\|b\|(C_p C_q)^{1/2} \end{aligned}$$

by Hölder's inequality. Therefore, $\|E[p_1 q_1']\| \leq (C_p C_q)^{1/2}$. Further, denote $S_i := p_i q_i' - E[p_i q_i']$ for $i = 1, \ldots, n$. By the triangle inequality and calculations above,

$$\begin{aligned} \|S_1\| &\leq \|p_1 q_1'\| + \|E[p_1 q_1']\| \\ &\leq \xi_n^2 + (C_p C_q)^{1/2} \leq \xi_n^2(1 + (C_p C_q)^{1/2}) =: R. \end{aligned}$$

Now, denote $Z_n := \sum_{i=1}^n S_i$. Then

$$\begin{aligned} E[Z_n Z_n'] &\leq n\|E[S_1 S_1']\| \\ &\leq n\|E[p_1 q_1' q_1 p_1']\| + n\|E[p_1 q_1']E[q_1 p_1']\| \leq n\|E[p_1 q_1' q_1 p_1']\| + nC_p C_q. \end{aligned}$$

For any $a \in \mathbb{R}^K$,

$$a'E[p_1 q_1' q_1 p_1']a \leq \xi_n^2 E[(a'p_1)^2] \leq \xi_n^2 \|a\|^2 C_p.$$

Therefore, $\|E[p_1 q_1' q_1 p_1']\| \leq \xi_n^2 C_p$, and so

$$\|E[Z_n Z_n']\| \leq nC_p(\xi_n^2 + C_q) \leq n\xi_n^2(1 + C_p)(1 + C_q).$$

Similarly, $\|E[Z_n' Z_n]\| \leq n\xi_n^2(1 + C_p)(1 + C_q)$, and so

$$\sigma^2 := \max(\|E[Z_n Z_n']\|, \|E[Z_n' Z_n]\|) \leq n\xi_n^2(1 + C_p)(1 + C_q).$$

Hence, by Corollary 6.2.1 in Tropp (2012),

$$\begin{aligned} P\left(\|n^{-1} Z_n\| \geq t\right) &\leq (K + J) \exp\left(-\frac{n^2 t^2/2}{\sigma^2 + Rt/3}\right) \\ &\leq \exp\left(\log(K + J) - \frac{Ant^2}{\xi_n^2(1 + t)}\right). \end{aligned}$$

This completes the proof of the lemma. Q.E.D.

References

Abrevaya, J., J. A. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model With Endogenous Regressors," *Econometrica*, 78(6), 2043–2061.

Andrews, D. W. K., and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721–762.

Bejenaru, I., and T. Tao (2006): "Sharp well-posedness and ill-posedness results for a quadratic non-linear Schrödinger equation," *Journal of Functional Analysis*, 233(1), 228–259.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2014): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," Discussion paper.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75(6), 1613–1669.

Blundell, R., J. Horowitz, and M. Parey (2013): "Nonparametric Estimation of a Heterogeneous Demand Function under the Slutsky Inequality Restriction," Working Paper CWP54/13, cemmap.

Blundell, R., J. L. Horowitz, and M. Parey (2012): "Measuring the price responsiveness of gasoline demand: Economic shape restrictions and nonparametric demand estimation," *Quantitative Economics*, 3(1), 29–51.

Brunk, H. D. (1955): "Maximum Likelihood Estimates of Monotone Parameters," *The Annals of Mathematical Statistics*, 26(4), 607–616.

Canay, I. A., A. Santos, and A. M. Shaikh (2013): "On the Testability of Identification in Some Nonparametric Models With Endogeneity," *Econometrica*, 81(6), 2535–2559.

Chatterjee, S., A. Guntuboyina, and B. Sen (2013): "Improved Risk Bounds in Isotonic Regression," Discussion paper.

Chen, X., and T. M. Christensen (2013): "Optimal Uniform Convergence Rates for Sieve Nonparametric Instrumental Variables Regression," Discussion paper.

Chen, X., and M. Reiss (2011): "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics," *Econometric Theory*, 27(Special Issue 03), 497–521.

