CESifo WORKING PAPERS

9046
2021

April 2021

Conditional Quantile Estimators: A Small Sample Theory

Grigory Franguridi, Bulat Gafarov, Kaspar Wüthrich

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9046*

Conditional Quantile Estimators: A Small Sample Theory

Abstract

We study the small sample properties of conditional quantile estimators such as classical and IV quantile regression. First, we propose a higher-order analytical framework for comparing competing estimators in small samples and assessing the accuracy of common inference procedures. Our framework is based on a novel approximation of the discontinuous sample moments by a Hölder-continuous process with a negligible error. For any consistent estimator, this approximation leads to asymptotic linear expansions with nearly optimal rates. Second, we study the higher-order bias of exact quantile estimators up to $O(1/n)$.
Using a novel non-smooth calculus technique, we uncover previously unknown non-negligible bias components that cannot be consistently estimated and depend on the employed estimation algorithm. To circumvent this problem, we propose a "symmetric" bias correction, which admits a feasible implementation. Our simulations confirm the empirical importance of bias correction.

JEL-Codes: C210, C260.

Keywords: non-smooth estimators, KMT coupling, Hungarian construction, higher-order asymptotic distribution, higher-order stochastic expansion, order statistic, bias correction, mixed integer linear programming (MILP), exact estimators, k-step estimators, quantile regression, instrumental variable quantile regression.

Grigory Franguridi
Department of Economics
University of Southern California
Los Angeles / CA / USA
franguri@usc edu

Bulat Gafarov
Department of Agricultural and Resource Economics, University of California, Davis
Davis / CA / USA
bgafarov@ucdavis.edu

Kaspar Wüthrich
Department of Economics
University of California, San Diego
San Diego / CA / USA
kwuthrich@ucsd.edu

April 9, 2021

1 Introduction

Many interesting empirical applications of classical quantile regression (QR) (Koenker and Bassett, 1978) and instrumental variable quantile regression (IVQR) (Chernozhukov and Hansen, 2005) feature small samples sizes, which can arise either as a result of a limited number of observations or when estimating tail quantiles, or both (e.g., Elsner et al., 2008; Chernozhukov and Fernández-Val, 2011; Adrian and Brunnermeier, 2016; Adrian et al., 2019). While many existing QR and IVQR estimators exhibit comparable statistical properties in sufficiently large samples, they differ dramatically with respect to their computational performance. As with other nonlinear estimators, they also differ with respect to their small sample bias and mean squared error (MSE) and the size accuracy of the corresponding inference procedures. However, there is no small sample theory providing guidance for choosing between the many available estimators in applications. In this paper, we develop a higher-order analytical framework allowing us to compare the competing estimators in small samples and provide new methods for improving *efficiency* and reducing their *bias*.

The main challenge when analyzing the most popular QR and IVQR estimators is that they solve optimization problems with non-smooth objective functions and *discontinuous* sample moments.[1] To formally study the properties of these estimators, we develop a novel approximation of their sample moments by a Hölder-continuous process with an *explicit modulus of continuity* and a *negligible* approximation error. This approximation allows us to characterize the key factors that determine the performance of these estimators: (i) the modulus of asymptotic stochastic continuity of the sample moments and (ii) the magnitude of the sample moment error evaluated at the estimators.

An important practical implication is that exact estimators such as linear programming (LP) estimators of QR and mixed integer programming (MIP) estimators of IVQR models (Chen and Lee, 2018; Zhu, 2019) are more efficient than estimators that do not exactly minimize the sample moment error. However, computing exact minima of non-smooth functions may be computationally prohibitive, and hence approximate but tractable estimators are often unavoidable.

Our higher-order analytical framework directly suggests an approach for alleviating the trade-off between computational feasibility and small-sample performance. We show that applying a 1-step Newton correction to any approximate solution reduces the bias and MSE to nearly $O\left(\frac{1}{n^{3/4}}\right)$. This analysis further allows us to explicitly establish the precision, also nearly $O_p\left(\frac{1}{n^{3/4}}\right)$, of asymptotic normal and bootstrap approximations of general QR and IVQR estimators.

[1] We discuss approaches based on smoothed objective functions and estimation equations in Section 2.3.

