# Supplement to “Nonparametric Instrumental Variable Estimation Under Monotonicity”

Denis Chetverikov* Daniel Wilhelm†

**Abstract**

This supplement provides proofs, additional results, more detailed discussions, an empirical application, and additional simulations referenced in the main text.

## Contents



*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) and the European Research Council (ERC-2014-CoG-646917-ROMIA and ERC-2015-CoG-682349).

# A Discussion of Theorem 1

## A.1 Examples

We provide two examples of distributions of $(X, W)$ that satisfy Assumptions 1 and 2, and show two possible ways in which the instrument $W$ can shift the conditional distribution of $X$ given $W$. Figure 1 displays the corresponding conditional distributions.

**Example A.1** (Normal density)**.** Let $(\widetilde{X}, \widetilde{W})$ be jointly normal with mean zero, variance one, and correlation $0 < \rho < 1$. Let $\Phi(\cdot)$ denote the distribution function of a $N(0,1)$ 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 $(X, W)$ satisfies condition (7) of our monotone IV Assumption 1. Lemma G.3 below verifies that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied. □

**Example A.2** (Two-dimensional unobserved heterogeneity)**.** 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 $W \sim U[0,1]$. Since $U_2$ is positive, it is straightforward to see that the stochastic dominance condition (7) is satisfied. Lemma G.4 below shows that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied. □

Figure 1 shows that, in Example A.1, the conditional distribution at two different values of the instrument is shifted to the right at every value of $X$, whereas, in Example A.2, the conditional support of $X$ given $W = w$ changes with $w$, but the positive shift in the cdf of $X|W = w$ occurs only for values of $X$ in a subinterval of $[0,1]$.

## A.2 Remarks

Consider the linear equation (5). By Assumption 2(i), the operator $T$ is compact, and so

$$\frac{\|h_k\|_2}{\|Th_k\|_2} \to \infty \text{ as } k \to \infty \text{ for some sequence } \{h_k, k \geq 1\} \subset L^2[0,1]. \tag{19}$$

Property (19) 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, (5) is ill-posed in Hadamard's sense[1]. Lemma F.1, on the other hand, implies that, under Assumptions 1 and 2, (19) is not possible if $h_k$ belongs to the set $\mathcal{M}$ of monotone functions

[1] Well- and ill-posedness in Hadamard's sense are defined as follows. Let $A : D \to R$ be a continuous mapping between metric spaces $(D, \rho_D)$ and $(R, \rho_R)$. 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.

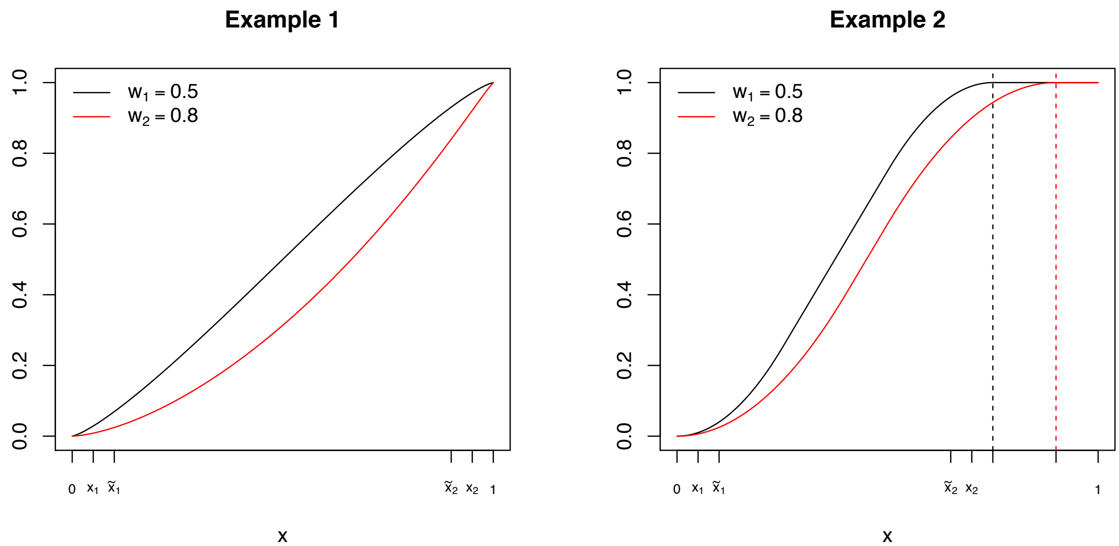


**Figure 1:** Plots of $F_{X|W}(x|w_1)$ and $F_{X|W}(x|w_2)$ in Examples A.1 and A.2, respectively.

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 (19) by the truncated $L^2$-norm $\|\cdot\|_{2,t}$.

In Remark A.1, we show that truncating the norm in the numerator is not a significant modification in the sense that for most ill-posed problems, and in particular for all severely ill-posed problems, (19) holds even if we replace the $L^2$-norm $\|\cdot\|_2$ in the numerator of the left-hand side of (19) by the truncated $L^2$-norm $\|\cdot\|_{2,t}$.

**Remark A.1** (Ill-posedness is preserved by norm truncation)**.** Under Assumptions 1 and 2, the integral operator $T$ satisfies (19). Here we demonstrate that, in many cases, and in particular in all severely ill-posed cases, (19) 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 (19), that is, there exists a sequence $\{l_k, k \geq 1\}$ in $L^2[0,1]$ such that

$$\frac{\|l_k\|_{2,t}}{\|Tl_k\|_2} \to \infty \text{ as } k \to \infty. \tag{20}$$

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 G.2 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 (20). 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 Lemma F.1 that $h$ belongs to the space $\mathcal{M}$ of *monotone* functions in $L^2[0,1]$ plays a crucial role for the result (25) to hold. On the other hand, whether the result (25) can be obtained for all $h \in \mathcal{M}$ without imposing our monotone IV Assumption 1 appears to be an open (and interesting) question. □

**Remark A.2** (Severe ill-posedness is preserved by norm truncation)**.** One might wonder whether our monotone IV Assumption 1 excludes all severely ill-posed problems, and whether the norm truncation significantly changes these problems. Here we show that there do exist severely ill-posed problems that satisfy our monotone IV Assumption 1, and also that severely ill-posed problems remain severely ill-posed even if we replace the $L^2$-norm $\|\cdot\|_2$ by the truncated $L^2$-norm $\|\cdot\|_{2,t}$. Indeed, consider Example A.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. More specifically, 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 G.2, 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 Lemma F.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 A.3** (Monotone IV assumption does not imply control function approach)**.** Our monotone IV Assumption 1 does not imply the applicability of a control function approach to the estimation of the function $g$. Consider Example A.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 $\varepsilon$ conditional on $V = C(X, W)$. As a consequence, our monotone IV 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.[2] Since multidimensional unobserved heterogeneity is

[2]It is easy to show that the existence of a control function does not imply our monotone IV condition either, so our and the control function approach rely on conditions that are non-nested.

common in economic applications (see Imbens (2007), Kasy (2014), and Hoderlein, Holzmann, Kasy, and Meister (2016)), we view our approach to avoiding ill-posedness as complementary to the control function approach. □

# B Discussion of Theorem 2

## B.1 Asymptotic Equivalence of Estimators for Strictly Increasing Functions

If the function $g$ is strictly increasing and the sample size $n$ is sufficiently large, then the constrained estimator $\widehat{g}^c$ coincides with the unconstrained estimator $\widehat{g}^u$, and the two estimators therefore share the same rate of convergence.

**Lemma B.1** (Asymptotic equivalence of constrained and unconstrained estimators). *Let Assumptions 2 and 4-8 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 we have $\tau_n^2 \xi_n^2 \log n/n \to 0$, $\sup_{x\in[0,1]} \|Dp(x)\|(\tau_n(K/n)^{1/2} + K^{-s}) \to 0$, and $\sup_{x\in[0,1]} |Dg(x) - Dg_n(x)| \to 0$ as $n \to \infty$, then*

$$\mathrm{P}\Big(\widehat{g}^c(x) = \widehat{g}^u(x) \text{ for all } x \in [0,1]\Big) \to 1 \text{ as } n \to \infty. \tag{21}$$

The result in Lemma B.1 is similar to that in Theorem 1 of Mammen (1991), which shows equivalence, in the sense of (21), of the constrained and unconstrained estimators of conditional mean functions. Lemma B.1 implies that imposing the monotonicity constraint cannot lead to improvements in the rate of convergence of the estimator if $g$ is strictly increasing. However, the result in Lemma B.1 is asymptotic and does not rule out significant performance gains in finite samples, which we find in our simulations and in Theorem 2.

## B.2 Fast Convergence Rate Under Local-to-Flat Asymptotics

**Corollary B.1** (Fast convergence rate of the constrained estimator under local 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 $\|Dg\|_\infty = O((K \log n/n)^{1/2})$, then*

$$\|\widehat{g}^c - g\|_{2,t} = O_p((K \log n/n)^{1/2} + K^{-s}). \tag{22}$$

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

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

This result means 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 $\|Dg\|_\infty = O((K \log n/n)^{1/2})$ 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 function $g$ is not too steep relative to the sample size.

