Nonparametric Instrumental Variable Estimation
Under Monotonicity*

Denis Chetverikov† Daniel Wilhelm‡

Abstract

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

*First version: January 2014. This version: October 10, 2014. We thank Richard Blundell, Stéphane Bonhomme, Moshe Buchinsky, Xiaohong Chen, Victor Chernozhukov, Jinyong Hahn, Simon Lee, Zhipeng Liao, and Rosa Matzkin for useful comments and discussions.
†Department of Economics, University of California at Los Angeles, 315 Portola Plaza, Bunche Hall, Los Angeles, CA 90024, USA; E-Mail address: chetverikov@econ.ucla.edu.
‡Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom; E-Mail address: d.wilhelm@ucl.ac.uk. The author gratefully acknowledges financial support from the ESRC Centre for Microdata Methods and Practice at IFS (RES-589-28-0001).
1 Introduction

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

In this paper, we show that augmenting the NPIV model by two monotonicity conditions, on the regression function and on the relationship between the endogenous covariate and the instrument, significantly changes the structure of the NPIV problem. First, we prove that, in the absence of point-identifying completeness assumptions, our two monotonicity assumptions contain non-trivial identifying power. Second, we demonstrate that the monotonicity assumptions significantly weaken the inherent ill-posedness of the problem, the property that, in the absence of the monotonicity assumptions, causes the slow convergence rate. Third, we derive a non-asymptotic risk bound for the constrained estimator that imposes monotonicity of the regression function and establish that, in finite samples, this estimator may perform much better than expected from its asymptotic rate of convergence, especially when the regression function is not too steep and the variability of the unrestricted estimator is high.

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

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

Our interest focuses on identification and estimation of the function \( g \). We assume that \( g \) is smooth and monotone, but do not impose any parametric restrictions. In addition, we assume that the relationship between the endogenous covariate \( X \) and the instrument \( W \) is also monotone in the sense that the conditional distribution of \( X \) given \( W \) corresponding to higher values of \( W \) first-order stochastically dominates the same conditional distribution corresponding to lower values of \( W \). We refer to this condition as monotone instrumental variable (MIV) assumption. To simplify the presentation, we assume that all variables are scalar.

The foundation of our main results in this paper consists of establishing implications of our monotonicity assumptions on the structure of the NPIV problem. We show that, under both monotonicity assumptions, the NPIV problem becomes locally quantitatively well-posed at constant functions, a concept that we introduce in this paper. Intuitively, this means that the NPIV problem is well-posed locally at constant functions. We then extend
this observation in two directions. First, we need to invert a conditional expectation operator to solve for the regression function $g$ and we establish that the operator’s inverse is not only continuous at constant functions, but we are also able to quantitatively control its modulus of continuity. Second, we show that, under the MIV assumption, a certain measure of ill-posedness of the problem is uniformly bounded when defined over the set of monotone functions even though the same measure is infinitely large when defined over the set of all functions. The boundedness of our measure of ill-posedness plays a central role in our analysis of identification and estimation of the regression function $g$.

The first application of our basic implications consists of deriving non-trivial identification bounds for the regression function. In particular, the identified set contains only functions that intersect each other and whose pairwise differences are not too close to monotone functions. Finally, a simple consequence of our two monotonicity assumptions is that the sign of the slope of $g$ is identified by the sign of the slope of the function $w \mapsto E[Y|W = w]$.

Second, we derive non-asymptotic risk bounds for the constrained NPIV estimator that imposes monotonicity of the regression function. These bounds depend on the sample size, the variability of the unconstrained estimator, and the magnitude of the regression function’s slope. When the sample size is not too large, then the monotonicity constraint is binding with non-trivial probability so that imposing the constraint has large informational content and leads to a low risk of the constrained estimator. When the sample size is large enough, then the monotonicity constraint is binding only with small probability so that imposing the monotonicity constraint has little informational content and leads to risk bounds comparable to those of the unconstrained estimator in the standard NPIV problem. The threshold sample size at which the regime switch occurs depends positively on the variability of the unconstrained estimator and negatively on the slope of the function $g$. In NPIV problems, especially in those with severe ill-posedness, the unconstrained estimator is very unstable and thus has large variability. In consequence, the threshold sample size is very large so that the constrained estimator performs well in terms of its risk even for moderately large sample sizes. In fact, the risk of the constrained estimator may even be comparable to that of standard conditional mean estimators. Our simulation experiment confirms the theoretical findings and shows dramatic finite-sample gains from imposing monotonicity of the regression function.

We regard both monotonicity assumptions as natural in many economic applications. Consider a generic example very similar to one of the examples in Kasy (2014), in which a firm produces log output $Y$ from log labor input $X$. Denote by $W$ the price of log output and let $U$ be log wage. Suppose the production function is $Y = g(X) + \varepsilon$, where $\varepsilon$ summarizes other determinants of output such as capital and total factor productivity,
and profits are $\pi(X) = e^W e^Y - e^U e^X$. Labor is chosen optimally so as to maximize profits. If $g$ is increasing and strictly concave, and the elasticity of output with respect to labor is strictly smaller than one, then a straightforward calculation shows that $\partial^2 \pi / \partial X \partial W \geq 0$ and $\partial^2 \pi / \partial X^2 < 0$. Therefore,

$$\frac{\partial X}{\partial W} = -\frac{\partial^2 \pi}{\partial X \partial W} / \frac{\partial^2 \pi}{\partial X^2} \geq 0$$

which implies our MIV condition.

**Related literature.** Newey and Powell (2003) prove consistency of their estimator of the regression function $g$ in the model (1). Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007), and Darolles, Fan, Florens, and Renault (2011) introduce other estimators and establish (optimal) rates of convergence of these estimators. See Horowitz (2011) for a recent survey and further references. In the mildly ill-posed case, Hall and Horowitz (2005) derive the minimax risk lower bound in $L^2$-norm loss and show that their estimator achieves this lower bound. Chen and Reiß (2011) derive a similar bound for the mildly and the severely ill-posed case and show that the estimator by Blundell, Chen, and Kristensen (2007) achieves this bound. Chen and Christensen (2013) establish minimax risk bounds in the sup-norm, again for both, the mildly and the severely ill-posed case. The optimal convergence rates in the severely ill-posed case are shown to be logarithmic which means that the slow convergence rate of existing estimators are not a deficiency of those estimators but rather an intrinsic feature of the statistical inverse problem.

Economic theory often provides restrictions on functions of interest, such as monotonicity, concavity, and Slutsky symmetry, that may be imposed to facilitate identification or to improve the performance of nonparametric estimators (Matzkin (1994)). Freyberger and Horowitz (2013) show how shape restrictions may yield informative bounds on functionals of the NPIV regression function when regressors and instruments are discrete. Blundell, Horowitz, and Parey (2013) impose Slutsky inequalities in a quantile NPIV model for gasoline demand and, in simulations, find that these restrictions improve the finite sample properties of the NPIV estimator. In the somewhat different ill-posed inverse problem of deconvolving a density, Carrasco and Florens (2011) show that forcing the estimator to be a density may improve its convergence rate, but do not obtain well-posedness. Similarly, Grasmair, Scherzer, and Vanhems (2013) derive the convergence rate of a NPIV estimator under general, non-convex constraints, but also do not obtain well-posedness. In fact, unlike our results, both of these papers significantly restrict the sense in which the inverse problem can be severely ill-posed.

Since at least Brunk (1955) the statistics literature on nonparametric estimation of monotone functions has developed into a vast area of research. Yatchew (1998), Dele-
croix and Thomas-Agnan (2000), and Gijbels (2004) provide recent surveys and further references. For the case in which the regression function is both, smooth and monotone, many different ways of imposing monotonicity on the estimator have been studied, for example Mukerjee (1988), Cheng and Lin (1981), Wright (1981), Friedman and Tibshirani (1984), Ramsay (1988), Mammen (1991), Ramsay (1998), Mammen and Thomas-Agnan (1999), Hall and Huang (2001), Mammen, Marron, Turlach, and Wand (2001), and Dette, Neumeyer, and Pilz (2006). Importantly, the standard, unrestricted nonparametric regression estimators are known to be monotone with probability approaching one when the regression function is strictly increasing under mild assumption that these estimators consistently estimate the derivative of the regression function. Therefore, such an estimator converges at the same rate as monotone estimators (Mammen (1991)). As a consequence, we expect gains from imposing monotonicity only when the monotonicity constraint is binding in the sense that the regression function has flat parts or, in the extreme case, drifts towards a constant function (local-to-constant asymptotics). Zhang (2002) and Chatterjee, Guntuboyina, and Sen (2013) formalize this intuition by deriving risk bounds of the isotonic regression estimator and showing that these bounds imply fast convergence rates when the regression function has flat parts. We refer to these two papers for a more detailed review of the statistics literature analyzing the sense in which isotonic estimators may adapt to flatness of their respective estimands.

**Notation.** For a function $f : \mathbb{R} \to \mathbb{R}$, let $Df(x)$ denote the derivative of a function $f$. For a function $f : \mathbb{R}^2 \to \mathbb{R}$, let $D_w f(w, u)$ and $D_u f(w, u)$ denote the partial derivatives with respect to the first and the second arguments, respectively. For random variables $A$ and $B$, denote by $f_{A,B}(a, b)$, $f_{A\mid B}(a, b)$, and $f_{A}(a)$ the joint, conditional and marginal densities of $(A, B)$, $A\mid B$, and $A$, respectively. Similarly, let $F_{A,B}(a, b)$, $F_{A\mid B}(a, b)$, and $F_{A}(a)$ refer to the corresponding cumulative distribution functions. For an operator $T : L^2[0, 1] \to L^2[0, 1]$, let $\|T\|_2$ denote the operator norm defined as

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

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

**Outline.** The remainder of the paper is organized as follows. In the next section, we introduce our MIV condition and show that, together with some regularity conditions, it implies that the conditional expectation operator defined on the set of monotone functions has a bounded inverse. Section 3 and Section 4 discuss the implications of our monotonicity assumptions for identification and estimation, respectively. In particular, we show
that the rate of convergence of our estimator is always not worse than that of unrestricted estimators but may be much faster in local-to-constant asymptotics. Section 5 provides adaptive tests of our monotonicity assumptions. In Section 6, we present results of a Monte Carlo simulation study. All proofs are contained in the appendix.

2 Local Quantitative Well-Posedness under Monotonicity

In this section, we study the properties of the NPIV model (1) when we impose monotonicity constraints on the regression function \( g \) and on the relationship between the covariate \( X \) and the instrument \( W \). The NPIV model requires solving the equation \( E[Y|W] = E[g(X)|W] \) for the function \( g \). Letting \( T : L^2[0, 1] \to L^2[0, 1] \) be the linear operator \((Th)(w) := E[h(X)|W = w]f_W(w)\) and denoting \( m(w) := E[Y|W = w]f_W(w)\), we can express this equation as

\[
Tg = m.
\] (2)

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

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

\[
\|g' - g''\|_{2,t} \leq \bar{C}\|m' - m''\|_{2},
\]

where \( \|\cdot\|_{2,t} \) is a truncated \( L^2 \)-norm defined below. This result is central to the derivation of useful bounds on the measure of ill-posedness, of identification bounds, and of fast convergence rates of a monotone NPIV estimator studied in this paper.

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

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

$$w' \leq w'' \implies F_{X|W}(x|w') \geq F_{X|W}(x|w'').$$

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

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

and

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

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

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

Condition (3) can be equivalently stated in terms of monotonicity with respect to the instrument $W$ of the reduced form first stage function. Indeed, by the Skorohod representation, it is always possible to construct a random variable $U$ distributed uniformly on $[0, 1]$ such that $U$ is independent of $W$, and $X = r(W, U)$ holds for the reduced form first stage function $r(w, u) := F_{X|W}^{-1}(u|w) := \inf\{x : F_{X|W}(x|w) \geq u\}$. Therefore, condition (3) is equivalent to the assumption that the function $w \mapsto r(w, u)$ is increasing for all $u \in [0, 1]$.

Notice, however, that our condition (3) allows for general unobserved heterogeneity of dimension larger than one, for instance as in Example 2 below. Condition (3) is related to but weaker than a corresponding condition in Kasy (2014) who assumes that the (structural) first stage has the form $X = \tilde{r}(W, \tilde{U})$ where $\tilde{U}$, representing (potentially multidimensional) unobserved heterogeneity, is independent of $W$, and the function $w \mapsto \tilde{r}(w, \tilde{u})$ is increasing for all values $\tilde{u}$. Kasy employs his condition for identification of (nonseparable) triangular systems with multidimensional unobserved heterogeneity whereas we use
our condition (3) to derive useful bounds on the measure of ill-posedness and to obtain a fast rate of convergence of a monotone NPIV estimator of \( g \) in the (separable) model (1).

Condition (3) is not related to the MIV assumption in the influential work by Manski and Pepper (2000) which requires the function \( w \mapsto \mathbb{E}[\varepsilon | W = w] \) to be increasing. Instead, we maintain the mean independence condition \( \mathbb{E}[\varepsilon | W] = 0 \).