Cheng, K.-F., and P.-E. Lin (1981): "Nonparametric estimation of a regression function," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(2), 223–233.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013a): "Anti-Concentration and Honest, Adaptive Confidence Bands," *The Annals of Statistics*, forthcoming.

——— (2013b): "Gaussian Approximation of Suprema of Empirical Processes," Discussion paper.

——— (2013c): "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *The Annals of Statistics*, 41(6), 2786–2819.

Chetverikov, D. (2012): "Testing Regression Monotonicity in Econometric Models," Discussion paper.

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

De Vore, R. A. (1977a): "Monotone approximation by polynomials," *SIAM Journal on Mathematical Analysis*, 8(5), 906–921.

——— (1977b): "Monotone approximation by splines," *SIAM Journal on Mathematical Analysis*, 8(5), 891–905.

Delecroix, M., and C. Thomas-Agnan (2000): "Spline and Kernel Regression under Shape Restrictions," in *Smoothing and Regression*, pp. 109–133. John Wiley and Sons, Inc.

Delgado, M. A., and J. C. Escanciano (2012): "Distribution-free tests of stochastic monotonicity," *Journal of Econometrics*, 170(1), 68–75.

Dette, H., N. Neumeyer, and K. F. Pilz (2006): "A simple nonparametric estimator of a strictly monotone regression function," *Bernoulli*, 12(3), 469–490.

Dudley, R. M. (2002): *Real Analysis and Probability*. Cambridge University Press, Cambridge.

——— (2014): *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge.

Freyberger, J., and J. Horowitz (2013): "Identification and shape restrictions in nonparametric instrumental variables estimation," Working Paper CWP31/13, cemmap.

Friedman, J., and R. Tibshirani (1984): "The Monotone Smoothing of Scatterplots," *Technometrics*, 26(3), 243–250.

Ghosal, S., A. Sen, and A. W. v. d. Vaart (2000): "Testing Monotonicity of Regression," *The Annals of Statistics*, 28(4), 1054–1082.

Gijbels, I. (2004): "Monotone Regression," in *Encyclopedia of Statistical Sciences*. John Wiley and Sons, Inc.

Grasmair, M., O. Scherzer, and A. Vanhems (2013): "Nonparametric instrumental regression with non-convex constraints," *Inverse Problems*, 29(3), 1–16.

Hadamard, J. (1923): *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Yale University Press, New Haven.

Hall, P., and J. L. Horowitz (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *The Annals of Statistics*, 33(6), 2904–2929.

Hall, P., and L.-S. Huang (2001): "Nonparametric kernel regression subject to monotonicity constraints," *The Annals of Statistics*, 29(3), 624–647.

Hausman, J. A., and W. K. Newey (1995): "Nonparametric Estimation of Exact Consumers Surplus and Deadweight Loss," *Econometrica*, 63(6), 1445–1476.

Horowitz, J. L. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79(2), 347–394.

——— (2012): "Specification Testing in Nonparametric Instrumental Variable Estimation," *Journal of Econometrics*, 167(2), 383–396.

——— (2014): "Ill-Posed Inverse Problems in Economics," *Annual Review of Economics*, 6, 21–51.

Horowitz, J. L., and S. Lee (2012): "Uniform confidence bands for functions estimated nonparametrically with instrumental variables," *Journal of Econometrics*, 168(2), 175–188.

Horowitz, J. L., and V. G. Spokoiny (2001): "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative," *Econometrica*, 69(3), 599–631.

Imbens, G. W. (2007): "Nonadditive Models with Endogenous Regressors," in *Advances in Economics and Econometrics*, ed. by R. Blundell, W. Newey, and T. Persson, vol. 3, pp. 17–46. Cambridge University Press.

Imbens, G. W., and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77(5), 1481–1512.

Kasy, M. (2011): "Identification in Triangular Systems Using Control Functions," *Econometric Theory*, 27, 663–671.

——— (2014): "Instrumental variables with unrestricted heterogeneity and continuous treatment," *The Review of Economic Studies*, forthcoming.

Lee, S., O. Linton, and Y.-J. Whang (2009): "Testing for Stochastic Monotonicity," *Econometrica*, 77(2), 585–602.

Lee, S., K. Song, and Y.-J. Whang (2014): "Testing for a general class of functional inequalities," Working Paper CWP 09/14, cemmap.