## B.3 Remarks

**Remark B.1** (On robustness of the constrained estimator)**.** Notice that the fast convergence rates in the local asymptotics derived in Corollary B.1 are obtained under two monotonicity conditions, Assumptions 1 and 3, but the estimator imposes only the monotonicity of the regression function, not that of the first stage. Therefore, our proposed constrained estimator consistently estimates the regression function $g$ even when the monotone IV assumption is violated. □

**Remark B.2** (Imposing Monotonicity by Re-arrangement)**.** Chernozhukov, Fernández-Val, and Galichon (2009) show that re-arranging any unconstrained estimator so that it becomes monotone decreases the estimation error of the estimator. However, their argument does not quantify this improvement, so that it does not seem possible to conclude from their argument whether and when the improvement is large, even qualitatively. In contrast, the main contribution of our paper is to show that enforcing the monotonicity constraint in the NPIV model yields substantial performance improvements even in large samples and for steep regression functions $g$ as long as the NPIV model is severely ill-posed. □

**Remark B.3** (Estimating partially flat functions)**.** 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 B.4** (Computational aspects)**.** The implementation of the constrained estimator in (14) 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. The estimator $\widehat{\beta}^c$ 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, 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 $f(x) \geq 0$ can then be written as $\widetilde{p}(x)'M\widetilde{p}(x) \geq 0$ for some matrix $M$ that depends on $b$. This condition is equivalent to requiring the matrix

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

$M$ to be positive semi-definite. $\widehat{\beta}^c$ 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 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 B.5** (Penalization and shape constraints)**.** Recall that the estimators $\widehat{g}^u$ and $\widehat{g}^c$ require setting the constraint $\|b\| \leq C_b$ in the optimization problems (13) and (14). In practice, this constraint, or similar constraints in terms of Sobolev norms, which also impose bounds on derivatives of $g$, are typically not enforced in the implementation of 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) and Gagliardini and Scaillet (2012), for example, argue that penalizing the norm of $g$ and of its derivatives may stabilize the estimator by reducing its variance. In this sense, penalizing the norm of $g$ and of its derivatives may have a similar effect as imposing monotonicity. However, there are at least two important differences between penalization and imposing monotonicity. First, penalization increases bias of the estimators. In fact, especially in severely ill-posed problems, even small amount of penalization may lead to large bias (otherwise severely ill-posed problems could lead to estimators with fast convergence rates). In contrast, the monotonicity constraint on the estimator does not increase bias much when the function $g$ itself satisfies the monotonicity constraint. Second, penalization requires selecting a tuning parameter that governs the strength of penalization, which is a difficult statistical problem. In contrast, imposing monotonicity does not require such choices and can often be motivated directly from economic theory. □

# C Identification Bounds under Monotonicity

In Section 3, we derived non-asymptotic error bounds on the constrained estimator in the NPIV model (1) assuming that $g$ is point-identified, or equivalently, that the linear operator $T$ is invertible. Newey and Powell (2003) linked point-identification of $g$ to completeness of the conditional distribution of $X$ given $W$, but this completeness condition has been argued to be strong (Santos (2012)) and non-testable (Canay, Santos, and Shaikh (2013)). In this section, we therefore discard the completeness condition and explore the identification power of our monotonicity conditions, which appear natural in many economic applications.

By a slight abuse of notation, we define the sign of the slope of a differentiable, monotone function $f \in \mathcal{M}$ by

$$sign(Df) := \begin{cases} 1, & Df(x) \geq 0 \, \forall x \in [0,1] \text{ and } Df(x) > 0 \text{ for some } x \in [0,1] \\ 0, & Df(x) = 0 \, \forall x \in [0,1] \\ -1, & Df(x) \leq 0 \, \forall x \in [0,1] \text{ and } Df(x) < 0 \text{ for some } x \in [0,1] \end{cases}$$

and the sign of a scalar $b$ by $sign(b) := 1\{b > 0\} - 1\{b < 0\}$. We first show that if the function $g$ is monotone, the sign of its slope is identified under our monotone IV assumption (and some other technical conditions):

**Theorem C.1** (Identification of the sign of the slope)**.** *Suppose Assumptions 1 and 2 hold and $f_{X,W}(x,w) > 0$ for all $(x,w) \in (0,1)^2$. If $g$ is monotone and continuously differentiable, then $sign(Dg)$ is identified.*

This theorem shows that, under certain regularity conditions, the monotone IV assumption and monotonicity of the regression function $g$ imply identification of the sign of the regression function's slope, even though the regression function itself is, in general, not point-identified. This result is useful because in many empirical applications it is natural to assume a monotone relationship between outcome variable $Y$ and the endogenous regressor $X$, 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. The discussions in Abrevaya, Hausman, and Khan (2010) and Kline (2016), for example, provide examples and motivations for why one may be interested in the sign of a marginal effect.

**Remark C.1** (A test for the sign of the slope of $g$)**.** In fact, Theorem C.1 yields a surprisingly simple way to test the sign of the slope of the function $g$. Indeed, the proof of Theorem C.1 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 G.1), the latter assertions are equivalent to the coefficient $\beta$ in the linear regression model

$$Y = \alpha + \beta W + U, \;\; \mathrm{E}[UW] = 0 \tag{23}$$

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

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

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 the corresponding hypotheses about the sign of the slope coefficient $\beta$ in the linear regression model (23). In particular, under our two monotonicity assumptions, one can test the hypothesis of "no effect" of $X$ on $Y$, i.e. that $g$ is a constant, by testing whether $\beta = 0$ or not using the usual t-statistic. The asymptotic theory for this statistic is exactly the same as in the standard regression case with exogenous regressors, yielding the standard normal limiting distribution and, therefore, completely avoiding the ill-posed inverse problem of recovering $g$. □

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

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

The following theorem provides necessary conditions for observational equivalence.

**Theorem C.2** (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 $C_1 := (x_2 - x_1)^{1/2} \,/ \min\{x_1 - \delta_2, 1 - \delta_2 - x_2\}$ and $c_p := \min\{1 - w_2, w_1\}\min\{(C_F - 1)/2, 1\}c_w c_f/4$. 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.*

Under Assumption 3 that $g$ is increasing, Theorem C.2 suggests the construction of a set $\Theta'$ that includes the identified set $\Theta$ by $\Theta' := \mathcal{M}_+ \backslash \Delta$, where $\mathcal{M}_+ := \mathcal{H}(0)$ denotes all increasing functions in $\mathcal{M}$ and

$$\Delta := \Big\{ g' \in \mathcal{M}_+ : \text{there exists } h \in L^2[0,1] \text{ such that} \\ g' - g + h \in \mathcal{M} \;\text{ and }\; \|h\|_{2,t} + \bar{C}\|T\|_2\|h\|_2 < \|g' - g\|_{2,t} \Big\}. \quad (24)$$

We emphasize that $\Delta$ is not empty, which means that our Assumptions 1–3 possess identifying power leading to nontrivial bounds on $g$. Notice that the constant $\bar{C}$ depends only on the estimable quantities $c_w$, $c_f$, and $C_F$ from Assumptions 1–2, and on the known constants $x_1$, $x_2$, $w_1$, $w_2$, and $\delta_2$. Therefore, the set $\Theta'$ could, in principle, be estimated.

**Remark C.2** (Further insight on identification bounds)**.** It is possible to provide more insight into which functions are in $\Delta$ and thus not in $\Theta'$. First, under the additional minor condition that $f_{X,W}(x,w) > 0$ for all $(x,w) \in (0,1)^2$, all functions in $\Theta'$ have to intersect $g$; otherwise they are not observationally equivalent to $g$. 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 (24) is satisfied if $\|h\|_2$ is not too large relative to $\|g' - g\|_{2,t}$. In the extreme case, setting $h = 0$ shows that $\Theta'$ does not contain elements $g'$ that disagree with $g$ on $[x_1, x_2]$ and such that $g' - g$ is monotone. More generally, $\Theta'$ does not contain elements $g'$ whose difference with $g$ is too close to a monotone function. Therefore, for example, functions $g'$ that are much steeper than $g$ are excluded from $\Theta'$. □

# D Additional Simulations

In this section, we report simulations for other choices of the parameters $\rho$ and $\eta$, which govern the strength of the instrument and the degree of endogeneity, respectively. We vary the parameters by setting $\rho = 0.3$ or $0.5$ and $\eta = 0.3$ or $0.7$. Tables 1–6 show the results.

