# Nonparametric Instrumental Variable Estimation Under Monotonicity*

Denis Chetverikov† Daniel Wilhelm‡

**Abstract**

The ill-posedness of the 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 function to be estimated is monotone and consider a sieve estimator that enforces this monotonicity constraint. We define a constrained measure of ill-posedness that is relevant for the constrained estimator and show that, under a monotone IV assumption and certain other mild regularity conditions, this measure is bounded uniformly over the dimension of the sieve space. This finding is in stark contrast to the well-known result that the unconstrained sieve measure of ill-posedness that is relevant for the unconstrained estimator 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 estimator. The bound gives a set of data-generating processes for which the monotonicity constraint has a particularly strong regularization effect and considerably improves the performance of the estimator. The form of the bound implies that the regularization effect can be strong even in large samples and even if the function to be estimated is steep, particularly so if the NPIV model is severely ill-posed. Our simulation study confirms these findings and reveals the potential for large performance gains from imposing the monotonicity constraint.

*First version: January 2014. This version: April 14, 2017. 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. We are also thankful for excellent research assistance by Dongwoo Kim who implemented the NPIV estimators in STATA (http://github.com/danielwilhelm/STATA-NPIV).

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

‡Department of Economics, University College London, Gower Street, London WC1E 6BT, United Kingdom; E-Mail address: d.wilhelm@ucl.ac.uk. The author gratefully acknowledges financial support from the ESRC Centre for Microdata Methods and Practice at IFS (RES-589-28-0001) and the European Research Council (ERC-2014-CoG-646917-ROMIA and ERC-2015-CoG-682349).

# 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 NPIV estimators. 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 explanatory variable, and $W$ an instrumental variable (IV). We are interested in the estimation of the nonparametric 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 explanatory variables. Departing from the existing literature on the estimation of the NPIV model, we assume that the function $g$ is 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 first stage 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 $g$ is increasing with probability approaching one, in which case the corresponding constrained estimator is 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 estimator often outperforms, sometimes substantially, the unconstrained one even if the sample size $n$ is rather large and the function $g$ is strictly increasing; see Figure 1 for an example. Hence, it follows that the rate result misses an important finite-sample phenomenon.

In this paper, we derive a novel non-asymptotic error bound for the constrained estimator that captures this finite-sample phenomenon. For each sample size $n$, the bound gives a set of data-generating processes for which the monotonicity constraint has a particularly strong regularization effect, thereby considerably improving the performance of the estimator. The