Mammen, E. (1991): "Estimating a Smooth Monotone Regression Function," *The Annals of Statistics*, 19(2), 724–740.

Mammen, E., J. S. Marron, B. A. Turlach, and M. P. Wand (2001): "A General Projection Framework for Constrained Smoothing," *Statistical Science*, 16(3), 232–248.

Mammen, E., and C. Thomas-Agnan (1999): "Smoothing Splines and Shape Restrictions," *Scandinavian Journal of Statistics*, 26(2), 239–252.

Manski, C. F., and J. V. Pepper (2000): "Monotone Instrumental Variables: With an Application to the Returns to Schooling," *Econometrica*, 68(4), 997–1010.

Matzkin, R. L. (1994): "Restrictions of Economic Theory in Nonparametric Methods," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2523–2558. Elsevier Science B.V.

Mikusheva, A. (2007): "Uniform Inference in Autoregressive Models," *Econometrica*, 75(5), 1411–1452.

Mukerjee, H. (1988): "Monotone Nonparametric Regression," *The Annals of Statistics*, 16(2), 741–750.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565–1578.

Newey, W. K., J. L. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67(3), 565–603.

Piterbarg, V. (1996): *Asymptotic Methods in the Theory of Gaussian Processes and Fields*. American Mathematical Society, Providence, RI.

Powers, V., and B. Reznick (2000): "Polynomials That Are Positive on an Interval," *Transactions of the American Mathematical Society*, 352(10), 4677–4692.

Ramsay, J. O. (1988): "Monotone Regression Splines in Action," *Statistical Science*, 3(4), 425–441.

——— (1998): "Estimating smooth monotone functions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2), 365–375.

Reznick, B. (2000): "Some Concrete Aspects of Hilbert's 17th Problem," in *Contemporary Mathematics*, vol. 253, pp. 251–272. American Mathematical Society.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954.

Santos, A. (2012): "Inference in Nonparametric Instrumental Variables With Partial Identification," *Econometrica*, 80(1), 213–275.

Stroock, D. W. (1999): *A Concise introduction to the theory of integration*. Birkhäuser, 3rd edn.

Tropp, J. A. (2012): *User-friendly tools for random matrices: an introduction*.

van de Geer, S. (2000): *Empirical Processes in M-Estimation*. Cambridge University Press, 1st edn.

Wright, F. T. (1981): "The Asymptotic Behavior of Monotone Regression Estimates," *The Annals of Statistics*, 9(2), 443–448.

Yatchew, A. (1998): "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 36(2), 669–721.

Zhang, C.-H. (2002): "Risk Bounds in Isotonic Regression," *Annals of Statistics*, 30(2), 528–555.