| | Constrained estimator | | | | |
|---|---|---|---|---|---|
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 7.33 | 4.46 | 13.19 | 17.54 | 4.46 |
| $n=1,000$ | 6.46 | 2.07 | 9.40 | 11.33 | 2.07 |
| $n=5,000$ | 5.74 | 0.50 | 2.97 | 5.36 | 0.50 |
| $n=10,000$ | 5.66 | 0.34 | 1.48 | 3.70 | 0.34 |
| $n=50,000$ | 5.58 | 0.09 | 0.38 | 2.39 | 0.09 |
| $n=100,000$ | 5.57 | 0.04 | 0.19 | 1.98 | 0.04 |
| $n=500,000$ | 5.56 | 0.01 | 0.08 | 0.48 | 0.01 |
| | Unconstrained estimator | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 7.33 | 10.40 | 105.00 | 573.43 | 7.33 |
| $n=1,000$ | 6.46 | 4.55 | 58.69 | 354.45 | 4.55 |
| $n=5,000$ | 5.74 | 0.92 | 8.80 | 152.28 | 0.92 |
| $n=10,000$ | 5.66 | 0.45 | 4.09 | 96.93 | 0.45 |
| $n=50,000$ | 5.58 | 0.09 | 1.10 | 22.50 | 0.09 |
| $n=100,000$ | 5.57 | 0.04 | 0.48 | 10.35 | 0.04 |
| $n=500,000$ | 5.56 | 0.01 | 0.08 | 1.71 | 0.01 |
| | Ratio | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 1.00 | 0.43 | 0.13 | 0.03 | 0.61 |
| $n=1,000$ | 1.00 | 0.45 | 0.16 | 0.03 | 0.45 |
| $n=5,000$ | 1.00 | 0.54 | 0.34 | 0.04 | 0.54 |
| $n=10,000$ | 1.00 | 0.75 | 0.36 | 0.04 | 0.75 |
| $n=50,000$ | 1.00 | 1.00 | 0.34 | 0.11 | 1.00 |
| $n=100,000$ | 1.00 | 1.00 | 0.40 | 0.19 | 1.00 |
| $n=500,000$ | 1.00 | 1.00 | 0.96 | 0.28 | 1.00 |

**Table 1:** simulation results for the case $g(x) = x^2 + 0.2x$, $\rho = 0.5$, and $\eta = 0.3$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of $n$ and $K$. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of $n$ and $K$. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over $K$. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function $n$ and $K$. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

| | Constrained estimator | | | | |
|---|---|---|---|---|---|
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 9.32 | 13.06 | 26.14 | 39.35 | 9.32 |
| $n=1,000$ | 7.48 | 6.32 | 23.72 | 34.13 | 6.32 |
| $n=5,000$ | 5.94 | 1.40 | 10.65 | 13.77 | 1.40 |
| $n=10,000$ | 5.75 | 0.81 | 8.62 | 9.85 | 0.81 |
| $n=50,000$ | 5.59 | 0.29 | 2.51 | 5.36 | 0.29 |
| $n=100,000$ | 5.57 | 0.26 | 1.46 | 4.48 | 0.26 |
| $n=500,000$ | 5.56 | 0.05 | 0.36 | 3.16 | 0.05 |
| | Unconstrained estimator | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 9.32 | 61.71 | 499.02 | 1036.88 | 9.32 |
| $n=1,000$ | 7.48 | 31.47 | 325.31 | 904.59 | 7.48 |
| $n=5,000$ | 5.94 | 6.23 | 168.48 | 516.09 | 5.94 |
| $n=10,000$ | 5.75 | 2.72 | 107.17 | 574.64 | 2.72 |
| $n=50,000$ | 5.59 | 0.52 | 26.89 | 320.54 | 0.52 |
| $n=100,000$ | 5.57 | 0.29 | 11.77 | 229.12 | 0.29 |
| $n=500,000$ | 5.56 | 0.05 | 1.98 | 129.61 | 0.05 |
| | Ratio | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 1.00 | 0.21 | 0.05 | 0.04 | 1.00 |
| $n=1,000$ | 1.00 | 0.20 | 0.07 | 0.04 | 0.85 |
| $n=5,000$ | 1.00 | 0.22 | 0.06 | 0.03 | 0.24 |
| $n=10,000$ | 1.00 | 0.30 | 0.08 | 0.02 | 0.30 |
| $n=50,000$ | 1.00 | 0.56 | 0.09 | 0.02 | 0.56 |
| $n=100,000$ | 1.00 | 0.92 | 0.12 | 0.02 | 0.92 |
| $n=500,000$ | 1.00 | 1.00 | 0.18 | 0.02 | 1.00 |

**Table 2:** simulation results for the case $g(x) = x^2 + 0.2x$, $\rho = 0.3$, and $\eta = 0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of $n$ and $K$. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of $n$ and $K$. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over $K$. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function $n$ and $K$. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

| | Constrained estimator | | | | |
|---|---|---|---|---|---|
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 7.38 | 4.37 | 12.84 | 17.90 | 4.37 |
| $n=1,000$ | 6.45 | 2.13 | 9.07 | 11.33 | 2.13 |
| $n=5,000$ | 5.74 | 0.51 | 2.99 | 5.24 | 0.51 |
| $n=10,000$ | 5.65 | 0.34 | 1.66 | 3.79 | 0.34 |
| $n=50,000$ | 5.57 | 0.08 | 0.34 | 2.44 | 0.08 |
| $n=100,000$ | 5.57 | 0.05 | 0.20 | 1.95 | 0.05 |
| $n=500,000$ | 5.56 | 0.01 | 0.08 | 0.40 | 0.01 |
| | Unconstrained estimator | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 7.38 | 10.21 | 110.96 | 552.21 | 7.38 |
| $n=1,000$ | 6.45 | 4.76 | 61.47 | 403.53 | 4.76 |
| $n=5,000$ | 5.74 | 0.93 | 7.97 | 167.60 | 0.93 |
| $n=10,000$ | 5.65 | 0.44 | 4.64 | 107.45 | 0.44 |
| $n=50,000$ | 5.57 | 0.08 | 1.07 | 25.08 | 0.08 |
| $n=100,000$ | 5.57 | 0.05 | 0.46 | 9.28 | 0.05 |
| $n=500,000$ | 5.56 | 0.01 | 0.09 | 1.63 | 0.01 |
| | Ratio | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 1.00 | 0.43 | 0.12 | 0.03 | 0.59 |
| $n=1,000$ | 1.00 | 0.45 | 0.15 | 0.03 | 0.45 |
| $n=5,000$ | 1.00 | 0.54 | 0.37 | 0.03 | 0.54 |
| $n=10,000$ | 1.00 | 0.77 | 0.36 | 0.04 | 0.77 |
| $n=50,000$ | 1.00 | 1.00 | 0.32 | 0.10 | 1.00 |
| $n=100,000$ | 1.00 | 1.00 | 0.44 | 0.21 | 1.00 |
| $n=500,000$ | 1.00 | 1.00 | 0.96 | 0.25 | 1.00 |

**Table 3:** simulation results for the case $g(x) = x^2 + 0.2x$, $\rho = 0.5$, and $\eta = 0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of $n$ and $K$. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of $n$ and $K$. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over $K$. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function $n$ and $K$. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

| | Constrained estimator | | | | |
|---|---|---|---|---|---|
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 8.12 | 5.36 | 9.66 | 13.58 | 5.36 |
| $n=1,000$ | 7.26 | 2.77 | 5.67 | 9.63 | 2.77 |
| $n=5,000$ | 6.55 | 1.18 | 1.81 | 4.74 | 1.18 |
| $n=10,000$ | 6.46 | 1.03 | 1.43 | 3.51 | 1.03 |
| $n=50,000$ | 6.37 | 0.87 | 0.66 | 1.45 | 0.66 |
| $n=100,000$ | 6.36 | 0.86 | 0.42 | 0.88 | 0.42 |
| $n=500,000$ | 6.35 | 0.85 | 0.08 | 0.48 | 0.08 |
| | Unconstrained estimator | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 8.12 | 11.43 | 105.00 | 563.75 | 8.12 |
| $n=1,000$ | 7.26 | 5.55 | 58.69 | 353.28 | 5.55 |
| $n=5,000$ | 6.55 | 1.70 | 8.80 | 156.02 | 1.70 |
| $n=10,000$ | 6.46 | 1.29 | 4.09 | 95.33 | 1.29 |
| $n=50,000$ | 6.37 | 0.89 | 1.10 | 22.51 | 0.89 |
| $n=100,000$ | 6.36 | 0.84 | 0.48 | 10.10 | 0.48 |
| $n=500,000$ | 6.35 | 0.81 | 0.08 | 1.75 | 0.08 |
| | Ratio | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 1.00 | 0.47 | 0.09 | 0.02 | 0.66 |
| $n=1,000$ | 1.00 | 0.50 | 0.10 | 0.03 | 0.50 |
| $n=5,000$ | 1.00 | 0.69 | 0.21 | 0.03 | 0.69 |
| $n=10,000$ | 1.00 | 0.80 | 0.35 | 0.04 | 0.80 |
| $n=50,000$ | 1.00 | 0.98 | 0.60 | 0.06 | 0.74 |
| $n=100,000$ | 1.00 | 1.02 | 0.87 | 0.09 | 0.87 |
| $n=500,000$ | 1.00 | 1.05 | 1.00 | 0.27 | 1.00 |