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

This is a mild regularity assumption. The first part of the assumption implies that the operator \( T \) is compact. The second and the third parts of the assumption require the conditional distribution of \( X \) given \( W = w_1 \) or \( w_2 \) and the marginal distribution of \( W \) to be bounded away from zero over some intervals. Recall that we have \( 0 \leq x_1 < x_2 \leq 1 \) and \( 0 \leq w_1 < w_2 \leq 1 \). We could simply set \([x_1, x_2] = [w_1, w_2] = [0, 1]\) in the second part of the assumption but having \( 0 < x_1 < x_2 < 1 \) and \( 0 < w_1 < w_2 < 1 \) is required to allow for densities such as the normal, which, even after a transformation to the interval \([0, 1]\), may not yield a conditional density \( f_{X|W}(x|w) \) bounded away from zero; see Example 1 below. Therefore, we allow for the general case \( 0 \leq x_1 < x_2 \leq 1 \) and \( 0 \leq w_1 < w_2 \leq 2 \). The restriction \( f_W(w) \leq C_W \) for all \( w \in [0, 1] \) imposed in Assumption 2 is not actually required by the results in this section, but rather those of Section 4.

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

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

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

where \( U \) is independent of \( W \). Therefore, the pair \((X, W)\) clearly satisfies condition (3) of our MIV Assumption 1. Lemma 9 in the appendix verifies that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied. \( \square \)
Example 2 (Two-dimensional unobserved heterogeneity). Let \( X = U_1 + U_2W \), where \( U_1, U_2, W \) are mutually independent, \( U_1, U_2 \sim U[0, 1/2] \) and \( W \sim U[0, 1] \). Since \( U_2 \) is positive, it is straightforward to see that the stochastic dominance condition (3) is satisfied. Lemma 10 in the appendix shows that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied.

We are now ready to state our first main result in this section. Define the truncated \( L^2 \)-norm \( \| h \|_{2,t} \) for \( h \in L^2[0, 1] \) by

\[
\| h \|_{2,t} := \left( \int_{\tilde{x}_1}^{\tilde{x}_2} h(x)^2 \, dx \right)^{1/2}.
\]

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

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

\[
\| h \|_{2,t} \leq \bar{C} \| Th \|_2
\]

for any function \( h \in \mathcal{M} \).

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

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

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

Property (7) means that \( \| Th \|_2 \) being small does not necessarily imply that \( \| h \|_2 \) is small and, therefore, the inverse of the operator \( T : L^2[0, 1] \to L^2[0, 1] \), when it exists, cannot...
be continuous. Therefore, (2) is ill-posed in Hadamard’s sense\(^1\). Theorem 1, on the other hand, implies that, under Assumptions 1 and 2, (7) is not possible if \(h_k\) belongs to the set \(\mathcal{M}\) of monotone functions in \(L^2[0,1]\) for all \(k \geq 1\) and we replace the \(L^2\)-norm \(\| \cdot \|_2\) in the numerator of the left-hand side of (7) by the truncated \(L^2\)-norm \(\| \cdot \|_{2,t}\).\(^2\)

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

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

Well-posedness in Hadamard’s sense is a useful property to establish consistency of an estimator of the solution of equation (2), but it does not provide any quantitative control on the modulus of continuity of the inverse \(\mathcal{T}^{-1}\). This latter quantitative control is crucial for our subsequent derivation of fast convergence rates, which is why we now explore the implications of Theorem 1 in this direction. To this end, we introduce the concept of (local) quantitative well-posedness, which is inspired by the definitions in Bejenaru and Tao (2006), who studied the Cauchy problem for the quadratic non-linear Schrödinger equation.

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

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

---

\(^1\)Well- and ill-posedness in Hadamard’s sense are defined as follows. Let \(A: D \rightarrow R\) be a continuous mapping between pseudo-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) \(A^{-1}: R \rightarrow D\) 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 continuous in “the data” \(r\). Otherwise, the equation is called “ill-posed” in Hadamard’s sense.

\(^2\)In Remark 1 below, we argue that replacing the norm in the numerator is not a significant modification in the sense that most ill-posed problems, and in particular all severely ill-posed problems, imply (7) under either norm.
Well-posedness in Hadamard’s sense is useful for establishing consistency of the solution to the equation \( Ad = r \): if, for a sequence of estimators \( \hat{r}_n \) of \( r \), we have \( \rho_R(\hat{r}_n, r) \to 0 \) as \( n \to \infty \), then \( \rho_D(\hat{d}_n, d) \to 0 \) as \( n \to \infty \) where \( \hat{d}_n = A^{-1}\hat{r}_n \) is an estimator of \( d \). The concept of quantitative well-posedness is stronger than well-posedness in Hadamard’s sense because it requires a quantitative control on the modulus of continuity of the inverse map \( A^{-1} : R \to D \). Therefore, quantitative well-posedness guarantees that the convergence rate of \( \hat{d}_n \) to \( d \) is not slower than that of \( \hat{r}_n \) to \( r \) and in turn allows to establish not only consistency, but also a fast convergence rate. Local quantitative well-posedness is a weaker concept since it applies only to convergence to some particular value \( d = d_0 \).

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

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

Next, we show that Theorem 1 implies an upper bound on the measure of ill-posedness in equation (2). Importantly, the bound that we derive below is valid not only on the set of monotone functions \( \mathcal{M} \) but also on the set of functions that are not too far from \( \mathcal{M} \). It is exactly this corollary of Theorem 1 that allows us to obtain a fast convergence rate of the monotone NPIV estimator not only when the regression function \( g(x) \) is constant but, more generally, when \( g(x) \) is not too far away from constant.

For \( a \in \mathbb{R} \), let

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

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

\[
\tau(a) := \sup_{\|h\|_{2,t} = 1} \frac{\|h\|_{2,t}}{\|Th\|_2}.
\]

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

**Corollary 3 (Bound for the Measure of Ill-Posedness).** Let Assumptions 1 and 2 be satisfied. Then there exist constants $c_{\tau} > 0$ and $0 < C_{\tau} < \infty$ depending only on $\zeta$ such that
\[
\tau(a) \leq C_{\tau}
\] for all $a \leq c_{\tau}$.

**Remark 1.** Under Assumptions 1 and 2, the integral operator $T$ satisfies (7). Moreover, in many cases, and in particular in all severely ill-posed cases, there exists a sequence $\{h_k, k \geq 1\}$ such that
\[
\frac{\|h_k\|_{2,t}}{\|Th_k\|_2} \to \infty \text{ as } k \to \infty.
\] (10)

Indeed, under Assumptions 1 and 2, $T$ is compact, and so the spectral theorem implies that there exists a spectral decomposition of operator $T$, $\{(h^0_j, \varphi_j), j \geq 1\}$, where $\{h^0_j, j \geq 1\}$ 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 $\|Th^0_j\|_2 = \varphi_j \|h^0_j\|_2 = \varphi_j$. Also, Lemma 8 in the appendix shows that if $\{h^0_j, j \geq 1\}$ is an orthonormal basis in $L^2[0,1]$, then for any $\alpha > 0$, $\|h^0_j\|_{2,t} > j^{-1/2-\alpha}$ for infinitely many $j$, and so there exists a subsequence $\{h^0_{j_k}, k \geq 1\}$ such that $\|h^0_{j_k}\|_{2,t} > j_k^{-1/2-\alpha}$. Therefore, if $j^{1/2+\alpha}\varphi_j \to 0$ as $j \to \infty$, using $\|h^0_{j_k}\|_2 = 1$ for all $k \geq 1$, we conclude that for the subsequence $h_k = h^0_{j_k}$,
\[
\frac{\|h_k\|_{2,t}}{\|Th_k\|_2} \geq \frac{\|h^0_{j_k}\|_2}{j_k^{1/2+\alpha}\|Th^0_{j_k}\|_2} = \frac{1}{j_k^{1/2+\alpha}\varphi_{j_k}} \to \infty \text{ as } k \to \infty
\]
leading to (10). Note also that for severely ill-posed problems, there exists a constant $c > 0$ such that $\varphi_j \leq e^{-cj}$ for all large $j$, so that the condition $j^{1/2+\alpha}\varphi_j \to 0$ as $j \to \infty$ necessarily holds. Thus, under our Assumptions 1 and 2, the restriction in Theorem 1 that $h$ belongs to the subspace $\mathcal{M}$ of monotone functions in $L^2[0,1]$ plays a crucial role for the result (6) to hold. On the other hand, whether the result (6) can be obtained for all $h \in \mathcal{M}$ without imposing our MIV Assumption 1 appears to be an open question. □

**Remark 2.** In Example 1, it is well known that the integral operator $T$ corresponding to the bivariate normal distribution has singular values decreasing exponentially fast. Thus, the spectral decomposition $\{(h_k, \varphi_k), k \geq 1\}$ of the operator $T$ satisfies $\varphi_k = \rho^k$ for all $k$ and some $\rho < 1$, so that
\[
\frac{\|h_k\|_2}{\|Th_k\|_2} = \left(\frac{1}{\rho}\right)^k.
\]
Since \((1/\rho)^k \to \infty\) as \(k \to \infty\) exponentially fast, the normal density leads to a severely ill-posed problem. Moreover, by Lemma 8, for any \(\alpha > 0\) and \(\rho' \in (\rho, 1)\),
\[
\frac{\|h_k\|_{2,t}}{\|Th_k\|_2} > \frac{1}{k^{1/2+\alpha}} \left(\frac{1}{\rho}\right)^k \geq \left(\frac{1}{\rho'}\right)^k
\]
for infinitely many \(k\). Thus, replacing the \(L^2\) norm \(\|\cdot\|_2\) by the truncated \(L^2\) norm \(\|\cdot\|_{2,t}\) preserves the severe ill-posedness of the problem. However, it follows from Theorem 1 that uniformly over all \(h \in \mathcal{M}\), \(\|h\|_{2,t}/\|Th\|_2 \leq \bar{C}\). Therefore, in Example 1, as well as in all other severely ill-posed problems, imposing monotonicity on the function \(h \in L^2[0, 1]\) significantly changes the properties of the ratio \(\|h\|_{2,t}/\|Th\|_2\).

Remark 3. In Example 2, the first-stage relationship has a two-dimensional vector \((U_1, U_2)\) of unobserved heterogeneity, a feature that is common in economic applications; Imbens (2007) and Kasy (2014) provide examples. By Proposition 4 of Kasy (2011), there does not exist any control function \(C : [0, 1]^2 \to \mathbb{R}\) such that \(C\) is invertible in its second argument and \(X \perp \varepsilon|V\) with \(V = C(X, W)\). One consequence of this observation is that our MIV Assumption 1 does not imply any of the existing control function conditions such as those in Newey, Powell, and Vella (1999) and Imbens and Newey (2009), for example. Therefore, we view the control function approach as a complementary approach to avoiding ill-posedness under a set of assumptions that is neither weaker nor stronger than ours.

Remark 4. Finally, let us briefly comment on the role of the truncated norm \(\|\cdot\|_{2,t}\) in (6). There are two reasons why we need the truncated norm \(\|\cdot\|_{2,t}\) rather than the usual \(\|\cdot\|_2\)-norm. First, we want to allow for the normal density as in Example 1, which violates condition (ii) of Assumption 2 if we set \([x_1, x_2] = [0, 1]\). Second, when \([x_1, x_2] = [\bar{x}_1, \bar{x}_2] = [w_1, w_2] = [0, 1]\) and Assumptions 1 and 2 hold, we can show (see Lemma 3 in the appendix) that there exists a constant \(0 < C_2 < \infty\) such that
\[
\|h\|_1 \leq C_2\|Th\|_1
\]
for any increasing and continuously differentiable function \(h \in L^1[0, 1]\). To extend this result to \(L^2[0, 1]\)-norms we need to introduce a positive, but arbitrarily small, amount of trimming at the boundaries, so that we have a control \(\|\cdot\|_{2,t} \leq C\|\cdot\|_1\) for some constant \(C\).

3 Identification Bounds under Monotonicity

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

**Assumption 3** (Monotone Regression). *The function \( g(x) \) is monotone.*

This is a mild assumption that we expect to hold in many empirical applications such
as the production example in the introduction. Note also that our monotone regression
assumption is the same as the monotone treatment response assumption of Manski (1997).

**Lemma 1** (Identification of the sign of the slope). *Suppose Assumptions 1, 2, and 3 hold
and that \( g \) is continuously differentiable. Then \( \text{sign}(Dg(x)) \) is identified.*

This lemma is very useful in the sense that, under the regularity conditions of Assumption 2, monotone instruments and a monotone regression function suffice to identify
the sign of the regression function’s slope, even though the regression function itself is,
in general, not point-identified. In many empirical applications it is natural to assume a
monotone relationship between outcome variable \( Y \) and endogenous covariate \( X \), given by
the function \( g \), but the main question of interest concerns not the exact shape of \( g \) itself,
but whether the effect of \( X \) on \( Y \), given by the slope of \( g \), is positive, zero, or negative; see, for example, the discussion in Abrevaya, Hausman, and Khan (2010)). By Lemma 1,
this question can be answered in large samples under our conditions.

