Chetverikov, Denis; Wilhelm, Daniel

Working Paper

Nonparametric instrumental variable estimation under monotonicity

cemmap working paper, No. CWP48/16

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Chetverikov, Denis; Wilhelm, Daniel (2016) : Nonparametric instrumental variable estimation under monotonicity, cemmap working paper, No. CWP48/16, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2016.4816

This Version is available at:
https://hdl.handle.net/10419/149794

cemmap
centre for microdata methods and practice

Nonparametric instrumental variable estimation under monotonicity

Denis Chetverikov
Daniel Wilhelm

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP48/16

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

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 (NPIV) model leads to estimators that may suffer from poor statistical performance. In this paper, we explore the possibility of imposing shape restrictions to improve the performance of the NPIV estimators. We assume that the regression function is monotone and consider sieve estimators that enforce the monotonicity constraint. We define a restricted measure of ill-posedness that is relevant for the constrained estimators and show that under the monotone IV assumption and certain other conditions, our measure of ill-posedness is bounded uniformly over the dimension of the sieve space, in stark contrast with a well-known result that the unrestricted sieve measure of ill-posedness that is relevant for the unconstrained estimators grows to infinity with the dimension of the sieve space. Based on this result, we derive a novel non-asymptotic error bound for the constrained estimators. The bound gives a set of data-generating processes where the monotonicity constraint has a particularly strong regularization effect and considerably improves the performance of the estimators. The bound shows that the regularization effect can be strong even in large samples and for steep regression functions if the NPIV model is severely ill-posed – a finding that is confirmed by our simulation study. We apply the constrained estimator to the problem of estimating gasoline demand from U.S. data.

*First version: January 2014. This version: September 19, 2016. We thank Alex Belloni, Richard Blundell, Stéphane Bonhomme, Moshe Buchinsky, Matias Cattaneo, Xiaohong Chen, Victor Chernozhukov, Andrew Chesher, Joachim Freyberger, Jerry Hausman, Jinyong Hahn, Joel Horowitz, Dennis Kristensen, Simon Lee, Zhipeng Liao, Rosa Matzkin, Eric Mbakop, Matthew Kahn, Ulrich Müller, Whitney Newey, Markus Reiß, Andres Santos, Susanne Schennach, Azeem Shaikh, Vladimir Spokoiny, and three referees 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

Nonparametric instrumental variable (NPIV) methods have received a lot of attention in the recent econometric theory literature, but they are still far from the popularity that linear IV and nonparametric conditional mean estimation methods enjoy in empirical work. One of the main reasons for this originates from the fact that the NPIV model is ill-posed, which may cause nonparametric estimators in this model to suffer from poor statistical performance.

In this paper, we explore the possibility of imposing shape constraints to improve the performance of the NPIV estimator. We study the NPIV model

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

where Y is a dependent variable, X an endogenous regressor, and W an instrumental variable (IV). We are interested in the estimation of the nonparametric regression function g based on a random sample of size n from the distribution of the triple (Y, X, W). To simplify the presentation we assume that X is a scalar, although the results can be easily extended to the case where X is a vector containing one endogenous and several exogenous regressors (in this case we assume that W is a vector consisting of all exogenous regressors and an instrument for the endogenous regressor). Departing from the existing literature on the estimation of the NPIV model, we assume that the function g is monotone increasing[1] and consider a constrained estimator $\widehat{g}^c$ of g that is similar to the unconstrained sieve estimators of Blundell, Chen, and Kristensen (2007) and Horowitz (2012) but that enforces the monotonicity constraint. In addition to the monotonicity of g, we also assume a monotone reduced form relationship between X and W 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 (the monotone IV assumption).

We start our analysis from the observation that as long as the function g is strictly increasing, as the sample size n gets large, any appropriate unconstrained estimator of g will be increasing with probability approaching one, in which case the corresponding constrained estimator will be numerically identical to the unconstrained one. Thus, the constrained estimator must have the same, potentially very slow, rate of convergence as that of the unconstrained estimator. In simulations, however, we find that the constrained estimators often outperform, sometimes substantially, the unconstrained ones even if the sample size n is rather large; see Figure 1 for an example. Hence, it follows that the rate result above misses an important finite-sample phenomenon.

[1] All results in the paper hold also when g is decreasing. In fact, as we show in Section 4 the sign of the slope of g is identified under our monotonicity conditions.

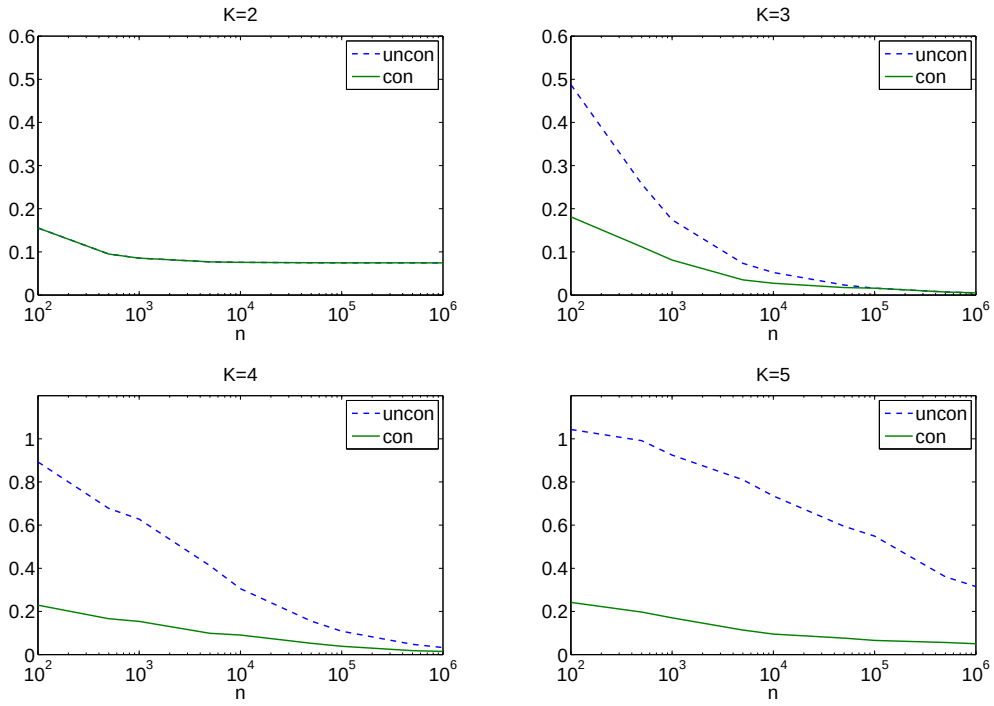

Figure 1: an example demonstrating performance gains from imposing the monotonicity constraint. In this example, $g(x) = x^2 + 0.2x$, $W = \Phi(\zeta)$, $X = \Phi(\rho\zeta + \sqrt{1-\rho^2}\epsilon)$, $\varepsilon = \sigma(\eta\epsilon + \sqrt{1-\eta^2}\nu)$, where (ζ, ϵ, ν) is a triple of independent $N(0,1)$ random variables, $\rho = 0.3$, $\eta = 0.3$, $\sigma = 0.5$, and $\Phi(\cdot)$ is the cdf of the $N(0,1)$ distribution. Four panels of the figure show the square root of the MISE of the constrained (con) and the unconstrained (uncon) sieve estimators defined in Section 3 as a function of the sample size n depending on the dimension of the sieve space K. We use the sieve estimators based on the quadratic regression splines, so that the sieve space is spanned by $(1, x)$ if $K = 2$, by $(1, x, x^2)$ if $K = 3$, by $(1, x, x^2, (x-1/2)_+^2)$ if $K = 4$, and by $(1, x, x^2, (x-1/3)_+^2, (x-2/3)_+^2)$ if $K = 5$. The figure shows that the constrained estimator substantially outperforms the unconstrained one as long as $K \geq 3$ even in large samples.

In this paper, we derive a novel non-asymptotic error bound for the constrained estimators that captures this finite-sample phenomenon. For each sample size n, the bound gives a set of data-generating processes where the monotonicity constraint has a particularly strong regularization effect thus considerably improving the performance of the estimators. The bound shows that the regularization effect can be strong even in large samples and for steep functions g if the NPIV model is severely ill-posed.

To establish our non-asymptotic error bound, we define a restricted sieve measure of ill-posedness that is relevant for the constrained estimators. We demonstrate that as long as the monotone IV assumption is satisfied, under certain conditions, this measure is bounded uniformly over the dimension of the sieve space. This should be contrasted with a well-known result that the unrestricted sieve measure of ill-posedness that is relevant for

the unconstrained estimators grows to infinity, potentially very fast, with the dimension of the sieve space; see Blundell, Chen, and Kristensen (2007). Thus, if the dimension of the sieve space is large enough, the ratio of the restricted and unrestricted sieve measures of ill-posedness can be arbitrarily small, which explains a substantial regularization effect of the monotonicity constraint.

More specifically, our non-asymptotic error bound for the constrained estimator $\widehat{g}^c$ of g has the following structure: for each sample size n, uniformly over a certain large class of data-generating processes,

$$\|\widehat{g}^c - g\|_{2,t} \leq C\Big(\min\{\|Dg\|_\infty + V_n,\ \tau_n V_n\} + B_n \Big), \tag{2}$$

holds with large probability, where C is a constant independent of n and g, $\|Dg\|_\infty$ the maximum slope of g, and $\|\cdot\|_{2,t}$ a norm defined below. Further, B_n on the right-hand side of this bound is a bias term that behaves similarly to that of the unconstrained NPIV estimators, and $\min\{\|Dg\|_\infty + V_n, \tau_n V_n\}$ is the variance term, where V_n is of the same order as the variance of a nonparametric conditional mean estimator up to a log-term, i.e. of a well-posed problem, and τ_n is the unrestricted sieve measure of ill-posedness. Without the monotonicity constraint, the variance term would be $\tau_n V_n$, up to a log-term, but because of the monotonicity constraint, we can replace $\tau_n V_n$ by $\min\{\|Dg\|_\infty + V_n, \tau_n V_n\}$.

The main implications of the bound (2) are the following. First, note that the right-hand side of the bound becomes smaller as the maximum slope of g decreases. Second, because of ill-posedness, τ_n may be large, in which case

$$\min\{\|Dg\|_\infty + V_n, \tau_n V_n\} = \|Dg\|_\infty + V_n \ll \tau_n V_n, \tag{3}$$

and it is the scenario when the monotonicity constraint has a strong regularization effect. If the NPIV model is severely ill-posed, τ_n may be particularly large, in which case (3) holds even if the maximum slope $\|Dg\|_\infty$ is relatively far away from zero, i.e. the function g is steep.

As the sample size n gets large, the bound eventually switches to the regime when $\tau_n V_n$ becomes small relative to $\|Dg\|_\infty$, and the regularization effect of the monotonicity constraint disappears. Asymptotically, the ill-posedness of the model, therefore, undermines the statistical properties of the constrained estimator $\widehat{g}^c$ just as it does for the unconstrained estimator, and may lead to slow, logarithmic convergence rates. However, when ill-posedness is severe, the switch to this regime may occur only at extremely large sample sizes.

Finally, we use the error bound to derive the constrained estimator's convergence rate under a sequence of data-generating processes indexed by the sample size n in which the maximum slope $\|Dg\|_\infty$ drifts to zero with the sample size (we refer to this sequence of

data-generating processes as a local-to-flat asymptotics). In this case, the error bound is always in the regime with $\|Dg\|_\infty + V_n \leq \tau_n V_n$, and the ill-posedness of the NPIV model, expressed by τ_n, does not affect the convergence rate. In fact, we show that as long as $\|Dg\|_\infty$ drifts to zero fast enough, the convergence rate of the constrained estimator is equal to that of nonparametric conditional mean estimators up to a log-term. This asymptotic theory may be viewed as an approximation to the finite-sample situation in which the regression function g is not too steep relative to the sample size, so that the monotonicity constraint is binding with a non-trivial probability, and therefore offers an alternative explanation for the good performance of the constrained NPIV estimator in finite samples.

Our simulation experiments confirm the theoretical findings and demonstrate possibly large finite-sample performance improvements of the constrained estimators relative to the unconstrained ones when the monotone IV assumption is satisfied. The estimates show that imposing the monotonicity constraint on g removes the estimator's non-monotone oscillations due to sampling noise, which in the NPIV model can be particularly pronounced because of its ill-posedness. Therefore, imposing the monotonicity constraint significantly reduces variance while only slightly increasing bias.

We regard both monotonicity conditions as natural in many economic applications. In fact, both of these conditions often directly follow from economic theory. Consider the following generic example. Suppose an agent chooses input X (e.g. schooling) to produce an outcome Y (e.g. life-time earnings) such that $Y = g(X) + \varepsilon$, where ε summarizes determinants of outcome other than X. The cost of choosing a level $X = x$ is $C(x, W, \eta)$, where W is a cost-shifter (e.g. distance to college) and η represents (possibly vector-valued) unobserved heterogeneity in costs (e.g. family background, a family's taste for education, variation in local infrastructure). The agent's optimization problem can then be written as

$$X = \arg\max_x \{g(x) + \varepsilon - c(x, W, \eta)\}$$

so that, from the first-order condition of this optimization problem,

$$\frac{\partial X}{\partial W} = \frac{\frac{\partial^2 c}{\partial X \partial W}}{\frac{\partial^2 g}{\partial X^2} - \frac{\partial^2 c}{\partial X^2}} \geq 0 \tag{4}$$

if marginal cost are decreasing in W (i.e. $\partial^2 c / \partial X \partial W \leq 0$), marginal cost are increasing in X (i.e. $\partial^2 c / \partial X^2 > 0$), and the production function is concave (i.e. $\partial^2 g / \partial X^2 \leq 0$). As long as W is independent of the pair (ε, η), condition (4) implies our monotone IV assumption and g increasing corresponds to the assumption of a monotone regression function. Dependence between η and ε generates endogeneity of X, and independence of W from (ε, η) implies that W can be used as an instrument for X.