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

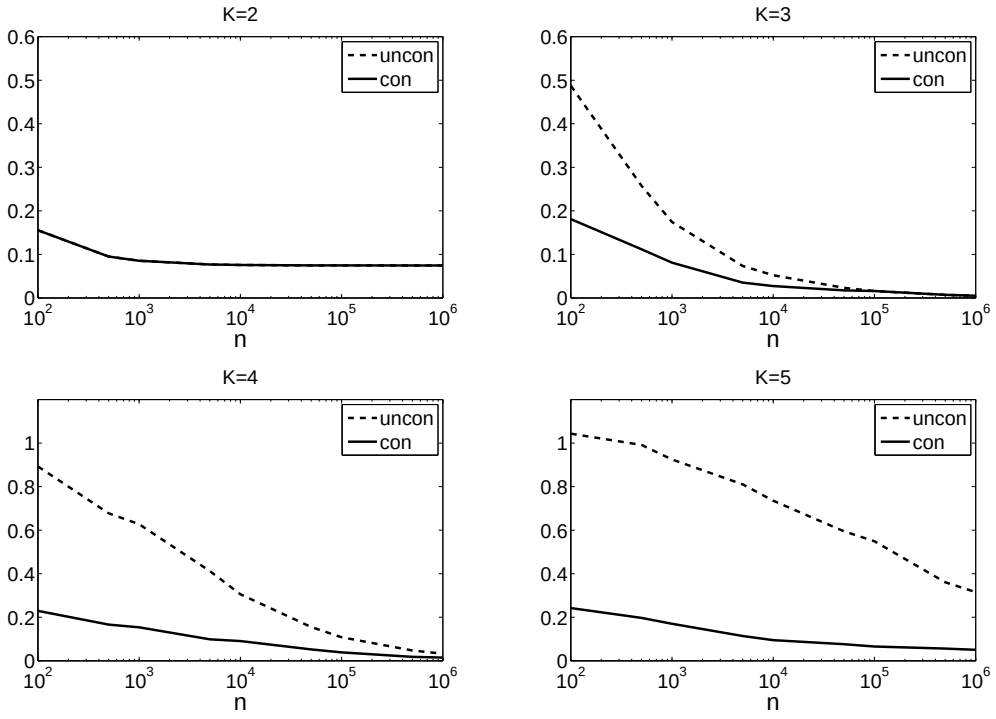


**Figure 1:** an example demonstrating performance gains from imposing the monotonicity constraint. In this example, $g(x) = x^2 + 0.2x$, $W = \Phi(\zeta)$, $X = \Phi(\rho\zeta + \sqrt{1-\rho^2}\epsilon)$, $\varepsilon = \sigma(\eta\epsilon + \sqrt{1-\eta^2}\nu)$, where $(\zeta, \epsilon, \nu)$ is a 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. The 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.

form of the bound implies that the regularization effect can be strong even in large samples and for steep functions $g$, particularly so if the NPIV model is severely ill-posed.

To establish our non-asymptotic error bound, we define a constrained sieve measure of ill-posedness that is relevant for the constrained estimator. We demonstrate that as long as the monotone IV assumption and certain other regularity conditions are satisfied, this measure is bounded uniformly over the dimension of the sieve space. This should be contrasted with the well-known result that the unconstrained sieve measure of ill-posedness that is relevant for the unconstrained estimator grows to infinity, potentially very fast, with the dimension of the sieve space; see Blundell, Chen, and Kristensen (2007).

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 certain $L^2$-norm defined below. $B_n$ on the right-hand side of this bound is a bias term that behaves similarly to that of the unconstrained NPIV estimator, 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 term of a nonparametric conditional mean estimator up to a log-term, i.e. of a well-posed problem, and $\tau_n$ is the unconstrained sieve measure of ill-posedness. Without the monotonicity constraint, the variance term would be $\tau_n V_n$, 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, $\tau_n$ 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 this scenario in which the monotonicity constraint has a strong regularization effect. If the NPIV model is severely ill-posed, $\tau_n$ 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 a slow, logarithmic convergence rate. However, when ill-posedness is severe, the switch to this regime may occur only at extremely large sample sizes.

Our simulation experiments confirm the theoretical findings and demonstrate possibly large finite-sample performance improvements of the constrained estimator relative to the unconstrained one 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.

Both of our monotonicity assumptions can be tested in the data. Perhaps more importantly though, we regard both assumptions as natural in many economic applications. In fact, both of these conditions often directly follow from economic theory. To see this 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 $\varepsilon$ summarizes determinants of the 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 $\eta$ 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, the function $X(W, \eta)$ satisfies

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

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 $(\varepsilon, \eta)$, condition (4) implies our monotone IV assumption and $g$ increasing corresponds to our monotonicity assumption on the function of interest. Dependence between $\eta$ and $\varepsilon$ generates endogeneity of $X$, and independence of $W$ from $(\varepsilon, \eta)$ implies that $W$ can be used as an instrument for $X$. Other examples are the estimation of Engel and demand curves.

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 (2015) show that, in the absence of point-identification, shape restrictions may yield informative bounds on functionals of $g$ and develop inference procedures when the explanatory variable $X$ 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 $g$, particularly so 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 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 a 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.

**STATA code.** We provide STATA code implementing both the unconstrained and constrained sieve NPIV estimators at `http://github.com/danielwilhelm/STATA-NPIV`.

**Supplement.** The supplement to this paper consists of two parts. Sections A and B are available as "online supplement" and contain the proofs of the two main theorems. Sections C–I are part of an additional supplement within "data and programs" on the *Econometrica* website. These sections provide more detailed discussions of our results through examples, additional results, and simulations.

**Notation.** For a differentiable function $f : \mathbb{R} \to \mathbb{R}$, we use $Df(x)$ to denote its derivative. 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.