**Remark 5.** In fact, Lemma 1 yields a surprisingly simple way to test the sign of the slope
of the function \( g \). Indeed, the proof of Lemma 1 reveals that \( g \) is increasing, constant,
or decreasing if the function \( w \mapsto E[Y|W = w] \) is increasing, constant, or decreasing,
respectively. By Chebyshev’s association inequality (Lemma 7 in the appendix), the
latter assertions are equivalent to the coefficient \( \beta \) in the linear regression model

\[
Y = \alpha + \beta W + U, \quad E[UV] = 0
\]

(11)

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

\[
\text{cov}(W,Y) = E[XY] - E[W]E[Y] = E[W|Y|W] - E[W]E[Y|W] = \text{cov}(W, E[Y|W])
\]

by the law of iterated expectations. Therefore, under our conditions, hypotheses about
the sign of the slope of the function \( g \) can be tested by testing corresponding hypotheses
about the sign of the slope coefficient \( \beta \) in the linear regression model (11).
It turns out that our two monotonicity assumptions possess identifying power beyond the slope of the regression function.

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

The following lemma provides necessary conditions for observational equivalence.

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

Lemma 2 suggests the construction of bounds on the regression function as $\Theta' := \mathcal{M}\setminus\Delta$ with

$$\Delta := \{g' \in \mathcal{M} : \text{there exists } h \in L^2[0, 1] \text{ such that } g' - g + h \in \mathcal{M} \text{ and } \|h\|_2 + \bar{C}\|T\|_2\|h\|_2 < \|g' - g''\|_2\}.$$ (12)

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

It is possible to provide more insight into which functions are in $\Delta$ and thus not in $\Theta'$. First, all functions in $\Theta'$ have to intersect, otherwise they are not observationally equivalent. Second, for a given $g' \in \mathcal{M}$ and $h \in L^2[0, 1]$ such that $g' - g + h$ is monotone, the inequality in condition (12) is satisfied if $\|h\|_2$ is not too large relative to $\|g' - g\|_2$. In the extreme case, consider setting $h = 0$ to see that $\Theta'$ does not contain elements $g'$ 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, functions $g'$ that are much steeper than $g$ are excluded from $\Theta'$. Finally, since by Lemma 1 the sign of $g$ is identified, the set $\Theta'$ can only contain increasing or decreasing functions, but not both.

4 Fast Convergence Rate under Monotonicity

The rate at which unrestricted NPIV estimators converge to their probability limit depends crucially on the so-called sieve measure of ill-posedness, which, unlike $\tau(a)$, does
not measure ill-posedness over the space $\mathcal{H}(a)$, but rather over a finite-dimensional approximation $\mathcal{H}_n(a)$ to $\mathcal{H}(a)$. In particular, the convergence rate is slower the faster the sieve measure of ill-posedness grows with the dimensionality of the sieve space $\mathcal{H}_n(a)$. The convergence rates can be as slow as logarithmic in the severely ill-posed case. Since by Corollary 3, our monotonicity assumptions imply boundedness of $\tau(a)$ for some range of finite values $a$, we expect the monotonicity restrictions to translate into favorable large sample properties of an estimator of $g$ that imposes those monotonicity constraints. In fact, we show below that imposing only monotonicity on the regression function suffices to produce significant improvements in the convergence rate relative to the rate of the unrestricted estimator in the absence of the monotonicity constraints.

Suppose we observe an i.i.d. sample $(Y_i, X_i, W_i)$, $i = 1, \ldots, n$, from the distribution of $(Y, X, W)$. To define our estimator, we introduce the following the notation. Let $p_1(x), p_2(x), \ldots$ and $q_1(w), q_2(w), \ldots$ be two orthonormal bases in $L^2[0,1]$. For $K = K_n \geq 1$ and $J = J_n \geq K_n$, denote

$$p(x) := (p_1(x), \ldots, p_K(x))'$$

and

$$q(w) := (q_1(w), \ldots, q_J(w))'.$$

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

$$\hat{\beta}^u := \arg\min_{b \in \mathbb{R}^K : \|b\| \leq C_b} (Y - Pb)'Q(Q'Q)^{-1}Q'(Y - Pb)$$

(13)

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

$$\hat{\beta}^r := \arg\min_{b \in \mathbb{R}^K : p(\cdot)'b \in \mathcal{H}_n(0), \|b\| \leq C_b} (Y - Pb)'Q(Q'Q)^{-1}Q'(Y - Pb)$$

(14)

which imposes the constraint that the estimator is a nondecreasing function.

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

$$\mathcal{H}_n(a) := \left\{ h \in L^2[0,1] : \exists b_1, \ldots, b_{K_n} \in \mathbb{R} \text{ with } h = \sum_{j=1}^{K_n} b_j p_j \text{ and } \inf_{x \in [0,1]} Dh(x) \geq -a \right\}$$

that become dense in $\mathcal{H}(a)$ as $n \to \infty$. Define

$$\tau_{n,t}(a) := \sup_{h \in \mathcal{H}_n(a), \|h\|_2 = 1} \frac{\|h\|_{L^2}}{\|Th\|_2} \text{ and } \tau_n := \sup_{h \in \mathcal{H}_n(\infty)} \frac{\|h\|_2}{\|Th\|_2}.$$
The sieve measure of ill-posedness defined in Blundell, Chen, and Kristensen (2007) and also used, for example, in Horowitz (2012) is $\tau_n$. Blundell, Chen, and Kristensen (2007) show that $\tau_n$ is related to the eigenvalues of $T^* T$, where $T^*$ is the adjoint of $T$. If the eigenvalues converge to zero at the rate $K^{-2r}$ as $K \to \infty$, then the measure of ill-posedness diverges at a polynomial rate, $\tau_n = O(K^r_n)$. This case is typically called the “mildly ill-posed” case. On the other hand, when the eigenvalues decrease at a fast exponential rate, then $\tau_n = O(e^{cK_n})$, for some constant $c > 0$, and this case is typically called “severely ill-posed”.

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

We now specify the conditions that we need to derive non-asymptotic risk bounds for the estimators $\hat{g}^u(x)$ and $\hat{g}^r(x).

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

This is a mild moment condition. Let $s > 0$ be some constant.

**Assumption 5** (Approximation of $g$). There exist $\beta_n \in \mathbb{R}^K$ and a constant $C_g < \infty$ such that the function $g_n(x) := p(x)' \beta_n$, defined for all $x \in [0, 1]$, satisfies (i) $g_n \in H_n(0)$, (ii) $\|g - g_n\|_2 \leq C_g K^{-s}$, and (iii) $\|T(g - g_n)\|_2 \leq C_g \tau_n^{-1} K^{-s}$.

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

**Assumption 6** (Approximation of $m$). There exist $\gamma_n \in \mathbb{R}^J$ and a constant $C_m < \infty$ such that the function $m_n(w) := q(w)' \gamma_n$, defined for all $w \in [0, 1]$, satisfies $\|m - m_n\|_2 \leq C_m \tau_n^{-1} J^{-s}$.
This condition is similar to Assumption 3(iii) in Horowitz (2012). Finally, define the operator $T_n : L^2[0,1] \to L^2[0,1]$ by
\[
(T_n h)(w) := q(w)\mathbb{E}[q(W)p(X)']\mathbb{E}[p(U)h(U)]
\]
for all $w \in [0,1]$ where $U \sim U[0,1]$.

**Assumption 7 (Operator $T$).** (i) The operator $T$ is injective and (ii) for some constant $C_a < \infty$, $\|(T - T_n)h\|_2 \leq C_a \tau_n^{-1}K^{-s}\|h\|_2$ for all $h \in \mathcal{H}_n(\infty)$.

This condition is similar to Assumption 5 in Horowitz (2012). Finally, let
\[
\xi_{K,p} := \sup_{x \in [0,1]} \|p(x)\|, \quad \xi_{J,q} := \sup_{w \in [0,1]} \|q(w)\|, \quad \xi_n := \max(\xi_{K,p}, \xi_{J,q}).
\]
The following theorem states non-asymptotic risk bounds for the estimators $\hat{g}^u$ and $\hat{g}^r$.

**Theorem 2 (Non-asymptotic risk bounds).** Let Assumptions 1-7 be satisfied and $\delta \geq 0$ be some constant. If $\xi_n^2 \log n/n \leq c$ for sufficiently small $c > 0$, then
\[
\|\hat{g}^u - g\|_{2,t} \leq C \left\{ \tau_{n,t}(\infty) \left( \left( \frac{K}{\alpha n} \right)^{1/2} + \left( \frac{\xi_n^2 \log n}{n} \right)^{1/2} \right) + K^{-s} \right\}, \quad (15)
\]
\[
\|\hat{g}^r - g\|_{2,t} \leq C \left\{ \max \left[ \delta, \tau_{n,t}(\infty) \left( \left( \frac{K}{\alpha n} \right)^{1/2} + \left( \frac{\xi_n^2 \log n}{n} \right)^{1/2} \right), \frac{1}{\delta} \right] \right\} + K^{-s}, \quad (16)
\]
\[
\|\hat{g}^r - g\|_{2,t} \leq C \left\{ \max \left[ \|Dg\|_{\infty}, \left( \frac{K}{\alpha n} \right)^{1/2} + \left( \frac{\xi_n^2 \log n}{n} \right)^{1/2} \right] + K^{-s} \right\} \quad \text{with probability at least } 1 - \alpha. \quad (17)
\]

In addition, if $\tau_n^2 \xi_n^2 \log n/n \leq c$ for sufficiently small $c > 0$, then
\[
\|\hat{g}^u - g\|_{2,t} \leq C \left\{ \tau_{n,t}(\infty) \left( \left( \frac{K}{\alpha n} \right)^{1/2} + K^{-s} \right) \right\}, \quad (18)
\]
\[
\|\hat{g}^r - g\|_{2,t} \leq C \left\{ \max \left[ \delta, \tau_{n,t}(\infty) \left( \frac{K}{\alpha n} \right)^{1/2} + K^{-s} \right] \right\}, \quad (19)
\]
\[
\|\hat{g}^r - g\|_{2,t} \leq C \left\{ \max \left[ \|Dg\|_{\infty}, \left( \frac{K}{\alpha n} \right)^{1/2} + K^{-s} \right] \right\}, \quad (20)
\]
with probability at least $1 - \alpha$. Here the constants $c, C < \infty$ can be chosen to depend only on the constants appearing in Assumptions 1-7.

The bound (18) implies that $\|\hat{g}^u - g\|_{2,t} = O_P(\tau_{n,t}(\infty)(K/n)^{1/2} + K^{-s})$, which is a standard rate obtained in the literature on the NPIV model up to the modification that we replace $\tau_n$ usually appearing in the literature by $\tau_{n,t}(\infty)$; see, for example, Blundell, Chen, and Kristensen (2007).

Setting $\delta = 0$ shows that our bounds on $\|\hat{g}^r - g\|_{2,t}$ in (16) and (19) are always not worse than the bounds on $\|\hat{g}^u - g\|_{2,t}$ in (15) and (18), at least up-to a constant $C$, which may vary across (15)-(20).
When \( g \) is sufficiently flat, that is, \( \|Dg\|_{\infty} \) is sufficiently small, the bounds on \( \|\hat{g}^r - g\|_{2,t} \) in (17) and (20) are much better than the bounds on \( \|\hat{g}^u - g\|_{2,t} \) in (15) and (18). In particular, in the local-to-constant asymptotics, where the regression function \( g = g_n \) is allowed to depend on \( n \) and belongs to the shrinking (as \( n \to \infty \)) neighborhood of constant functions, we obtain, for example, from (17) the following result:

**Corollary 4** (Fast convergence rate of \( \hat{g}^r \) in the local-to-constant asymptotics). Assume that the function \( g = g_n \) is such that \( \sup_{x \in [0,1]} Dg(x) = O((K \log n/n)^{1/2}) \). In addition, let Assumptions 1-7 be satisfied with the same constants for all \( n \). Finally, assume that \( \xi_n^2 \log n/n \leq c \) for sufficiently small \( c \). Then

\[
\|\hat{g}^r - g\|_{2,t} = O_p((K \log n/n)^{1/2} + K^{-s}).
\]

Therefore,

\[
\|\hat{g}^r - g\|_{2,t} = O_p(n^{-s/(1+2s)} \sqrt{\log n})
\]

if we set \( K = K_n = C_K n^{1/(1+2s)} \), for some \( 0 < C_K < \infty \).

The local-to-constant asymptotics considered in this corollary captures the finite sample situation in which the regression function is not too steep relative to the sample size. The convergence rate in this corollary is the standard polynomial rate of nonparametric conditional mean regression estimators up to a \( (\log n)^{1/2} \) factor, regardless of whether the original NPIV problem without monotonicity is mildly or severely ill-posed. One way to interpret this result is that the monotone estimator is able to recover regression functions in the shrinking neighborhood of flat functions at a very fast polynomial rate. Therefore, in finite samples, we expect the estimator to perform better the smaller the upper bound on the derivative of the regression function relative to the sample size.

**Remark 6.** Notice that the fast convergence rates derived in this section are obtained under both monotonicity assumptions, Assumptions 1 and 3, but the estimator imposes only the monotonicity of the regression function, not that of the instrument. Therefore, our proposed restricted estimator consistently estimates the regression function even when the monotone IV assumption is violated.