Another example is the estimation of Engel curves. In this case, the outcome variable Y is the budget share of a good, the endogenous variable X is total expenditure, and the instrument W is gross income. Our monotonicity conditions are plausible in this example because for normal goods such as food-in, the budget share is decreasing in total expenditure, and total expenditure increases with gross income. Finally, consider the estimation of (Marshallian) demand curves. The outcome variable Y is quantity of a consumed good, the endogenous variable X is the price of the good, and W could be some variable that shifts production cost of the good. For a normal good, the Slutsky inequality predicts Y to be decreasing in price X as long as income effects are not too large. Furthermore, price is increasing in production cost and, thus, increasing in the instrument W, and so our monotonicity conditions are plausible in this example as well.

Both of our monotonicity assumptions are testable. In the appendix, we provide a new *adaptive* test of the monotone IV condition, with the value of the involved smoothness parameter chosen in a data-driven fashion.

Matzkin (1994) advocates the use of shape restrictions in econometrics and argues that economic theory often provides restrictions on functions of interest, such as monotonicity, concavity, and/or Slutsky symmetry. In the context of the NPIV model (1), Freyberger and Horowitz (2013) show that, in the absence of point-identification, shape restrictions may yield informative bounds on functionals of g and develop inference procedures when the regressor X and the instrument W are discrete. Blundell, Horowitz, and Parey (2013) demonstrate via simulations that imposing Slutsky inequalities in a quantile NPIV model for gasoline demand improves finite-sample properties of the NPIV estimator. Grasmair, Scherzer, and Vanhems (2013) study the problem of demand estimation imposing various constraints implied by economic theory, such as Slutsky inequalities, and derive the convergence rate of a constrained NPIV estimator under an abstract projected source condition. Our results are different from theirs because we focus on non-asymptotic error bounds, we derive our results under easily interpretable, low-level conditions, and we show that the regularization effect of the monotonicity constraint can be strong even in large samples and for steep functions g if the NPIV model is severely ill-posed.

Other related literature. The NPIV model has received substantial attention in the recent econometrics literature. Newey and Powell (2003), Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007), and Darolles, Fan, Florens, and Renault (2011) study identification of the NPIV model (1) and propose estimators of the regression function g. See Horowitz (2011, 2014) for recent surveys and further references. In the mildly ill-posed case, Hall and Horowitz (2005) derive the minimax risk lower bound in L^2-norm and show that their estimator achieves this lower bound. Under different conditions,

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 both for the mildly and the severely ill-posed case. The optimal convergence rates in the severely ill-posed case were shown to be logarithmic, which means that the slow convergence rate of existing estimators is not a deficiency of those estimators but rather an intrinsic feature of the statistical inverse problem.

There is also large statistics literature on nonparametric estimation of monotone functions when the regressor is exogenous, i.e. $W = X$, so that g is a conditional mean function. This literature can be traced back at least to Brunk (1955). Surveys of this literature and further references can be found in Yatchew (1998), Delecroix and Thomas-Agnan (2000), and Gijbels (2004). For the case in which the regression function is both smooth and monotone, many different ways of imposing monotonicity on the estimator have been studied; see, for example, Mukerjee (1988), Cheng and Lin (1981), Wright (1981), Friedman and Tibshirani (1984), Ramsay (1988), Mammen (1991), Ramsay (1998), Mammen and Thomas-Agnan (1999), Hall and Huang (2001), Mammen, Marron, Turlach, and Wand (2001), and Dette, Neumeyer, and Pilz (2006). Importantly, under the mild assumption that the estimators consistently estimate the derivative of the regression function, the standard unconstrained nonparametric regression estimators are known to be monotone with probability approaching one when the regression function is strictly increasing. Therefore, such estimators have the same rate of convergence as the corresponding constrained estimators that impose monotonicity (Mammen (1991)). As a consequence, gains from imposing a monotonicity constraint can only be expected when the regression function is close to the boundary of the constraint and/or in finite samples. Zhang (2002) and Chatterjee, Guntuboyina, and Sen (2013) formalize this intuition by deriving risk bounds of the isotonic (monotone) regression estimators and showing that these bounds imply fast convergence rates when the regression function has flat parts. Our results are different from theirs because we focus on the endogenous case with $W \neq X$ and study the impact of monotonicity constraints in the presence of ill-posedness which is absent in the standard regression problem.

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

corresponding cumulative distribution functions. For an operator $T : L^2[0,1] \to L^2[0,1]$, we let $\|T\|_2$ denote the operator norm defined as

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

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

Outline. The remainder of the paper is organized as follows. In the next section, we analyze ill-posedness of the model (1) under our monotonicity conditions and derive a useful bound on a restricted measure of ill-posedness for the model (1). Section 3 discusses the implications of our monotonicity assumptions for estimation of the regression function g. In particular, we show that the rate of convergence of our estimator is always not worse than that of unconstrained estimators but may be much faster in a large, but slowly shrinking, neighborhood of constant functions. Section 4 shows that our monotonicity conditions have non-trivial identification power. In Section 5, we present results of a Monte Carlo simulation study that demonstrates large gains in performance of the constrained estimator relative to the unconstrained one. Finally, Section 6 applies the constrained estimator to the problem of estimating gasoline demand functions. All proofs are collected in the appendix.

2 Boundedness of the Measure of Ill-posedness under Monotonicity

In this section, we introduce a restricted measure of ill-posedness for the NPIV model (1) that is relevant for the constrained estimator and study its properties.

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

$$Tg = m. \tag{5}$$

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 the same fast convergence rate. In infinite-dimensional models, however, T is an operator that, under weak conditions, 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 (5) ill-posed and introduces challenges in the estimation of the NPIV model (1) that are not present in parametric regressions nor in nonparametric regressions with exogenous regressors; see Horowitz (2011, 2014) for a detailed discussion.

In this section, we show that, under our conditions, there exists a finite constant $\bar{C}$ such that for any monotone function g' and any constant function g'', with $m' = Tg'$ and $m'' = Tg''$, we have

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

where $\|\cdot\|_{2,t}$ is a truncated L^2-norm defined below. This result plays a central role in the remainder of the paper.

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

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

$$w' \leq w'' \qquad \Rightarrow \qquad F_{X|W}(x|w') \geq F_{X|W}(x|w''). \tag{6}$$

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

$$F_{X|W}(x|w_1) \geq C_F F_{X|W}(x|w_2), \qquad \forall x \in (0, x_2) \tag{7}$$

and

$$C_F(1 - F_{X|W}(x|w_1)) \leq 1 - F_{X|W}(x|w_2), \qquad \forall x \in (x_1, 1) \tag{8}$$

Assumption 1 is crucial for our analysis. The first part, condition (6), requires first-order stochastic dominance of the conditional distribution of the endogenous regressor X given the instrument W as we increase the value of the instrument W. This condition (6) is testable; see, for example, Lee, Linton, and Whang (2009). In Appendix D, we extend the results of Lee, Linton, and Whang (2009) by providing an *adaptive* test of the first-order stochastic dominance condition (6).

The second and third parts of Assumption 1, conditions (7) and (8), strengthen the stochastic dominance condition (6) 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 (7) and (8) are rather weak as they require such a shift only in some intervals $(0, x_2)$ and $(x_1, 1)$, respectively.

Condition (6) 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 the equation $X = r(W,U)$ holds for the reduced form first stage function $r(w,u) := F^{-1}_{X|W}(u|w) := \inf\{x : F_{X|W}(x|w) \geq u\}$. Therefore, condition (6) 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 (6) allows for general unobserved heterogeneity of dimension larger than one, for instance as in Example 2 below.

Condition (6) is related to a corresponding condition in Kasy (2014) who assumes that the (structural) first stage has the form $X = \widetilde{r}(W,\widetilde{U})$ where $\widetilde{U}$, representing (potentially multidimensional) unobserved heterogeneity, is independent of W, and the function $w \mapsto \widetilde{r}(w,\widetilde{u})$ is increasing for all values $\widetilde{u}$. Kasy employs his condition for identification of (nonseparable) triangular systems with multidimensional unobserved heterogeneity whereas we use our condition (6) to derive a useful bound on the restricted measure of ill-posedness and to obtain a fast rate of convergence of a monotone NPIV estimator of g in the (separable) model (1). Condition (6) is not related to the monotone IV assumption in the influential work by Manski and Pepper (2000) which requires the function $w \mapsto \mathrm{E}[\varepsilon|W = w]$ to be increasing. Instead, we maintain the mean independence condition $\mathrm{E}[\varepsilon|W] = 0$.

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

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

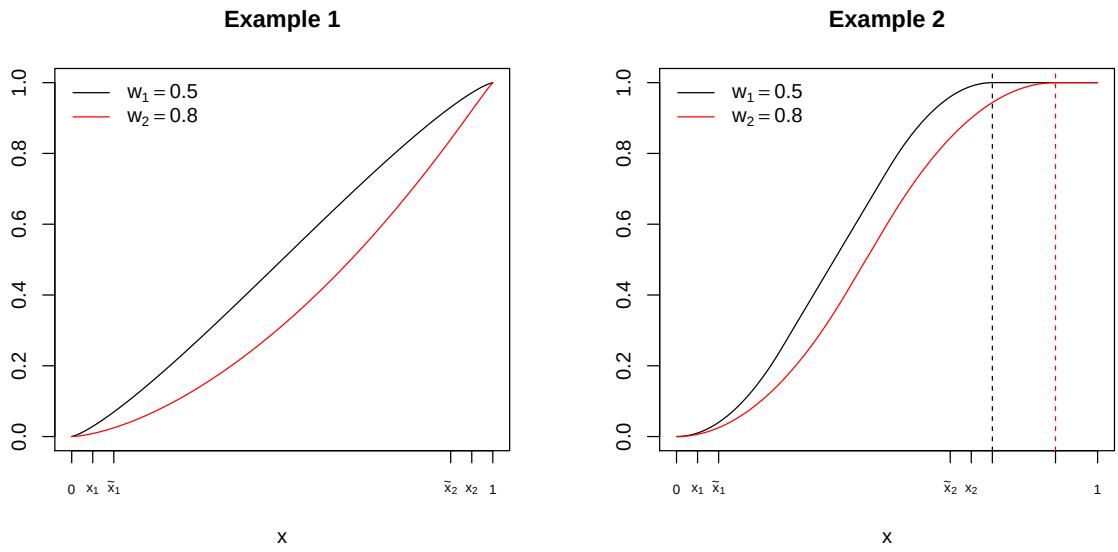

Figure 2: Plots of $F_{X|W}(x|w_1)$ and $F_{X|W}(x|w_2)$ in Examples 1 and 2, respectively.

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

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

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

where U is independent of W. Therefore, the pair (X, W) satisfies condition (6) of our monotone IV Assumption 1. Lemma 7 in the appendix verifies that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied. □

Example 2 (Two-dimensional unobserved heterogeneity)**.** Let $X = U_1 + U_2 W$, where U_1, U_2, W are mutually independent, $U_1, U_2 \sim U[0, 1/2]$ and $W \sim U[0,1]$. Since U_2 is positive, it is straightforward to see that the stochastic dominance condition (6) is satisfied. Lemma 8 in the appendix shows that the remaining conditions of Assumption 1 as well as Assumption 2 are also satisfied. □

Figure 2 shows that, in Example 1, the conditional distribution at two different values of the instrument is shifted to the right at every value of X, whereas, in Example 2, the

conditional support of X given $W = w$ changes with w, but the positive shift in the cdf of $X|W = w$ occurs only for values of X in a subinterval of $[0, 1]$.

Before stating our results in this section, we introduce some additional notation. Define the truncated L^2-norm $\|\cdot\|_{2,t}$ by

$$\|h\|_{2,t} := \left(\int_{\widetilde{x}_1}^{\widetilde{x}_2} h(x)^2 dx\right)^{1/2}, \quad h \in L^2[0, 1]. \tag{9}$$

Also, let $\mathcal{M}$ denote the set of all monotone functions in $L^2[0, 1]$. Finally, define $\zeta := (c_f, c_W, C_F, C_T, w_1, w_2, x_1, x_2, \widetilde{x}_1, \widetilde{x}_2)$. Below is our first main result in this section.

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

$$\|h\|_{2,t} \leq \bar{C}\|Th\|_2 \tag{10}$$

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 monotone IV Assumption 1, the function $w \mapsto \mathrm{E}[h(X)|W = w]$ is monotone whenever h is. Together with non-redundancy of the instrument W implied by conditions (7) and (8) of Assumption 1, this allows us to show that $\mathrm{E}[h(X)|W = w_1]$ and $\mathrm{E}[h(X)|W = w_2]$ cannot both be close to zero so that $\|\mathrm{E}[h(X)|W = \cdot]\|_2$ is bounded from below by a strictly positive constant from the values of $\mathrm{E}[h(X)|W = w]$ in the neighborhood of either w_1 or w_2. By Assumption 2, $\|Th\|_2$ must then also be bounded away from zero.