We also characterize the higher-order bias of exact QR and IVQR estimators, which is a key aspect of their small-sample performance. Our simulations show that quantile estimators can exhibit substantial biases even in simple location models (see the blue dots in Figure 1). The bias is non-negligible even at the median and can be arbitrarily large in the tails (see the results for DGP3 in Section 5.3). To study the bias, we develop a novel non-smooth calculus technique, which allows us to compute the asymptotic bias of exact QR and IVQR estimators up to even higher order than that of the MSE — namely $O\left(\frac{1}{n}\right)$. An important implication of our results is that the bias contains previously unknown and non-negligible components that cannot be consistently estimated and depend on the specifics of the employed numerical estimation algorithm.[2] To overcome this challenge, we propose a feasible bias correction procedure based on a novel *symmetric correction*, which eliminates the components that cannot be consistently estimated.

The trade-off between the computational cost of QR and IVQR estimators and their higher-order bias and MSE, which we document, has important implications for empirical practice. For small samples, we recommend the exact moment estimators: they achieve the smallest bias, which can be further reduced using our bias correction approach. Bias reduction is particularly important for valid inference and enables applying QR and IVQR estimators in small samples. In larger samples, where exact estimators may be computationally prohibitive, estimators can be chosen based on their computational performance. We show that a 1-step Newton correction can reduce both bias and MSE of any $\sqrt{n}$-consistent approximate estimator (e.g., inverse quantile regression (Chernozhukov and Hansen, 2006), Quasi-Bayesian MCMC (Chernozhukov and Hong, 2003), or fixed-point estimators (Kaido and Wüthrich, 2019)).

Higher-order analysis has been successfully applied to compare the small sample performance of alternative estimators and inference procedures in a wide range of statistical problems, including linear IV (e.g., Nagar, 1959; Rothenberg, 1984), smooth moment estimators (e.g., Rilstone et al., 1996; Newey and Smith, 2004; Anatolyev, 2005), HAR inference in time series (e.g., Andrews, 1991; Sun et al., 2008; Lazarus et al., 2021), incidental parameters in panel data (e.g., Hahn and Newey, 2004; Dhaene and Jochmans, 2015), among others. By its very nature, this approach relies on differentiability of the corresponding empirical processes. Extending this approach to the case of non-smooth quantile estimators has proven to be particularly challenging (e.g., Kiefer, 1967; Jurečková and Sen, 1987).

The existing literature on the higher-order properties of conditional quantile models with

[2] We emphasize that our approach is different from the generalized function heuristic or "shortcut" used in the existing literature (Phillips, 1991; Lee et al., 2017, 2018). It is rigorous and captures previously unknown non-negligible bias components.

non-smooth objective functions has only considered specific *convex* M-estimators that satisfy subgradient optimality conditions, which coincide with the sample moment equations up to a negligible error (e.g., Zhou and Portnoy, 1996; Knight, 2002; Portnoy, 2012). Here our goal is to study and compare general quantile estimators based on non-smooth moment conditions, including IVQR estimators with endogenous covariates. Such general estimators solve non-convex (often discrete) optimization problems that do not admit optimality conditions in finite samples. To overcome this obstacle, we developed novel proof techniques for bounding the asymptotic error of the sample moment conditions.

Moreover, the existing literature provides no complete characterization of the asymptotic bias of QR and IVQR estimators.[3] In fact, such a result is not even available for unconditional quantiles. An analytical characterization of the bias is particularly important in our context because alternative bootstrap-based methods for higher-order bias correction are not applicable (Knight, 2003). We believe that our higher-order methods can be useful for studying other estimators that have a linear index structure and are based on discontinuous sample moments or loss functions (e.g., censored QR and maximum score). Other directions for future research are outlined in Section 6.

2 Detailed discussion of the theoretical results

Here we provide a detailed discussion of the theoretical results of the paper and relate them to the existing literature. We make four main theoretical contributions, each described in more detail in the subsections below. First, we establish a novel coupling of the weighted empirical cumulative distribution function (CDF) with a tight Hölder continuous process with a uniform remainder bound. Second, based on this coupling result, we derive remainder rates for asymptotic linear expansions and MSE of exact (moment) estimators and k-step estimators. Third, we derive an asymptotic bias formula for exact estimators up to order $O\left(\frac{1}{n}\right)$. Fourth, we propose a feasible bias correction procedure that eliminates the bias components that cannot be consistently estimated.

Throughout the paper, we will often refer to QR and IVQR estimators collectively as conditional quantile estimators.

[3] Lee et al. (2017, 2018) provided a partial characterization of the bias based on the generalized functions heuristic of Phillips (1991). As we discuss below, this approach does not capture all terms of the bias.