**Table 4:** simulation results for the case $g(x) = 2(x-1/2)_+^2 + 0.5x$, $\rho = 0.5$, and $\eta = 0.3$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of $n$ and $K$. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of $n$ and $K$. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over $K$. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function $n$ and $K$. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

| | Constrained estimator | | | | |
|---|---|---|---|---|---|
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 10.19 | 14.21 | 18.48 | 27.35 | 10.19 |
| $n=1,000$ | 8.28 | 7.27 | 13.89 | 20.89 | 7.27 |
| $n=5,000$ | 6.76 | 2.10 | 7.48 | 10.45 | 2.10 |
| $n=10,000$ | 6.57 | 1.49 | 4.10 | 7.93 | 1.49 |
| $n=50,000$ | 6.38 | 0.95 | 1.52 | 4.94 | 0.95 |
| $n=100,000$ | 6.37 | 0.90 | 1.36 | 4.62 | 0.90 |
| $n=500,000$ | 6.36 | 0.86 | 0.67 | 2.58 | 0.67 |
| | Unconstrained estimator | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 10.19 | 60.82 | 499.02 | 1053.94 | 10.19 |
| $n=1,000$ | 8.28 | 32.95 | 325.31 | 927.74 | 8.28 |
| $n=5,000$ | 6.76 | 6.88 | 168.48 | 523.23 | 6.76 |
| $n=10,000$ | 6.57 | 3.52 | 107.17 | 577.86 | 3.52 |
| $n=50,000$ | 6.38 | 1.32 | 26.89 | 318.74 | 1.32 |
| $n=100,000$ | 6.37 | 1.11 | 11.77 | 229.59 | 1.11 |
| $n=500,000$ | 6.36 | 0.86 | 1.98 | 132.66 | 0.86 |
| | Ratio | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 1.00 | 0.23 | 0.04 | 0.03 | 1.00 |
| $n=1,000$ | 1.00 | 0.22 | 0.04 | 0.02 | 0.88 |
| $n=5,000$ | 1.00 | 0.31 | 0.04 | 0.02 | 0.31 |
| $n=10,000$ | 1.00 | 0.42 | 0.04 | 0.01 | 0.42 |
| $n=50,000$ | 1.00 | 0.72 | 0.06 | 0.02 | 0.72 |
| $n=100,000$ | 1.00 | 0.81 | 0.12 | 0.02 | 0.81 |
| $n=500,000$ | 1.00 | 1.00 | 0.34 | 0.02 | 0.79 |

**Table 5:** simulation results for the case $g(x) = 2(x-1/2)_+^2 + 0.5x$, $\rho = 0.3$, and $\eta = 0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of $n$ and $K$. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of $n$ and $K$. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over $K$. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function $n$ and $K$. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

| | Constrained estimator | | | | |
|---|---|---|---|---|---|
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 8.15 | 5.34 | 9.59 | 13.52 | 5.34 |
| $n=1,000$ | 7.26 | 2.91 | 5.96 | 9.74 | 2.91 |
| $n=5,000$ | 6.54 | 1.20 | 1.84 | 4.67 | 1.20 |
| $n=10,000$ | 6.45 | 1.04 | 1.48 | 3.75 | 1.04 |
| $n=50,000$ | 6.37 | 0.87 | 0.61 | 1.33 | 0.61 |
| $n=100,000$ | 6.36 | 0.86 | 0.42 | 0.83 | 0.42 |
| $n=500,000$ | 6.35 | 0.85 | 0.09 | 0.46 | 0.09 |
| | Unconstrained estimator | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 8.15 | 10.90 | 110.96 | 543.05 | 8.15 |
| $n=1,000$ | 7.26 | 5.80 | 61.47 | 416.74 | 5.80 |
| $n=5,000$ | 6.54 | 1.77 | 7.97 | 170.99 | 1.77 |
| $n=10,000$ | 6.45 | 1.23 | 4.64 | 105.77 | 1.23 |
| $n=50,000$ | 6.37 | 0.89 | 1.07 | 24.87 | 0.89 |
| $n=100,000$ | 6.36 | 0.85 | 0.46 | 9.24 | 0.46 |
| $n=500,000$ | 6.35 | 0.81 | 0.09 | 1.68 | 0.09 |
| | Ratio | | | | |
| | $K=2$ | $K=3$ | $K=4$ | $K=5$ | optimal $K$ |
| $n=500$ | 1.00 | 0.49 | 0.09 | 0.02 | 0.65 |
| $n=1,000$ | 1.00 | 0.50 | 0.10 | 0.02 | 0.50 |
| $n=5,000$ | 1.00 | 0.68 | 0.23 | 0.03 | 0.68 |
| $n=10,000$ | 1.00 | 0.85 | 0.32 | 0.04 | 0.85 |
| $n=50,000$ | 1.00 | 0.98 | 0.57 | 0.05 | 0.69 |
| $n=100,000$ | 1.00 | 1.01 | 0.91 | 0.09 | 0.91 |
| $n=500,000$ | 1.00 | 1.05 | 1.00 | 0.28 | 1.00 |

**Table 6:** simulation results for the case $g(x) = 2(x-1/2)_+^2 + 0.5x$, $\rho = 0.5$, and $\eta = 0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of $n$ and $K$. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of $n$ and $K$. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over $K$. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function $n$ and $K$. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

## E 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:

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

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 $\varepsilon$ 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[4] $\partial g / \partial z_1$, the Slutsky condition implies that the uncompensated (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 constrained NPIV estimator appears to be an attractive estimator of demand functions in this setting.

We consider three benchmark estimators. First, we compute the unconstrained nonparametric ("uncon. 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 $\gamma$ 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 ("con. NP") repeats the same steps (i)–(iii) except that it imposes monotonicity (in price) of $g$ in steps (i) and (iii).

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

The third benchmark estimator is the unconstrained NPIV estimator ("uncon. NPIV") that accounts for the covariates $Z_2$ in similar fashion as the first, unconstrained 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 constrained NPIV estimator ("con. 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 unconstrained 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 $\mathbf{Z}_2$.[5]

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 (as discussed in Remark B.5). 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 2 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 (7). We formally test this monotone IV assumption by applying the test proposed in Chetverikov and Wilhelm (2017). We find a test statistic value of 0.139 and 95%-critical value of 1.720.[6] Therefore, we fail to reject the monotone IV assumption.

Figure 3 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 unconstrained NPIV estimator as proposed in Horowitz and Lee (2012).[7] The black lines correspond to the estimators assuming

[5]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.

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

[7]Critical values are computed from 1,000 bootstrap samples and the bands are computed on a grid of 100

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 unconstrained and constrained 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 constrained estimates are very similar, indicating that endogeneity of prices may not be important in this problem, but they are both significantly flatter than the unconstrained estimates across all three income groups, which implies that households appear to be less sensitive to price changes than the unconstrained estimates suggest. The small maximum slope of the constrained NPIV estimator also suggests that the error bound in Theorem 2 may be small and therefore we expect the constrained NPIV estimate to be precise for this data set.

equally-spaced points in the support of the data for $X$.