Theorem 1 has an important consequence. Indeed, consider the linear equation (5). By Assumption 2(i), the operator T is compact, and so

$$\frac{\|h_k\|_2}{\|Th_k\|_2} \to \infty \text{ as } k \to \infty \text{ for some sequence } \{h_k, k \geq 1\} \subset L^2[0, 1]. \tag{11}$$

Property (11) means that $\|Th\|_2$ being small does not necessarily imply that $\|h\|_2$ is small and, therefore, the inverse of the operator $T : L^2[0, 1] \to L^2[0, 1]$, when it exists, cannot be continuous. Therefore, (5) is ill-posed in Hadamard's sense[2]. Theorem 1, on the other

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

hand, implies that, under Assumptions 1 and 2, (11) 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 (11) by the truncated L^2-norm $\|\cdot\|_{2,t}$. Even though this result may appear surprising and is important for studying the finite-sample behavior of the constrained estimator we present in the next section, it does not imply well-posedness of the constrained NPIV problem; see Scaillet (2016) for more details.

In Remark 1, we show that truncating the norm in the numerator is not a significant modification in the sense that for most ill-posed problems, and in particular for all severely ill-posed problems, (11) holds even if we replace the L^2-norm $\|\cdot\|_2$ in the numerator of the left-hand side of (11) by the truncated L^2-norm $\|\cdot\|_{2,t}$.

Next, we derive an implication of Theorem 1 for the (quantitative) measure of ill-posedness of the model (1). We first define the restricted measure of ill-posedness. For $a \in \mathbb{R}$, let

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

be the space containing all functions in $L^2[0,1]$ with lower derivative bounded from below by $-a$ uniformly over the interval $[0,1]$. Note that $\mathcal{H}(a') \subset \mathcal{H}(a'')$ whenever $a' \leq a''$ and that $\mathcal{H}(0)$ is 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 *restricted measure of ill-posedness*:

$$\tau(a) := \sup_{\substack{h \in \mathcal{H}(a) \\ \|h\|_{2,t}=1}} \frac{\|h\|_{2,t}}{\|Th\|_2}. \tag{12}$$

We 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 (12). However, Theorem 1 implies that $\tau(0)$ is bounded from above by $\bar{C}$ and, by definition, $\tau(a)$ is increasing in a, i.e. $\tau(a') \leq \tau(a'')$ for $a' \leq a''$. It turns out that $\tau(a)$ is bounded from above even for some positive values of a:

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

$$\tau(a) \leq C_\tau \tag{13}$$

for all $a \leq c_\tau$.

This is our second main result in this section. It is exactly this corollary of Theorem 1 that allows us to obtain the novel non-asymptotic error bound for the constrained estimator proposed in the next section.

Remark 1 (Ill-posedness is preserved by norm truncation)**.** Under Assumptions 1 and 2, the integral operator T satisfies (11). Here we demonstrate that, in many cases, and in particular in all severely ill-posed cases, (11) continues to hold if we replace the L^2-norm $\|\cdot\|_2$ by the truncated L^2-norm $\|\cdot\|_{2,t}$ in the numerator of the left-hand side of (11), that is, there exists a sequence $\{l_k, k \geq 1\}$ in $L^2[0,1]$ such that

$$\frac{\|l_k\|_{2,t}}{\|Tl_k\|_2} \to \infty \text{ as } k \to \infty. \tag{14}$$

Indeed, under Assumptions 1 and 2, T is compact, and so the spectral theorem implies that there exists a spectral decomposition of operator T, $\{(h_j, \varphi_j), j \geq 1\}$, where $\{h_j, j \geq 1\}$ is an orthonormal basis of $L^2[0,1]$ and $\{\varphi_j, j \geq 1\}$ is a decreasing sequence of positive numbers such that $\varphi_j \to 0$ as $j \to \infty$, and $\|Th_j\|_2 = \varphi_j \|h_j\|_2 = \varphi_j$. Also, Lemma 6 in the appendix shows that if $\{h_j, j \geq 1\}$ is an orthonormal basis in $L^2[0,1]$, then for any $\alpha > 0$, $\|h_j\|_{2,t} > j^{-1/2-\alpha}$ for infinitely many j, and so there exists a subsequence $\{h_{j_k}, k \geq 1\}$ such that $\|h_{j_k}\|_{2,t} > {j_k}^{-1/2-\alpha}$. Therefore, under a weak condition that $j^{1/2+\alpha}\varphi_j \to 0$ as $j \to \infty$, using $\|h_{j_k}\|_2 = 1$ for all $k \geq 1$, we conclude that for the subsequence $l_k = h_{j_k}$,

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

leading to (14). Note also that the condition that $j^{1/2+\alpha}\varphi_j \to 0$ as $j \to \infty$ necessarily holds if there exists a constant $c > 0$ such that $\varphi_j \leq e^{-cj}$ for all large j, that is, if the problem is severely ill-posed. Thus, under our Assumptions 1 and 2, the restriction in Theorem 1 that h belongs to the space $\mathcal{M}$ of *monotone* functions in $L^2[0,1]$ plays a crucial role for the result (10) to hold. On the other hand, whether the result (10) can be obtained for all $h \in \mathcal{M}$ without imposing our monotone IV Assumption 1 appears to be an open (and interesting) question. □

Remark 2 (Severe ill-posedness is preserved by norm truncation)**.** One might wonder whether our monotone IV Assumption 1 excludes all severely ill-posed problems, and whether the norm truncation significantly changes these problems. Here we show that there do exist severely ill-posed problems that satisfy our monotone IV Assumption 1, and also that severely ill-posed problems remain severely ill-posed even if we replace the L^2-norm $\|\cdot\|_2$ by the truncated L^2-norm $\|\cdot\|_{2,t}$. Indeed, consider Example 1 above. Because, in this example, the pair (X, W) is a transformation of the normal distribution, it is well known that the integral operator T in this example has singular values decreasing exponentially fast. More specifically, the spectral decomposition $\{(h_k, \varphi_k), k \geq 1\}$ of the operator T satisfies $\varphi_k = \rho^k$ for all k and some $\rho < 1$. Hence,

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

Since $(1/\rho)^k \to \infty$ as $k \to \infty$ exponentially fast, this example leads to a severely ill-posed problem. Moreover, by Lemma 6, for any $\alpha > 0$ and $\rho' \in (\rho, 1)$,

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

for infinitely many k. Thus, replacing the L^2 norm $\|\cdot\|_2$ by the truncated L^2 norm $\|\cdot\|_{2,t}$ preserves the severe ill-posedness of the problem. However, it follows from Theorem 1 that uniformly over all $h \in \mathcal{M}$, $\|h\|_{2,t}/\|Th\|_2 \leq \bar{C}$. Therefore, in this example, as well as in all other severely ill-posed problems satisfying Assumptions 1 and 2, imposing monotonicity on the function $h \in L^2[0,1]$ significantly changes the properties of the ratio $\|h\|_{2,t}/\|Th\|_2$. □

Remark 3 (Monotone IV assumption does not imply control function approach)**.** Our monotone IV Assumption 1 does not imply the applicability of a control function approach to the estimation of the function g. Consider Example 2 above. In this example, the relationship between X and W has a two-dimensional vector (U_1, U_2) of unobserved heterogeneity. Therefore, by Proposition 4 of Kasy (2011), there does not exist any control function $C : [0,1]^2 \to \mathbb{R}$ such that (i) C is invertible in its second argument, and (ii) X is independent of ε conditional on $V = C(X, W)$. As a consequence, our monotone IV Assumption 1 does not imply any of the existing control function conditions such as those in Newey, Powell, and Vella (1999) and Imbens and Newey (2009), for example.[3] Since multidimensional unobserved heterogeneity is common in economic applications (see Imbens (2007) and Kasy (2014)), we view our approach to avoiding ill-posedness as complementary to the control function approach. □

Remark 4 (On the role of norm truncation)**.** Let us also briefly comment on the role of the truncated norm $\|\cdot\|_{2,t}$ in (10). The main reason for truncating the norm is that we want to cover the case of X and W being jointly normal, which gives an example of a severely ill-posed problem. To avoid the truncation, we would have to set $\widetilde{x}_1 = x_1 = 0$ and $\widetilde{x}_2 = x_2 = 1$, and then the part (ii) of Assumption 2 would exclude the case of X and W being jointly normal. If we do set $\widetilde{x}_1 = x_1 = 0$ and $\widetilde{x}_2 = x_2 = 1$, however, it then follows from Lemma 2 in the appendix that, under Assumptions 1 and 2, there exists a constant $0 < C_2 < \infty$ such that

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

for any increasing and continuously differentiable function $h \in L^1[0,1]$. This result does not require any truncation of the norms and implies boundedness of a measure of ill-posedness defined in terms of $L^1[0,1]$-norms: $\sup_{h \in L^1[0,1], h \text{ increasing}} \|h\|_1/\|Th\|_1 \leq C_2$. □

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

3 Non-asymptotic Risk Bounds Under Monotonicity

The rate at which unconstrained NPIV estimators converge to g 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 the space $\mathcal{H}_n(\infty)$, a finite-dimensional (sieve) approximation to $\mathcal{H}(\infty)$. In particular, the convergence rate is slower the faster the sieve measure of ill-posedness grows with the dimensionality of the sieve space $\mathcal{H}_n(\infty)$. The convergence rates can be as slow as logarithmic in the severely ill-posed case. Since by Corollary 1, our monotonicity assumptions imply boundedness of $\tau(a)$ for some range of finite values a, we expect these assumptions to translate into favorable performance of a constrained estimator that imposes monotonicity of g. In this section, we derive a novel non-asymptotic bound on the estimation error of the constrained estimator that imposes monotonicity of g (Theorem 2), which gives a set of data-generating processes where the monotonicity constraint has a strong regularization effect and substantially improves finite-sample properties of the estimator.

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

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

Let $\mathbf{P} := (p(X_1), \ldots, p(X_n))'$ and $\mathbf{Q} := (q(W_1), \ldots, q(W_n))'$. Similarly, stack all observations on Y in $\mathbf{Y} := (Y_1, \ldots, Y_n)'$. Let $\mathcal{H}_n(a)$ be a sequence of finite-dimensional spaces defined by

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

which become dense in $\mathcal{H}(a)$ as $n \to \infty$. Throughout the paper, we assume that $\|g\|_2 \leq C_b$ where C_b is a large but finite constant known by the researcher. We define two estimators of g: the unconstrained estimator $\widehat{g}^u(x) := p(x)'\widehat{\beta}^u$ with

$$\widehat{\beta}^u := \text{argmin}_{b \in \mathbb{R}^K : \|b\| \leq C_b} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b) \tag{15}$$

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

$$\widehat{\beta}^c := \text{argmin}_{b \in \mathbb{R}^K : p(\cdot)'b \in \mathcal{H}_n(0), \|b\| \leq C_b} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b), \tag{16}$$

which imposes the monotonicity of g through the constraint $p(\cdot)'b \in \mathcal{H}_n(0)$.

To study properties of the two estimators we introduce a finite-dimensional, or sieve, counterpart of the restricted measure of ill-posedness $\tau(a)$ defined in (12) and also recall the definition of the (unrestricted) sieve measure of ill-posedness. Specifically, define the *restricted* and *unrestricted* sieve measures of ill-posedness $\tau_{n,t}(a)$ and τ_n as

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

The sieve measure of ill-posedness defined in Blundell, Chen, and Kristensen (2007) and also used, for example, in Horowitz (2012) is τ_n. Blundell, Chen, and Kristensen (2007) show that τ_n is related to the singular values of T.[4] If the singular values converge to zero at the rate K^{-r} as $K \to \infty$, then, under certain conditions, τ_n diverges at a polynomial rate, that is $\tau_n = O(K_n^r)$. This case is typically referred to as "mildly ill-posed". On the other hand, when the singular values decrease at a fast exponential rate, then $\tau_n = O(e^{cK_n})$, for some constant $c > 0$. This case is typically referred to as "severely ill-posed".

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

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

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

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

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

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

[4] In fact, Blundell, Chen, and Kristensen (2007) talk about the eigenvalues of T^*T, where T^* is the adjoint of T but there is a one-to-one relationship between eigenvalues of T^*T and singular values of T.

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

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

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

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

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

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

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

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

$$\xi_{K,p} := \sup_{x \in [0,1]} \|p(x)\|, \quad \xi_{J,q} := \sup_{w \in [0,1]} \|q(w)\|, \quad \xi_n := \max(\xi_{K,p}, \xi_{J,q}). \tag{17}$$

The constant ξ_n satisfies the bound $\xi_n^2 \leq C_\xi K$ for some $0 < C_\xi < \infty$ independent of K if the sequences $\{p_k(x), k \geq 1\}$ and $\{q_k(w), k \geq 1\}$ consist of commonly used bases such as Fourier, spline, wavelet, or local polynomial partition series; see Belloni, Chernozhukov, Chetverikov, and Kato (2014) for details.

We start our analysis in this section with a simple observation that, if the function g is strictly increasing and the sample size n is sufficiently large, then the constrained estimator $\widehat{g}^c$ coincides with the unconstrained estimator $\widehat{g}^u$, and the two estimators therefore share the same rate of convergence.