2.1 Coupling of sample moments

In Theorem 1, we develop a novel coupling of the stochastically weighted empirical CDF process with a tight Hölder continuous process with exponent $\frac{1}{2} - \gamma$ for any small $\gamma > 0$ and a uniformly bounded $O_p\left(\frac{\log n}{n}\right)$ error. The expectation of the absolute approximation error depends only on the support of the random weights and the sample size; the expected Hölder constant depends both on the bound on the conditional probability density function (PDF) of the outcome and the support of the weights. Our result is a generalization of the celebrated KMT coupling of the uniform empirical process (Komlós et al., 1975, 1976), which additionally exploits the smoothness of the conditional CDF. Hölder continuity of the approximating process follows from the Kolmogorov-Chentsov theorem (e.g., Schilling and Partzsch, 2014, Theorem 10.6). Under some conditions on the regressors and the instruments, this result translates into a coupling of the sample moments with a Hölder continuous process with optimal remainder rates.[4]

Both properties of the approximation we derive, Hölder continuity and rate optimality, are essential for bounding the error of the sample moment conditions in the higher-order representations and equivalence results. By contrast, the existing generic strong approximations have suboptimal remainder rates that only imply very crude bounds on the sample moments. These generic bounds dominate all other higher-order terms in the expansion of conditional quantile estimators and, thus, are not useful for our purposes.[5]

2.2 Bahadur-Kiefer representations

Conditional quantile estimators are non-linear. Nevertheless, it is well-known that such estimators can be approximately represented as sample averages of score functions. Such asymptotic linear expansions are called *Bahadur-Kiefer representations* (Bahadur, 1966; Kiefer, 1967). To make statements about the precision of normal and bootstrap approximations of different estimators, it is important to derive explicit bounds on the non-linear remainder terms.

Using our coupling result, we derive Bahadur-Kiefer expansions for $\sqrt{n}$-consistent conditional quantile estimators under some support restrictions on the regressors (see Theorem 2). After a 1-step *Newton correction*, any $\sqrt{n}$-consistent estimator has an asymptotic linear representation with nearly $O_p\left(\frac{1}{n^{3/4}}\right)$ remainder rate.[6] This rate is *nearly-optimal* in the sense

[4] See Rio (1994) for a discussion of the optimality of remainder rates.

[5] Examples include: Massart (1989), Koltchinskii (1994), and Berthet et al. (2006). Chernozhukov et al. (2020) apply these results to study nonparametric IVQR models.

[6] 1-step estimators have a long tradition in statistics and econometrics, starting with Fisher, Neyman, and others; see, for example, Bickel (1975), which was one of the first papers to study non-smooth M-estimators,

that it matches (up to a logarithmic factor) the classical Bahadur-Kiefer remainder rate in the univariate order statistics case.

To our knowledge, this paper is the first to provide an explicit, nearly-optimal rate of the remainder in the Bahadur-Kiefer representation for general IVQR estimators with possibly endogenous regressors.[7] The available explicit results are limited to order statistics (Bahadur, 1966; Kiefer, 1967), classical QR with exogenous regressors (e.g., Zhou and Portnoy, 1996; Knight, 2002), and nonparametric series QR with exogenous regressors (e.g., Belloni et al., 2019). We show that the availability of a Bahadur-Kiefer expansion with nearly $O_p\left(\frac{1}{n^{3/4}}\right)$ rate does not depend on the specific structure of the estimator, but rather on the (asymptotic stochastic) Hölder continuity of the sample moments.

Our results have important implications for the higher-order properties of k-step estimators (e.g., Zhu, 2019) and complement the findings in Robinson (1988) for estimators based on smooth sample moment conditions. Unlike with smooth extremum estimators, additional Newton steps may not result in convergence of k-step estimators to the exact minimizer of the GMM objective function. However, we show that $\sqrt{n}$-consistent estimators become equivalent up to nearly $O_p\left(\frac{1}{n^{3/4}}\right)$ after a single Newton correction (see Theorems 3 and 4). This is in contrast to Andrews (2002a,b) who shows that 1-step corrected smooth estimators are equivalent with higher precision, $O_p\left(\frac{1}{n^{3/2}}\right)$.