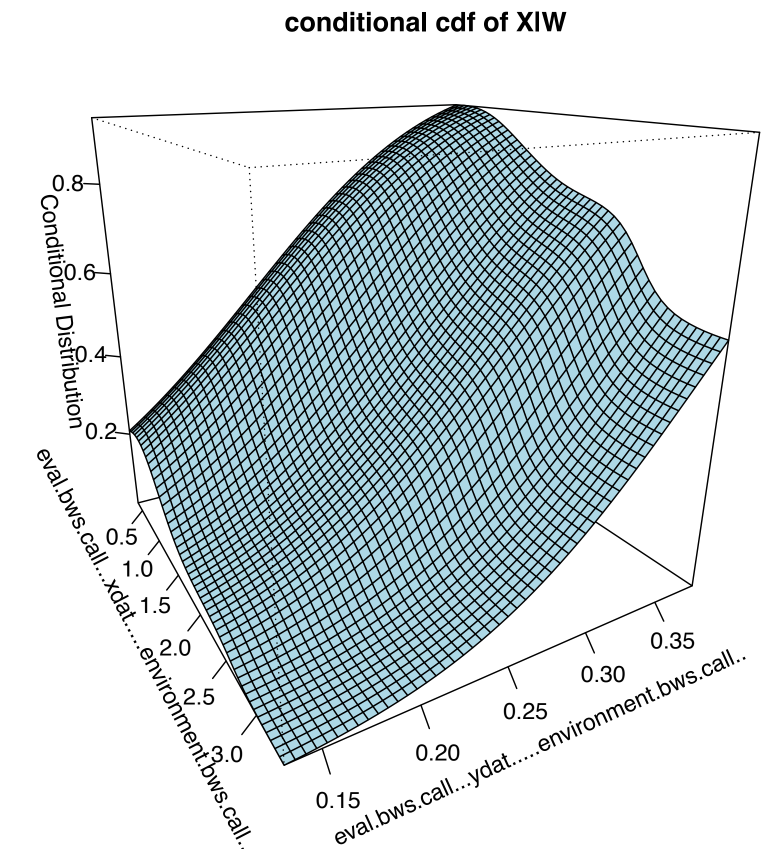


**Figure 2:** Nonparametric kernel estimate of the conditional cdf $F_{X|W}(x|w)$.

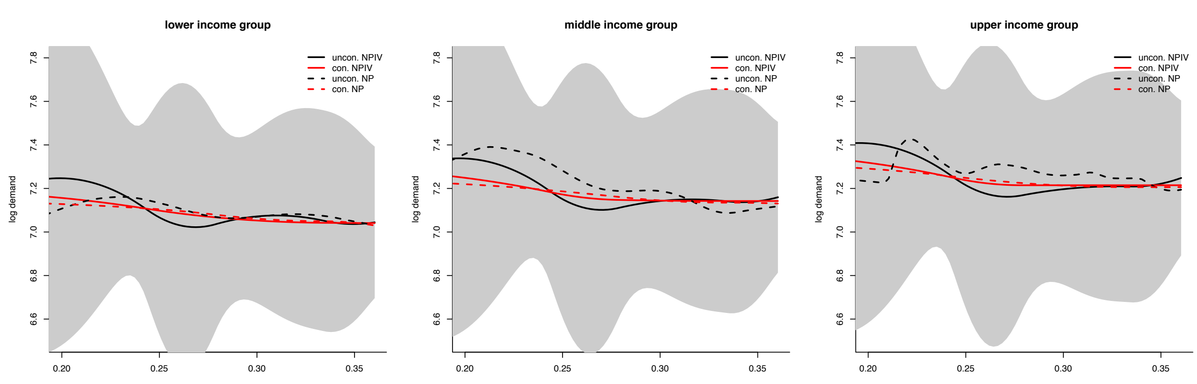


**Figure 3:** Estimates of $g(x, z_1)$ plotted as a function of price $x$ for $z_1$ fixed at three income levels.

# F Proofs

## F.1 Proof of Theorem 1

For any $h \in L^1[0,1]$, let $\|h\|_1 := \int_0^1 |h(x)|dx$, $\|h\|_{1,t} := \int_{\delta_2}^{1-\delta_2} |h(x)|dx$ and define the operator norm by $\|T\|_2 := \sup_{h \in L^2[0,1]:\, \|h\|_2 > 0} \|Th\|_2 / \|h\|_2$. Note that $\|T\|_2^2 \le \int_0^1 \int_0^1 f_{X,W}^2(x,w)dxdw$, and so under Assumption 2, $\|T\|_2 \le \sqrt{C_T}$. Also, let $\mathcal{M}$ denote the set of all monotone functions in $L^2[0,1]$. To prove Theorem 1 from the main text, we first establish some auxiliary results.

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

$$\|h\|_{2,t} \le \bar{C}\|Th\|_2 \tag{25}$$

*for any function $h \in \mathcal{M}$. Here $\bar{C}$ depends only on the constants appearing in Assumptions 1, 2, and on $x_1$, $x_2$.*

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

$$\|h\|_{2,t} \le C_1\|h\|_{1,t} \tag{26}$$

for $C_1 := (x_2 - x_1)^{1/2} / \min\{x_1 - \delta_2, 1 - \delta_2 - x_2\}$. Indeed, by monotonicity of $h$,

$$\begin{aligned}
\|h\|_{2,t} &= \left(\int_{x_1}^{x_2} h(x)^2 dx\right)^{1/2} \le \sqrt{x_2 - x_1}\max\{|h(x_1)|, |h(x_2)|\} \\
&\le \sqrt{x_2 - x_1}\,\frac{\int_{\delta_2}^{1-\delta_2} |h(x)|dx}{\min\{x_1 - \delta_2, 1 - \delta_2 - x_2\}}
\end{aligned}$$

so that (26) follows. Therefore, for any increasing continuously differentiable $h \in \mathcal{M}$,

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

where the first inequality follows from (26), the second from Lemma F.2 below (which is the main step in the proof of Theorem 1), and the third by Jensen's inequality. Hence, conclusion (25) of Lemma F.1 holds for increasing continuously differentiable $h \in \mathcal{M}$ with $\bar{C} := C_1 C_2$ and $C_2$ as defined in Lemma F.2.

Next, for any increasing function $h \in \mathcal{M}$, it follows from Lemma G.5 that one can find a sequence of increasing continuously differentiable functions $h_k \in \mathcal{M}$, $k \ge 1$, such that $\|h_k - h\|_2 \to 0$ as $k \to \infty$. Therefore, by the triangle inequality,

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

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 (25) of Lemma F.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 (25) of Lemma F.1 also holds for all decreasing $h \in \mathcal{M}$, and thus for all $h \in \mathcal{M}$. This completes the proof of the lemma. Q.E.D.

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

$$\|h\|_{1,t} = \int_{\delta_2}^{1-\delta_2} |h(x)|dx \leq C_2 \|Th\|_1$$

*where* $C_2 := ((c_W c_f/4)\min\{1-w_2, w_1\}\min\{(C_F-1)/2, 1\})^{-1}$.

*Proof.* Take any increasing continuously differentiable function $h \in L^1[0,1]$ such that $\|h\|_{1,t} = 1$. Define $M(w) := \mathrm{E}[h(X)|W=w]$ 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 $\int_0^1 |M(w)|dw \geq (c_W C_2)^{-1}$.

To do so, 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 (7) of Assumption 1 and $Dh(x) \geq 0$ for all $x$ imply that the function $M(w)$ is increasing.

Next, 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,

$$\begin{aligned}\int_0^1 |M(w)|dw &\geq \int_{w_2}^1 |M(w)|dw \geq (1-w_2)M(w_2) = (1-w_2)\int_0^1 h(x) f_{X|W}(x|w_2)dx \\ &\geq (1-w_2)\int_{\delta_2}^{1-\delta_2} h(x) f_{X|W}(x|w_2)dx \geq (1-w_2)c_f \int_{\delta_2}^{1-\delta_2} h(x)dx \\ &= (1-w_2)c_f\|h\|_{1,t} = (1-w_2)c_f \geq (c_W C_2)^{-1}\end{aligned}$$

by Assumption 2(ii). Similarly,

$$\int_0^1 |M(w)|dw \geq w_1 c_f \geq (c_W C_2)^{-1}$$

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

$$M(w) = \int_0^{x^*} h(x) f_{X|W}(x|w)dx + \int_{x^*}^1 h(x) f_{X|W}(x|w)dx$$
$$= -\int_0^{x^*} Dh(x) F_{X|W}(x|w)dx + \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w))dx. \tag{27}$$

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

$$M(w_k) = A_k - B_k, \quad k = 1, 2.$$

Consider the following five cases separately, depending on where $x^*$ lies relative to $\delta_2$, $\delta_1$, $1 - \delta_1$, and $1 - \delta_2$ (note that we have $0 \leq \delta_2 \leq \delta_1 < 1 - \delta_1 \leq 1 - \delta_2 \leq 1$).

**Case I ($\delta_1 < x^* < 1 - \delta_1$):** First, we have

$$\begin{aligned} A_1 + B_2 &= \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx + \int_0^{x^*} Dh(x) F_{X|W}(x|w_2)dx \\ &= \int_{x^*}^1 h(x) f_{X|W}(x|w_1)dx - \int_0^{x^*} h(x) f_{X|W}(x|w_2)dx \\ &\geq \int_{x^*}^{1-\delta_2} h(x) f_{X|W}(x|w_1)dx - \int_{\delta_2}^{x^*} h(x) f_{X|W}(x|w_2)dx \\ &\geq c_f \int_{x^*}^{1-\delta_2} h(x)dx + c_f \int_{\delta_2}^{x^*} |h(x)|dx = c_f \int_{\delta_2}^{1-\delta_2} |h(x)|dx \\ &= c_f \|h\|_{1,t} = c_f \end{aligned} \tag{28}$$

where the fourth line follows from Assumption 2(ii). Second, by (7) and (8) 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

$$M(w_2) - M(w_1) \geq (C_F - 1)B_2. \tag{29}$$

Similarly, by (7) and (9) 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

$$M(w_2) - M(w_1) \geq (C_F - 1)A_1. \tag{30}$$

In conclusion, equations (28), (29), and (30) yield

$$M(w_2) - M(w_1) \geq (C_F - 1)(A_1 + B_2)/2 \geq (C_F - 1)c_f/2. \tag{31}$$

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

$$\int_0^1 |M(w)|dw \geq \int_{w_2}^1 |M(w)|dw \geq (1 - w_2)M(w_2) \geq (1 - w_2)(C_F - 1)c_f/2 \geq (c_W C_2)^{-1}. \tag{32}$$

Similarly,

$$\int_0^1 |M(w)|dw \geq \int_0^{w_1} |M(w)|dw \geq w_1|M(w_1)| \geq w_1(C_F - 1)c_f/2 \geq (c_W C_2)^{-1} \tag{33}$$

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 (32) shows that

$$\int_0^1 |M(w)|dw \geq (1 - w_2)(C_F - 1)c_f/4 \geq (c_W C_2)^{-1}.$$

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

$$\int_0^1 |M(w)|dw \geq \int_0^{w_1} |M(w)|dw \geq w_1(C_F - 1)c_f/4 \geq (c_W C_2)^{-1}.$$

This completes the proof of Case I.