Lemma 1 (Asymptotic equivalence of constrained and unconstrained estimators)**.** *Let Assumptions 2 and 4-8 be satisfied. In addition, assume that g is continuously differentiable and $Dg(x) \geq c_g$ for all $x \in [0,1]$ and some constant $c_g > 0$. If $\tau_n^2 \xi_n^2 \log n / n \to 0$,*

$\sup_{x\in[0,1]} \|Dp(x)\|(\tau_n(K/n)^{1/2}+K^{-s}) \to 0$, *and* $\sup_{x\in[0,1]} |Dg(x)-Dg_n(x)| \to 0$ *as* $n \to \infty$, *then*

$$\mathrm{P}\Big(\widehat{g}^c(x) = \widehat{g}^u(x) \text{ for all } x \in [0,1]\Big) \to 1 \text{ as } n \to \infty. \tag{18}$$

The result in Lemma 1 is similar to that in Theorem 1 of Mammen (1991), which shows equivalence, in the sense of (18), of the constrained and unconstrained estimators of conditional mean functions. Lemma 1 implies that imposing the monotonicity constraint cannot lead to improvements in the rate of convergence of the estimator if g is strictly increasing. However, the result in Lemma 1 is asymptotic and does not rule out significant performance gains in finite samples, which we find in our simulations presented in Section 5. Therefore, we next derive a *non-asymptotic* estimation error bound for the constrained estimator $\widehat{g}^c$ and study the impact of the monotonicity constraint on this bound.

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

$$\|\widehat{g}^c - g\|_{2,t} \leq C\Big\{\delta + \tau_{n,t}\Big(\frac{\|Dg_n\|_\infty}{\delta}\Big)V_n + K^{-s}\Big\} \tag{19}$$

and

$$\|\widehat{g}^c - g\|_{2,t} \leq C \min\Big\{\|Dg\|_\infty + V_n, \tau_n V_n\Big\} + CK^{-s}, \tag{20}$$

where $V_n := \sqrt{K/(\alpha n) + (\xi_n^2 \log n)/n}$*. Here the constants* $c, C < \infty$ *depend only on the constants appearing in Assumptions 1-8.*

This is the main result of this section. An important feature of the bounds (19) and (20) in this result is that since the constant C depends only on the constants appearing in Assumptions 1-8, these bounds hold uniformly over all data-generating processes that satisfy those assumptions with the same constants.[5] In particular, for any two data-generating processes in this set, the same finite-sample bounds (19) and (20) hold with the same constant C, even though the unrestricted sieve measure of ill-posedness τ_n may be of different order of magnitude for these two data-generating processes.

Another important feature of the bound (19) is that it depends on the restricted sieve measure of ill-posedness that we know to be smaller than the unrestricted sieve measure of ill-posedness, appearing in the analysis of the unconstrained estimator. In

[5]The dependence of C on those constants can actually be traced from the proof of the theorem. We do not trace this dependence, however, because the main purpose of the theorem is not to calculate the regularization effect of the monotonicity constraint exactly, but rather to show that this effect can be strong even in large samples and for relatively steep functions g.

particular, we know from Section 2 that $\tau_{n,t}(a) \leq \tau(a)$ and that, by Corollary 1, $\tau(a)$ is uniformly bounded if a is not too large. Employing this result, we obtain the bound (20) of Theorem 2.[6]

The right-hand side of the bound (20) consists of two parts, the bias term CK^{-s} that vanishes with the number of series terms K and the variance term $C\min\{\|Dg\|_\infty + V_n, \tau_n V_n\}$. The variance term depends on the maximum slope of the regression function $\|Dg\|_\infty$, the unrestricted sieve measure of ill-posedness τ_n, and V_n that, for many commonly used bases, is of order $\sqrt{K \log n/n}$, the order of the variance in well-posed problems such as conditional mean estimation (up to the log-factor); see our discussion after equation (17) above.

The bound (20) has several interesting features. First, the right-hand side of the bound weakly decreases with the magnitude of the maximum slope of g, so that the bound is tighter for flatter functions. Also, the higher the desired level of confidence α with which we want to bound the estimation error, the larger the bound.

Second, and more importantly, the variance term in the bound (20) is determined by the minimum of two regimes. For a given sample size n, the minimum is attained in the first regime if

$$\|Dg\|_\infty \leq (\tau_n - 1)V_n. \tag{21}$$

In this regime, the right-hand side of the bound (20) is independent of the (unrestricted) sieve measure of ill-posedness τ_n, and so is independent of whether the original NPIV model (1) is mildly or severely ill-posed. This is the regime in which the bound relies upon the monotonicity constraint imposed on the estimator $\widehat{g}^c$ and in which the regularization effect of the monotonicity constraint is strong as long as $\|Dg\|_\infty \ll (\tau_n - 1)V_n$. This regime is important since even though V_n is expected to be small, because of ill-posedness, τ_n can be large, or even very large in severely ill-posed problems, and so this regime may be active even in relatively large samples and for relatively steep functions g.

Third, as the sample size n gets large, the right-hand side of the inequality (21) decreases (if $K = K_n$ grows slowly enough) and eventually becomes smaller than the left-hand side, and the bound (20) switches to its second regime, in which it depends on the (unrestricted) sieve measure of ill-posedness τ_n. This is the regime in which the monotonicity constraint imposed on $\widehat{g}^c$ has no impact on the error bound. However, when the problem is sufficiently ill-posed, this regime switch may occur only at extremely large sample sizes. Panel (a) in Figure 3 illustrates this point. Lines A and B denote $\|Dg\|_\infty + V_n$ (first regime) and $\tau_n V_n$ (second regime), respectively. A converges to the maximum slope $\|Dg\|_\infty$ as $n \to \infty$, but is of smaller order than B because of the multiplication by the

[6] Ideally, it would be of great interest to have a tight bound on the restricted sieve measure of ill-posedness $\tau_{n,t}(a)$ for all $a \geq 0$, so that it would be possible to optimize (19) over δ.

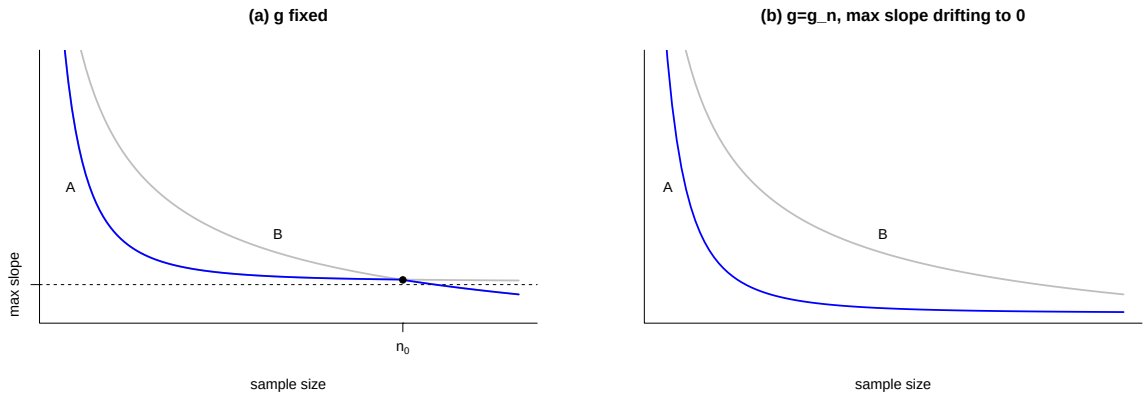

Figure 3: Stylized graphs showing the relationship of the two regimes determining the minimum of the estimation error bound in (19).

possibly large factor τ_n. As n grows sufficiently large, i.e. larger than n_0, B becomes smaller than A. Therefore, the error bound (blue line) is in its first regime, in which the monotonicity constraint has an impact, up to the sample size n_0 and then switches to the second regime in which the constraint becomes irrelevant and the ill-posedness of the problem determines the speed of convergence to zero.

We note also that the truncation of the L^2-norm on the left-hand side of the bounds (19) and (20) does not change the meaning of the bounds from the practical point of view. This truncation does imply that there might be complicated boundary phenomena affecting the precision of the constrained estimator $\widehat{g}^c(x)$ for the values of x that are close to either 0 or 1 (boundaries of the support of X) but in most applications, researchers are typically not interested in these values and instead are interested in the values of x in the interior of the support of X.

Another way to understand the improvements from the monotonicity constraint is via local-to-flat asymptotics. Specifically, consider a sequence of data-generating processes indexed by the sample size n in which the maximum slope of g drifts to zero. The following corollary is a straightforward consequence of the bound (20):

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

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

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

$0 < C_K < \infty$, *then*

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

The local asymptotic theory considered in this corollary captures the finite-sample situation in which the regression function is not too steep relative to the sample size, i.e. $\|Dg\|_\infty = O((K \log n/n)^{1/2})$. In this shrinking neighborhood, the constrained estimator's convergence rate 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 our monotonicity assumptions is mildly or severely ill-posed. The reason why this is possible is illustrated in panel (b) of Figure 3. In the local neighborhood of functions g such that $\|Dg\|_\infty = O((K \log n/n)^{1/2})$, the minimum in (20) is always attained in the first regime because A is always below B.

This result means that the constrained estimator $\widehat{g}^c$ is able to recover regression functions in the shrinking neighborhood of constant functions at a fast polynomial rate. Notice that the neighborhood of functions g that satisfy $\|Dg\|_\infty = O((K \log n/n)^{1/2})$ is shrinking at a slow rate because $K \to \infty$, in particular the rate is much slower than $n^{-1/2}$. Therefore, in finite samples, we expect the estimator to perform well for a wide range of (non-constant) regression functions g as long as the function g is not too steep relative to the sample size.

In conclusion, in general, the convergence rate of the constrained estimator is the same as the standard minimax optimal rate, which depends on the degree of ill-posedness and may, in the worst-case, be logarithmic. This case occurs in the interior of the monotonicity constraint when g is strictly monotone. On the other hand, under the monotone IV assumption, the constrained estimator converges at a very fast rate, independently of the degree of ill-posedness, in a large but slowly shrinking neighborhood of constant functions, a part of the boundary of the monotonicity constraint. In finite samples, we expect to experience cases between the two extremes, and the bounds (19) and (20) show how the performance of the constrained estimator depends on the degree of ill-posedness, the sample size, and how close the monotonicity constraint is to binding in the population. Since the first regime of bound (20) is active in a large set of data-generating processes and sample size combinations, and since the fast convergence rate in Corollary 2 is obtained in a large but slowly shrinking neighborhood of constant functions, we expect the boundary effect due to the monotonicity constraint to be strong even far away from the boundary and for relatively large sample sizes.

Remark 5 (On robustness of the constrained estimator, I)**.** Implementation of the estimators $\widehat{g}^c$ and $\widehat{g}^u$ requires selecting the number of series terms $K = K_n$ and $J = J_n$. This is a difficult problem because the measure of ill-posedness $\tau_n = \tau(K_n)$, appearing in the

convergence rate of both estimators, depends on $K = K_n$ and can blow up quickly as we increase K. Therefore, setting K higher than the optimal value may result in a severe deterioration of the statistical properties of $\widehat{g}^u$. The problem is alleviated, however, in the case of the constrained estimator $\widehat{g}^c$ because $\widehat{g}^c$ satisfies the bound (20) of Theorem 2, which is independent of τ_n for sufficiently large K. In this sense, the constrained estimator $\widehat{g}^c$ possesses some robustness against setting K too high. □

Remark 6 (On robustness of the constrained estimator, II)**.** Notice that the fast convergence rates in the local asymptotics derived in this section are obtained under two monotonicity conditions, Assumptions 1 and 3, but the estimator imposes only the monotonicity of the regression function, not that of the first stage. Therefore, our proposed constrained estimator consistently estimates the regression function g even when the monotone IV assumption is violated. □

Remark 7 (Imposing Monotonicity by Re-arrangement)**.** Chernozhukov, Fernández-Val, and Galichon (2009) show that re-arranging any unconstrained estimator so that it becomes monotone decreases the estimation error of the estimator. However, their argument does not quantify this improvement, so that it does not seem possible to conclude from their argument whether and when the improvement is large, even qualitatively. In contrast, the main contribution of our paper is to show that enforcing the monotonicity constraint in the NPIV model yields substantial performance improvements even in large samples and for steep regression functions g as long as the NPIV model is severely ill-posed. □

Remark 8 (Estimating partially flat functions)**.** Since the inversion of the operator T is a global inversion in the sense that the resulting estimators $\widehat{g}^c(x)$ and $\widehat{g}^u(x)$ depend not only on the shape of $g(x)$ locally at x, but on the shape of g over the whole domain, we do not expect convergence rate improvements from imposing monotonicity when the function g is partially flat. However, we leave the question about potential improvements from imposing monotonicity in this case for future research. □

Remark 9 (Computational aspects)**.** The implementation of the constrained estimator in (16) is particularly simple when the basis vector $p(x)$ consists of polynomials or B-splines of order 2. In that case, $Dp(x)$ is linear in x and, therefore, the constraint $Dp(x)'b \geq 0$ for all $x \in [0,1]$ needs to be imposed only at the knots or endpoints of $[0,1]$, respectively. The estimator $\widehat{\beta}^c$ thus minimizes a quadratic objective function subject to a (finite-dimensional) linear inequality constraint. When the order of the polynomials or B-splines in $p(x)$ is larger than 2, imposing the monotonicity constraint is slightly more complicated, but it can still be transformed into a finite-dimensional constraint using a

representation of non-negative polynomials as a sum of squared polynomials:[7] one can represent any non-negative polynomial $f : \mathbb{R} \to \mathbb{R}$ as a sum of squares of polynomials (see the survey by Reznick (2000), for example), i.e. $f(x) = \widetilde{p}(x)'M\widetilde{p}(x)$ where $\widetilde{p}(x)$ is the vector of monomials up to some order and M a matrix of coefficients. Letting $f(x) = Dp(x)'b$, our monotonicity constraint $f(x) \geq 0$ can then be written as $\widetilde{p}(x)'M\widetilde{p}(x) \geq 0$ for some matrix M that depends on b. This condition is equivalent to requiring the matrix M to be positive semi-definite. $\widehat{\beta}^c$ thus minimizes a quadratic objective function subject to a (finite-dimensional) semi-definiteness constraint.