**Remark 7.** In the local-to-constant asymptotic framework where \( \sup_{x \in [0,1]} Dg(x) = O((K \log n/n)^{1/2}) \), the rate of convergence in (21) can also be obtained by simply fitting a constant. However, such an estimator, unlike our monotone estimator, is not consistent when the regression function \( g(x) \) does not drift towards a constant.
5 Testing Monotonicity of the Instruments

In this section, we consider the problem of testing the main condition (3) of our Monotone IV Assumption 1. More precisely, we test the null hypothesis,

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

against the alternative,

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

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

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

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

Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function satisfying the following conditions:
Assumption 8 (Kernel). The kernel function $K : \mathbb{R} \to \mathbb{R}$ is such that (i) $K(w) > 0$ for all $w \in (-1, 1)$, (ii) $K(w) = 0$ for all $w \notin (-1, 1)$, (iii) $K$ is continuous, and (iv) $\int_{-\infty}^{\infty} K(w)dw = 1$.

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

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

$$E[(1\{X_i \leq x\} - 1\{X_j \leq x\})\text{sign}(W_i - W_j)K_{h}(W_i - w)K_{h}(W_j - w)] \leq 0 \quad (22)$$

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

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

taking the sum of the left-hand side in (22) over $i, j = 1, \ldots, n$, and rearranging give

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

or, equivalently,

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

(23)

where

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

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

Assumption 9 (Bandwidth values). The collection of bandwidth values is $\mathcal{H}_n := \{h \in \mathbb{R} : h = u^l/2, l = 0, 1, 2, \ldots, h \geq h_{\min}\}$ for some $u \in (0, 1)$ where $h_{\min} := h_{\min,n}$ is such that $1/(nh_{\min}) \leq C_h n^{-c_h}$ for some constants $c_h, C_h > 0$.

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

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

$$T := \max_{(x,w,h) \in X_n \times W_n \times \mathcal{H}_n} \frac{\sum_{i=1}^{n} k_{i,h}(w)1\{X_i \leq x\}}{\left(\sum_{i=1}^{n} k_{i,h}(w)^2\right)^{1/2}}. \quad (24)$$

21
The statistic $T$ is most closely related to that in Lee, Linton, and Whang (2009). The main difference is that we take the maximum with respect to the set of bandwidth values $h \in \mathcal{H}_n$ to achieve adaptiveness of the test.

We now discuss the construction of a critical value for the test. Suppose that we would like to have a test of level (approximately) $\alpha$. As succinctly demonstrated by Lee, Linton, and Whang (2009), the derivation of the asymptotic distribution of $T$ is complicated even when $\mathcal{H}_n$ is a singleton. Moreover, when $\mathcal{H}_n$ is not a singleton, it is generally unknown whether $T$ converges to some nondegenerate asymptotic distribution after an appropriate normalization. We avoid these complications by employing the nonasymptotic approach developed in CCK and using a multiplier bootstrap critical value for the test. Let $e_1, \ldots, e_n$ be an i.i.d. sequence of $N(0, 1)$ random variables that are independent of the data. Also, let $\hat{F}_{X|W}(x|w)$ be an estimator of $F_{X|W}(x|w)$ satisfying the following conditions:

**Assumption 10 (Estimator of $F_{X|W}(x|w)$).** The estimator $\hat{F}_{X|W}(x|w)$ of $F_{X|W}(x|w)$ is such that (i) $P\left( \max_{(x,w) \in \mathcal{X}_n \times \mathcal{W}_n} |\hat{F}_{X|W}(x|w) - F_{X|W}(x|w)| > C_F n^{-c_F} |\{\mathcal{W}_n\} \right) \leq C_F n^{-c_F}$ for some constants $c_F, C_F > 0$, and (ii) $|\hat{F}_{X|W}(x|w)| \leq C_F$ for all $(x, w) \in \mathcal{X}_n \times \mathcal{W}_n$.

This is a mild assumption implying uniform consistency of an estimator $\hat{F}_{X|W}(x|w)$ of $F_{X|W}(x|w)$ over $(x, w) \in \mathcal{X}_n \times \mathcal{W}_n$. For completeness, the Supplemental Material provides sufficient conditions for Assumption 10 when $\hat{F}_{X|W}(x|w)$ is a series estimator.

Define a bootstrap test statistic by

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

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

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

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

We reject $H_0$ if and only if $T > c(\alpha)$. To prove validity of this test, we assume that the conditional distribution function $F_{X|W}(x|w)$ satisfies the following condition:
**Assumption 11** (Conditional Distribution Function $F_{X|W}(x|w)$). The conditional distribution function $F_{X|W}(x|w)$ is such that $c_\epsilon \leq F_{X|W}(\epsilon|w) \leq F_{X|W}(1-\epsilon|w) \leq C_\epsilon$ for all $w \in (0,1)$ and some constants $0 < c_\epsilon < C_\epsilon < 1$.

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

**Theorem 3** (Polynomial Size Control). Let Assumptions 2, 8, 9, and 10 be satisfied. If $H_0$ holds, then

\[
P(T > c(\alpha)) \leq \alpha + Cn^{-c}.
\]

(25)

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

\[
|P(T > c(\alpha)) - \alpha| \leq Cn^{-c}.
\]

(26)

In both, (25) and (26), the constants $c$ and $C$ depend only on $c_W, C_W, c_h, C_h, c_F, C_F, c_\epsilon, C_\epsilon$, and the kernel $K$.

**Remark 8** (Weak Condition on the Bandwidth Values). Our theorem requires

\[
\frac{1}{nh} \leq C_h n^{-c_h}
\]

(27)

for all $h \in \mathcal{H}_n$, which is considerably weaker than the analogous condition in Lee, Linton, and Whang (2009) who require $1/(nh^3) \to 0$, up-to logs. This is achieved by using a conditional test and by applying the results of CCK. As follows from the proof of the theorem, the multiplier bootstrap distribution approximates the conditional distribution of the test statistic given $W_n = \{W_1, \ldots, W_n\}$. Conditional on $W_n$, the denominator in the definition of $T$ is fixed, and does not require any approximation. Instead, we could try to approximate the denominator of $T$ by its probability limit. This is done in Ghosal, Sen, and Vaart (2000) using the theory of Hoeffding projections but they require the condition $1/nh^2 \to 0$. Our weak condition (27) also crucially relies on the fact that we use the results of CCK. Indeed, it has already been demonstrated (see Chernozhukov, Chetverikov, and Kato (2013a,b), and Belloni, Chernozhukov, Chetverikov, and Kato (2014)) that, in typical nonparametric problems, the techniques of CCK often lead to weak conditions on the bandwidth value or the number of series terms. Our theorem is another instance of this fact.

**Remark 9** (Polynomial Size Control). Note that, by (25) and (26), the probability of rejecting $H_0$ when $H_0$ is satisfied can exceed the nominal level $\alpha$ only by a term that is polynomially small in $n$. We refer to this phenomenon as a *polynomial size control*. As explained in Lee, Linton, and Whang (2009), when $\mathcal{H}_n$ is a singleton, convergence of
to the limit distribution is logarithmically slow. Therefore, Lee, Linton, and Whang (2009) used higher-order corrections derived in Piterbarg (1996) to obtain polynomial size control. Here we show that the multiplier bootstrap also gives higher-order corrections and leads to polynomial size control. This feature of our theorem is also inherited from the results of CCK.

Remark 10 (Uniformity). The constants $c$ and $C$ in (25) and (26) depend on the data generating process only via constants (and the kernel) appearing in Assumptions 2, 8, 9, and 10. Therefore, inequalities (25) and (26) hold uniformly over all data generating processes satisfying Assumptions 2, 8, 9, and 10 with the same constants. The issue of uniformity has been studied intensively in the recent econometric literature and several techniques have been developed to prove uniformity (for instance, Mikusheva (2007) and Andrews and Guggenberger (2009)). Here we obtain uniformity of the result essentially for free since the distributional approximation theorems of CCK, which we employ, are nonasymptotic, and do not rely on convergence arguments.

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

Theorem 4 (Consistency). Let Assumptions 2, 8, 9, and 10 be satisfied and assume that $F_{X|W}(x|w)$ is continuously differentiable. If $H_a$ holds with $D_wF_{X|W}(x|w) > 0$ for some $x \in (\epsilon, 1-\epsilon)$ and $w \in (0,1)$, then

$$P(T > c(\alpha)) \to 1 \text{ as } n \to \infty.$$ (28)

This theorem shows that our test is consistent against any model in $H_a$ (with smooth $F_{X|W}(x|w)$) whose deviation from $H_0$ is not on the boundary, so that the deviation $D_wF_{X|W}(x|w) > 0$ occurs for $x \in (\epsilon, 1-\epsilon)$. It is also possible to extend our results and to show that Theorems 3 and 4 hold with $\epsilon = 0$ at the expense of additional technicalities.

6 Simulations

In this section, we study the finite sample behavior of our restricted estimator that imposes monotonicity and compare its performance to that of the unrestricted estimator. We consider the NPIV model $Y = g(X) + \varepsilon, E[\varepsilon|W] = 0$, for two different regression functions:

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

Model 2: $g(x) = 10\kappa \left[-(x - 0.25)^2 \mathbf{1}\{x \in [0, 0.25]\} + (x - 0.75)^2 \mathbf{1}\{x \in [0.75, 1]\}\right]$
where \( \varepsilon = \kappa \sigma \tilde{\varepsilon} \) and \( \bar{\varepsilon} = \eta \varepsilon + \sqrt{1 - \eta^2} \nu \). The regressors and instruments are generated by \( X = \Phi(\xi) \) and \( W = \Phi(\zeta) \), respectively, where \( \Phi \) is the standard normal cdf and \( \xi = \rho \zeta + \sqrt{1 - \rho^2} \varepsilon \). The errors are generated by \( (\nu, \zeta, \epsilon) \sim N(0, I) \).

We vary the parameter \( \kappa \) in \{1, 0.5, 0.1\} to study how the restricted and unrestricted estimators’ performance compares when the regression function becomes flatter. \( \eta \) governs the dependence of \( X \) on the regression error \( \varepsilon \) and \( \rho \) the strength of the first stage. All results are based on 1,000 MC samples and the normalized B-spline basis for \( p(x) \) and \( q(w) \) of degree 3 and 4, respectively.

Tables 1–4 report the squared bias, and variance of the two estimators, averaged over a grid on the interval \([0, 1]\), together with the mean integrated square error (“MISE”) and the ratio of the restricted estimator’s MISE divided by the unrestricted estimator’s MISE. \( k_X \) and \( k_W \) denote, respectively, the number of knots used for the basis \( p(x) \) and \( q(w) \). The first two tables vary the number of knots, and the remaining two the dependence parameters \( \rho \) and \( \eta \). Different sample sizes and different values for the error variance \( \sigma_{\tilde{\varepsilon}}^2 \) yield qualitatively similar results. Figures 2 and 3 show the two estimators for a particular combination of the simulation parameters. The dashed lines represent confidence bands, computed as two times the (pointwise) empirical standard deviation of the estimators across simulation samples.

7 Conclusion

In this paper, we provide a theoretical explanation for the dramatic gains in finite sample performance that are possible when imposing monotonicity in the NPIV estimation procedure. In particular, we show that monotone instruments together with a monotone regression function lead to a so-called locally quantitatively well-posed problem. This feature of the restricted problem significantly reduces the statistical difficulty in nonparametric estimation of the regression function. We show that the restricted NPIV estimator may possess finite-sample risk much lower than the unrestricted estimator, especially when the regression function is not too steep and the unrestricted estimator exhibits high variability. In fact, the constrained estimator’s risk may be comparable to that of standard conditional mean estimators.
A Proofs for Section 2

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

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

\[
\|h\|_{2,t} \leq C_1 \|h\|_{1,t} \tag{29}
\]

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

\[
\|h\|_{2,t} = \left( \int_{\bar{x}_1}^{\bar{x}_2} h(x)^2dx \right)^{1/2} \leq \sqrt{\bar{x}_2 - \bar{x}_1} \max\{|h(\bar{x}_1)|, |h(\bar{x}_2)|\}
\]

\[
\leq \sqrt{\bar{x}_2 - \bar{x}_1} \frac{\int_{\bar{x}_1}^{\bar{x}_2} |h(x)|dx}{\min\{\bar{x}_1 - x_1, x_2 - \bar{x}_2\}}
\]

so that (29) follows. Therefore, for any increasing continuously differentiable \( h \in \mathcal{M} \),

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

where the first inequality follows from (29), the second from Lemma 3 below (which is the main step in the proof of the theorem), and the third by Jensen’s inequality. Hence, conclusion (6) of Theorem 1 holds for increasing continuously differentiable \( h \in \mathcal{M} \) with \( \bar{C} := C_1 C_2 \) and \( C_2 \) as defined in Lemma 3.