Our results also have important implications for exact estimators, which exactly minimize an ℓ_p norm of the sample moments. Such estimators can be implemented using MIP techniques (Chen and Lee, 2018; Zhu, 2019), or, in the case of classical QR, exact LP algorithms. We use the coupling result to prove that the norm of the sample moments computed at the exact estimators attains a nearly $O_p\left(\frac{1}{n}\right)$ asymptotic rate and establish a higher-order expansion of exact estimators of just-identified models up to the nearly $O_p\left(\frac{1}{n^{5/4}}\right)$ order (see Theorem 5). It implies, in particular, that any two exact minimizers (possibly corresponding to different norms) are equivalent up to nearly $O_p\left(\frac{1}{n}\right)$. This equivalence applies to the exact estimators proposed by Chen and Lee (2018) and Zhu (2019), which minimize the ℓ_2 and ℓ_∞ norm, respectively.[8]

In related work, Portnoy (2012) developed a nearly $\sqrt{n}$-Gaussian approximation for classical QR estimators based on exact LP formulations using a saddlepoint approximation.[9]

and the references therein.

[7] Bahadur-Kiefer expansions with suboptimal (non-explicit) $o_p\left(\frac{1}{\sqrt{n}}\right)$ remainder rates can be obtained using standard VC class arguments (e.g., Chernozhukov and Hansen, 2006, Theorem 3).

[8] Pouliot (2019) proposed an alternative MIP approach based on the inverse quantile regression estimator of Chernozhukov and Hansen (2006). We further discuss this approach in Section 4.4.

[9] Ronchetti and Sabolová (2016) developed a saddlepoint inference procedures for classical QR, which are shown to be more accurate than existing methods in simulations.

His results imply the presence of a nearly $O\left(\frac{1}{n}\right)$ bias of the classical QR estimator. Unfortunately, the saddlepoint approach is difficult to extend to general conditional quantile estimators such as IVQR, which do not admit convex programming formulations. The main technical challenge is that the sample moment error and the resulting bias of general estimators can be nearly as large as $O_p\left(\frac{1}{\sqrt{n}}\right)$ as opposed to $O_p\left(\frac{1}{n}\right)$ for classical QR estimators based on exact LP algorithms. Therefore, this approach cannot be used to compare different estimators and obtain higher-order equivalence results in our setup. Moreover, while the saddlepoint approximation is convenient for analyzing the approximate density of the estimator and the order of the bias, it is less suitable for deriving a feasible bias correction procedure (cf. Section 2.3–2.4). We take a different approach based on our coupling of the sample moments with a continuous process admitting the nearly optimal remainder rate. This approach allows us to show higher-order improvements of k-step estimators and provides an improved bound on the sample moments evaluated at the exact IVQR estimator.

2.3 Higher-order bias formula

We provide a formula for the bias of the exact QR and IVQR estimators in just-identified conditional quantile models up to an error of $O\left(\frac{1}{n}\right)$. The bias has four components: (1) the bias from non-zero sample moments at the optimum; (2) the bias from the covariance of the linear influence of a single observation and the sample moments; (3) the bias from the residuals having point mass at 0; (4) the bias from the typical higher-order quadratic component of the population moment conditions (e.g., Rilstone et al., 1996).

To obtain the formula, we develop a novel non-smooth calculus argument to compute the expectation of the discontinuous sample moments evaluated at the estimated parameter value. In particular, we exploit the fact that the sample moments admit a directional Taylor expansion after conditioning on the estimator. This novel approach could be used to study the bias of other estimators based on discontinuous sample moments and loss functions (e.g. censored QR and the maximum score estimator).

To our knowledge, there is no complete characterization of the higher-order bias of conditional quantile estimators up to $O\left(\frac{1}{n}\right)$, not even in the sample quantile case. Unlike the approach taken in the existing literature (Lee et al., 2017, 2018), which is based on the generalized functions heuristic or "shortcut" of Phillips (1991), our argument is rigorous and leads to the discovery of additional non-negligible bias terms. We discuss their work in more detail in Section 5.

A popular approach for studying the higher-order properties of QR and IVQR models is to consider smoothed estimating equations or objective functions (e.g., Horowitz, 1998; Kaplan

and Sun, 2017; Fernandes et al., 2021). This approach has several important limitations. First, as we show in this paper, optimal smoothing introduces a bias that is larger than that of exact estimators with 1-step correction (see Section 5). Second, this approach requires choosing a smoothing bandwidth, which is often difficult in small samples. This appears to be one major reason why smoothed estimators have not been more popular in practice. Finally, smoothing may also unintentionally create multiple local optima, which complicates the search for the actual global optimum. For these reasons, we prefer to work directly with the original sample moment conditions.