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

Remark 10 (Penalization and shape constraints)**.** Recall that the estimators $\widehat{g}^u$ and $\widehat{g}^c$ require setting the constraint $\|b\| \leq C_b$ in the optimization problems (15) and (16). In practice, this constraint, or similar constraints in terms of Sobolev norms, which also impose bounds on derivatives of g, are typically not enforced in the implementation of an NPIV estimator. Horowitz (2012) and Horowitz and Lee (2012), for example, observe that imposing the constraint does not seem to have an effect in their simulations. On the other hand, especially when one includes many series terms in the computation of the estimator, Blundell, Chen, and Kristensen (2007) and Gagliardini and Scaillet (2012), for example, argue that penalizing the norm of g and of its derivatives may stabilize the estimator by reducing its variance. In this sense, penalizing the norm of g and of its derivatives may have a similar effect as imposing monotonicity. However, there are at least two important differences between penalization and imposing monotonicity. First, penalization increases bias of the estimators. In fact, especially in severely ill-posed problems, even small amount of penalization may lead to large bias (otherwise severely ill-posed problems could lead to estimators with fast convergence rates). In contrast, the monotonicity constraint on the estimator does not increase bias much when the function g itself satisfies the monotonicity constraint. Second, penalization requires selecting a tuning parameter that governs the strength of penalization, which is a difficult statistical problem. In contrast, imposing monotonicity does not require such choices and can often be motivated directly from economic theory. □

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

4 Identification Bounds under Monotonicity

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

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

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

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

Theorem 3 (Identification of the sign of the slope)**.** *Suppose Assumptions 1 and 2 hold and $f_{X,W}(x,w) > 0$ for all $(x,w) \in (0,1)^2$. If g is monotone and continuously differentiable, then $sign(Dg)$ is identified.*

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

Remark 11 (A test for the sign of the slope of g)**.** In fact, Theorem 3 yields a surprisingly simple way to test the sign of the slope of the function g. Indeed, the proof of Theorem 3 reveals that g is increasing, constant, or decreasing if the function $w \mapsto \mathrm{E}[Y|W = w]$ is increasing, constant, or decreasing, respectively. By Chebyshev's association inequality

(Lemma 5 in the appendix), the latter assertions are equivalent to the coefficient β in the linear regression model

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

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

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

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

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

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

The following theorem provides necessary conditions for observational equivalence.

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

Under Assumption 3 that g is increasing, Theorem 4 suggests the construction of a set Θ' that includes the identified set Θ by $\Theta' := \mathcal{M}_+ \backslash \Delta$, where $\mathcal{M}_+ := \mathcal{H}(0)$ denotes all increasing functions in $\mathcal{M}$ and

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

We emphasize that Δ is not empty, which means that our Assumptions 1–3 possess identifying power leading to nontrivial bounds on g. Notice that the constant $\bar{C}$ depends

only on the observable quantities c_w, c_f, and C_F from Assumptions 1–2, and on the known constants $\widetilde{x}_1$, $\widetilde{x}_2$, x_1, x_2, w_1, and w_2. Therefore, the set Θ' could, in principle, be estimated, but that may be difficult based on the above representation of the identified set. In this section, we mainly want to point out that that our two monotonicity assumptions possess identification power in the sense that the identified set for g is not equal to the set of all monotone functions in $L^2[0,1]$.

Remark 12 (Further insight on identification bounds)**.** It is possible to provide more insight into which functions are in Δ and thus not in Θ'. First, under the additional minor condition that $f_{X,W}(x,w) > 0$ for all $(x,w) \in (0,1)^2$, all functions in Θ' have to intersect g; otherwise they are not observationally equivalent to g. Second, for a given $g' \in \mathcal{M}_+$ and $h \in L^2[0,1]$ such that $g' - g + h$ is monotone, the inequality in condition (24) is satisfied if $\|h\|_2$ is not too large relative to $\|g' - g\|_{2,t}$. In the extreme case, setting $h = 0$ shows that Θ' does not contain elements g' that disagree with g on $[\widetilde{x}_1, \widetilde{x}_2]$ and such that $g' - g$ is monotone. More generally, Θ' does not contain elements g' whose difference with g is too close to a monotone function. Therefore, for example, functions g' that are much steeper than g are excluded from Θ'. □

5 Simulations

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

$$\begin{aligned} &\text{Model 1: } g(x) = x^2 + 0.2x, && x \in [0,1], \\ &\text{Model 2: } g(x) = 2(x - 1/2)_+^2 + 0.5x, && x \in [0,1], \end{aligned}$$

where for any $a \in \mathbb{R}$, we denote $(a)_+ = a1\{a > 0\}$. We set $W = \Phi(\zeta)$, $X = \Phi(\rho\zeta + \sqrt{1-\rho^2}\epsilon)$, and $\varepsilon = \sigma(\eta\epsilon + \sqrt{1-\eta^2}\nu)$, where ρ, η, and σ are parameters and ζ, ϵ, and ν are independent $N(0,1)$ random variables. We always set $\sigma = 0.5$ but depending on the experiment, we set $\rho = 0.3$ or 0.5 and $\eta = 0.3$ or 0.7. Thus, we consider $2 \cdot 2 \cdot 2 = 8$ different designs in total.

In all cases, we use the same sieve spaces for X and W both for the constrained and unconstrained estimators, that is, $p(x) = q(x)$ for all $x \in [0,1]$. Depending on the experiment, the dimension of the sieve space K is 2, 3, 4, or 5, and we always use regression splines, so that $p(x) = (1,x)'$ if $K = 2$, $p(x) = (1,x,x^2)'$ if $K = 3$, $p(x) = (1,x,x^2,(x-1/2)_+^2)'$ if $K = 4$, and $p(x) = (1,x,x^2,(x-1/3)_+^2,(x-2/3)_+^2)'$ if $K = 5$.

The results of our experiments are presented in Tables 1-8, with one table corresponding to one simulation design. Each table shows the MISE of the constrained estimator (top panel), the MISE of the unconstrained estimator (middle panel), and their ratio (bottom panel) as a function of the sample size n and the dimension of the sieve space K for the corresponding simulation design. Specifically, the top and middle panels show the empirical median of $1,000 \cdot \int_0^1 (\widehat{g}^c(x) - g(x))^2 dx$ and $1,000 \cdot \int_0^1 (\widehat{g}^u(x) - g(x))^2 dx$, respectively, over 500 simulations (we have also calculated the empirical means but we prefer to report the empirical medians because the empirical mean for the unconstrained estimator is often unstable due to outliers arising when some singular values of the matrix $\mathbf{P}'\mathbf{Q}/n$ are too close to zero; reporting the empirical means would be even more favorable for the constrained estimator). The bottom panel reports the ratio of these two quantities. Both for the constrained and unconstrained estimators, we also report in the last column of the top and middle panels the optimal value of the corresponding MISE that is obtained by optimization over the dimension of the sieve space K. Finally, the last column of the bottom panel reports the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

The results indicate that the constrained estimator often outperforms, sometimes substantially, the unconstrained one even if the sample size n is rather large. For example, in the design with $g(x) = x^2 + 0.2x$, $\rho = 0.3$, and $\eta = 0.3$ (Table 1), when $n = 5,000$ and K is chosen optimally both for the constrained and unconstrained estimators, the ratio of the MISE of the constrained estimator to the MISE of the unconstrained one is equal to remarkable 0.2, so that the constrained estimator is 5 times more efficient than the unconstrained one. The reason for this efficiency gain is that using the unconstrained estimator with $K = 2$ yields a large bias but increasing K to 3 leads to a large variance, whereas using the constrained estimator with $K = 3$ gives a relatively small variance, with the bias being relatively small as well. In addition, in the design with $g(x) = 2(x - 1/2)_+^2 + 0.5x$, $\rho = 0.3$, and $\eta = 0.7$ (Table 7), when K is chosen optimally both for the constrained and unconstrained estimators, the ratio of the MISE of the constrained estimator to the MISE of the unconstrained one does not exceed 0.8 even if $n = 500,000$, which is a very large sample size for a typical dataset in economics.

Our simulation results also show that imposing the monotonicity of g on the estimator sometimes may not lead to efficiency gains in small samples (see the case $n = 500$ in Tables 1, 3, 5, and 7). This happens because in small samples, it is optimal to set $K = 2$, so that $p(x) = (1, x)'$, even for the constrained estimator, in which case the monotonicity constraint is not binding with large probability. However, in some cases the gain can be substantial even when $n = 500$; see design with $g(x) = x^2 + 0.2x$, $\rho = 0.5$, and $\eta = 0.7$ (Table 4), for example.

Finally, it is interesting to note that whenever K is set to be larger than optimal, the growth of the MISE of the constrained estimator as we further increase K is much slower than that of the MISE of the unconstrained estimator. For example, in the design with $g(x) = x^2 + 0.2$, $\rho = 0.5$, and $\eta = 0.3$ (Table 2) with $n = 1,000$, it is optimal to set $K = 3$ both for the constrained and unconstrained estimators, but when we increase K from 3 to 4, the MISE of the constrained estimator grows from 2.07 to 9.40 and the MISE of the constrained estimator grows from 4.55 to 58.69. This shows that the constrained estimator is more robust than the unconstrained one to incidental mistakes in the choice of K.

6 Gasoline Demand in the United States

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

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

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

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

positive income derivative[8] $\partial g/\partial z_1$, the Slutsky condition implies that the uncompensated (Marshallian) demand curves are also downward-sloping, i.e. $g(\cdot, z_1)$ should be monotone for any z_1, as long as income effects do not completely offset price effects. Finally, we expect the cost shifter W to monotonically increase cost of producing gasoline and thus satisfy our monotone IV condition. In conclusion, our constrained NPIV estimator appears to be an attractive estimator of demand functions in this setting.

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

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

We report results for the following choice of bases. All estimators employ a quadratic B-spline basis with 3 knots for price X and a cubic B-spline with 10 knots for the instrument W. Denote these two bases by $\mathbf{P}$ and $\mathbf{Q}$, using the same notation as in Section 3. In step (i), the NPIV estimators include the additional exogenous covariates (Z_1, Z_2) in the respective bases for X and W, so they use the estimator defined in Section 3 except that the bases $\mathbf{P}$ and $\mathbf{Q}$ are replaced by $\widetilde{\mathbf{P}} := [\mathbf{P}, \mathbf{P} \times \mathbf{Z}_1, \mathbf{Z}_2]$ and $\widetilde{\mathbf{Q}} := [\mathbf{Q}, \mathbf{Q} \times (\mathbf{Z}_1, \mathbf{Z}_2)]$, respectively, where $\mathbf{Z}_k := (Z_{k,1}, \ldots, Z_{k,n})'$, $k = 1, 2$, stacks the observations $i = 1, \ldots, n$ and $\mathbf{P} \times \mathbf{Z}_1$ denotes the tensor product of the columns of the two matrices. Since, in the basis $\widetilde{\mathbf{P}}$, we include interactions of $\mathbf{P}$ with $\mathbf{Z}_1$, but not with $\mathbf{Z}_2$, the resulting estimator allows for a nonlinear, nonseparable dependence of Y on X and Z_1, but imposes additive separability in Z_2. The conditional expectation of Y given W, Z_1, and Z_2 does not have to be additively separable in Z_2, so that, in the basis $\widetilde{\mathbf{Q}}$, we include interactions of $\mathbf{Q}$ with both $\mathbf{Z}_1$ and $\mathbf{Z}_2$.[9]

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

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

We estimated the demand functions for many different combinations of the order of B-spline for W, the number of knots in both bases, and even with various penalization terms (as discussed in Remark 10). While the shape of the unconstrained NPIV estimate varied slightly across these different choices of tuning parameters (mostly near the boundary of the support of X), the constrained NPIV estimator did not exhibit any visible changes at all.

Figure 4 shows a nonparametric kernel estimate of the conditional distribution of the price X given the instrument W. Overall the graph indicates an increasing relationship between the two variables as required by our stochastic dominance condition (6). We formally test this monotone IV assumption by applying our new test proposed in Section D. We find a test statistic value of 0.139 and 95%-critical value of 1.720.[10] Therefore, we fail to reject the monotone IV assumption.

Figure 5 shows the estimates of the demand function at three income levels, at the lower quartile ($\$42,500$), the median ($\$57,500$), and the upper quartile ($\$72,500$). The area shaded in grey represents the 90% uniform confidence bands around the unconstrained NPIV estimator as proposed in Horowitz and Lee (2012).[11] The black lines correspond to the estimators assuming exogeneity of price and the red lines to the NPIV estimators that allow for endogeneity of price. The dashed black line shows the kernel estimate of Blundell, Horowitz, and Parey (2012) and the solid black line the corresponding series estimator that imposes monotonicity. The dashed and solid red lines similarly depict the unconstrained and constrained NPIV estimators, respectively.

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

The two constrained estimates are very similar, indicating that endogeneity of prices may not be important in this problem, but they are both significantly flatter than the

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

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

unconstrained estimates across all three income groups, which implies that households appear to be less sensitive to price changes than the unconstrained estimates suggest. The small maximum slope of the constrained NPIV estimator also suggests that the error bound in Theorem 2 may be small and therefore we expect the constrained NPIV estimate to be precise for this data set.