Next, for any increasing function \( h \in \mathcal{M} \), it follows from Lemma 11 that one can find a sequence of increasing continuously differentiable functions \( h_k \in \mathcal{M} \), \( k \geq 1 \), such that \( \|h_k - h\|_2 \to 0 \) as \( k \to \infty \). Therefore, by the triangle inequality,

\[
\|h\|_{2,t} \leq \|h_k\|_{2,t} + \|h_k - h\|_{2,t} \leq \bar{C} \|Th_k\|_2 + \|h_k - h\|_{2,t}
\]

\[
\leq \bar{C} \|Th\|_2 + \bar{C} \|T(h_k - h)\|_2 + \|h_k - h\|_{2,t}
\]

\[
\leq \bar{C} \|Th\|_2 + \bar{C} \|T\|_2 \|h_k - h\|_2 + \|h_k - h\|_{2,t}
\]

\[
\leq \bar{C} \|Th\|_2 + (\bar{C} \|T\|_2 + 1) \|h_k - h\|_2
\]

\[
\leq \bar{C} \|Th\|_2 + (\bar{C} C_T + 1) \|h_k - h\|_2
\]

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

Finally, since for any decreasing \( h \in \mathcal{M} \), we have that \( -h \in \mathcal{M} \) is increasing, \( \|-h\|_{2,t} = \|h\|_{2,t} \) and \( \|Th\|_2 = \|T(-h)\|_2 \), conclusion (6) of Theorem 1 also holds for all decreasing \( h \in \mathcal{M} \), and thus for all \( h \in \mathcal{M} \). This completes the proof of the theorem. Q.E.D.
Lemma 3. Let Assumptions 1 and 2 hold. Then for any increasing continuously differentiable $h \in L^1[0,1]$, 
\[
\|h\|_{1,t} \leq C_2\|Th\|_1
\]
where $C_2 := 1/c_p$ and $c_p := c_wc_f/2\min\{1-w_2,w_1\}\min\{(C_F - 1)/2,1\}$.

Proof. Take any increasing continuously differentiable function $h \in L^1[0,1]$ such that $\|h\|_{1,t} = 1$. Define $M(w) := E[h(X)|W = w]$ for all $w \in [0,1]$ and note that 
\[
\|Th\|_1 = \int_0^1 |M(w)f_W(w)|dw \geq c_w\int_0^1 |M(w)|dw
\]
where the inequality follows from Assumption 2(iii). Therefore, the asserted claim follows if we can show that $\int_0^1 |M(w)|dw$ is bounded away from zero by a constant that depends only on $\zeta$.

First, note that $M(w)$ is increasing. This is because, by integration by parts,
\[
M(w) = \int_0^1 h(x)f_X|W(x|w)dx = h(1) - \int_0^1 Dh(x)F_X|W(x|w)dx,
\]
so that condition (3) of Assumption 1 and $Dh(x) \geq 0$ for all $x$ imply that the function $M(w)$ is increasing.

Consider the case in which $h(x) \geq 0$ for all $x \in [0,1]$. Then $M(w) \geq 0$ for all $w \in [0,1]$. Therefore,
\[
\int_0^1 |M(w)|dw = \int_{w_2}^{w_1} |M(w)|dw \geq (1 - w_2)M(w_2) = (1 - w_2)\int_0^1 h(x)f_X|W(x|w_2)dx
\]
\[
\geq (1 - w_2)\int_{x_1}^{x_2} h(x)f_X|W(x|w_2)dx \geq (1 - w_2)c_f\int_{x_1}^{x_2} h(x)dx
\]
\[
= (1 - w_2)c_f\|h\|_{1,t} = (1 - w_2)c_f > 0
\]
by Assumption 2(ii). Similarly,
\[
\int_0^1 |M(w)|dw \geq w_1c_f > 0
\]
when $h(x) \leq 0$ for all $x \in [0,1]$. Therefore, it remains to consider the case in which there exists $x^* \in (0,1)$ such that $h(x) \leq 0$ for $x \leq x^*$ and $h(x) \geq 0$ for $x > x^*$. Since $h(x)$ is continuous, $h(x^*) = 0$, and so integration by parts yields
\[
M(w) = \int_0^{x^*} h(x)f_X|W(x|w)dx + \int_{x^*}^1 h(x)f_X|W(x|w)dx
\]
\[
= -\int_0^{x^*} Dh(x)F_X|W(x|w)dx + \int_{x^*}^1 Dh(x)(1 - F_X|W(x|w))dx. \quad (30)
\]
For \( k = 1, 2 \), let \( A_k := \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_k)) \) and \( B_k := \int_0^{x^*} Dh(x)F_{X|W}(x|w_k)dx \), so that \( M(w_k) = A_k - B_k \).

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

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

\[
A_1 + B_2 = \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx + \int_{x^*}^1 Dh(x)F_{X|W}(x|w_2)dx
\]

\[
= \int_{x^*}^1 h(x)f_{X|W}(x|w_1)dx - \int_0^{x^*} h(x)f_{X|W}(x|w_2)dx
\]

\[
\geq \int_{x^*}^{x_2} h(x)f_{X|W}(x|w_1)dx - \int_{x_1}^{x^*} h(x)f_{X|W}(x|w_2)dx
\]

\[
\geq c_1 \int_{x^*}^{x_2} h(x)dx + c_f \int_{x_1}^{x^*} h(x)|dx = c_f \int_{x_1}^{x_2} h(x)|dx
\]

\[
= c_f\|h\|_{1,t} = c_f > 0
\]

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

\[
M(w_1) = \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx - \int_{x^*}^{x_2} Dh(x)f_{X|W}(x|w_1)dx
\]

\[
\leq \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_2))dx - C_F \int_{x^*}^{x_2} Dh(x)f_{X|W}(x|w_2)dx
\]

\[
= A_2 - C_F B_2
\]

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

\[
M(w_2) - M(w_1) \geq (C_F - 1)B_2. \quad (32)
\]

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

\[
M(w_2) = \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_2))dx - \int_{x^*}^{x_2} Dh(x)f_{X|W}(x|w_2)dx
\]

\[
\geq C_F \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx - \int_{x^*}^{x_2} Dh(x)f_{X|W}(x|w_1)dx
\]

\[
= C_F A_1 - B_1
\]

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

\[
M(w_2) - M(w_1) \geq (C_F - 1)A_1. \quad (33)
\]
In conclusion, equations (31), (32), and (33) yield

\[ M(w_2) - M(w_1) \geq (C_F - 1)(A_1 + B_2)/2 \geq (C_F - 1)c_f/2 > 0. \]  

(34)

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

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

(35)

Similarly,

\[ \int_0^1 |M(w)|dw \geq w_1(C_F - 1)c_f/2 > 0 \]  

(36)

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 (35) shows that

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

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

\[ \int_0^1 |M(w)|dw \geq \int_{w_1}^{w_2} |M(w)|dw \geq w_1(C_F - 1)c_f/4 > 0. \]

This completes the proof of Case I.

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

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

With this inequality we proceed as in Case I to show that \( \int_0^1 |M(w)|dw \) is bounded from below by a positive constant that depends only on \( \zeta \). On the other hand, when \( M(w_1) \leq -c_f/2 \) we bound \( \int_0^1 |M(w)|dw \) as in (36), and the proof of Case II is complete.
Case III \((x^* \leq x_1)\): Similarly as in Case II, suppose first that \(M(w_2) \leq c_f/2\). As in Case I we have \(M(w_1) \leq A_2 - C_FB_2\) so that together with \(M(w_2) = A_2 - B_2\),

\[
M(w_2) - M(w_1) = M(w_2) - C_FM(w_2) + C_FM(w_2) - M(w_1) \\
\geq (1 - C_F)M(w_2) + (C_F - 1)A_2 \\
= (C_F - 1) \left( \int_{x^*}^{1} Dh(x)(1 - F_X|W(x|w_2)dx - M(w_2) \right) \\
= (C_F - 1) \left( \int_{x^*}^{1} h(x)f_X|W(x|w_2)dx - M(w_2) \right) \\
\geq (C_F - 1) \left( \int_{x_1}^{2w} h(x)f_X|W(x|w_2)dx - M(w_2) \right) \\
\geq (C_F - 1) \left( \int_{x_1}^{2w} h(x)dx - \frac{c_f}{2} \right) = \frac{(C_F - 1)c_f}{2} > 0
\]

and we proceed as in Case I to bound \(\int_0^1 |M(w)|dw\) from below by a positive constant that depends only on \(\zeta\). On the other hand, when \(M(w_2) > c_f/2\), we bound \(\int_0^1 |M(w)|dw\) as in \((35)\), and the proof of Case III is complete. The lemma is proven. Q.E.D.

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

\[
\sup_{x \in [\bar{x}_1, \bar{x}_2]} |g_k(x)| \leq C\|g\|_{1,t} \leq CC_2\|Tg_k\|_1 = CC_2\|h_k\|_1 \leq CC_2\|h_k\|_2 \to CC_2\|h\|_2
\]

as \(k \to \infty\), where the first inequality follows for some \(C > 0\) depending only on \(x_1, x_2, \bar{x}_1, \bar{x}_2\) by an argument similar to that used in the proof of Theorem 1 since \(g_k\) is monotone, and the second inequality follows from Lemma 3. Therefore, there exists some \(k_0\) such that for all \(k \geq k_0\), \(\sup_{x \in [\bar{x}_1, \bar{x}_2]} |g_k(x)| \leq CC_2(\|h\|_2 + 1) < \infty\). This means that for all \(k \geq k_0\), the functions \(g_k(x)\) belong to the space of monotone functions in \(L^2[0,1]\) uniformly bounded by \(CC_2(\|h\|_2 + 1) < \infty\) for \(x \in [\bar{x}_1, \bar{x}_2]\). Since this space is compact under \(\|\cdot\|_{2,t}\) (see, for example, discussion on p. 18 in van de Geer (2000)), it follows from Lemma 6 that \(\|g_k - g\|_{2,t} \to 0\) as \(k \to \infty\) as desired. Q.E.D.

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

\[
\|T^{-1}m\|_{2,t} = \|h\|_{2,t} \leq \tilde{C}\|Th\|_2 = \tilde{C}\|m\|_2
\]

by Theorem 1. Also, for any function \(m'\) such that \(m' = Tg'\) for some monotone \(g' \in \mathcal{M}\), \(g' - g = T^{-1}(m' - m)\) and \(g' - g \in \mathcal{M}\). Therefore,

\[
\|g' - g\|_{2,t} \leq \tilde{C}\|m' - m\|_2
\]
implying the asserted claim. Q.E.D.

Proof of Corollary 3. Note that since $\tau(a') \leq \tau(a'')$ whenever $a' \leq a''$, the claim for $a \leq 0$, follows from $\tau(a) \leq \tau(0) \leq \bar{C}$. 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 $h(x) + ax$ is increasing and so belongs to the class $\mathcal{M}$. Also, $\|h'\|_{2,t} \leq \|h\|_{2,t} \leq a/\sqrt{3}$. Thus, the bound (37) in Lemma 4 below applies whenever $(1 + \bar{C}\|T\|_2)a/\sqrt{3} \leq \alpha$. Therefore, for all $a$ satisfying inequality

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

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

Lemma 4. 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

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

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

Proof. Define

$$\tilde{h}(x) := \frac{h(x) + h'(x)}{\|h\|_{2,t} - \|h'\|_{2,t}}$$

for all $x \in [0, 1]$. By assumption, $\|h'\|_{2,t} < \|h\|_{2,t}$, and so the triangle inequality yields

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

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

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

Hence, applying the triangle inequality once again yields

$$\|Th\|_2 \geq (\|h\|_{2,t} - \|h'\|_{2,t})\|T\tilde{h}\|_2 - \|Th'\|_2 \geq (\|h\|_{2,t} - \|h'\|_{2,t})\|T\tilde{h}\|_2 - \|T\|_2\|h'\|_2 \geq \frac{\|h\|_{2,t} - \|h'\|_{2,t}}{\bar{C}} - \|T\|_2\|h'\|_2 \geq \frac{\|h\|_{2,t}}{\bar{C}} \left(1 - \frac{\|h'\|_{2,t} + \bar{C}\|T\|_2\|h'\|_2}{\|h\|_{2,t}}\right)$$

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

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

which is equivalent to (37). Q.E.D.
B Proofs for Section 3

Proof of Lemma 1. From the proof of Lemma 3 we know that \( g \) being increasing, constant, or decreasing implies that \( M(w) := E[YW = w] \) is increasing, constant, or decreasing, respectively. Therefore, the sign of \( Dg(x) \) is equal to the sign of \( DM(w) \), which is identified from the observed distribution of \( (Y, W) \). Q.E.D.

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

C Proofs for Section 4

Proof of Theorem 2. Let us define a function \( \hat{m} \in L^2[0, 1] \) and an operator \( \hat{T} : L^2[0, 1] \to L^2[0, 1] \) by

\[
\hat{m}(w) = q(w)'E_n[q(W_i)Y_i],
\]

\[
(\hat{T}h)(w) = q(w)'E_n[q(W_i)p(X_i)']\log n/n E[Uh(U)]
\]

for all \( w \in [0, 1] \).

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

\[
||E_n[q(W_i)p(X_i)'] - E[q(W)p(X)']|| \leq C(\xi_2^2 \log n/n)^{1/2},
\]

\[
||E_n[q(W_i)q(W_i)'] - E[q(W)q(W)']|| \leq C(\xi_2^2 \log n/n)^{1/2},
\]

\[
||\hat{m} - m||_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1}J^{-s})
\]

for some sufficiently large constant \( 0 < C < \infty \). It follows from Lemmas 5 and 12 that all three events hold jointly with probability at least \( 1 - \alpha - n^{-c} \) where \( c > 0 \) can be made arbitrarily large by increasing \( C > 0 \).