## 2 Boundedness of the Constrained Measure of Ill-posedness

In this section, we introduce a constrained measure of ill-posedness for the NPIV model (1), which is relevant for studying the behavior of the constrained estimator. We show that, unlike the standard, unconstrained measure of ill-posedness introduced by Blundell, Chen, and Kristensen (2007), our constrained measure is bounded, a result that plays a fundamental role in the derivation of our non-asymptotic error bound in Section 3.

In the NPIV model (1), the function $g$ solves the equation $\mathrm{E}[Y|W] = \mathrm{E}[g(X)|W]$. 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 $m(w) := \mathrm{E}[Y|W = w] f_W(w)$, we can express this equation as

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

Let $0 < x_1 < x_2 < 1$ be some constants and define the truncated $L^2$-norm $\|\cdot\|_{2,t}$ by $\|h\|_{2,t} := (\int_{x_1}^{x_2} h(x)^2 dx)^{1/2}$, $h \in L^2[0,1]$. For $a \in \mathbb{R}$, let

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

be the space containing all functions in $L^2[0,1]$ with lower derivative bounded from below by $-a$ uniformly over the interval $[0,1]$. Then, the *constrained measure of ill-posedness* is

$$\tau(a) := \sup_{\substack{h \in \mathcal{H}(a) \\ \|h\|_{2,t} = 1}} \frac{\|h\|_{2,t}}{\|Th\|_2}. \tag{6}$$

To study properties of $\tau(a)$, we impose two assumptions. Let $0 < \delta_1 < 1/2$, $0 \le \delta_2 \le \delta_1$, $\delta_2 < x_1 < x_2 < 1 - \delta_2$, and $0 < w_1 < w_2 < 1$ be some constants.

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

$$w' \le w'' \qquad \Rightarrow \qquad F_{X|W}(x|w') \ge F_{X|W}(x|w''). \tag{7}$$

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

$$F_{X|W}(x|w_1) \ge C_F F_{X|W}(x|w_2), \qquad \forall x \in (0, 1 - \delta_1) \tag{8}$$

*and*

$$C_F(1 - F_{X|W}(x|w_1)) \le 1 - F_{X|W}(x|w_2), \qquad \forall x \in (\delta_1, 1) \tag{9}$$

The first part of this assumption, condition (7), requires first-order stochastic dominance of the conditional distribution of the endogenous explanatory variable $X$ given the instrument $W$ as we increase the value of the instrument $W$. Standard tests for stochastic dominance (e.g. Lee,

Linton, and Whang, 2009) can be employed to test condition (7). If the first stage relationship can be written as $X = r(W, U)$ for some vector of unobservables $U$ that is independent of $W$, then condition (7) is equivalent to $r(w, u)$ being monotone in its first argument for every $u$. Finally, note that this condition 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 $\mathrm{E}[\varepsilon|W] = 0$.

Conditions (8) and (9) strengthen the stochastic dominance in (7) 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 (8) and (9) are rather weak as they require such a shift to occur only in some intervals $(0, 1 - \delta_1)$ and $(\delta_1, 1)$, respectively.

**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 $x \in [\delta_2, 1 - \delta_2]$ 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. Examples C.1 and C.2 in the supplement show that Assumptions 1 and 2 are satisfied when $(X, W)$ is a transformation of a bivariate normal random vector with positive correlation to $[0,1]^2$ or when the first-stage relationship between $X$ and $W$ has random coefficients.

Our first result gives a bound on $\tau(a)$.

**Theorem 1** (Bound for the constrained Measure of Ill-Posedness)**.** *Let Assumptions 1 and 2 be satisfied. Then there exist constants $c_\tau > 0$ and $0 < C_\tau < \infty$ such that, for all $a \leq c_\tau$,*

$$\tau(a) \leq C_\tau.$$

*Here, $c_\tau$ and $C_\tau$ depend only on the constants appearing in Assumptions 1, 2, and on $x_1$, $x_2$.*

This theorem shows that $\tau(a)$ must be finite when $a$ is not too large. It implies in particular that the constrained measure of ill-posedness is finite when defined over the set of increasing functions $h \in \mathcal{H}(0)$, i.e. $\tau(0) \leq C_\tau < \infty$. This result is important because, as we show in Section C.2 of the supplement, $\tau(\infty)$ is infinite for many ill-posed and, in particular, for all severely ill-posed problems. This suggests that imposing shape constraints, like monotonicity, may have a substantial regularization effect.