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 $\|T\|_2 \leq \int_0^1 \int_0^1 f^2_{X,W}(x,w)dxdw$, and so under Assumption 2, $\|T\|_2 \leq C_T$.

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

$$\|h\|_{2,t} \leq C_1 \|h\|_{1,t} \tag{25}$$

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

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

so that (25) 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 (25), the second from Lemma 2 below (which is the main step in the proof of the theorem), and the third by Jensen's inequality. Hence, conclusion (10) of Theorem 1 holds for increasing continuously differentiable $h \in \mathcal{M}$ with $\bar{C} := C_1 C_2$ and C_2 as defined in Lemma 2.

Next, for any increasing function $h \in \mathcal{M}$, it follows from Lemma 9 that one can find a sequence of increasing continuously differentiable functions $h_k \in \mathcal{M}$, $k \geq 1$, such that $\|h_k - h\|_2 \to 0$ as $k \to \infty$. Therefore, by the triangle inequality,

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

where the third line follows from the Cauchy-Schwarz inequality, the fourth from $\|h_k - h\|_{2,t} \leq \|h_k - h\|_2$, and the fifth from Assumption 2(i). Taking the limit as $k \to \infty$ of both the left-hand and the right-hand sides of this chain of inequalities yields conclusion (10) 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 (10) of Theorem 1 also holds for all decreasing $h \in \mathcal{M}$, and thus for all $h \in \mathcal{M}$. This completes the proof of the theorem. Q.E.D.

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

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

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

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

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

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

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

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

so that condition (6) of Assumption 1 and $Dh(x) \geq 0$ for all x imply that the function $M(w)$ is increasing.

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

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

by Assumption 2(ii). Similarly,

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

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

$$\begin{aligned}
M(w) &= \int_0^{x^*} h(x) f_{X|W}(x|w) dx + \int_{x^*}^1 h(x) f_{X|W}(x|w) dx \\
&= -\int_0^{x^*} Dh(x) F_{X|W}(x|w) dx + \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w)) dx. \qquad (26)
\end{aligned}$$

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

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

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

$$\begin{aligned} A_1 + B_2 &= \int_{x^*}^1 Dh(x)(1 - F_{X|W}(x|w_1))dx + \int_0^{x^*} Dh(x)F_{X|W}(x|w_2)dx \\ &= \int_{x^*}^1 h(x)f_{X|W}(x|w_1)dx - \int_0^{x^*} h(x)f_{X|W}(x|w_2)dx \\ &\geq \int_{x^*}^{x_2} h(x)f_{X|W}(x|w_1)dx - \int_{x_1}^{x^*} h(x)f_{X|W}(x|w_2)dx \\ &\geq c_1 \int_{x^*}^{x_2} h(x)dx + c_f \int_{x_1}^{x^*} |h(x)|dx = c_f \int_{x_1}^{x_2} |h(x)|dx \\ &= c_f \|h\|_{1,t} = c_f > 0 \end{aligned} \tag{27}$$

where the fourth line follows from Assumption 2(ii). Second, by (6) and (7) of Assumption 1,

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

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

$$M(w_2) - M(w_1) \geq (C_F - 1)B_2. \tag{28}$$

Similarly, by (6) and (8) of Assumption 1,

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

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

$$M(w_2) - M(w_1) \geq (C_F - 1)A_1. \tag{29}$$

In conclusion, equations (27), (28), and (29) yield

$$M(w_2) - M(w_1) \geq (C_F - 1)(A_1 + B_2)/2 \geq (C_F - 1)c_f/2 > 0. \tag{30}$$

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

$$\int_0^1 |M(w)|dw \geq \int_{w_2}^1 |M(w)|dw \geq (1 - w_2)M(w_2) \geq (1 - w_2)(C_F - 1)c_f/2 > 0. \tag{31}$$

Similarly,

$$\int_0^1 |M(w)|dw \geq w_1(C_F - 1)c_f/2 > 0 \tag{32}$$

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 (31) shows that

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

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

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

This completes the proof of Case I.

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

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

With this inequality we proceed as in Case I to show that $\int_0^1 |M(w)|dw$ is bounded from below by a positive constant that depends only on ζ. On the other hand, when $M(w_1) \leq -c_f/2$ we bound $\int_0^1 |M(w)|dw$ as in (32), and the proof of Case II is complete.

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

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

and we proceed as in Case I to bound $\int_0^1 |M(w)|dw$ from below by a positive constant that depends only on ζ. On the other hand, when $M(w_2) > c_f/2$, we bound $\int_0^1 |M(w)|dw$ as in (31), and the proof of Case III is complete. The lemma is proven. Q.E.D.

Proof of Corollary 1. Note that since $\tau(a') \le \tau(a'')$ whenever $a' \le a''$, the claim for $a \le 0$, follows from $\tau(a) \le \tau(0) \le \bar{C}$, where the second inequality holds by Theorem 1. Therefore, assume that $a > 0$. Fix any $\alpha \in (0, 1)$. Take any function $h \in \mathcal{H}(a)$ such that $\|h\|_{2,t} = 1$. Set $h'(x) = ax$ for all $x \in [0, 1]$. Note that the function $x \mapsto h(x) + ax$ is increasing and so belongs to the class $\mathcal{M}$. Also, $\|h'\|_{2,t} \le \|h'\|_2 \le a/\sqrt{3}$. Thus, the bound (33) in Lemma 3 below applies whenever $(1 + \bar{C}\|T\|_2)a/\sqrt{3} \le \alpha$. Therefore, for all a satisfying the inequality

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

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

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

$$\|h\|_{2,t} \le \frac{\bar{C}}{1 - \alpha}\|Th\|_2 \tag{33}$$

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

Proof. Define

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

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

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

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

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

Hence, applying the triangle inequality once again yields

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

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

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

which is equivalent to (33). Q.E.D.

B Proofs for Section 3

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

Proof of Lemma 1. Observe that if $D\widehat{g}^u(x) \geq 0$ for all $x \in [0,1]$, then $\widehat{g}^c$ coincides with $\widehat{g}^u$, so that to prove (18), it suffices to show that

$$\mathrm{P}\Big(D\widehat{g}^u(x) \geq 0 \text{ for all } x \in [0,1]\Big) \to 1 \text{ as } n \to \infty. \tag{34}$$

In turn, (34) follows if

$$\sup_{x\in[0,1]} |D\widehat{g}^u(x) - Dg(x)| = o_p(1) \tag{35}$$

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

To prove (35), define a function $\widehat{m} \in L^2[0,1]$ by

$$\widehat{m}(w) = q(w)'\mathrm{E}_n[q(W_i)Y_i], \qquad w \in [0,1], \tag{36}$$

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

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

Throughout the proof, we assume that the events

$$\|\mathrm{E}_n[q(W_i)p(X_i)'] - \mathrm{E}[q(W)p(X)']\| \leq C(\xi_n^2 \log n/n)^{1/2}, \tag{37}$$
$$\|\mathrm{E}_n[q(W_i)q(W_i)'] - \mathrm{E}[q(W)q(W)']\| \leq C(\xi_n^2 \log n/n)^{1/2}, \tag{38}$$
$$\|\mathrm{E}_n[q(W_i)g_n(X_i)] - \mathrm{E}[q(W)g_n(X)]\| \leq C(J/(\alpha n))^{1/2}, \tag{39}$$
$$\|\widehat{m} - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1} J^{-s}) \tag{40}$$

hold for some sufficiently large constant $0 < C < \infty$. It follows from Markov's inequality and Lemmas 4 and 10 that all four events hold jointly with probability at least $1 - \alpha - n^{-1}$ since the constant C is large enough.

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

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

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

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

by the triangle inequality. The bound on $\|\widehat{m} - m\|_2$ is given in (40). Also, since $\|\widehat{g}^u\|_2 \leq C_b$ by construction,

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

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

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

Moreover,

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

by (37), and

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

by (39).

Further, by Assumption 2(iii), all eigenvalues of $\mathrm{E}[q(W)q(W)']$ are bounded from below by c_w and from above by C_w, and so it follows from (38) that for large n, all eigenvalues

of $Q_n := \mathrm{E}_n[q(W_i)q(W_i)']$ are bounded below from zero and from above. Therefore,

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

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

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

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

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

Combining the inequalities above shows that the inequality

$$\|\widehat{g}^u - g_n\|_2 \leq C\Big(\tau_n(J/(\alpha n))^{1/2} + K^{-s} + \tau_n(\xi_n^2 \log n/n)^{1/2}\|\widehat{g} - g_n\|_2\Big) \tag{41}$$

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

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

and so by the triangle inequality,

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

uniformly over $x \in [0,1]$ since $J \leq C_J K$ by Assumption 5. Since by the conditions of the lemma, $\sup_{x\in[0,1]} \|Dp(x)\|(\tau_n(K/n)^{1/2} + K^{-s}) \to 0$, (35) follows by taking $\alpha = \alpha_n \to 0$ slowly enough. This completes the proof of the lemma. Q.E.D.

Proof of Theorem 2. Consider the event that inequalities (37)-(40) hold for some sufficiently large constant C. As in the proof of Lemma 1, this events occurs with probability at least $1 - \alpha - n^{-1}$. Also, applying the same arguments as those in the proof of Lemma 1 with $\widehat{g}^c$ replacing $\widehat{g}^u$ and using the bound

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

instead of the bound for $\|(T_n - \widehat{T})\widehat{g}^u\|_2$ used in the proof of Lemma 1, it follows that on this event,

$$\|T(\widehat{g}^c - g_n)\|_2 \leq C\Big((K/(\alpha n))^{1/2} + (\xi_n^2 \log n/n)^{1/2} + \tau_n^{-1}K^{-s}\Big). \tag{42}$$

Further,

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

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

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

Combining these inequalities gives the asserted claim (19).

To prove (20), observe that combining (42) and Assumption 6(iii) and applying the triangle inequality shows that with probability at least $1 - \alpha - n^{-1}$,

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

which, by the same argument as that used to prove (19), gives

$$\|\widehat{g}^c - g\|_{2,t} \leq C\Big\{\delta + \tau\Big(\frac{\|Dg\|_\infty}{\delta}\Big)\Big(\frac{K}{\alpha n} + \frac{\xi_n^2 \log n}{n}\Big)^{1/2} + K^{-s}\Big\}. \tag{43}$$

The asserted claim (20) now follows by applying (19) with $\delta = 0$ and (43) with $\delta = \|Dg\|_\infty/c_\tau$ and using Corollary 1 to bound $\tau(c_\tau)$. This completes the proof of the theorem. Q.E.D.

Lemma 4. *Suppose that Assumptions 2, 4, and 7 hold. Then* $\|\widehat{m} - m\|_2 \leq C((J/(\alpha n))^{1/2} + \tau_n^{-1}J^{-s})$ *with probability at least* $1 - \alpha$ *where* $\widehat{m}$ *is defined in* (36).

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

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

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

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

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

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

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

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

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

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

Hence, by the triangle inequality,

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

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

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

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

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

C Proofs for Section 4

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

Proof of Theorem 3. Assume that g is increasing but not constant, that is, $g \in \mathcal{M}_\uparrow$. Define $M(w) := \mathrm{E}[Y|W = w]$, $w \in [0,1]$. Below we show that $M \in \mathcal{M}_\uparrow$. To prove it, observe that, as in the proof of Lemma 2, integration by parts gives

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

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

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

Since g is not constant and is continuously differentiable, there exists $\bar{x} \in (0,1)$ such that $Dg(\bar{x}) > 0$. Also, since $0 \leq x_1 < x_2 \leq 1$ (the constants x_1 and x_2 appear in Assumption 1), we have $\bar{x} \in (0, x_2)$ or $\bar{x} \in (x_1, 1)$. In the first case,

$$M(w_2) - M(w_1) \geq \int_0^{x_2} (C_F - 1)Dg(x)F_{X|W}(x|w_2)dx > 0.$$

In the second case,

$$M(w_2) - M(w_1) \geq \int_{x_1}^1 (C_F - 1)Dg(x)(1 - F_{X|W}(x|w_1))dx > 0.$$

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

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

D Testing the Monotonicity Assumptions

In this section, we propose tests of our two monotonicity assumptions based on an i.i.d. sample (X_i, W_i), $i = 1, \ldots, n$, from the distribution of (X, W). First, we discuss an adaptive procedure for testing the stochastic dominance condition (6) in our monotone IV Assumption 1. The null and alternative hypotheses are

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

respectively. 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), to the best of our knowledge there does not exist any procedure that adapts to the unknown smoothness level of $F_{X|W}(x|w)$. We provide a test that is adaptive in this sense, a feature that is not only 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 the selection of an appropriate value of h_n in a given data set to the researcher (see, for example, Lee, Linton, and Whang (2009) and Lee, Song, and Whang (2014)). 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. Our construction 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 on high dimensional distributional approximations in Chernozhukov, Chetverikov, and Kato (2013c) (CCK). Note that $F_{X|W}(x|w) = \mathrm{E}[1\{X \leq x\}|W = w]$, so that for a fixed $x \in (0,1)$, the hypothesis that $F_{X|W}(x|w') \geq F_{X|W}(x,w'')$ for all $0 \leq w' \leq w'' \leq 1$ is equivalent to the hypothesis that the regression function $w \mapsto \mathrm{E}[1\{X \leq x\}|W = w]$ is decreasing. An adaptive test of this hypothesis was developed in Chetverikov (2012). In our case, H_0 requires the regression function $w \mapsto \mathrm{E}[1\{X \leq x\}|W = w]$ to be decreasing not only for a particular value $x \in (0,1)$ but for all $x \in (0,1)$, and so we need to extend the results obtained in Chetverikov (2012).