We first derive a bound on \( ||T(\hat{g} - g)||_2 \) where \( \hat{g} \) is either \( \hat{g}^a \) or \( \hat{g}^r \). Since \( m = Tg \),

\[
||T(\hat{g} - g)||_2 \leq ||(T - T_n)\hat{g}||_2 + \|(T_n - \hat{T})\hat{g}||_2 + \|\hat{T}\hat{g} - \hat{m}||_2 + ||\hat{m} - m||_2
\]

by the triangle inequality. The bound on \( ||\hat{m} - m||_2 \) is given in (41). Also, since \( ||\hat{g}||_2 \leq C_b \) by construction,

\[
||\hat{g}||_2 \leq C_b\tau_n^{-1}K^{-s}
\]

by Assumption 7(ii). In addition,

\[
||T_n - \hat{T}|| \leq C_b||T_n - \hat{T}|| = C_b||E_n[q(W_i)p(X_i)'] - E[q(W)p(X)']|| \leq C(\xi_2^2 \log n/n)^{1/2}
\]

32
by (39). Further, by Assumption 2(iii), all eigenvalues of $E[q(W)q(W)']$ are bounded from below by $c_w$ and from above by $C_w$, and so it follows from (40) that for large $n$, all eigenvalues of $Q_n := E_n[q(W_i)q(W_i)']$ are bounded below from zero and from above. Therefore, introducing $\beta_n \in \mathbb{R}^K$ such that $g_n(x) = p(x)'\beta_n$ for all $x \in [0, 1]$ and letting $\hat{\beta}$ be either $\hat{\beta}^u$ or $\hat{\beta}^r$, we obtain

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

by optimality of $\hat{\beta}$. Moreover,

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

by the triangle inequality. The term $\|(\hat{T} - T)g_n\|_2$ is bounded by $\|(\hat{T} - T_n)g_n\|_2 + \|(T_n - T)g_n\|_2$ and both of these terms can be bounded as $\|(\hat{T} - T_n)\hat{g}\|_2$ and $\|(T_n - T)\hat{g}\|_2$. The term $\|\hat{m} - m\|_2$ has been bounded above. Finally, by Assumption 5(iii),

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

Conclude that

$$
\|T(\hat{g} - g)\|_2 \leq C(\frac{J}{\alpha n})^{1/2} + (\xi^2 \log n/n)^{1/2} + \tau_n^{-1}K^{-s} \quad (42)
$$

for all large $n$ with probability at least $1 - \alpha - n^{-c}$.

Now, to prove (15), observe that by the triangle inequality,

$$
\|\hat{g}^u - g\|_{2,t} \leq \|\hat{g}^u - g_n\|_{2,t} + \|g_n - g\|_{2,t} \\
\leq \tau_{n,t}(\infty)\|T(\hat{g}^u - g_n)\|_2 + C_gK^{-s} \\
\leq \tau_{n,t}(\infty)(\|T(\hat{g}^u - g)\|_2 + \|T(g - g_n)\|_2) + C_gK^{-s} \\
\leq \tau_{n,t}(\infty)(\|T(\hat{g}^u - g)\|_2 + 2C_gK^{-s}
$$

where the second line follows from the definition of $\tau_{n,t}(\infty)$ and Assumption 5(ii) and the fourth line from Assumption 5(ii) and $\tau_{n,t}(\infty) \leq \tau_n$. So (15) follows from (42).

To prove (16), observe that

$$
\|\hat{g}^r - g_n\|_{2,t} \leq \max \left( \delta, \tau_{n,t} \left( \frac{\|Dg_n\|_{\infty}}{\delta} \right) \|T(\hat{g}^r - g_n)\|_2 \right)
$$

since $\hat{g}^r$ is increasing (indeed, if $\|\hat{g}^r - g\|_{2,t} \leq \delta$, the bound is trivial; otherwise, apply the definition of $\tau_{n,t}$ to the function $(\hat{g}^r - g_n)/\|\hat{g}^r - g_n\|_{2,t}$ and use inequality $\tau_{n,t}(\|Dg_n\|_{\infty}/\|\hat{g}^r - g_n\|_{2,t}) \leq \tau_{n,t}(\|Dg_n\|_{\infty}/\delta)$). The rest of the proof of (16) is similar to that given for (15).
To prove (17), apply (16) with $\delta = \|Dg_n\|/c_\tau$ and use Corollary 3 to show that $\tau_{n,t}(c_\tau) \leq C_\tau$.

Next, to prove (18)-(20), we need a different bound on $\|(T_n - \hat{T})\hat{g}\|_2$. To derive the required bound, throughout the rest of the proof, we will assume that the event
$$
\| (T_n - \hat{T})g_n \|_2 \leq C(K/(\alpha n))^{1/2}
$$
holds in addition to (39)-(41). Since
$$
\| (T_n - \hat{T})g_n \|_2 = \| E_n[q(W_i)p(X_i)'\beta_n] - E[q(W)p(X)'\beta_n] \|,
$$
this event holds with probability at least $1 - \alpha$ by the same argument as that used in the proof of Lemma 5. Then by the triangle inequality,
$$
\| (T_n - \hat{T})g_n \|_2 \leq \| (T_n - \hat{T})(\hat{g} - g_n) \|_2 + \| (T_n - \hat{T})g_n \|_2
\leq \| T_n - \hat{T} \| \| \hat{g} - g_n \|_2 + C(K/(\alpha n))^{1/2}
\leq C((\xi_n^2 \log n/n)^{1/2}(\| \hat{g} - g \|_2 + K^{-s}) + (K/(\alpha n))^{1/2}).
$$
Since $\tau_n^2 \xi_n^2 \log n/n \leq c$ for sufficiently small $c > 0$, using the same argument as that to derive (15), we obtain
$$
\| \hat{g} - g \|_2 \leq C(\tau_n(K/(\alpha n))^{1/2} + K^{-s}).
$$
Substituting this inequality back to the bound on $\|(T_n - \hat{T})\hat{g}\|_2$, we obtain
$$
\| (T_n - \hat{T})\hat{g} \|_2 \leq C((\xi_n^2 \log n/n)^{1/2}(\tau_n(K/(\alpha n))^{1/2} + K^{-s}) + (K/(\alpha n))^{1/2})
\leq C((K/(\alpha n))^{1/2} + \tau_n^{-1}K^{-s}).
$$
The rest of the proof is the same as that given for (15)-(17). This completes the proof of the theorem.

**Q.E.D.**

**Lemma 5.** Under conditions of Theorem 2, $\| \hat{m} - m \|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1}J^{-s})$ with probability at least $1 - \alpha$ where $\hat{m}$ is defined in (38).

**Proof.** Using the triangle inequality and an elementary inequality $(a + b)^2 \leq 2a^2 + 2b^2$ for all $a, b \geq 0$,
$$
\| E_n[q(W_i)Y_i] - E[q(W)g(X)] \|^2 \leq 2\| E_n[q(W_i)\varepsilon_i] \|^2 + 2\| E_n[q(W_i)g(X_i)] - E[q(W)g(X)] \|^2.
$$
To bound the first term on the right-hand side of this inequality, we have
$$
E \left[ \| E_n[q(W_i)\varepsilon_i] \|^2 \right] = n^{-1}E[\| q(W)\varepsilon \|^2] \leq (C_B/n)E[\| q(W) \|^2] \leq CJ/n
$$
where the first and the second inequalities follow from Assumptions 4 and 2, respectively. Similarly,
$$
E \left[ \| E_n[q(W_i)g(X_i)] - E[q(W)g(X)] \|^2 \right] \leq n^{-1}E[\| q(W)g(X) \|^2]
\leq (C_B/n)E[\| q(W) \|^2] \leq CJ/n
$$

34
by Assumption 4. Therefore, denoting \( \bar{m}_n(w) := q(w)E[q(W)g(X)] \) for all \( w \in [0, 1] \), we obtain
\[
E[\|\hat{m} - \bar{m}_n\|^2] \leq CJ/n,
\]
and so by Markov’s inequality, \( \|\hat{m} - \bar{m}_n\|_2 \leq C(J/(\alpha n))^{1/2} \) with probability at least \( 1 - \alpha \).

Further, introducing \( \gamma_n \in \mathbb{R}^J \) such 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
\[
\bar{m}_n(w) = q(w)' \int_0^1 \int_0^1 q(t)g(x)f_{X,W}(x,t)dxdt
= q(w)' \int_0^1 q(t)m(t)dt = q(w)' \int_0^1 q(t)(q(t)\gamma_n + r_n(t))dt
= q(w)'\gamma_n + q(w)' \int_0^1 q(t)r_n(t)dt = m(w) - r_n(w) + q(w)' \int_0^1 q(t)r_n(t)dt.
\]

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 6. Applying the triangle inequality one more time, we obtain
\[
\|\hat{m} - m\|_2 \leq \|\hat{m} - \bar{m}_n\| + \|\bar{m}_n - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1}J^{-s})
\]
with probability at least \( 1 - \alpha \). This completes the proof of the lemma. Q.E.D.

D Proofs for Section 5

Proof of Theorem 3. In this proof, \( c \) and \( C \) are understood as sufficiently small and large constants, respectively, whose values may change at each appearance but can be chosen to depend only on \( c_W, C_W, c_h, C_H, c_F, C_F, c_c, C_c, \) and the kernel \( K \).

To prove the asserted claims, we apply Corollary 3.1, Case (E.3), from CCK conditional on \( W_n = \{W_1, \ldots, W_n\} \). Under \( H_0 \),
\[
T \leq \max_{(x,w,h) \in X_n \times W_n \times H_n} \sum_{i=1}^n k_{i,h}(w)(1\{X_i \leq x\} - F_{X|W}(x|W_i)) / (\sum_{i=1}^n k_{i,h}(w)^2)^{1/2} =: T_0
\]
with equality if the functions \( w \mapsto F_{X|W}(x|w) \) are constant for all \( x \in (0, 1) \). Using notation of CCK,
\[
T_0 = \max_{1 \leq j \leq p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij}
\]
where $p = |\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n|$, the number of elements in the set $\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n$, $x_{ij} = z_{ij}e_{ij}$ with $z_{ij}$ having the form $\sqrt{n}k_{i,h}(w)/(\sum_{i=1}^{n} k_{i,h}(w))^2)^{1/2}$, and $e_{ij}$ having the form $1\{X_i \leq x\} - F_{X|W}(x|W_i)$ for some $(x, w, h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n$. The dimension $p$ satisfies $\log p \leq C \log n$. Also, $n^{-1}\sum_{i=1}^{n} z_{ij}^2 = 1$. Further, since $0 \leq 1\{X_i \leq x\} \leq 1$, we have $|\varepsilon_{ij}| \leq 1$, and so $E[|\varepsilon_{ij}|/2]\{|W_n| \leq 2$. In addition, $E[\varepsilon_{ij}^2|W_n] \geq c_\epsilon (1 - C) > 0$ by Assumption 11. Thus, $T_0$ satisfies conditions of Case (E.3) in CCK with a sequence of constants $B_n$ as long as $|z_{ij}| \leq B_n$ for all $j = 1, \ldots, p$. In turn, Proposition B.2 in Chetverikov (2012) shows that under Assumptions 2, 8, and 9, with probability at least $1 - Cn^{-c}$, $z_{ij} \leq C/\sqrt{\hat{h}_{\min}} =: B_n$ uniformly over all $j = 1, \ldots, p$ (Proposition B.2 in Chetverikov (2012) is stated with “w.p.a.1” replacing “$1 - Cn^{-c}$”); however, inspecting the proof of Proposition B.2 (and supporting Lemma H.1) shows that the result applies with “$1 - Cn^{-c}$” instead of “w.p.a.1”). Let $B_{1,n}$ denote the event that $|z_{ij}| \leq C/\sqrt{\hat{h}_{\min}} = B_n$ for all $j = 1, \ldots, p$. As we just established, $P(B_{1,n}) \geq 1 - Cn^{-c}$. Since $(\log n)^7/(nh_{\min}) \leq C_{n^{-c}}$ by Assumption 9, we have that $B_n^2/(\log n)^7/n \leq C_{n^{-c}}$, and so condition (i) of Corollary 3.1 in CCK is satisfied on the event $B_{1,n}$.