2.4 Feasible bias correction

Bias component (1), the bias from non-zero sample moments at the optimum, and bias component (3), the bias from residuals having point mass at 0, appear because of the discontinuity of the sample moments. These two components depend on the realization of the estimator and, to our knowledge, there exists no approach to consistently estimate them. To overcome this issue, we suggest a *symmetric 1-step correction* of exact estimators, which is the average of two Newton corrections based on the sample moments corresponding to (τ, Y_i) and $(1-\tau, -Y_i)$, where τ is the quantile level and Y_i is the outcome variable. This correction eliminates bias components (1) and (3). The remaining components involve the Jacobian of the moment functions and weighted average derivatives, which can estimated using standard plug-in methods (e.g., Powell, 1986; Powell et al., 1989; Angrist et al., 2006; Kato, 2012; Hong et al., 2015).

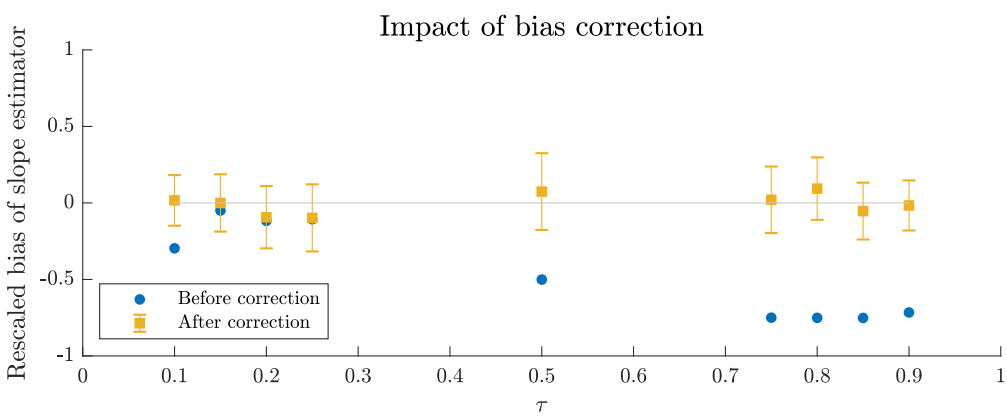

Figure 1: Bias of slope coefficient before (blue circle) and after correction (gold squares), both scaled by n with $n = 50$. The true model is a bivariate location model with uniform errors (DGP1 in Section 5.3). Based on Monte Carlo simulations with 20,000 repetitions. The error bands correspond to $3 \times$ Monte Carlo error. Note that the scale of the y-axis for DGP3 is different.

We validate our bias correction procedure in a Monte Carlo study. The simulations show that the asymptotic bias correction removes a substantial portion of the finite sample bias for sample sizes as small as $n = 50$. Figure 1 shows the actual bias and the impact of bias correction in a simple bivariate location model with uniform errors. Before bias correction, the bias can be substantial and varies considerably across quantiles. Unaccounted for, this bias will lead to size distortions of standard analytical inference methods. After the bias correction, the Monte Carlo error bands for the remaining bias contain zero for all quantile levels.

Our general bias formula has implications for order statistics. Specifically, since order statistics can be represented as exact moment estimators, our results also apply in that case. We show that our asymptotic bias formula matches the well-known exact bias formula for the case of uniformly distributed data up to $O\left(\frac{1}{n^2}\right)$.

3 Setup and background

In this section, we introduce the model, review existing approaches for deriving stochastic expansions of estimating equations estimators, and discuss the complications in deriving higher-order results arising from discontinuous sample moments.

3.1 The model

Consider a setting with a continuous outcome variable Y, a $(k \times 1)$ vector of covariates W, and a $(k' \times 1)$ vector of instruments Z. Every observation (Y_i, W_i, Z_i), $i = 1, \ldots, n$, is jointly drawn from a distribution P. We assume that (Y_i, W_i, Z_i) is iid. The parameter of interest $\theta_0(\tau) \in \mathbb{R}^k$ is defined as a solution to the following unconditional quantile moment restrictions

$$\mathbb{E}[(1\{Y \leq W'\theta_0(\tau)\} - \tau)Z] = 0, \tag{1}$$

for a fixed quantile level $\tau \in (0, 1)$. These moment restrictions arise from linear IVQR model (Chernozhukov and Hansen, 2006, 2008) when $W \neq Z$, from classical QR (Koenker and Bassett, 1978) when $W = Z$, and from unconditional quantiles when $W = Z = 1$.[10] We