Even though Theorem 1 may seem 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 (Scaillet, 2016).

**Remark 1** (Reasons for norm truncation)**.** There are two reasons for using the truncated $L^2$-norm $\|\cdot\|_{2,t}$ in the numerator on the right-hand side of (6) instead of the usual $L^2$-norm $\|\cdot\|_2$.

First, the main argument in the proof of Theorem 1, Lemma A.2 in the supplement, shows that for any increasing continuously differentiable $h \in L^1[0,1]$, we have

$$\int_{\delta_2}^{1-\delta_2} |h(x)|dx \leq C\|Th\|_1, \tag{10}$$

where $C$ is finite if $c_f > 0$. For some distributions of $(X, W)$, like a transformation of a bivariate normal random vector with positive correlation to $[0,1]^2$, Assumption 2 holds with $c_f > 0$ only if $\delta_2 > 0$, which introduces the norm truncation on the left-hand side of (10). For many other distributions of $(X, W)$, however, Assumption 2 holds with $c_f > 0$ even if $\delta_2 = 0$. In this case, (10) becomes

$$\|h\|_1 \leq C\|Th\|_1, \tag{11}$$

and we can avoid the norm truncation in the $L^1$-norm bound. Second, the norm truncation is required to transform the $L^1$-norm bounds (10) and (11) into the desired $L^2$-norm bound. To see this, consider the case $\delta_2 = 0$. Since $\|Th\|_1 \leq \|Th\|_2$ and, for any increasing $h$,

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

the inequality (11) implies

$$\|h\|_{2,t} \leq \frac{C\sqrt{x_2 - x_1}}{\min\{x_1, 1 - x_2\}}\|Th\|_2; \tag{12}$$

see Lemma A.1 in the supplement for details. This explains the reasons for the norm truncation in our arguments and also how the bound (12) changes as we send $x_1$ to 0 and $x_2$ to 1. □

## 3 Non-asymptotic Risk Bounds

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. Since by Theorem 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 for which 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 $J = J_n \geq K_n$, let $p(x) := (p_1(x), \ldots, p_K(x))'$

and $q(w) := (q_1(w), \ldots, q_J(w))'$ be vectors of basis functions. Define $\mathbf{P} := (p(X_1), \ldots, p(X_n))'$, $\mathbf{Q} := (q(W_1), \ldots, q(W_n))'$, and $\mathbf{Y} := (Y_1, \ldots, Y_n)'$. 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 $\|g\|_2 < C_b$ 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

$$\widehat{\beta}^u := \text{argmin}_{b \in \mathbb{R}^K : \|b\| \leq C_b} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b), \tag{13}$$

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

$$\widehat{\beta}^c := \text{argmin}_{b \in \mathbb{R}^K : p(\cdot)'b \in \mathcal{H}_n(0), \|b\| \leq C_b} (\mathbf{Y} - \mathbf{P}b)'\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'(\mathbf{Y} - \mathbf{P}b), \tag{14}$$

which imposes the monotonicity of $g$ through the constraint $p(\cdot)'b \in \mathcal{H}_n(0)$. Define the *constrained* and *unconstrained sieve measures of ill-posedness* $\tau_{n,t}(a)$ and $\tau_n$ as

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

It is well-known that the unconstrained measure $\tau_n \to \infty$ as $n \to \infty$ and the rate at which this happens is related to the rate at which the singular values of $T$ converge to zero (e.g. Blundell, Chen, and Kristensen (2007), Horowitz (2012)). Since $\tau(a) \leq C_\tau$ for all $a \leq c_\tau$ by Theorem 1 and since by definition $\tau_{n,t}(a) \leq \tau(a)$, we also have $\tau_{n,t}(a) \leq C_\tau$ for all $a \leq c_\tau$. Thus, for all values of $a$ that are not too large, the constrained measure $\tau_{n,t}(a)$ remains bounded uniformly over all $n$, irrespectively of how fast the singular values of $T$ converge to zero.

Let $s > 0$ be some constant. Also, define the operator $T_n : L^2[0,1] \to L^2[0,1]$ by $(T_n h)(w) := q(w)'\text{E}[q(W)p(X)']\text{E}[p(U)h(U)]$ with $w \in [0,1]$ and $U \sim U[0,1]$. Finally, denote $\xi_{K,p} := \sup_{x \in [0,1]} \|p(x)\|$, $\xi_{J,q} := \sup_{w \in [0,1]} \|q(w)\|$, and $\xi_n := \max(\xi_{K,p}, \xi_{J,q})$.