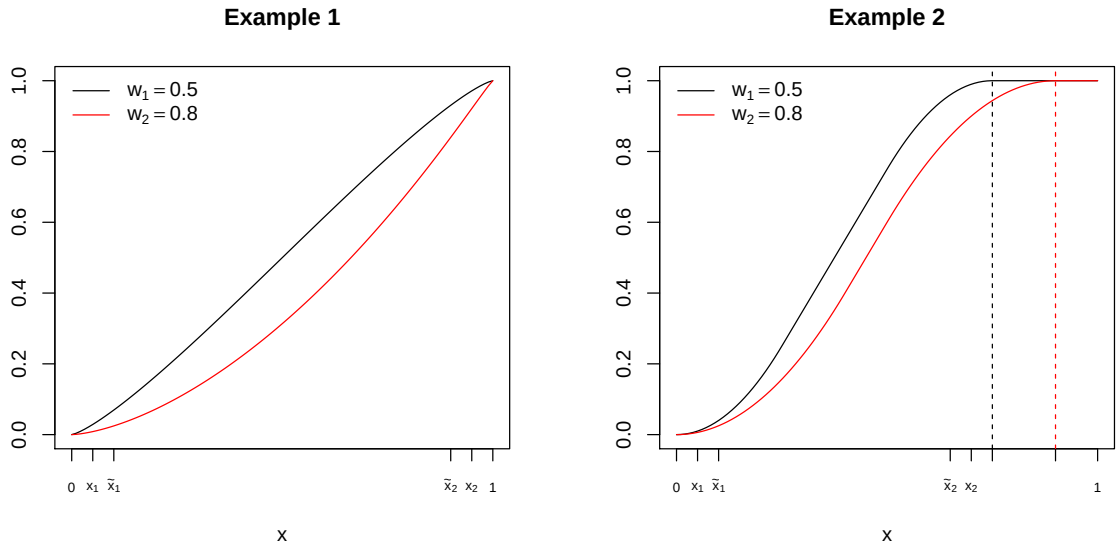

Figure 1: Plots of $F_{X|W}(x|w_1)$ and $F_{X|W}(x|w_2)$ in Examples 1 and 2, respectively.

				Model 1					
				$\kappa = 1$		$\kappa = 0.5$		$\kappa = 0.1$	
σ_ε	k_X	k_W		unrestr.	restr.	unrestr.	restr.	unrestr.	restr.
0.1	2	3	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
			var	0.021	0.007	0.005	0.002	0.000	0.000
			MSE	0.021	0.009	0.005	0.002	0.000	0.000
			MSE ratio		0.406		0.409		0.347
	2	5	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
			var	0.009	0.004	0.002	0.001	0.000	0.000
			MSE	0.009	0.005	0.002	0.001	0.000	0.000
			MSE ratio		0.529		0.510		0.542
	3	4	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
			var	0.026	0.009	0.005	0.002	0.000	0.000
			MSE	0.026	0.009	0.005	0.002	0.000	0.000
			MSE ratio		0.355		0.412		0.372
	3	7	bias sq.	0.000	0.000	0.000	0.000	0.000	0.000
			var	0.013	0.005	0.003	0.001	0.000	0.000
			MSE	0.013	0.005	0.003	0.001	0.000	0.000
			MSE ratio		0.405		0.486		0.605
	5	8	bias sq.	0.000	0.000	0.000	0.000	0.000	0.000
			var	0.027	0.007	0.005	0.002	0.000	0.000
			MSE	0.027	0.007	0.005	0.002	0.000	0.000
			MSE ratio		0.266		0.339		0.411
0.7	2	3	bias sq.	0.001	0.020	0.000	0.005	0.000	0.000
			var	0.857	0.097	0.263	0.024	0.012	0.001
			MSE	0.857	0.118	0.263	0.029	0.012	0.001
			MSE ratio		0.137		0.110		0.101
	2	5	bias sq.	0.001	0.015	0.000	0.004	0.000	0.000
			var	0.419	0.080	0.102	0.020	0.004	0.001
			MSE	0.420	0.095	0.102	0.024	0.004	0.001
			MSE ratio		0.227		0.235		0.221
	3	4	bias sq.	0.001	0.016	0.000	0.004	0.000	0.000
			var	0.763	0.104	0.223	0.026	0.010	0.001
			MSE	0.763	0.121	0.223	0.030	0.010	0.001
			MSE ratio		0.158		0.133		0.119
	3	7	bias sq.	0.001	0.011	0.000	0.003	0.000	0.000
			var	0.350	0.083	0.104	0.020	0.004	0.001
			MSE	0.351	0.094	0.104	0.023	0.004	0.001
			MSE ratio		0.267		0.218		0.229
	5	8	bias sq.	0.001	0.011	0.000	0.003	0.000	0.000
			var	0.433	0.094	0.131	0.023	0.006	0.001
			MSE	0.434	0.105	0.131	0.025	0.006	0.001
			MSE ratio		0.243		0.193		0.170

Table 1: Model 1: Performance of the unrestricted and restricted estimators for $N = 500$, $\rho = 0.3$, $\eta = 0.3$.