Let $B_{2,n}$ denote the event that

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

By Assumption 10, $P(B_{2,n}) \geq 1 - C_{n^{-c}}$. We will apply Corollary 3.1 from CCK conditional on $W_n$ on the event $B_{1,n} \cap B_{2,n}$. For this, we need to show that on the event $B_{2,n}$, $\zeta_1, n \sqrt{\log n} + \zeta_2, n \leq C_{n^{-c}}$ where $\zeta_1, n$ and $\zeta_2, n$ are positive sequences such that

$$P (P_e(|T^b - T^b_0| > \zeta_1, n) > \zeta_2, n|W_n) < \zeta_2, n $$

(44)

where

$$T^b_0 := \max_{(x,w,h)\in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} \sum_{i=1}^{n} e_i (k_{i,h}(w)(1\{X_i \leq x\} - F_{X|W}(x|W_i))) \left(\sum_{i=1}^{n} k_{i,h}(w)^2\right)^{1/2}$$

and where $P_e(\cdot)$ denotes the probability distribution with respect to the distribution of $e_1, \ldots, e_n$ and keeping everything else fixed. To find such sequences $\zeta_1, n$ and $\zeta_2, n$, note that $\zeta_1, n \sqrt{\log n} + \zeta_2, n \leq C_{n^{-c}}$ follows from $\zeta_1, n + \zeta_2, n \leq C_{n^{-c}}$ (with different constants $c, C > 0$), so that it suffices to verify the latter condition. Also,

$$|T^b - T^b_0| \leq \max_{(x,w,h)\in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} \left| \sum_{i=1}^{n} e_i k_{i,h}(w)(\hat{F}_{X|W}(x|W_i) - F_{X|W}(x|W_i)) \left(\sum_{i=1}^{n} k_{i,h}(w)^2\right)^{1/2} \right|.$$

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

$$P_e \left( |T^b - T^b_0| > C \sqrt{\log n} \max_{(x,w)\in \mathcal{X}_n \times \mathcal{W}_n} \left| \hat{F}_{X|W}(x|w) - F_{X|W}(x|w) \right| \right) \leq C_{n^{-c}}.$$
Hence, on the event that
\[
\max_{(x,w) \in \mathcal{X}_n \times \mathcal{W}_n} \left| \hat{F}_{X|W}(x|w) - F_{X|W}(x|w) \right| \leq C_F n^{-c_F},
\]
whose conditional probability given \(\mathcal{W}_n\) on \(\mathcal{B}_{2,n}\) is at least \(1 - C_F n^{-c_F}\) by the definition of \(\mathcal{B}_{2,n}\),
\[
P_e \left( |T^b - T_0| > C n^{-c} \right) \leq C n^{-c}
\]
implicating that (44) holds for some \(\zeta_{1,n}\) and \(\zeta_{2,n}\) satisfying \(\zeta_{1,n} + \zeta_{2,n} \leq C n^{-c}\).

Thus, applying Corollary 3.1, Case (E.3), from CCK conditional on \(\{W_1, \ldots, W_n\}\) on the event \(\mathcal{B}_{1,n} \cap \mathcal{B}_{2,n}\) gives
\[
\alpha - C n^{-c} \leq P(T_0 > c(\alpha)|\mathcal{W}_n) \leq \alpha + C n^{-c}.
\]
Since \(P(\mathcal{B}_{1,n} \cap \mathcal{B}_{2,n}) \geq 1 - C n^{-c}\), integrating this inequality over the distribution of \(\mathcal{W}_n = \{W_1, \ldots, W_n\}\) gives (26). Combining this inequality with (43) gives (25). This completes the proof of the theorem. Q.E.D.

Proof of Theorem 4. Conditional on the data, the random variables
\[
T^b(x, w, h) := \frac{\sum_{i=1}^n c_i \left( k_{i,h}(w)(1\{X_i \leq x\} - \hat{F}_{X|W}(x|W_i)) \right)}{\left( \sum_{i=1}^n k_{i,h}(w)^2 \right)^{1/2}}
\]
for \((x, w, h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n\) are normal with zero mean and variances bounded from above by
\[
\frac{\sum_{i=1}^n \left( k_{i,h}(w)(1\{X_i \leq x\} - \hat{F}_{X|W}(x|W_i)) \right)^2}{\sum_{i=1}^n k_{i,h}(w)^2} \leq \max_{(x, w, h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} \max_{1 \leq i \leq n} \left( 1\{X_i \leq x\} - \hat{F}_{X|W}(x|W_i) \right)^2 \leq (1 + C_h)^2
\]
by Assumption 10. Therefore, \(c(\alpha) \leq C(\log n)^{1/2}\) for some constant \(C > 0\) since \(c(\alpha)\) is the \((1 - \alpha)\) conditional quantile of \(T^b\) given the data, \(T^b = \max_{(x, w, h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n} T^b(x, w, h)\), and \(p := |\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n|\), the number of elements of the set \(\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{H}_n\), satisfies \(\log p \leq C \log n\) (with a possibly different constant \(C > 0\)). Thus, the growth rate of the critical value \(c(\alpha)\) satisfies the same upper bound \((\log n)^{1/2}\) as if we were testing monotonicity of one particular regression function \(w \mapsto E[1\{X \leq x_0\}|W = w]\) with \(\mathcal{X}_n\) replaced by \(x_0\) for some \(x_0 \in (0, 1)\) in the definition of \(T\) and \(T^b\). Hence, the asserted claim follows from the same arguments as those given in the proof of Theorem 4.2 in Chetverikov (2012). This completes the proof of the theorem. Q.E.D.
E Technical tools

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

**Lemma 6** (Tikhonov). Let $(D, \rho_D)$ and $(R, \rho_R)$ be two pseudo-metric spaces and assume that $D$ is compact. Further, suppose there is a one-to-one continuous operator $A : D \to R$. Then the inverse operator $A^{-1}$ exists and is continuous over the range $A(D)$ of $A$.

**Remark 11.** This Tikhonov’s lemma is essentially well known but it is typically presented for metric spaces whereas we require pseudo-metric spaces. Following Dudley (2002), we define a pseudo-metric $\rho_D$ on the space $D$ as a function $\rho_D : D \times D \to \mathbb{R}$ that satisfies for any $d_1, d_2, d_3 \in D$, (i) $\rho_D(d_1, d_2) \geq 0$, (ii) $\rho_D(d_1, d_1) = 0$, (iii) $\rho_D(d_1, d_2) = \rho_D(d_2, d_1)$, and (iv) $\rho_D(d_1, d_3) \leq \rho_D(d_1, d_2) + \rho_D(d_2, d_3)$. Importantly for our application of the above lemma, the pseudo-metric $\rho_D$ allows for the case that $\rho_D(d_1, d_2) = 0$, but $d_1 \neq d_2$ and thus $A(d_1) \neq A(d_2)$.

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

**Lemma 7.** 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 some function. If $M$ is constant, then $\text{cov}(W, M(W)) = 0$. 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 $\text{cov}(W, M(W)) > 0$.

**Remark 12.** We slightly changed the assertion of the inequality from the conventional one since it is more convenient for our purposes.

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

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

which holds for increasing $M$ since $(M(W) - M(W'))(W - W') \geq 0$ almost surely and $(M(W) - M(W'))(W - W') > 0$ with strictly positive probability. This completes the proof of the lemma. Q.E.D.
Lemma 8. 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
\[
h = \sum_{j=1}^{\infty} \beta_j h_j \quad \text{in} \quad L^2[0, 1], \quad \beta_j := \frac{\int_{t_{m-1}}^{t_m} h_j(x) dx}{(t_m - t_{m-1})^{1/2}}, \quad \text{and} \quad \sum_{j=1}^{\infty} \beta_j^2 = 1.
\]
Therefore, by the Cauchy-Schwartz inequality,
\[
1 = \sum_{j=1}^{\infty} \beta_j^2 \geq \frac{1}{t_m - t_{m-1}} \sum_{j=1}^{\infty} \left( \int_{t_{m-1}}^{t_m} h_j(x) dx \right)^2 \leq \sum_{j=1}^{\infty} \int_{t_{m-1}}^{t_m} (h_j(x))^2 dx.
\]
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 9. Let \( (X, W) \) be a pair of random variables defined as in Example 1. Then
Assumptions 1 and 2 of Section 2 are satisfied if \( 0 < x_1 < x_2 < 1 \) and \( 0 < w_1 < w_2 < 1 \).

Proof. As noted in Example 1, we have
\[
X = \Phi(\rho \Phi^{-1}(W) + (1 - \rho^2)^{1/2} U)
\]
where \( \Phi(x) \) is the distribution function of a \( N(0, 1) \) random variable and \( U \) is a \( N(0, 1) \) random variable that is independent of \( W \). Therefore, the conditional distribution function of \( X \) given \( W \) is
\[
F_{X|W}(x|w) := \Phi \left( \frac{\Phi^{-1}(x) - \rho \Phi^{-1}(w)}{\sqrt{1 - \rho^2}} \right).
\]
Since the function \( w \mapsto F_{X|W}(x|w) \) is decreasing for all \( x \in (0, 1) \), condition (3) of Assumption 1 follows. Further, to prove condition (4) of Assumption 1, it suffices to show that
\[
\frac{\partial \log F_{X|W}(x|w)}{\partial w} \leq c_F \quad \text{(45)}
\]
for some constant $c_F < 0$, all $x \in (0, x_2)$, and all $w \in (w_1, w_2)$ because, for every $x \in (0, x_2)$ and $w \in (w_1, w_2)$, there exists $\tilde{w} \in (w_1, w_2)$ such that

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

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

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

for all $x \in (0, x_2)$. To show (45), observe that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} = -\frac{\rho}{\sqrt{1-\rho^2}} \frac{\phi(y)}{\Phi(\Phi^{-1}(w))} = -\frac{\sqrt{2\pi}\rho}{\sqrt{1-\rho^2}} \phi(y)$$ (46)

where $y := (\Phi^{-1}(x) - \rho\Phi^{-1}(w))/(1-\rho^2)^{1/2}$. Thus, (45) holds for some $c_F < 0$ and all $x \in (0, x_2)$ and $w \in (w_1, w_2)$ such that $\Phi^{-1}(x) \geq \rho\Phi^{-1}(w)$ since $x_2 < 1$ and $0 < w_1 < w_2 < 1$. On the other hand, when $\Phi^{-1}(x) < \rho\Phi^{-1}(w)$, so that $y < 0$, it follows from Proposition 2.5 in Dudley (2014) that $\phi(y)/\Phi(y) \geq (2/\pi)^{1/2}$, and so (46) implies that

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

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

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

$$f_{X|W}(x|w) := \frac{1}{\sqrt{1-\rho^2}} \phi \left( \frac{\Phi^{-1}(x) - \rho\Phi^{-1}(w)}{\sqrt{1-\rho^2}} \right) \frac{1}{\phi(\Phi^{-1}(x))}.$$ (47)

Since $0 < x_1 < x_2 < 1$ and $0 < w_1 < w_2 < 1$, condition (ii) of Assumption 2 is satisfied as well. Finally, to prove condition (i) of Assumption 2, note that since $f_W(w) = 1$ for all $w \in [0, 1]$, (47) combined with the change of variables formula with $x = \Phi(\tilde{x})$ and $w = \Phi(\tilde{w})$ give

$$(1-\rho^2) \int_0^1 \int_0^1 f_{X,W}(x, w) dxdw = (1-\rho^2) \int_0^1 \int_0^1 f_{X|W}(x|w) dxdw$$

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \phi^2 \left( \frac{\tilde{x} - \rho\tilde{w}}{\sqrt{1-\rho^2}} \right) \frac{\phi(\tilde{w})}{\phi(\tilde{x})} d\tilde{x} d\tilde{w}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp \left[ \left( \frac{1}{2} - \frac{1}{1-\rho^2} \right) \tilde{x}^2 + \frac{2\rho}{1-\rho^2} \tilde{x} \tilde{w} - \left( \frac{\rho^2}{1-\rho^2} + \frac{1}{2} \right) \tilde{w}^2 \right] d\tilde{x} d\tilde{w}$$

$$= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp \left[ -\frac{1+\rho^2}{2(1-\rho^2)} \left( -\tilde{x}^2 + \frac{4\rho}{1+\rho^2} \tilde{x} \tilde{w} - \tilde{w}^2 \right) \right] d\tilde{x} d\tilde{w}.$$
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 10.** Let $X = U_1 + U_2W$ where $U_1, U_2, W$ are mutually independent, $U_1, U_2 \sim U[0, 1/2]$ and $W \sim U[0, 1]$. Then Assumptions 1 and 2 of Section 2 are satisfied if $0 < w_1 < w_2 < 1, 0 < x_1 < x_2 < 1$, and $w_1 > w_2 - \sqrt{w_2}/2$.