**Case II ($1 - \delta_1 \leq x^* \leq 1 - \delta_2$):** Note that since

$$\|h\|_{1,t} = \int_{\delta_2}^{x^*} |h(x)|dx + \int_{x^*}^{1-\delta_2} h(x)dx = 1,$$

it follows that either $\int_{\delta_2}^{x^*} |h(x)|dx \geq 1/2$ or $\int_{x^*}^{1-\delta_2} h(x)dx \geq 1/2$. We first consider the case $\int_{\delta_2}^{x^*} |h(x)|dx \geq 1/2$. Suppose that $M(w_1) \geq -c_f/4$. 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

$$\begin{aligned}
M(w_2) - M(w_1) &= M(w_2) - C_F M(w_1) + C_F M(w_1) - M(w_1) \\
&\geq (C_F - 1)B_1 + (C_F - 1)M(w_1) \\
&= (C_F - 1)\left(\int_0^{x^*} Dh(x)F_{X|W}(x|w_1)dx + M(w_1)\right) \\
&= (C_F - 1)\left(\int_0^{x^*} |h(x)|f_{X|W}(x|w_1)dx + M(w_1)\right) \\
&\geq (C_F - 1)\left(\int_{\delta_2}^{x^*} |h(x)|f_{X|W}(x|w_1)dx - \frac{c_f}{4}\right) \\
&\geq (C_F - 1)\left(c_f \int_{\delta_2}^{x^*} |h(x)|dx - \frac{c_f}{4}\right) \geq \frac{(C_F - 1)c_f}{4}.
\end{aligned}$$

We then proceed as in Case I using this inequality to replace (31) to show that $\int_0^1 |M(w)|dw \geq (c_W C_2)^{-1}$. On the other hand, when $M(w_1) < -c_f/4$ we bound $\int_0^1 |M(w)|dw$ as in (33), and the proof of Case II with $\int_{\delta_2}^{x^*} |h(x)|dx \geq 1/2$ is complete.

Next, we consider the case $\int_{x^*}^{1-\delta_2} h(x)dx \geq 1/2$. As above, we have $M(w_2) \geq C_F A_1 - B_1$ and $M(w_1) = A_1 - B_1$. Hence,

$$\begin{aligned}
M(w_2) - M(w_1) &\geq (C_F - 1)A_1 \\
&= (C_F - 1)\int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx \\
&= (C_F - 1)\int_{x^*}^1 h(x) f_{X|W}(x|w_1)dx \\
&\geq (C_F - 1)\int_{x^*}^{1-\delta_2} h(x) f_{X|W}(x|w_1)dx \\
&\geq (C_F - 1)c_f \int_{x^*}^{1-\delta_2} h(x)dx \geq \frac{(C_F - 1)c_f}{2}.
\end{aligned}$$

We then again proceed as in Case I to show that $\int_0^1 |M(w)|dw \geq (c_W C_2)^{-1}$. The proof of Case II with $\int_{x^*}^{1-\delta_2} h(x)dx \geq 1/2$ is complete.

**Case III ($1-\delta_2 < x^*$):** 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

$$\begin{aligned}
M(w_2) - M(w_1) &= M(w_2) - C_F M(w_1) + C_F M(w_1) - M(w_1) \\
&\geq (C_F - 1)B_1 + (C_F - 1)M(w_1) \\
&= (C_F - 1)\left(\int_0^{x^*} Dh(x)F_{X|W}(x|w_1)dx + M(w_1)\right) \\
&= (C_F - 1)\left(\int_0^{x^*} |h(x)| f_{X|W}(x|w_1)dx + M(w_1)\right) \\
&\geq (C_F - 1)\left(\int_{\delta_2}^{1-\delta_2} |h(x)| f_{X|W}(x|w_1)dx - \frac{c_f}{2}\right) \\
&\geq (C_F - 1)\left(c_f \int_{\delta_2}^{1-\delta_2} |h(x)|dx - \frac{c_f}{2}\right) = \frac{(C_F - 1)c_f}{2}.
\end{aligned}$$

We then proceed as in Case I to show that $\int_0^1 |M(w)|dw \geq (c_W C_2)^{-1}$. On the other hand, when $M(w_1) < -c_f/2$ we bound $\int_0^1 |M(w)|dw$ as in (33), and the proof of Case III is complete.

**Case IV ($\delta_2 \leq x^* \leq \delta_1$):** Similar to Case II, we first consider the case $\int_{x^*}^{1-\delta_2} h(x)dx \geq 1/2$. Suppose first that $M(w_2) \leq c_f/4$. As in Case I we have $M(w_1) \leq A_2 - C_F B_2$ so that together with $M(w_2) = A_2 - B_2$,

$$
\begin{aligned}
M(w_2) - M(w_1) &= M(w_2) - C_F M(w_2) + C_F M(w_2) - M(w_1) \\
&\geq (1 - C_F)M(w_2) + (C_F - 1)A_2 \\
&= (C_F - 1)\left(\int_{x^*}^{1} Dh(x)(1 - F_{X|W}(x|w_2))dx - M(w_2)\right) \\
&= (C_F - 1)\left(\int_{x^*}^{1} h(x)f_{X|W}(x|w_2)dx - M(w_2)\right) \\
&\geq (C_F - 1)\left(\int_{x^*}^{1-\delta_2} h(x)f_{X|W}(x|w_2)dx - M(w_2)\right) \\
&\geq (C_F - 1)\left(c_f \int_{x^*}^{1-\delta_2} h(x)dx - \frac{c_f}{4}\right) \geq \frac{(C_F - 1)c_f}{4},
\end{aligned}
$$

and we proceed as in Case I to show that $\int_0^1 |M(w)|dw \geq (c_W C_2)^{-1}$. On the other hand, when $M(w_2) > c_f/4$, we bound $\int_0^1 |M(w)|dw$ as in (32), and the proof of Case IV with $\int_{x^*}^{1-\delta_2} h(x)dx \geq 1/2$ is complete.

Next, consider the case $\int_{\delta_2}^{x^*} |h(x)|dx \geq 1/2$. As above, we have $M(w_1) \leq A_2 - C_F B_2$ and $M(w_2) = A_2 - B_2$. Hence,

$$
\begin{aligned}
M(w_2) - M(w_1) &\geq (C_F - 1)B_2 \\
&= (C_F - 1)\int_0^{x^*} Dh(x)F_{X|W}(x|w_2)dx \\
&= (C_F - 1)\int_0^{x^*} |h(x)|f_{X|W}(x|w_2)dx \\
&\geq (C_F - 1)\int_{\delta_2}^{x^*} |h(x)|f_{X|W}(x|w_2)dx \\
&\geq (C_F - 1)c_f \int_{\delta_2}^{x^*} |h(x)|dx \geq \frac{(C_F - 1)c_f}{2}.
\end{aligned}
$$

We then again proceed as in Case I to show that $\int_0^1 |M(w)|dw \geq (c_W C_2)^{-1}$. The proof of Case IV with $\int_{\delta_2}^{x^*} |h(x)|dx \geq 1/2$ is complete.

**Case V ($x^* < \delta_2$):** Similar to Case III, 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 $M(w_2) = A_2 - B_2$,

$$
\begin{aligned}
M(w_2) - M(w_1) &= M(w_2) - C_F M(w_2) + C_F M(w_2) - M(w_1) \\
&\geq (1 - C_F)M(w_2) + (C_F - 1)A_2 \\
&= (C_F - 1)\left(\int_{x^*}^{1} Dh(x)(1 - F_{X|W}(x|w_2))dx - M(w_2)\right) \\
&= (C_F - 1)\left(\int_{x^*}^{1} h(x)f_{X|W}(x|w_2)dx - M(w_2)\right) \\
&\geq (C_F - 1)\left(\int_{\delta_2}^{1-\delta_2} h(x)f_{X|W}(x|w_2)dx - M(w_2)\right)
\end{aligned}
$$

$$\geq (C_F - 1)\left(c_f \int_{\delta_2}^{1-\delta_2} h(x)dx - \frac{c_f}{2}\right) = \frac{(C_F-1)c_f}{2},$$

and we proceed as in Case I to show that $\int_0^1 |M(w)|dw \geq (c_W C_2)^{-1}$. On the other hand, when $M(w_2) > c_f/2$, we bound $\int_0^1 |M(w)|dw$ as in (32), and the proof of Case V is complete. The lemma is proven. Q.E.D.