[10] As explained by Chernozhukov and Hansen (2006, Footnote 1), IVQR also nests the two-stage quantile regression (2SQR) model considered in Amemiya (1982), Powell (1983), Chen and Portnoy (1996) as a case with constant QTE. Our results do not cover instrumental variables approaches based on the local average treatment effects framework (e.g., Abadie et al., 2002; Frölich and Melly, 2013; Melly and Wüthrich, 2017; Wüthrich, 2020) and triangular models (e.g., Ma and Koenker, 2006; Lee, 2007; Imbens and Newey, 2009, among others). An alternative approach to conditional distribution modeling is based on distribution regression (Foresi and Peracchi, 1995; Chernozhukov et al., 2013; Rothe and Wied, 2013); distribution regression can also be used for estimating IVQR models (Wüthrich, 2019). Distribution and quantile regression are not

refer to Koenker (2005) for a comprehensive review of classical QR and to Chernozhukov and Hansen (2013) and Chernozhukov et al. (2017) for recent reviews of IVQR.

We will focus on a fixed quantile level $\tau \in (0,1)$ and omit the dependence on τ unless it causes confusion. We will use the following notation for the unconditional moment restrictions as a function of $\theta \in \Theta$,

$$g(\theta) \triangleq \mathbb{E}(1\{Y \leq W'\theta\} - \tau)Z. \tag{2}$$

To abstract from additional complications arising from the estimation of the optimal weighting matrix, we focus on the just-identified case where $k = k'$.

We will maintain the following assumptions, which rule out partial and weak identification.

Assumption 1 (Identification)**.**

1. *θ_0 is the unique solution to $g(\theta) = 0$ over Θ, where θ_0 is in the interior of the compact set Θ.*

2. *The Jacobian of the moment functions evaluated at θ_0, $G(\theta_0)$, is well-defined and has full rank.*

Assumption 1 is a high-level assumption commonly imposed in GMM settings and in the literature on general conditional quantile models (e.g., Chernozhukov and Hansen, 2006, 2008; Kaido and Wüthrich, 2019); see Chernozhukov and Hansen (2006) and Kaido and Wüthrich (2019) for primitive conditions for global identification. As noted by Chernozhukov and Hansen (2006), compactness of the parameter space Θ "is not restrictive in microeconometric applications" (p. 502). The full rank assumption is necessary for asymptotic normality of the estimator and underlies our higher-order expansion.

3.2 Complications due to discontinuous sample moments

Consider any estimator $\hat{\theta}$ of θ_0 that approximately solves

$$\hat{g}(\theta) = 0, \tag{3}$$

where $\hat{g}(\theta)$ is the sample analog of moment condition (1), $\hat{g}(\theta) \triangleq \mathbb{E}_n(1\{Y \leq W'\theta\} - \tau)Z$ for $\theta \in \Theta$. Here we used the notation $\mathbb{E}_n m$ as a shortcut for expectation with respect to the empirical measure, i.e., $\frac{1}{n}\sum_{i=1}^n m_i$. We denote the generic empirical process operator as

nested in general and, thus, the results in our paper do not directly apply to distribution regression models.

$\mathbb{G}_n \triangleq \sqrt{n}(\mathbb{E}_n - \mathbb{E})$, and introduce the following shortcut notation for the sample moment process,

$$\mathbb{G}_n(\theta) \triangleq \mathbb{G}_n(1\{Y \leq W'\theta\} - \tau)Z = \sqrt{n}(\hat{g}(\theta) - g(\theta)), \quad \theta \in \Theta. \tag{4}$$

The standard way to get an approximate solution to (3) is to minimize $||\hat{g}(\theta)||_p$, where $||\cdot||_p$ is an ℓ_p norm on $\mathbb{R}^k$.