				Model 2					
				$\kappa = 1$		$\kappa = 0.5$		$\kappa = 0.1$	
σ_ε	k_X	k_W		unrestr.	restr.	unrestr.	restr.	unrestr.	restr.
0.1	2	3	bias sq.	0.001	0.002	0.000	0.001	0.000	0.000
			var	0.024	0.003	0.007	0.001	0.000	0.000
			MSE	0.024	0.006	0.007	0.001	0.000	0.000
			MSE ratio		0.229		0.201		0.222
	2	5	bias sq.	0.001	0.002	0.000	0.000	0.000	0.000
			var	0.010	0.002	0.002	0.001	0.000	0.000
			MSE	0.011	0.004	0.002	0.001	0.000	0.000
			MSE ratio		0.405		0.475		0.446
	3	4	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
			var	0.022	0.003	0.006	0.001	0.000	0.000
			MSE	0.022	0.004	0.006	0.001	0.000	0.000
			MSE ratio		0.192		0.176		0.157
	3	7	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
			var	0.009	0.002	0.002	0.001	0.000	0.000
			MSE	0.009	0.003	0.002	0.001	0.000	0.000
			MSE ratio		0.325		0.292		0.323
	5	8	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
			var	0.014	0.003	0.003	0.001	0.000	0.000
			MSE	0.014	0.004	0.003	0.001	0.000	0.000
			MSE ratio		0.269		0.268		0.217
0.7	2	3	bias sq.	0.002	0.005	0.001	0.001	0.000	0.000
			var	1.102	0.032	0.321	0.008	0.012	0.000
			MSE	1.104	0.038	0.321	0.009	0.012	0.000
			MSE ratio		0.034		0.029		0.032
	2	5	bias sq.	0.001	0.006	0.000	0.002	0.000	0.000
			var	0.462	0.031	0.103	0.008	0.004	0.000
			MSE	0.463	0.037	0.104	0.009	0.004	0.000
			MSE ratio		0.080		0.088		0.088
	3	4	bias sq.	0.001	0.004	0.000	0.001	0.000	0.000
			var	0.936	0.036	0.255	0.009	0.012	0.000
			MSE	0.936	0.040	0.255	0.010	0.012	0.000
			MSE ratio		0.043		0.039		0.034
	3	7	bias sq.	0.001	0.005	0.000	0.001	0.000	0.000
			var	0.387	0.035	0.110	0.009	0.004	0.000
			MSE	0.388	0.040	0.110	0.010	0.004	0.000
			MSE ratio		0.103		0.089		0.092
	5	8	bias sq.	0.002	0.005	0.000	0.001	0.000	0.000
			var	0.508	0.041	0.144	0.010	0.007	0.000
			MSE	0.510	0.046	0.144	0.011	0.007	0.000
			MSE ratio		0.090		0.078		0.065

Table 2: Model 2: Performance of the unrestricted and restricted estimators for $N = 500$, $\rho = 0.3$, $\eta = 0.3$.

			Model 1					
			$\kappa = 1$		$\kappa = 0.5$		$\kappa = 0.1$	
ρ	η		unrestr.	restr.	unrestr.	restr.	unrestr.	restr.
0.3	0.3	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
		var	0.026	0.009	0.005	0.002	0.000	0.000
		MSE	0.026	0.009	0.005	0.002	0.000	0.000
		MSE ratio		0.355		0.412		0.372
0.3	0.7	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
		var	0.026	0.008	0.005	0.002	0.000	0.000
		MSE	0.026	0.009	0.005	0.002	0.000	0.000
		MSE ratio		0.342		0.395		0.449
0.7	0.3	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
		var	0.025	0.002	0.003	0.001	0.000	0.000
		MSE	0.025	0.003	0.003	0.001	0.000	0.000
		MSE ratio		0.125		0.248		0.266
0.7	0.7	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
		var	0.023	0.002	0.004	0.001	0.000	0.000
		MSE	0.023	0.003	0.004	0.001	0.000	0.000
		MSE ratio		0.136		0.212		0.259

Table 3: Model 1: Performance of the unrestricted and restricted estimators for $\sigma_\varepsilon = 0.1$, $k_X = 3$, $k_W = 4$, $N = 500$.