**Assumption 3** (Monotonicity of $g$)**.** *The function $g$ is increasing.*

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

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

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

**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 $\|m - m_n\|_2 \leq C_m \tau_n^{-1} J^{-s}$ .*

**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 $h \in \mathcal{H}_n(\infty)$.*

Assumption 3 is one of our two monotonicity conditions. The other assumptions are standard in the NPIV literature. Assumption 4 is a mild moment condition. Assumption 5 is satisfied when the dimension of the vector $q(w)$ is not much larger than the dimension of the vector $p(x)$. The first part of Assumption 6 restricts the approximating function $g_n$ for $g$ 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 $s$. The third part of this condition is similar to Assumption 6 in Blundell, Chen, and Kristensen (2007). Assumption 7 is similar to Assumption 3(iii) in Horowitz (2012) and Assumption 8 is similar to Assumption 5 in Horowitz (2012). Therefore, comments made there also apply here.

Lemma B.1 in the supplement formalizes the intuitive result that the unconstrained and constrained estimators must possess the same convergence rate when the function of interest, $g$, is strictly increasing. Therefore, imposing the monotonicity constraint cannot improve the convergence rate of the NPIV estimator in that case. On the other hand, our simulations in Section 4 show significant finite sample performance gains from imposing the constraint, even in very large samples and for functions $g$ that are strictly increasing and relatively steep. The following theorem, the main result of this paper, explains these seemingly conflicting findings:

**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*

$$\|\widehat{g}^c - g\|_{2,t} \leq C\Big\{\delta + \tau_{n,t}\Big(\frac{\|Dg_n\|_\infty}{\delta}\Big)V_n + K^{-s}\Big\} \tag{15}$$

*and*

$$\|\widehat{g}^c - g\|_{2,t} \leq C \min\Big\{\|Dg\|_\infty + V_n, \tau_n V_n\Big\} + CK^{-s}, \tag{16}$$

*where $V_n := \sqrt{K/(\alpha n) + (\xi_n^2 \log n)/n}$. Here, $c$ and $C$ depend only on the constants appearing in Assumptions 1-8, and on $x_1$, $x_2$.*

To explain the main features of this theorem, it is important to notice that $C$ in the bounds (15) and (16) depends only on the constants appearing in Assumptions 1-8, and on $x_1$, $x_2$, and so these bounds hold uniformly over all data-generating processes that satisfy those assumptions with the same constants.[2] In particular, for any two data-generating processes in this set, the

[2]The dependence of $C$ on those constants can actually be traced from the proof of the theorem, but we omit these expressions here to save space.

same non-asymptotic bounds (15) and (16) hold with the same constant $C$, even though the unconstrained sieve measure of ill-posedness $\tau_n$ may be of different order of magnitude for these two data-generating processes.

The bound (15) holds for any $\delta \geq 0$, which means, in principle, we could minimize the right-hand side of the bound over $\delta$. However, we do not know the explicit form of the function $\tau_{n,t}(\cdot)$, which makes it impossible to obtain an explicit minimal value of the right-hand side of (15). On the other hand, we know from the discussion above that $\tau_{n,t}(a) \leq C_\tau$ for all $a \leq c_\tau$. Using this inequality, we obtain the bound (16).

The right-hand side of (16) 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 $\|Dg\|_\infty$ of $g$, the unconstrained sieve measure of ill-posedness $\tau_n$, 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).

The bound (16) 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 $1 - \alpha$ with which we want to bound the estimation error, the larger the bound.