When the sample moment conditions $\hat{g}(\theta)$ are a.s. differentiable, the higher-order expansions of the exact and ℓ_2-approximate solutions of (3) were studied correspondingly in Rilstone et al. (1996) and Newey and Smith (2004) (unlike these papers, we abstract from the additional higher-order bias terms arising from overidentification and estimation of the GMM weighting matrix). One can understand the nature of the standard argument using the following tautology,

$$\underbrace{\hat{g}(\hat{\theta})}_{(i)} = \underbrace{g(\theta_0)}_{(ii)} + \underbrace{\frac{1}{\sqrt{n}}\mathbb{G}_n(\theta_0)}_{(iii)} + \underbrace{g(\hat{\theta}) - g(\theta_0)}_{(iv)} + \underbrace{\frac{1}{\sqrt{n}}(\mathbb{G}_n(\hat{\theta}) - \mathbb{G}_n(\theta_0))}_{(v)}. \tag{5}$$

Under standard assumptions (e.g., Rilstone et al., 1996, Assumptions B and C), including Lipschitz continuity of the sample moments and non-degeneracy of their Jacobian, the following results hold:[11]

(i) can be made arbitrarily small since (3) admits an exact solution;

(ii) is zero for correctly specified models;

(iii) is $O_p\left(\frac{1}{\sqrt{n}}\right)$ and is asymptotically normal (after rescaling) by the CLT;

(iv) is approximately linear in $(\hat{\theta} - \theta_0)$ with an error that is $O_p\left(\frac{1}{n}\right)$ by the Taylor theorem applied to $g(\theta)$ at θ_0;

(v) is $O_p\left(\frac{1}{n}\right)$ by a.s. Lipschitz continuity of $\hat{g}(\theta)$.

Moreover, the exact solution to (3) is unique by the assumed non-degeneracy of the Jacobian $\partial_\theta \hat{g}(\hat{\theta})$. However, the quantile sample moment conditions $\hat{g}(\theta)$ are a.s. discontinuous and do not admit a Jacobian, which gives rise to three main complications.

Non-existence of exact solutions to the estimating equations. Even in the just-identified setting ($k = k'$), equations (3) do not have an exact solution in most cases. As a result, we cannot assume that the error in the estimating equations is arbitrarily small. This issue is illustrated in the following example.

[11] See also Assumptions 2(b) and 2(c) in Newey and Smith (2004).

Example 1. *Suppose we are interested in the τ-quantile of a one-dimensional, continuous r.v. Y. The corresponding estimating equations take the form*

$$\mathbb{E}_n(1\{Y \leq \theta)\} - \tau) = 0. \tag{6}$$

Equation (6) can be satisfied exactly only if τn is an integer. In general, the best approximate solution $\hat{\theta}$ can take any value in $[Y_{(\lfloor \tau n \rfloor)}, Y_{(\lfloor \tau n \rfloor + 1)}]$ depending on τn.[12] *Then by definition,*

$$\mathbb{E}_n(1\{Y \leq \hat{\theta}\}) = \frac{\lfloor \tau n \rfloor}{n}, \tag{7}$$

which implies $\hat{g}(\hat{\theta}) = \mathbb{E}_n(1\{Y \leq \hat{\theta}\} - \tau) = O_p\left(\frac{1}{n}\right)$. □

Example 1 illustrates that even in the simplest case where $W = Z = 1$, the estimating equations can have an error that is bounded away from zero by $O_p\left(\frac{1}{n}\right)$ for most values of τ. In the general case with non-constant regressors W and instrumental variables Z, the error in the estimating equations may be bounded away from 0 even if τn is an integer. However, in Theorem 5 in Section 4, we show that in the general case there exists an approximate solution to (3) achieving nearly $O_p\left(\frac{1}{n}\right)$ value of the objective function, as in the univariate Example 1. This error is non-negligible and appears in the expansion up to order $O_p\left(\frac{1}{n}\right)$.

Non-uniqueness of the best approximate solution. In the case of sample moments with non-degenerate Jacobian, the exact solution to the estimating equations is unique by the implicit function theorem. By contrast, due to the presence of the indicator function, the quantile sample moment conditions are step functions that do not admit a Jacobian. As as result, the set of solutions has non-negligible diameter. This is true even if the estimating equations admit exact solutions. To illustrate, let us revisit Example 1.