**Lemma F.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*

$$\|h\|_{2,t} \leq \frac{\bar{C}}{1-\alpha}\|Th\|_2 \tag{34}$$

*for the constant $\bar{C}$ defined in Lemma F.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 $\widetilde{h} \in \mathcal{M}$, Lemma F.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 by $1-\alpha$ by assumption, we obtain the inequality

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

which is equivalent to (34). Q.E.D.

*Proof of Theorem 1.* 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 Lemma F.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 (34) in Lemma F.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 theorem. Q.E.D.

## F.2 Proof of Theorem 2 and Related Results

In this section, we use $C$ to denote a strictly positive constant, which value may change from place to place. Also, we use $\mathrm{E}_n[\cdot]$ to denote the average over index $i = 1, \dots, n$; for example, $\mathrm{E}_n[X_i] = n^{-1}\sum_{i=1}^n X_i$. Before proving Theorem 2, we prove Lemma B.1.

*Proof of Lemma B.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 (21), it suffices to show that

$$\mathrm{P}\Big(D\widehat{g}^u(x) \geq 0 \text{ for all } x \in [0,1]\Big) \to 1 \text{ as } n \to \infty. \tag{35}$$

In turn, (35) follows if

$$\sup_{x\in[0,1]} |D\widehat{g}^u(x) - Dg(x)| = o_p(1) \tag{36}$$

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

To prove (36), define a function $\widehat{m} \in L^2[0,1]$ by

$$\widehat{m}(w) = q(w)'\mathrm{E}_n[q(W_i)Y_i], \qquad w \in [0,1], \tag{37}$$

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

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

Throughout the proof, we assume that the events

$$\|\mathrm{E}_n[q(W_i)p(X_i)'] - \mathrm{E}[q(W)p(X)']\| \leq C(\xi_n^2 \log n/n)^{1/2}, \tag{38}$$

$$\|\mathrm{E}_n[q(W_i)q(W_i)'] - \mathrm{E}[q(W)q(W)']\| \leq C(\xi_n^2 \log n/n)^{1/2}, \tag{39}$$

$$\|\mathrm{E}_n[q(W_i)g_n(X_i)] - \mathrm{E}[q(W)g_n(X)]\| \leq C(J/(\alpha n))^{1/2}, \tag{40}$$

$$\|\widehat{m} - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1}J^{-s}) \tag{41}$$

hold for some sufficiently large constant $0 < C < \infty$. It follows from Markov's inequality and Lemmas F.4 and G.6 that all four events hold jointly with probability at least $1 - \alpha - n^{-1}$ since the constant $C$ is large enough.

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

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

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

$$\|T(\widehat{g}^u - g)\|_2 \leq \|(T - T_n)\widehat{g}^u\|_2 + \|(T_n - \widehat{T})\widehat{g}^u\|_2 + \|\widehat{T}\widehat{g}^u - \widehat{m}\|_2 + \|\widehat{m} - m\|_2 \tag{42}$$

by the triangle inequality. The bound on $\|\widehat{m} - m\|_2$ is given in (41). 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,

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

Moreover,

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

by (38), and

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

by (40).

Further, by Assumption 2(iii), all eigenvalues of $\mathrm{E}[q(W)q(W)']$ are bounded from below by $c_w$ and from above by $C_w$, and so it follows from (39) that for large $n$, all eigenvalues of $Q_n := \mathrm{E}_n[q(W_i)q(W_i)']$ are bounded below from zero and from above. Therefore,

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

by optimality of $\widehat{\beta}^u$ (note that $\beta_n$ is feasible in the optimization problem (13) for $n$ large enough since $\|g\|_2 < C_b$ and $g_n(\cdot) = p(\cdot)'\beta_n$ satisfies $\|g - g_n\|_2 \to 0$ as $n \to \infty$). Moreover,

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

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),

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

Combining the inequalities above shows that the inequality

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

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

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

and so by the triangle inequality,

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

uniformly over $x \in [0,1]$ since $J \leq C_J K$ by Assumption 5. Since by the conditions of the lemma, $\sup_{x\in[0,1]} \|Dp(x)\|(\tau_n(K/n)^{1/2} + K^{-s}) \to 0$, (36) follows by taking $\alpha = \alpha_n \to 0$ slowly enough. This completes the proof of the lemma. Q.E.D.

*Proof of Theorem 2.* Consider the event that inequalities (38)-(41) hold for some sufficiently large constant $C$. As in the proof of Lemma B.1, this events occurs with probability at least $1 - \alpha - n^{-1}$. Also, applying the same arguments as those in the proof of Lemma B.1, starting from (42), with $\widehat{g}^c$ replacing $\widehat{g}^u$ 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 B.1, it follows that on this event,

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

and so, by Assumption 6(iii),

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

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 $\|\widehat{g}^c - g_n\|_{2,t} \leq \delta$, 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 (15).

To prove (16), observe that combining (44) and Assumption 6(iii) and applying the triangle inequality shows that with probability at least $1 - \alpha - n^{-1}$,

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

which, by the same argument as that used to prove (15), gives

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

The asserted claim (16) now follows by applying (15) with $\delta = 0$ and (45) with $\delta = \|Dg\|_\infty / c_\tau$ and using Theorem 1 to bound $\tau(c_\tau)$. This completes the proof of the theorem. Q.E.D.

**Lemma F.4.** *Suppose that Assumptions 2, 4, and 7 hold. Then* $\|\widehat{m} - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1} J^{-s})$ *with probability at least* $1 - \alpha$ *where* $\widehat{m}$ *is defined in (37).*

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

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

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

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

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

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

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

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

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.

*Proof of Corollary B.1.* The corollary follows immediately from Theorem 2. Q.E.D.

### F.3 Proof of Theorems C.1 and C.2

Let $\mathcal{M}_\uparrow$ be the set of all functions in $\mathcal{M}$ that are increasing but not constant. Similarly, let $\mathcal{M}_\downarrow$ be the set of all functions in $\mathcal{M}$ that are decreasing but not constant, and let $\mathcal{M}_\rightarrow$ be the set of all constant functions in $\mathcal{M}$.

*Proof of Theorem C.1.* Assume that $g$ is increasing but not constant, that is, $g \in \mathcal{M}_\uparrow$. Define $M(w) := \mathrm{E}[Y|W=w]$, $w \in [0,1]$. Below we show that $M \in \mathcal{M}_\uparrow$. To prove it, observe that, as in the proof of Lemma F.2, integration by parts gives

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

and so Assumption 1 implies that $M$ is increasing. Let us show that $M$ is not constant. To this end, note that

$$M(w_2) - M(w_1) = \int_0^1 Dg(x)(F_{X|W}(x|w_1) - F_{X|W}(x|w_2))dx.$$

Since $g$ is not constant and is continuously differentiable, there exists $\bar{x} \in (0,1)$ such that $Dg(\bar{x}) > 0$. Also, since $0 < \delta_1 < 1 - \delta_1 < 1$ (the constant $\delta_1$ appears in Assumption 1), we have $\bar{x} \in (0, 1-\delta_1)$ or $\bar{x} \in (\delta_1, 1)$. In the first case,

$$M(w_2) - M(w_1) \geq \int_0^{1-\delta_1} (C_F - 1) Dg(x) F_{X|W}(x|w_2)dx > 0.$$

In the second case,

$$M(w_2) - M(w_1) \geq \int_{\delta_1}^1 (C_F - 1) Dg(x)(1 - F_{X|W}(x|w_1))dx > 0.$$

Thus, $M$ is not constant, and so $M \in \mathcal{M}_\uparrow$. Similarly, one can show that if $g \in \mathcal{M}_\downarrow$, then $M \in \mathcal{M}_\downarrow$, and if $g \in \mathcal{M}_\rightarrow$, then $M \in \mathcal{M}_\rightarrow$. However, the distribution of the triple $(Y, X, W)$ uniquely determines whether $M \in \mathcal{M}_\uparrow$, $\mathcal{M}_\downarrow$, or $\mathcal{M}_\rightarrow$, and so it also uniquely determines whether $g \in \mathcal{M}_\uparrow$, $\mathcal{M}_\downarrow$, or $\mathcal{M}_\rightarrow$ This completes the proof. Q.E.D.