Second, and more importantly, the variance term in the bound (16) is determined by the minimum of two regimes. For a given sample size $n$, the minimum is attained in the first regime if

$$\|Dg\|_\infty \leq (\tau_n - 1)V_n. \tag{17}$$

In this regime, the right-hand side of the bound (16) is independent of the (unconstrained) sieve measure of ill-posedness $\tau_n$, 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, $\tau_n$ 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 (17) decreases (if $K = K_n$ grows slowly enough) and eventually becomes smaller than the left-hand side, and the bound (16) switches to its second regime, in which it depends on the (unconstrained) sieve measure of ill-posedness $\tau_n$. 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 2 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, for moderate $n$, is of smaller order than B because of the multiplication by the possibly large factor $\tau_n$. As $n$ grows sufficiently large, i.e. larger than $n_0$, B becomes smaller than A. Therefore, the error

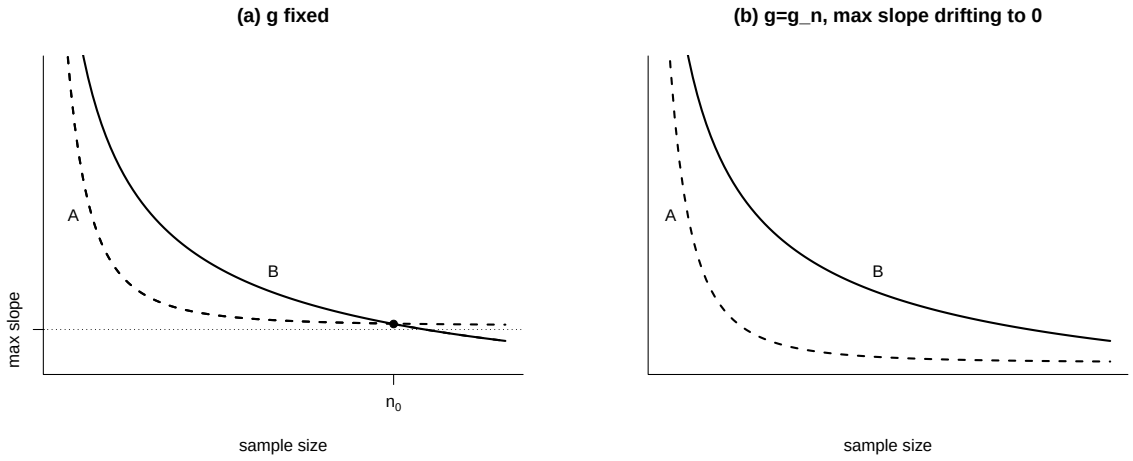


**Figure 2:** Stylized graphs showing the relationship of the two regimes determining the minimum of the estimation error bound in (15).

bound, which is the minimum of A and B, 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.

**Remark 2** (Truncated norm as performance measure)**.** The truncation of the $L^2$-norm on the left-hand side of the bounds (15) and (16) does not change the meaning of the bounds from a practical point of view because, in most applications, researchers are typically not interested in the values of the function $g$ arbitrarily close to the boundary but rather in the values of $g(x)$ for $x$ in the interior of the support of $X$. □

**Remark 3** (Robustness of the constrained estimator to tuning parameter choice)**.** 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 $\tau_n$, 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 (16) of Theorem 2, which is independent of $\tau_n$ for sufficiently large $K$. In this sense, the constrained estimator $\widehat{g}^c$ possesses some robustness against setting $K$ too high. This observation is confirmed by our simulations in Section 4 and in the supplement. □

**Remark 4** (Local-to-flat asymptotics)**.** An alternative explanation of the impact the monotonicity constraint has on the NPIV estimator comes from a "local-to-flat" asymptotic theory in which we consider a sequence of data-generating processes indexed by the sample size for which the maximum slope of $g$ drifts to zero. Corollary D.1 in the supplement, a straightforward implication of Theorem 2, shows that 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

$$\|\widehat{g}^c - g\|_{2,t} = O_p(n^{-s/(1+2s)}\sqrt{\log n}). \tag{18}$$

Therefore, in the shrinking neighborhood where $\|Dg\|_\infty = O(n^{-s/(1+2s)}\sqrt{\log n})$, the constrained estimator's convergence rate is the fast 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 2. In the shrinking neighborhood, the minimum in (16) is always attained in the first regime because A is always below B. Finally, we could consider letting the neighborhood for $\|Dg\|_\infty$ shrink at slower rates than $O(n^{-s/(1+2s)}\sqrt{\log n})$, thereby increasing the neighborhood, but would then obtain a convergence rate for the estimation error that lies between the fast one in (18) and the standard slow convergence rate for fixed $g$. □

# 4 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 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 $\rho$, $\eta$, and $\sigma$ are parameters and $\zeta$, $\epsilon$, and $\nu$ are independent $N(0,1)$ random variables. We set $\sigma = 0.5$ and $\rho = \eta = 0.3$. Simulations for other parameter choices yield similar findings and are reported in the supplement.