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

$$f_{X|W}(x|w) = \int_0^{1/2} f_{U_1}(x - u_2 w)f_{U_2}(u_2)du_2$$

$$= 4 \int_0^{1/2} \left\{ 0 \leq x - u_2 w \leq \frac{1}{2} \right\} du_2$$

$$= 4 \int_0^{1/2} \left\{ \frac{x}{w} - \frac{1}{2w} \leq u_2 \leq \frac{x}{w} \right\} du_2$$

$$= \left\{ \begin{array}{ll}
\frac{4x}{w}, & 0 \leq x < \frac{w}{2} \\
2, & \frac{w}{2} \leq x < \frac{1}{2} \\
\frac{2(1+w)}{w}, & \frac{1}{2} \leq x < \frac{1+w}{2} \\
0, & \frac{1+w}{2} \leq x \leq 1
\end{array} \right.$$  

and, thus,

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

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 (3) of Assumption 1 is satisfied. To check conditions (4) and (5), we proceed as in Lemma 9 and show $\partial \log F_{X|W}(x|w)/\partial w < 0$ uniformly for all $x \in [x_k, \bar{x}_1]$ and $w \in (\bar{w}_1, \bar{w}_2)$. First, notice that, as required by Assumption 2(iv), $[x_k, \bar{x}_k] = [0, (1 + \bar{w}_k)/2], k = 1, 2$. For $0 \leq x < w/2$ and $w \in (\bar{w}_1, \bar{w}_2)$,

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

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

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

Therefore, (4) holds uniformly over $x \in (x_k, 1/2)$ and (5) uniformly over $x \in (x_1, 1/2)$. Now, consider $1/2 \leq x < (1 + \bar{w}_1)/2$ and $w \in (\bar{w}_1, \bar{w}_2)$. Notice that, on this interval,
\[ \partial(F_{XW}(x|\bar{w}_1)/F_{XW}(x|\bar{w}_2))/\partial x \leq 0 \] so that

\[
\frac{F_{XW}(x|\bar{w}_1)}{F_{XW}(x|\bar{w}_2)} = \frac{1 - \frac{1}{w_1} (x - \frac{1 + \bar{w}_1}{2})^2}{1 - \frac{1}{w_2} (x - \frac{1 + \bar{w}_2}{2})^2} \geq \frac{1}{1 - \frac{1}{w_2} (\frac{1 + \bar{w}_1}{2} - \frac{1 + \bar{w}_2}{2})^2} = \frac{\bar{w}_2}{\bar{w}_2 - 2(\bar{w}_1 - \bar{w}_2)^2} > 1,
\]

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

\[
\frac{1 - F_{XW}(x|\bar{w}_2)}{1 - F_{XW}(x|\bar{w}_1)} = \frac{\frac{2}{w_2} (x - \frac{1 + \bar{w}_2}{2})^2}{\frac{2}{w_1} (x - \frac{1 + \bar{w}_1}{2})^2} = \frac{\bar{w}_2}{\bar{w}_1} > 1
\]

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

**Lemma 11.** 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:

\[
\bar{h}_a(x) := h(x)1\{|h(x)| \leq a\} + a1\{h(x) > a\} - a1\{h(x) < -a\}
\]

for all \( x \in [0, 1] \). Then \( \|\bar{h}_a - h\|_2 \to 0 \) 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 \( x \in \mathbb{R}\setminus[0, 1] \) 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 distribution function of a \( N(0, 1) \) random variable. Theorem 6.3.14 in Stroock (1999) shows that

\[
\|h_\sigma - h\|_2 = \left( \int_0^1 (h_\sigma(x) - h(x))^2 dx \right)^{1/2} \leq \left( \int_{-\infty}^\infty (h_\sigma(x) - h(x))^2 dx \right)^{1/2} \to 0
\]

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 $x \in [0, 1]$,

$$Dh_\sigma(x) = -\frac{1}{\sigma^2} \int_0^1 h(y) D\phi \left( \frac{y - x}{\sigma} \right) dy$$

$$= -\frac{1}{\sigma} \left( h(1) \phi \left( \frac{1 - x}{\sigma} \right) - h(0) \phi \left( -\frac{x}{\sigma} \right) - \int_0^1 \phi \left( \frac{y - x}{\sigma} \right) dh(y) \right)$$

$$\geq -\frac{1}{\sigma} \phi \left( \frac{1 - x}{\sigma} \right)$$

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

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

defined for all $x \in [0, 1]$ and some $\bar{x} \in (0, 1)$ is increasing and continuously differentiable for all $x \in (0, 1) \setminus \bar{x}$, 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

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

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

**Lemma 12.** Let $(p_1', q_1')', \ldots, (p_n', q_n')'$ be a sequence of i.i.d. random vectors where $p_i$’s are vectors in $\mathbb{R}^K$ and $q_i$’s are vectors in $\mathbb{R}^J$. Assume that $\|p_1\| \leq \xi_n$, $\|q_1\| \leq \xi_n$, $\|E[p_1p_1']\| \leq C_p$, and $\|E[q_1q_1']\| \leq C_q$ where $\xi_n \geq 1$. Then for all $t \geq 0$,

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

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

**Remark 13.** A closely related result can be found in Belloni, Chernozhukov, Chetverikov, and Kato (2014) who were the first to use it in econometrics literature. The current version of the result is more useful for our purposes.

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

$$a'E[p_1q_1']b = E[(a'p_1)(b'q_1)]$$

$$\leq (E[(a'p_1)^2]E[(b'q_1)^2])^{1/2} \leq \|a\| \|b\| (C_p C_q)^{1/2}$$
by Hölder’s inequality. Therefore, \( \|E[p_1q'_1]\| \leq (C_pC_q)^{1/2} \). Further, denote \( S_i := p_iq'_i - E[p_iq'_i] \) for \( i = 1, \ldots, n \). By the triangle inequality and calculations above,
\[ \|S_1\| \leq \|p_1q'_1\| + \|E[p_1q'_1]\| \leq \xi_n^2 + (C_pC_q)^{1/2} \leq \xi_n^2(1 + (C_pC_q)^{1/2}) =: R. \]

Now, denote \( Z_n := \sum_{i=1}^{n} S_i \). Then
\[ E[Z_n Z'_n] \leq n\|E[S_1S'_1]\| \leq n\|E[p_1q'_1q_1p'_1]\| + n\|E[p_1q'_1E[q_1p'_1]]\| \leq n\|E[p_1q'_1q_1p'_1]\| + nC_pC_q. \]

For any \( a \in \mathbb{R}^K \),
\[ a'E[p_1q'_1q_1p'_1]a \leq \xi_n^2E[(a'p_1)^2] \leq \xi_n^2\|a\|^2C_p. \]
Therefore, \( \|E[p_1q'_1q_1p'_1]\| \leq \xi_n^2C_p \), and so
\[ \|E[Z_n Z'_n]\| \leq nC_p(\xi_n^2 + C_q) \leq n\xi_n^2(1 + C_p)(1 + C_q). \]

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

Hence, by Corollary 6.2.1 in Tropp (2012),
\begin{align*}
P \left( \|n^{-1}Z_n\| \geq t \right) & \leq (K + J)\exp\left( -\frac{n^2t^2/2}{\sigma^2 + Rt/3} \right) \\
& \leq \exp\left( \log(K + J) - \frac{Ant^2}{\xi_n^2(1 + t)} \right).
\end{align*}

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

References

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

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

Bejenaru, I., and T. Tao (2006): “Sharp well-posedness and ill-posedness results for a quadratic non-linear Schrödinger equation,” Journal of Functional Analysis, 233(1), 228–259.
Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2014): “Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results,” Discussion paper.

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

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

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

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

Carrasco, M., and J.-P. Florens (2011): “A Spectral Method for Deconvolving a Density,” *Econometric Theory*, 7(Special Issue 03), 546–581.

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

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

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

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

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

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

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

45
Chetverikov, D. (2012): “Testing Regression Monotonicity in Econometric Models,” Discussion paper.

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

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

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

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

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

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

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

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

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

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

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

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

Grasmair, M., O. Scherzer, and A. Vanhems (2013): “Nonparametric instrumental regression with non-convex constraints,” *Inverse Problems*, 29(3), 1–16.
Hadamard, J. (1923): *Lectures on Cauchy’s Problem in Linear Partial Differential Equations*. Yale University Press, New Haven.

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

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

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

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

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

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

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

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

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

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

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

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

Mammen, E. (1991): “Estimating a Smooth Monotone Regression Function,” *The Annals of Statistics*, 19(2), 724–740.
Mammen, E., J. S. Marron, B. A. Turlach, and M. P. Wand (2001): “A General Projection Framework for Constrained Smoothing,” *Statistical Science*, 16(3), 232–248.

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

Manski, C. F. (1997): “Monotone Treatment Response,” *Econometrica*, 65(6), 1311–1334.

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

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

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

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

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

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

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

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

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

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*. 48
Van de Geer, S. (2000): *Empirical Processes in M-Estimation*. Cambridge University Press, 1st edn.

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

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

Zhang, C.-H. (2002): “Risk Bounds in Isotonic Regression,” *Annals of Statistics*, 30(2), 528–555.
Figure 1: Plots of $F_{X|W}(x|\tilde{w}_1)$ and $F_{X|W}(x|\tilde{w}_2)$ in Examples 1 and 2, respectively.
Table 1: Model 1: Performance of the unrestricted and restricted estimators for $N = 200$, $\rho = 0.7$, $\eta = 0.3$, $\sigma_\varepsilon = 0.7$. 
| $k_X$ | $k_W$ | $\kappa = 1$ | \(\kappa = 0.5\) | \(\kappa = 0.1\) |
|-------|-------|-------------|-------------|-------------|
|       |       | unrestr.    | restr.      | unrestr.    | restr.      | unrestr.    | restr.      |
| 2     | 2     | 0.001 0.004 | 0.016 0.001 | 0.003 0.001 |
| 2     | 3     | 0.001 0.004 | 0.000 0.002 | 0.000 0.001 |
| 2     | 5     | 0.001 0.004 | 0.001 0.002 | 0.000 0.002 |
| 2     | 7     | 0.001 0.004 | 0.001 0.002 | 0.000 0.002 |
| 3     | 5     | 0.000 0.002 | 0.000 0.001 | 0.001 0.002 |

Table 2: Model 2: Performance of the unrestricted and restricted estimators for $N = 200$, $\rho = 0.7$, $\eta = 0.3$, $\sigma_\varepsilon = 0.7$. 
| ρ | η |   |                      |                     |                      |                     |
|---|---|---|----------------------|---------------------|----------------------|---------------------|
|   |   |   | κ = 1                | κ = 0.5             | κ = 0.1              |
|   |   | unrestr. | restr. | unrestr. | restr. | unrestr. | restr. |
| 0.3| 0.3| 0.003 | 0.021 | 0.001 | 0.005 | 0.000 | 0.000 |
|   |   | 0.486 | 0.109 | 0.124 | 0.026 | 0.004 | 0.001 |
|   |   | 0.439 | 0.116 | 0.113 | 0.028 | 0.004 | 0.001 |
|   |   | 0.264 | 0.247 | 0.264 |       |       |       |
| 0.3| 0.7| 0.015 | 0.033 | 0.004 | 0.008 | 0.000 | 0.000 |
|   |   | 0.427 | 0.105 | 0.108 | 0.025 | 0.004 | 0.001 |
|   |   | 0.397 | 0.122 | 0.101 | 0.029 | 0.004 | 0.001 |
|   |   | 0.306 | 0.290 | 0.306 |       |       |       |
| 0.7| 0.3| 0.000 | 0.007 | 0.000 | 0.002 | 0.000 | 0.000 |
|   |   | 0.226 | 0.036 | 0.049 | 0.009 | 0.002 | 0.000 |
|   |   | 0.210 | 0.039 | 0.046 | 0.010 | 0.002 | 0.000 |
|   |   | 0.186 | 0.223 | 0.202 |       |       |       |
| 0.7| 0.7| 0.000 | 0.008 | 0.000 | 0.002 | 0.000 | 0.000 |
|   |   | 0.225 | 0.034 | 0.047 | 0.008 | 0.002 | 0.000 |
|   |   | 0.208 | 0.038 | 0.044 | 0.009 | 0.002 | 0.000 |
|   |   | 0.182 | 0.214 | 0.199 |       |       |       |

Table 3: Model 1: Performance of the unrestricted and restricted estimators for $\sigma_\varepsilon = 0.7$, $k_X = 2$, $k_W = 5$, $N = 200$. 

53
Table 4: Model 2: Performance of the unrestricted and restricted estimators for $\sigma_\varepsilon = 0.7$, $k_X = 2$, $k_W = 5$, $N = 200$. 

| $\rho$ | $\eta$ | $\kappa = 1$ | $\kappa = 0.5$ | $\kappa = 0.1$ |
|-------|--------|-------------|-------------|-------------|
|       |        | unrestricted | restricted  | unrestricted | restricted  |
| 0.3   | 0.3    | 0.004 0.013  | 0.004 0.013  | 0.004 0.014  | 0.004 0.014  |
|       |        | 0.488 0.055  | 0.496 0.040  | 0.449 0.028  | 0.449 0.028  |
|       |        | 0.443 0.061  | 0.451 0.048  | 0.409 0.037  | 0.409 0.037  |
|       |        | 0.137 0.106  | 0.106 0.091  |              |              |
| 0.3   | 0.7    | 0.016 0.029  | 0.016 0.029  | 0.015 0.029  | 0.015 0.029  |
|       |        | 0.430 0.056  | 0.433 0.042  | 0.395 0.032  | 0.395 0.032  |
|       |        | 0.401 0.077  | 0.404 0.063  | 0.370 0.054  | 0.370 0.054  |
|       |        | 0.191 0.157  | 0.157 0.145  |              |              |
| 0.7   | 0.3    | 0.001 0.004  | 0.001 0.002  | 0.000 0.002  | 0.000 0.002  |
|       |        | 0.227 0.017  | 0.197 0.012  | 0.217 0.007  | 0.217 0.007  |
|       |        | 0.211 0.018  | 0.183 0.012  | 0.200 0.008  | 0.200 0.008  |
|       |        | 0.087 0.067  | 0.067 0.040  |              |              |
| 0.7   | 0.7    | 0.001 0.004  | 0.001 0.002  | 0.000 0.002  | 0.000 0.002  |
|       |        | 0.225 0.016  | 0.190 0.011  | 0.207 0.007  | 0.207 0.007  |
|       |        | 0.209 0.018  | 0.176 0.012  | 0.191 0.008  | 0.191 0.008  |
|       |        | 0.086 0.069  | 0.069 0.043  |              |              |
Figure 2: Model 1: unrestricted and restricted estimates of $g(x)$ for $N = 200$, $\rho = 0.3$, $\eta = 0.3$, $\sigma_x = 0.3$, $k_X = 2$, $k_W = 5$.

Figure 3: Model 2: unrestricted and restricted estimates of $g(x)$ for $N = 200$, $\rho = 0.3$, $\eta = 0.3$, $\sigma_x = 0.3$, $k_X = 2$, $k_W = 5$. 