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

# G Technical tools

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

**Lemma G.1.** *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

$$\mathrm{E}[(M(W) - M(W'))(W - W')] > 0,$$

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 G.2.** *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

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

Therefore, by the Cauchy-Schwarz 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 G.3.** *Let $(X, W)$ be a pair of random variables defined as in Example A.1. Then Assumptions 1 and 2 of Section 2 are satisfied with $0 < \delta_1 < 1 - \delta_1 < 1$, $\delta_1 = \delta_2$, and $0 < w_1 < w_2 < 1$.*

*Proof.* As noted in Example A.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 (7) of Assumption 1 follows. Further, to prove condition (8) of Assumption 1, it suffices to show that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} \leq c_F \tag{46}$$

for some constant $c_F < 0$, all $x \in (0, 1-\delta_1)$, and all $w \in (w_1, w_2)$ because, for every $x \in (0, 1-\delta_1)$, 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 $x \in (0, 1-\delta_1)$ 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 $x \in (0, 1-\delta_1)$. To show (46), observe that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} = -\frac{\rho}{\sqrt{1-\rho^2}}\frac{\phi(y)}{\Phi(y)}\frac{1}{\phi(\Phi^{-1}(w))} \leq -\frac{\sqrt{2\pi}\rho}{\sqrt{1-\rho^2}}\frac{\phi(y)}{\Phi(y)} \tag{47}$$

where $y := (\Phi^{-1}(x) - \rho\Phi^{-1}(w))/(1-\rho^2)^{1/2}$. Thus, (46) holds for some $c_F < 0$ and all $x \in (0, 1-\delta_1)$ and $w \in (w_1, w_2)$ such that $\Phi^{-1}(x) \geq \rho\Phi^{-1}(w)$ since $1 - \delta_1 < 1$ 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 (47) 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 (8) of Assumption 1 is satisfied. Similar argument also shows that condition (9) 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

$$f_{X|W}(x|w) := \frac{1}{\sqrt{1-\rho^2}}\phi\left(\frac{\Phi^{-1}(x) - \rho\Phi^{-1}(w)}{\sqrt{1-\rho^2}}\right)\frac{1}{\phi(\Phi^{-1}(x))}. \tag{48}$$

Since $0 < \delta_1 < 1 - \delta_1 < 1$ 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 $w \in [0,1]$, (48) combined with the change of variables formula with $x = \Phi(\widetilde{x})$ and $w = \Phi(\widetilde{w})$ give

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

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 G.4.** *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 $0 < w_1 < w_2 < 1$, $0 < \delta_2 \leq \delta_1$, and $(1 - w_1)/2 < \delta_1 < 1/2$.*

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

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

and, thus,

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

It is easy to check that $\partial F_{X|W}(x|w)/\partial w \leq 0$ for all $x, w \in [0, 1]$ so that condition (7) of Assumption 1 is satisfied. We now check condition (8). Consider the case $0 \leq x < 1/2$ and $w \in (w_1, w_2)$. Then

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} = \left\{ \begin{array}{cc} -\frac{1}{w}, & 0 \leq x < \frac{w}{2} \\ \frac{-1/2}{2x - w/2}, & \frac{w}{2} \leq x < \frac{1}{2} \end{array} \right\} < -\frac{1}{w_1} < 0 \tag{49}$$

and thus, by the same reasoning as in the proof of Lemma G.3, (8) holds for all $0 \leq x < 1/2$. Now consider the case $x \in [1/2, 1 - \delta_1)$. On this interval, $\partial(F_{X|W}(x|w_1)/F_{X|W}(x|w_2))/\partial x \leq 0$ so that

$$\frac{F_{X|W}(x|w_1)}{F_{X|W}(x|w_2)} = \frac{1 - \frac{2}{w_1}\left(x - \frac{1+w_1}{2}\right)^2}{1 - \frac{2}{w_2}\left(x - \frac{1+w_2}{2}\right)^2} \geq \frac{1}{1 - \frac{2}{w_2}\left(\frac{1+w_1}{2} - \frac{1+w_2}{2}\right)^2} = \frac{w_2}{w_2 - (w_1 - w_2)^2/2} > 1,$$

where the first inequality substitutes in the upper bound $(1 + w_1)/2$ for $x$, and thus (8) holds also uniformly over $1/2 \leq x < 1 - \delta_1$.

We now show that (9) holds. By (49) and the same reasoning as in the proof of Lemma G.3, (9) holds for all $\delta_1 < x < 1/2$. Consider the case $1/2 \leq x < (1+w_1)/2$. Since, in that interval,

$$\frac{\partial}{\partial x}\left(\frac{1-F_{X|W}(x|w_2)}{1-F_{X|W}(x|w_1)}\right) = \frac{4w_1(w_2-w_1)(1+w_2-2x)}{w_2(1+w_1-2x)^3} > 0,$$

we have

$$\frac{1-F_{X|W}(x|w_2)}{1-F_{X|W}(x|w_1)} = \frac{\frac{2}{w_2}\left(x-\frac{1+w_2}{2}\right)^2}{\frac{2}{w_1}\left(x-\frac{1+w_1}{2}\right)^2} \geq \frac{\frac{2}{w_2}\left(\frac{w_2}{2}\right)^2}{\frac{2}{w_1}\left(\frac{w_1}{2}\right)^2} = \frac{w_2}{w_1} > 1,$$

where the first inequality substitutes in the lower bound $1/2$ for $x$. Therefore, (9) holds for all $1/2 \leq x < (1+w_1)/2$. Checking that (9) holds for all $(1+w_1)/2 \leq x < 1$ is trivial.

It is also clear that Assumption 2 holds because $\delta_2 > 0$. Q.E.D.

**Lemma G.5.** *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 $\phi$ 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 $h_\sigma$ 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

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

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 G.6.** *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 $\|p_1\| \leq \xi_n$, $\|q_1\| \leq \xi_n$, $\|\mathrm{E}[p_1p_1']\| \leq C_p$, and $\|\mathrm{E}[q_1q_1']\| \leq C_q$ where $\xi_n \geq 1$. Then for all $t \geq 0$,*

$$\mathrm{P}\left(\|\mathrm{E}_n[p_iq_i'] - \mathrm{E}[p_1q_1']\| \geq t\right) \leq \exp\left(\log(K+J) - \frac{Ant^2}{\xi_n^2(1+t)}\right)$$

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

**Remark G.1.** Closely related results have been used previously by Belloni, Chernozhukov, Chetverikov, and Kato (2015) and Chen and Christensen (2013).

*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 $b \in \mathbb{R}^J$,

$$\begin{aligned} a'\mathrm{E}[p_1q_1']b &= \mathrm{E}[(a'p_1)(b'q_1)] \\ &\leq \left(\mathrm{E}[(a'p_1)^2]\mathrm{E}[(b'q_1)^2]\right)^{1/2} \leq \|a\|\|b\|(C_pC_q)^{1/2} \end{aligned}$$

by Hölder's inequality. Therefore, $\|\mathrm{E}[p_1q_1']\| \leq (C_pC_q)^{1/2}$. Further, denote $S_i := p_iq_i' - \mathrm{E}[p_iq_i']$ for $i = 1, \ldots, n$. By the triangle inequality and calculations above,

$$\begin{aligned} \|S_1\| &\leq \|p_1q_1'\| + \|\mathrm{E}[p_1q_1']\| \\ &\leq \xi_n^2 + (C_pC_q)^{1/2} \leq \xi_n^2(1 + (C_pC_q)^{1/2}) =: R. \end{aligned}$$

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

$$\begin{aligned} \|\mathrm{E}[Z_nZ_n']\| &\leq n\|\mathrm{E}[S_1S_1']\| \\ &\leq n\|\mathrm{E}[p_1q_1'q_1p_1']\| + n\|\mathrm{E}[p_1q_1']\mathrm{E}[q_1p_1']\| \leq n\|\mathrm{E}[p_1q_1'q_1p_1']\| + nC_pC_q. \end{aligned}$$

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

$$a'\mathrm{E}[p_1q_1'q_1p_1']a \leq \xi_n^2\mathrm{E}[(a'p_1)^2] \leq \xi_n^2\|a\|^2C_p.$$

Therefore, $\|\mathrm{E}[p_1 q_1' q_1 p_1']\| \le \xi_n^2 C_p$, and so

$$\|\mathrm{E}[Z_n Z_n']\| \le n C_p(\xi_n^2 + C_q) \le n\xi_n^2(1 + C_p)(1 + C_q).$$

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

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

Hence, by Corollary 6.2.1 in Tropp (2012),

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

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.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some new asymptotic theory for least squares series: Pointwise and uniform results," *Journal of Econometrics*, 186(2), 345 – 366.

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.

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.

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

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2009): "Improving point and interval estimators of monotone functions by rearrangement," *Biometrika*, 96(3), 559–575.

Chetverikov, D., and D. Wilhelm (2017): "An Adaptive Test of Stochastic Dominance," Discussion paper.

Dudley, R. M. (2014): *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge.

Gagliardini, P., and O. Scaillet (2012): "Nonparametric Instrumental Variable Estimation of Structural Quantile Effects," *Econometrica*, 80(4), 1533–1562.

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

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

HODERLEIN, S., H. HOLZMANN, M. KASY, AND A. MEISTER (2016): "Erratum Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment," *The Review of Economic Studies*, forthcoming.

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

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

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*, 81(4), 1614.

KLINE, B. (2016): "Identification of the Direction of a Causal Effect by Instrumental Variables," *Journal of Business & Economic Statistics*, 34(2), 176–184.

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

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.

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

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*.