			Model 2					
			$\kappa = 1$		$\kappa = 0.5$		$\kappa = 0.1$	
ρ	η		unrestr.	restr.	unrestr.	restr.	unrestr.	restr.
0.3	0.3	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
		var	0.022	0.003	0.006	0.001	0.000	0.000
		MSE	0.022	0.004	0.006	0.001	0.000	0.000
		MSE ratio		0.192		0.176		0.157
0.3	0.7	bias sq.	0.000	0.001	0.000	0.000	0.000	0.000
		var	0.020	0.003	0.006	0.001	0.000	0.000
		MSE	0.020	0.004	0.006	0.001	0.000	0.000
		MSE ratio		0.209		0.163		0.160
0.7	0.3	bias sq.	0.000	0.000	0.000	0.000	0.000	0.000
		var	0.013	0.000	0.002	0.000	0.000	0.000
		MSE	0.013	0.001	0.002	0.000	0.000	0.000
		MSE ratio		0.040		0.063		0.047
0.7	0.7	bias sq.	0.000	0.000	0.000	0.000	0.000	0.000
		var	0.010	0.000	0.002	0.000	0.000	0.000
		MSE	0.011	0.001	0.002	0.000	0.000	0.000
		MSE ratio		0.051		0.060		0.050

Table 4: Model 2: Performance of the unrestricted and restricted estimators for $\sigma_\varepsilon = 0.1$, $k_X = 3$, $k_W = 4$, $N = 500$.

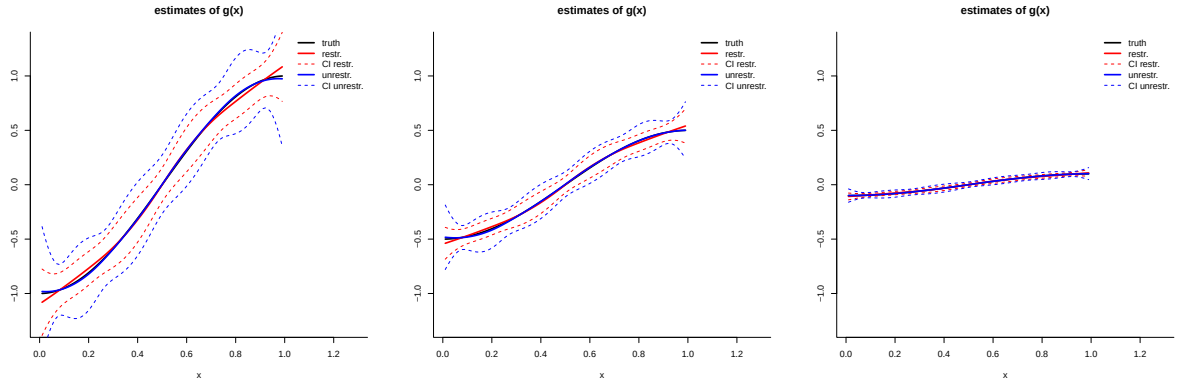

Figure 2: Model 1: unrestricted and restricted estimates of $g(x)$ for $N = 500$, $\rho = 0.3$, $\eta = 0.3$, $\sigma_\varepsilon = 0.1$, $k_X = 3$, $k_W = 4$.

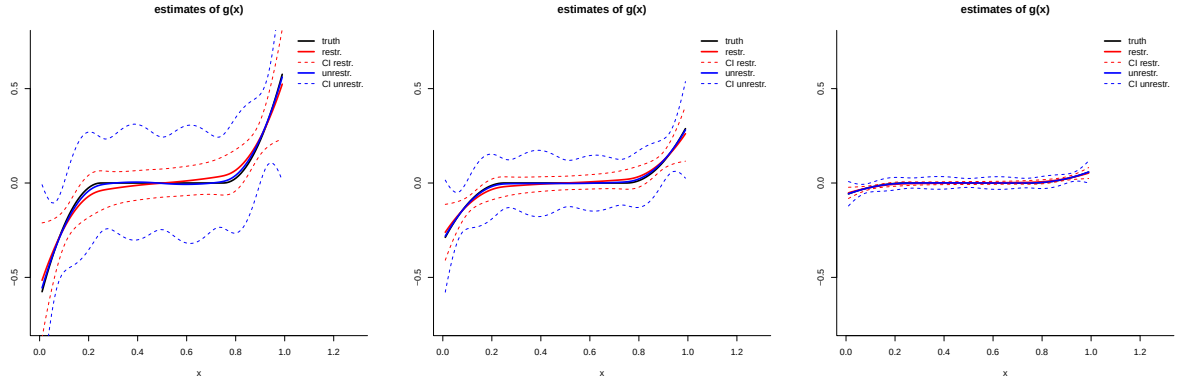

Figure 3: Model 2: unrestricted and restricted estimates of $g(x)$ for $N = 500$, $\rho = 0.3$, $\eta = 0.3$, $\sigma_\varepsilon = 0.1$, $k_X = 3$, $k_W = 4$.