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

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

We assume that the kernel function $K(w)$ has bounded support, is continuous, and is strictly positive on the support. The last condition excludes higher-order kernels. For a bandwidth value $h > 0$, define

$$K_h(w) := h^{-1}K(w/h), \quad w \in \mathbb{R}.$$

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

$$\mathrm{E}\left[(1\{X_i \leq x\} - 1\{X_j \leq x\})\mathrm{sign}(W_i - W_j)K_h(W_i - w)K_h(W_j - w)\right] \leq 0 \qquad (44)$$

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

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

taking the sum of the left-hand side in (44) over $i, j = 1, \ldots, n$, and rearranging give

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

or, equivalently,

$$\mathrm{E}\left[\sum_{i=1}^{n} k_{i,h}(w) 1\{X_i \leq x\}\right] \leq 0, \tag{45}$$

where

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

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

Assumption 10 (Bandwidth values)**.** *The collection of bandwidth values is* $\mathcal{B}_n := \{h \in \mathbb{R} : h = u^l/2, l = 0, 1, 2, \ldots, h \geq h_{\min}\}$ *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 $\mathcal{B}_n$ 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 $\mathcal{B}_n$ expands.

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

$$T := \max_{(x,w,h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{B}_n} \frac{\sum_{i=1}^{n} k_{i,h}(w) 1\{X_i \leq x\}}{\left(\sum_{i=1}^{n} k_{i,h}(w)^2\right)^{1/2}}. \tag{46}$$

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 $h \in \mathcal{B}_n$ to achieve adaptiveness of the test.

We now discuss the construction of a critical value for the test. Suppose that we would like to have a test of level (approximately) α. As succinctly demonstrated by Lee, Linton, and Whang (2009), the derivation of the asymptotic distribution of T is complicated even when $\mathcal{B}_n$ is a singleton. Moreover, when $\mathcal{B}_n$ 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 non-asymptotic approach developed in CCK and using a multiplier bootstrap critical value for the test. Let $e_1, \ldots, e_n$ be an i.i.d. sequence of $N(0, 1)$ random variables that are independent of the data. Also, let $\widehat{F}_{X|W}(x|w)$ be an estimator of $F_{X|W}(x|w)$ satisfying the following conditions:

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

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

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

This is a mild assumption implying uniform consistency of an estimator $\widehat{F}_{X|W}(x|w)$ of $F_{X|W}(x|w)$ over $(x,w) \in \mathcal{X}_n \times \mathcal{W}_n$. Define a bootstrap test statistic by

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

Then we define the critical value[12] $c(\alpha)$ for the test as

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

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

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

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

Theorem 5 (Polynomial Size Control)**.** *Let Assumptions 2, 9, 10, 11, and 12 be satisfied. If H_0 holds, then*

$$\mathrm{P}\left(T > c(\alpha)\right) \leq \alpha + Cn^{-c}. \tag{47}$$

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

$$\left|\mathrm{P}\left(T > c(\alpha)\right) - \alpha\right| \leq Cn^{-c}. \tag{48}$$

In both (47) and (48), 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.

[12] 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.

Remark 13 (Weak Condition on the Bandwidth Values)**.** Our theorem requires

$$\frac{1}{nh} \leq C_h n^{-c_h} \tag{49}$$

for all $h \in \mathcal{B}_n$, which is considerably weaker than the analogous condition in Lee, Linton, and Whang (2009) who require $1/(nh^3) \to 0$, up-to logs. This is achieved by using a conditional test and by applying the results of CCK. As follows from the proof of the theorem, the multiplier bootstrap distribution approximates the conditional distribution of the test statistic given $\mathcal{W}_n = \{W_1, \ldots, W_n\}$. Conditional on $\mathcal{W}_n$, the denominator in the definition of T is fixed, and does not require any approximation. Instead, we could try to approximate the denominator of T by its probability limit. This is done in Ghosal, Sen, and Vaart (2000) using the theory of Hoeffding projections but they require the condition $1/nh^2 \to 0$. Our weak condition (49) 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 14 (Polynomial Size Control)**.** Note that, by (47) and (48), the probability of rejecting H_0 when H_0 is satisfied can exceed the nominal level α only by a term that is polynomially small in n. We refer to this phenomenon as a *polynomial size control*. As explained in Lee, Linton, and Whang (2009), when $\mathcal{B}_n$ is a singleton, convergence of T to the limit distribution is logarithmically slow. Therefore, Lee, Linton, and Whang (2009) used higher-order corrections derived in Piterbarg (1996) to obtain polynomial size control. Here we show that the multiplier bootstrap also gives higher-order corrections and leads to polynomial size control. This feature of our theorem is also inherited from the results of CCK. □

Remark 15 (Uniformity)**.** The constants c and C in (47) and (48) depend on the data generating process only via constants (and the kernel) appearing in Assumptions 2, 9, 10, 11, and 12. Therefore, inequalities (47) and (48) hold uniformly over all data generating processes satisfying these assumptions with the same constants. We obtain uniformity directly from employing the distributional approximation theorems of CCK because they are non-asymptotic and do not rely on convergence arguments. □

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

Theorem 6 (Consistency). *Let Assumptions 2, 9, 10, 11, and 12 be satisfied and assume that $F_{X|W}(x|w)$ is continuously differentiable. If H_a holds with $D_w F_{X|W}(x|w) > 0$ for some $x \in (\epsilon, 1-\epsilon)$ and $w \in (0,1)$, then*

$$\mathrm{P}\left(T > c(\alpha)\right) \to 1 \text{ as } n \to \infty. \tag{50}$$

This theorem shows that our test is consistent against any model in H_a (with smooth $F_{X|W}(x|w)$) whose deviation from H_0 is not on the boundary, so that the deviation $D_w F_{X|W}(x|w) > 0$ occurs for $x \in (\epsilon, 1-\epsilon)$. It is also possible to extend our results to show that Theorems 5 and 6 hold with $\epsilon = 0$ at the expense of additional technicalities. Further, using the same arguments as those in Chetverikov (2012), it is possible to show that the test suggested here has minimax optimal rate of consistency against the alternatives belonging to certain Hölder classes for a reasonably large range of smoothness levels. We do not derive these results here for the sake of brevity of presentation.

We conclude this section by proposing a simple test of our second monotonicity assumption, that is, monotonicity of the regression function g. The null and alternative hypotheses are

$$\begin{aligned} H_0 &: g(x') \leq g(x'') \text{ for all } x', x'' \in (0,1) \text{ with } x' \leq x'' \\ H_a &: g(x') > g(x'') \text{ for some } x', x'' \in (0,1) \text{ with } x' \leq x'', \end{aligned}$$

respectively. The discussion in Remark 11 reveals that, under Assumptions 1 and 2, monotonicity of $g(x)$ implies monotonicity of $w \mapsto \mathrm{E}[Y|W=w]$. Therefore, under Assumptions 1 and 2, we can test H_0 by testing monotonicity of the conditional expectation $w \mapsto \mathrm{E}[Y|W=w]$ using existing tests such as Chetverikov (2012) and Lee, Song, and Whang (2014), among others. This procedure tests an implication of H_0 instead of H_0 itself and therefore may have low power against some alternatives. On the other hand, it does not require solving the model for $g(x)$ and therefore avoids the ill-posedness of the problem.

E Proofs for Section D

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

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

$$T \leq \max_{(x,w,h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{B}_n} \frac{\sum_{i=1}^n k_{i,h}(w)(1\{X_i \leq x\} - F_{X|W}(x|W_i))}{\left(\sum_{i=1}^n k_{i,h}(w)^2\right)^{1/2}} =: T_0 \tag{51}$$

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

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

where $p = |\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{B}_n|$, the number of elements in the set $\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{B}_n$, $x_{ij} = z_{ij}\varepsilon_{ij}$ with z_{ij} having the form $\sqrt{n}k_{i,h}(w)/(\sum_{i=1}^n k_{i,h}(w)^2)^{1/2}$, and ε_{ij} having the form $1\{X_i \le x\} - F_{X|W}(x|W_i)$ for some $(x,w,h) \in \mathcal{X}_n \times \mathcal{W}_n \times \mathcal{B}_n$. The dimension p satisfies $\log p \le C \log n$. Also, $n^{-1}\sum_{i=1}^n z_{ij}^2 = 1$. Further, since $0 \le 1\{X_i \le x\} \le 1$, we have $|\varepsilon_{ij}| \le 1$, and so $\mathrm{E}[\exp(|\varepsilon_{ij}|/2)|\mathcal{W}_n] \le 2$. In addition, $\mathrm{E}[\varepsilon_{ij}^2|\mathcal{W}_n] \ge c_\epsilon(1 - C_\epsilon) > 0$ by Assumption 12. Thus, T_0 satisfies the conditions of Case (E.3) in CCK with a sequence of constants B_n as long as $|z_{ij}| \le B_n$ for all $j = 1, \ldots, p$. In turn, Proposition B.2 in Chetverikov (2012) shows that under Assumptions 2, 9, and 10, with probability at least $1 - Cn^{-c}$, $z_{ij} \le C/\sqrt{h_{\min}} =: B_n$ uniformly over all $j = 1, \ldots, p$ (Proposition B.2 in Chetverikov (2012) is stated with "w.p.a.1" replacing "$1 - Cn^{-c}$"; however, inspecting the proof of Proposition B.2 (and supporting Lemma H.1) shows that the result applies with "$1 - Cn^{-c}$" instead of "w.p.a.1"). Let $\mathcal{B}_{1,n}$ denote the event that $|z_{ij}| \le C/\sqrt{h_{\min}} = B_n$ for all $j = 1, \ldots, p$. As we just established, $\mathrm{P}(\mathcal{B}_{1,n}) \ge 1 - Cn^{-c}$. Since $(\log n)^7/(nh_{\min}) \le C_h n^{-c_h}$ by Assumption 10, we have that $B_n^2(\log n)^7/n \le Cn^{-c}$, and so condition (i) of Corollary 3.1 in CCK is satisfied on the event $\mathcal{B}_{1,n}$.

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

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

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

$$\mathrm{P}\left(\mathrm{P}_e(|T^b - T_0^b| > \zeta_{1,n}) > \zeta_{2,n} | \mathcal{W}_n \right) < \zeta_{2,n} \tag{52}$$

where

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

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

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

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

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

Hence, on the event that

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

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

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

implying that (52) holds for some $\zeta_{1,n}$ and $\zeta_{2,n}$ satisfying $\zeta_{1,n} + \zeta_{2,n} \leq Cn^{-c}$.

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

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

Since $\mathrm{P}(\mathcal{B}_{1,n} \cap \mathcal{B}_{2,n}) \geq 1 - Cn^{-c}$, integrating this inequality over the distribution of $\mathcal{W}_n = \{W_1, \ldots, W_n\}$ gives (48). Combining this inequality with (51) gives (47). This completes the proof of the theorem. Q.E.D.

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

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

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

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

by Assumption 11. Therefore, $c(\alpha) \leq C(\log n)^{1/2}$ for some constant $C > 0$ since $c(\alpha)$ is the $(1-\alpha)$ conditional quantile of T^b given the data, $T^b = \max_{(x,w,h)\in\mathcal{X}_n\times\mathcal{W}_n\times\mathcal{B}_n} T^b(x, w, h)$, and $p := |\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{B}_n|$, the number of elements of the set $\mathcal{X}_n \times \mathcal{W}_n \times \mathcal{B}_n$, satisfies $\log p \leq C \log n$ (with a possibly different constant $C > 0$). Thus, the growth rate of

the critical value $c(\alpha)$ satisfies the same upper bound $(\log n)^{1/2}$ as if we were testing monotonicity of one particular regression function $w \mapsto \mathrm{E}[1\{X \leq x_0\}|W = w]$ with $\mathcal{X}_n$ replaced by x_0 for some $x_0 \in (0,1)$ in the definition of T and T^b. Hence, the asserted claim follows from the same arguments as those given in the proof of Theorem 4.2 in Chetverikov (2012). This completes the proof of the theorem. Q.E.D.

F Technical tools

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

Lemma 5. *Let W be a random variable with the density function bounded below from zero on its support $[0,1]$, and let $M : [0,1] \to \mathbb{R}$ be a monotone function. If M is constant, then $cov(W, M(W)) = 0$. If M is increasing in the sense that there exist $0 < w_1 < w_2 < 1$ such that $M(w_1) < M(w_2)$, then $cov(W, M(W)) > 0$.*

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

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

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

Lemma 6. *For any orthonormal basis $\{h_j, j \geq 1\}$ in $L^2[0,1]$, any $0 \leq x_1 < x_2 \leq 1$, and any $\alpha > 0$,*

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

for infinitely many j.

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

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

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

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

Therefore, by the Cauchy-Schwarz inequality,

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

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

Lemma 7. *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 (6) of Assumption 1 follows. Further, to prove condition (7) of Assumption 1, it suffices to show that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} \leq c_F \tag{53}$$

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

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

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

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

for all $x \in (0, x_2)$. To show (53), observe that

$$\frac{\partial \log F_{X|W}(x|w)}{\partial w} = -\frac{\rho}{\sqrt{1 - \rho^2}} \frac{\phi(y)}{\Phi(y)} \frac{1}{\phi(\Phi^{-1}(w))} \leq -\frac{\sqrt{2\pi}\rho}{\sqrt{1 - \rho^2}} \frac{\phi(y)}{\Phi(y)} \tag{54}$$

where $y := (\Phi^{-1}(x) - \rho \Phi^{-1}(w))/(1 - \rho^2)^{1/2}$. Thus, (53) 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 (54) 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 (7) of Assumption 1 is satisfied. Similar argument also shows that condition (8) of Assumption 1 is satisfied as well.