For both functions $g$ and both the constrained and unconstrained estimators, we use the same sieve spaces for $X$ and $W$, that is, $p(x) = q(x)$ for all $x \in [0,1]$. We vary the dimension of the sieve space, $K$, from 2 to 5 and choose the basis functions to be regression splines: $p(x) = (1, x)'$ 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 for models 1 and 2 are presented in Tables 1 and 2, respectively. 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 $K$. 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.[3]. The bottom panel reports the ratio of these two quantities. Both for

[3]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 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 $g(x) = x^2 + 0.2x$ (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 $g(x) = 2(x - 1/2)_+^2 + 0.5x$ (Table 2), 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 and 2). 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 $\eta = 0.3$ in Table 1 of the supplement, 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 $g(x) = 2(x-1/2)_+^2 + 0.5x$ (Table 2) with $n = 5,000$, 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 1.86 to 7.32 and the MISE of the unconstrained estimator grows from 6.49 to 149.46. This shows that the constrained estimator is more robust than the unconstrained one to incidental mistakes in the choice of $K$.

## 5 Concluding Remarks

In this paper, we develop a novel non-asymptotic bound on the estimation error of the constrained NPIV estimator that imposes the constraint that the function $g$ to be estimated is increasing. The bound is able to explain an empirical observation that the constrained estimator often substantially outperforms the unconstrained one even when the sample size is large and the function $g$ is strictly increasing and steep, which is difficult to explain because the

| | 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 $K$ |
| $n=500$ | 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 2:** 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 and unconstrained estimators are asymptotically equivalent as long as $g$ is strictly increasing.

In principle, assuming usual under-smoothing conditions, this bound can be used for constructing confidence bands for $g$ based on the constrained estimator. Indeed, from the proof of Theorem 2 in the supplement, we can trace back the dependence of the constant $C$ in the bound (16) on the constants appearing in the assumptions, and those constants can be estimated from the data. However, the resulting confidence bands would be rather wide because the bound (16) holds uniformly over a large class of data-generating processes, and the same constant $C$ applies to the whole class. Instead, confidence bands based on the constrained estimator can be constructed using the methods developed in Chernozhukov, Newey, and Santos (2015).

The main purpose of this paper is to investigate how the monotonicity constraints improve estimation of the point-identified NPIV model. However, point identification of the NPIV model requires completeness conditions, which are somewhat difficult to interpret, and so it is also of great interest to study how monotonicity or other shape restrictions help with identification of the NPIV model in the absence of these completeness conditions. Toward this goal, in Section E of the supplement, we show that the sign of the slope of $g$ is identified under our monotonicity conditions (when we assume that $g$ is monotone but do not specify whether it is increasing or decreasing). We also provide bounds on the identified set for $g$ that are implied by our monotonicity conditions, which complement the results of Freyberger and Horowitz (2015) for the case of discrete data. It would be interesting future work to tighten these bounds and to develop estimators for the resulting identified set.

In Section G of the supplement, we apply the constrained and unconstrained NPIV estimators to the estimation of gasoline demand in the U.S., allowing for endogeneity of prices. We find that imposing the monotonicity constraint on the estimator has a large impact by eliminating the implausible increasing parts of the unconstrained estimator. The constrained NPIV estimator is similar to the constrained conditional mean estimator that assumes exogeneity of prices, which confirms the findings of the exogeneity test in Blundell, Horowitz, and Parey (2012).

# References


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.

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., W. K. Newey, and A. Santos (2015): "Constrained Conditional Moment Restriction Models," Working Paper CWP 59/15, cemmap.

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.

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

Freyberger, J., and J. L. Horowitz (2015): "Identification and shape restrictions in nonparametric instrumental variables estimation," *Journal of Econometrics*, 189(1), 41 – 53.

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

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.

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

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

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

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

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

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

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.

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.

Scaillet, O. (2016): "On illposedness of nonparametric instrumental variable regression with convexity constraints," *The Econometrics Journal*, 19(2), 232–236.

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.