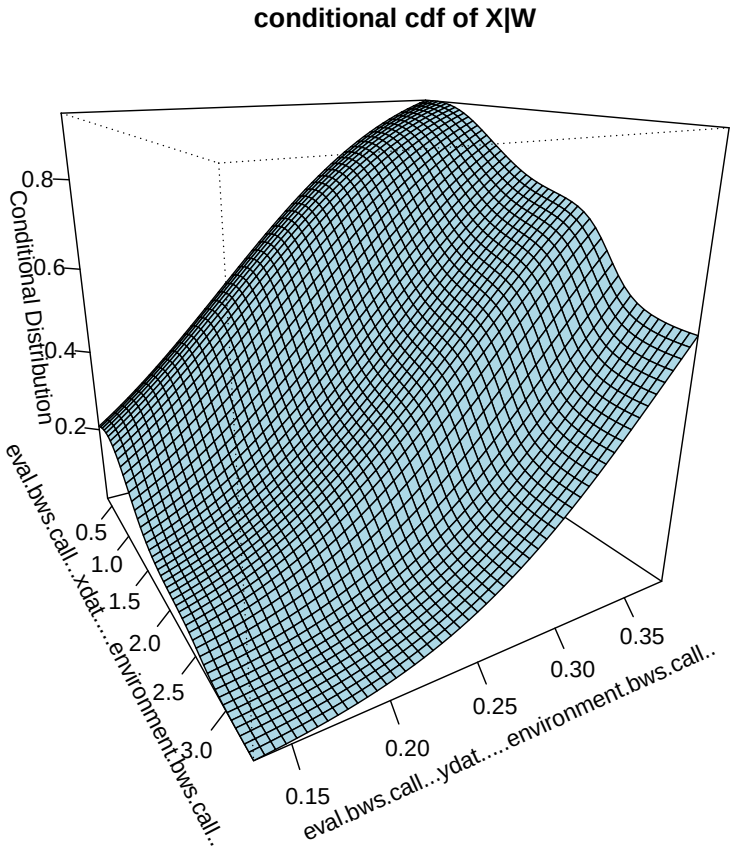

Figure 4: Nonparametric kernel estimate of the conditional cdf $F_{X|W}(x|w)$.

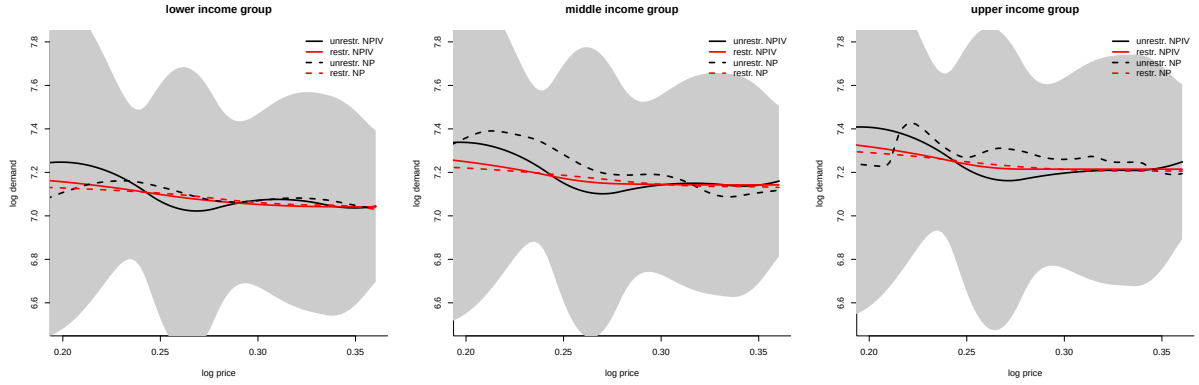

Figure 5: Estimates of $g(x, z_1)$ plotted as a function of price x for z_1 fixed at three income levels.