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

$$f_{X|W}(x|w) := \frac{1}{\sqrt{1-\rho^2}}\phi\left(\frac{\Phi^{-1}(x) - \rho\Phi^{-1}(w)}{\sqrt{1-\rho^2}}\right)\frac{1}{\phi(\Phi^{-1}(x))}. \tag{55}$$

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 $w \in [0,1]$, (55) combined with the change of variables formula with $x = \Phi(\widetilde{x})$ and $w = \Phi(\widetilde{w})$ give

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

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

Lemma 8. *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_2w,

$$\begin{aligned}
f_{X|W}(x|w) &= \int_0^{1/2} f_{U_1}(x - u_2w)f_{U_2}(u_2)du_2 \\
&= 4\int_0^{1/2} 1\left\{0 \leq x - u_2w \leq \frac{1}{2}\right\}du_2
\end{aligned}$$

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

and, thus,

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

It is easy to check that $\partial F_{X|W}(x|w)/\partial w \leq 0$ for all $x, w \in [0,1]$ so that condition (6) of Assumption 1 is satisfied. To check conditions (7) and (8), we proceed as in Lemma 7 and show $\partial \log F_{X|W}(x|w)/\partial w < 0$ uniformly for all $x \in [\underline{x}_2, \overline{x}_1]$ and $w \in (\widetilde{w}_1, \widetilde{w}_2)$. First, notice that, as required by Assumption 2(iv), $[\underline{x}_k, \overline{x}_k] = [0, (1+\widetilde{w}_k)/2]$, $k = 1, 2$. For $0 \leq x < w/2$ and $w \in (\widetilde{w}_1, \widetilde{w}_2)$,

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

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

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

Therefore, (7) holds uniformly over $x \in (\underline{x}_2, 1/2)$ and (8) uniformly over $x \in (x_1, 1/2)$. Now, consider $1/2 \leq x < (1+\widetilde{w}_1)/2$ and $w \in (\widetilde{w}_1, \widetilde{w}_2)$. Notice that, on this interval, $\partial(F_{X|W}(x|\widetilde{w}_1)/F_{X|W}(x|\widetilde{w}_2))/\partial x \leq 0$ so that

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

where the last inequality uses $\widetilde{w}_1 > \widetilde{w}_2 - \sqrt{\widetilde{w}_2/2}$, and thus (7) holds also uniformly over $1/2 \leq x < x_2$. Similarly,

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

so that (8) also holds uniformly over $1/2 \leq x < \overline{x}_1$. Assumption 2(i) trivially holds. Parts (ii) and (iii) of Assumption 2 hold for any $0 < \widetilde{x}_1 < \widetilde{x}_2 \leq \overline{x}_1 \leq 1$ and $0 \leq w_1 < \widetilde{w}_1 < \widetilde{w}_2 < w_2 \leq 1$ with $[\underline{x}_k, \overline{x}_k] = [0, (1+\widetilde{w}_k)/2]$, $k = 1, 2$. Q.E.D.

Lemma 9. *For any increasing function $h \in L^2[0,1]$, one can find a sequence of increasing continuously differentiable functions $h_k \in L^2[0,1]$, $k \geq 1$, such that $\|h_k - h\|_2 \to 0$ as $k \to \infty$.*

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

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

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

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

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

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

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

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

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

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

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

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

that $0 \leq h_\sigma(x) \leq 1$ for all $x \in [0,1]$, we obtain

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

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

Lemma 10. *Let $(p_1', q_1')', \ldots, (p_n', q_n')'$ be a sequence of i.i.d. random vectors where p_i's are vectors in $\mathbb{R}^K$ and q_i's are vectors in $\mathbb{R}^J$. Assume that $\|p_1\| \leq \xi_n$, $\|q_1\| \leq \xi_n$, $\|\mathrm{E}[p_1 p_1']\| \leq C_p$, and $\|\mathrm{E}[q_1 q_1']\| \leq C_q$ where $\xi_n \geq 1$. Then for all $t \geq 0$,*

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

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

Remark 16. Closely related results have been used previously by Belloni, Chernozhukov, Chetverikov, and Kato (2014) and Chen and Christensen (2013).

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

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

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

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

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

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

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

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

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

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

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

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

Hence, by Corollary 6.2.1 in Tropp (2012),

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chetverikov, D. (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. (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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Scaillet, O. (2016): "On Ill-Posedness of Nonparametric Instrumental Variable Regression With Convexity Constraints," Discussion paper.

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*.

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.

	Constrained estimator				
	$K=2$	$K=3$	$K=4$	$K=5$	optimal K
$n=500$	9.54	13.44	26.32	38.97	9.54
$n=1,000$	7.41	6.19	23.17	34.04	6.19
$n=5,000$	5.92	1.20	10.42	14.33	1.20
$n=10,000$	5.75	0.80	8.40	9.96	0.80
$n=50,000$	5.59	0.29	2.96	5.89	0.29
$n=100,000$	5.58	0.26	1.47	4.35	0.26
$n=500,000$	5.56	0.05	0.31	3.11	0.05
	Unconstrained estimator				
	$K=2$	$K=3$	$K=4$	$K=5$	optimal K
$n=500$	9.54	75.08	540.11	1069.93	9.54
$n=1,000$	7.41	32.65	389.19	839.24	7.41
$n=5,000$	5.92	6.10	149.46	515.16	5.92
$n=10,000$	5.75	2.68	104.85	546.66	2.68
$n=50,000$	5.59	0.54	30.41	382.70	0.54
$n=100,000$	5.58	0.28	11.14	248.70	0.28
$n=500,000$	5.56	0.05	1.92	125.80	0.05
	Ratio				
	$K=2$	$K=3$	$K=4$	$K=5$	optimal K
$n=500$	1.00	0.18	0.05	0.04	1.00
$n=1,000$	1.00	0.19	0.06	0.04	0.84
$n=5,000$	1.00	0.20	0.07	0.03	0.20
$n=10,000$	1.00	0.30	0.08	0.02	0.30
$n=50,000$	1.00	0.54	0.10	0.02	0.54
$n=100,000$	1.00	0.94	0.13	0.02	0.94
$n=500,000$	1.00	1.00	0.16	0.02	1.00

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

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

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

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

Table 3: simulation results for the case $g(x)=x^2+0.2x$, $\rho=0.3$, and $\eta=0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of n and K. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of n and K. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over K. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function n and K. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

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

Table 4: simulation results for the case $g(x)=x^2+0.2x$, $\rho=0.5$, and $\eta=0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of n and K. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of n and K. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over K. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function n and K. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

	Constrained estimator				
	$K=2$	$K=3$	$K=4$	$K=5$	optimal K
$n=500$	10.20	14.84	16.88	26.33	10.20
$n=1,000$	8.25	6.84	13.05	21.40	6.84
$n=5,000$	6.77	1.86	7.32	10.78	1.86
$n=10,000$	6.56	1.51	4.13	7.89	1.51
$n=50,000$	6.39	0.96	1.61	4.98	0.96
$n=100,000$	6.37	0.90	1.36	4.66	0.90
$n=500,000$	6.36	0.86	0.66	2.43	0.66
	Unconstrained estimator				
	$K=2$	$K=3$	$K=4$	$K=5$	optimal K
$n=500$	10.20	75.53	540.11	1069.94	10.20
$n=1,000$	8.25	34.95	389.19	850.47	8.25
$n=5,000$	6.77	6.49	149.46	510.83	6.49
$n=10,000$	6.56	3.69	104.85	554.15	3.69
$n=50,000$	6.39	1.35	30.41	375.25	1.35
$n=100,000$	6.37	1.08	11.14	248.41	1.08
$n=500,000$	6.36	0.86	1.92	128.26	0.86
	Ratio				
	$K=2$	$K=3$	$K=4$	$K=5$	optimal K
$n=500$	1.00	0.20	0.03	0.02	1.00
$n=1,000$	1.00	0.20	0.03	0.03	0.83
$n=5,000$	1.00	0.29	0.05	0.02	0.29
$n=10,000$	1.00	0.41	0.04	0.01	0.41
$n=50,000$	1.00	0.71	0.05	0.01	0.71
$n=100,000$	1.00	0.83	0.12	0.02	0.83
$n=500,000$	1.00	1.00	0.34	0.02	0.77

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

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

Table 6: simulation results for the case $g(x) = 2(x-1/2)_+^2 + 0.5x$, $\rho = 0.5$, and $\eta = 0.3$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of n and K. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of n and K. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over K. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function n and K. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

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

Table 7: simulation results for the case $g(x) = 2(x-1/2)_+^2 + 0.5x$, $\rho = 0.3$, and $\eta = 0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of n and K. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of n and K. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over K. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function n and K. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

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

Table 8: simulation results for the case $g(x) = 2(x-1/2)_+^2 + 0.5x$, $\rho = 0.5$, and $\eta = 0.7$. The top panel shows the MISE of the constrained estimator $\widehat{g}^c$, multiplied by 1000, as a function of n and K. The middle panel shows the MISE of the unconstrained estimator $\widehat{g}^u$, multiplied by 1000, as a function of n and K. Both in the top and in the middle panels, the last column shows the minimal value of the MISE of the corresponding estimator optimized over K. The bottom panel shows the ratio of the MISE of the constrained estimator to the MISE of the unconstrained estimator as a function n and K. The last column of the bottom panel shows the ratio of the optimal value of the MISE of the constrained estimator to the optimal value of the MISE of the unconstrained estimator.

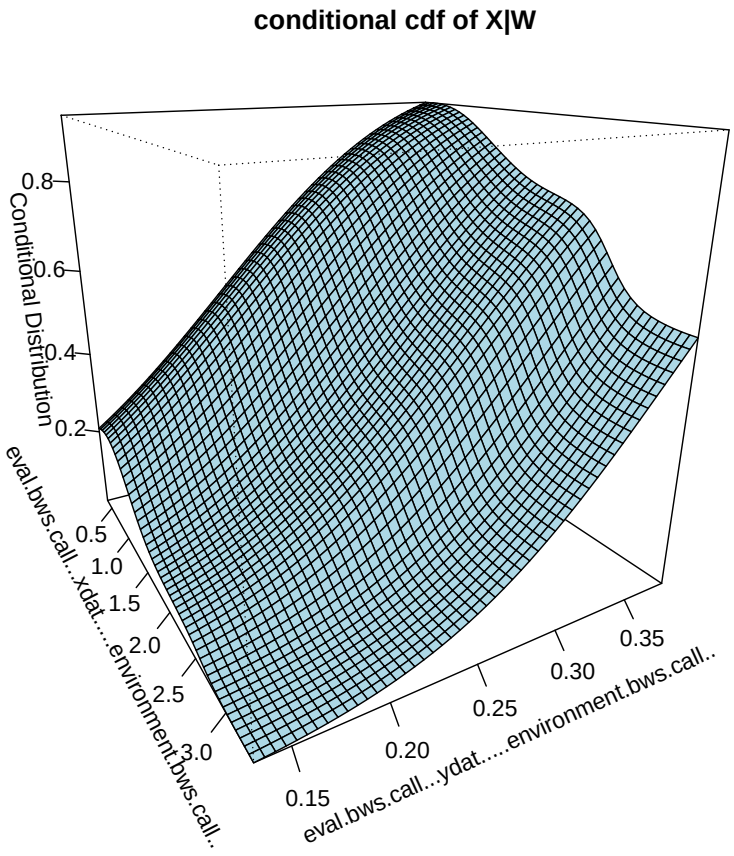

Figure 4: Nonparametric kernel estimate of the conditional cdf $F_{X|W}(x|w)$.

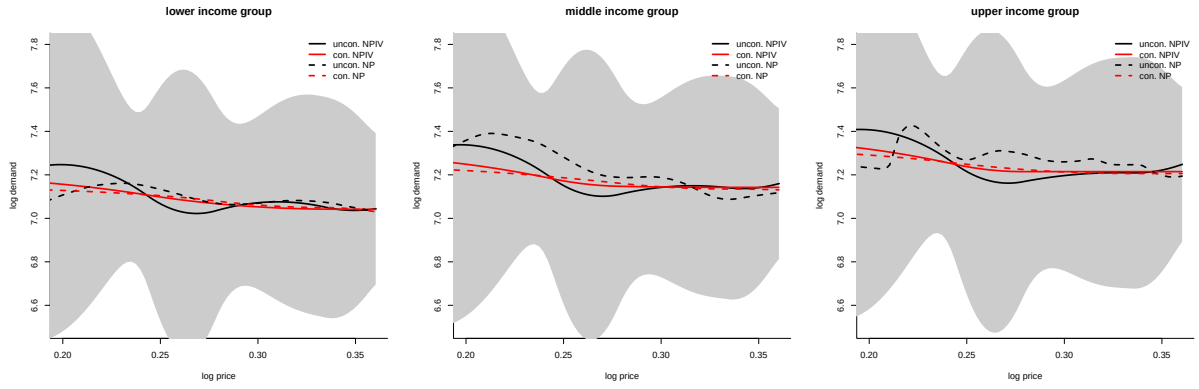

Figure 5: Estimates of $g(x, z_1)$ plotted as a function of price x for z_1 fixed at three income levels.