Example 2 (Example 1 cont.)**.** *Suppose that τn is an integer. Suppose further that Y has a uniform distribution on $[0, 1]$. Consider two solutions of (6), $\hat{\theta} = Y_{(\tau n)}$ and $\hat{\theta}^* = Y_{(\tau n)} + (1 - \epsilon)(Y_{(\tau n+1)} - Y_{(\tau n)})$ for some small $\epsilon > 0$. Both solutions will have the norm of the error exactly equal to 0. The well known formula for the order statistic in the case of uniform distribution is $\mathbb{E}Y_{(j)} = \frac{j}{n+1}$, so the difference in the means of the solutions is $\mathbb{E}(\hat{\theta}^* - \hat{\theta}) = \frac{(1-\epsilon)}{n+1}$.*[13] *As a result, $\hat{\theta}^* - \hat{\theta} \geq O_p(\frac{1}{n})$ since $\hat{\theta}^* \geq \hat{\theta}$ a.s.* □

This example shows that different solutions can have stochastic expansions that differ by $O_p\left(\frac{1}{n}\right)$. Corollary 2 of Theorem 4 in Section 4 shows that the equivalence of different exact estimators holds with nearly $\frac{1}{n}$-rate for general conditional quantile models.

[12] We use the notation $\lfloor x \rfloor$ for the integer part of a real number x and $Y_{(k)}$ for the k-th order statistic.

[13] See Ahsanullah et al. (2013, Example 8.1).

Non-Lipschitz sample moments. Typically, for models with smooth sample moments the empirical process $\mathbb{G}_n(\theta)$ has a Lipschitz constant K_n that is bounded in probability in a neighborhood of θ_0.[14] By this property, for any $\sqrt{n}$-consistent estimator $\hat{\theta}$,

$$\|\mathbb{G}_n(\hat{\theta}) - \mathbb{G}_n(\theta_0)\| \leq K_n \|\hat{\theta} - \theta_0\| = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{8}$$

This result is instrumental in showing that the precision of the linear approximation of the normalized estimating equations estimator $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ is $O_p\left(\frac{1}{\sqrt{n}}\right)$.

Unfortunately, such a high-precision linear approximation is not available for conditional quantile models since property (8) does not hold (see Example 3 below). To study the precision of this approximation, we exploit the special structure of the quantile moment conditions that resemble the empirical CDF. This structure is particularly evident in the special case of the sample quantile estimator, where the sample moment conditions can be directly represented using the empirical CDF $\hat{F}_Y(y) = \mathbb{E}_n 1\{Y \leq y\}$. When Y is uniformly distributed on $[0, 1]$, the empirical CDF admits the following strong approximation, which was first derived by Komlós et al. (1975, 1976).

Theorem (KMT coupling, Theorem <26> on p.252 in Pollard (2002))**.** *There exists a Brownian Bridge $\{B^\circ(y) : 0 \leq y \leq 1\}$ with continuous sample paths, and a uniform empirical process $\mathbb{F}_n(y) \triangleq \sqrt{n}(\hat{F}_Y(y) - F_Y(y))$, for which*

$$P\left\{\sup_{0 \leq y \leq 1} |\mathbb{F}_n(y) - B^\circ(y)| \geq C_1 \frac{x + \log n}{\sqrt{n}}\right\} \leq C_0 \exp(-x) \text{ for all } x \geq 0, \tag{9}$$

with constants C_1 and C_0 that depend on neither n nor x.[15]

The KMT theorem suggests that $O_p\left(\frac{1}{\sqrt{n}}\right)$ is unattainable even in the case of the sample quantile of a uniform r.v., as the following example shows.

Example 3 (Example 1 cont.)**.** *For a uniform Y, the CDF is $F_Y(y) = y$ for $y \in [0, 1]$. The KMT theorem implies the following a.s. representation for the sample moment conditions (cf. Pollard, 2002, p.255),*

$$\mathbb{G}_n(\theta) = \mathbb{F}_n(\theta) = B(\theta) - \theta B(1) + R_n(\theta), \tag{10}$$

where $\sup_{0 \leq \theta \leq 1} |R_n(\theta)| = O_p\left(\frac{\log n}{\sqrt{n}}\right)$, $B(1)$ is a standard Gaussian r.v. and $B(\theta) \triangleq B^\circ(\theta) + \theta B(1)$ is the standard Brownian motion process. It follows from the Kolmogorov-Chentsov

[14] See for example Assumption C in Rilstone et al. (1996) with $s \geq 2$.

[15] Theorems 1.7 and 1.8 in Dudley (2014) provide a specific choice of constants. By Theorem 1.8, one can choose $C_1 = 12$ and $C_0 = 2$ for $n \geq 2$.

theorem (see Lemma A.5 in Appendix A.2) that, for any small $\gamma > 0$,

$$B(\theta) - B(\theta_0) = O_p\left(\|\theta - \theta_0\|^{\frac{1}{2}-\gamma}\right). \tag{11}$$

Therefore, the increment of the empirical process $\mathbb{G}_n$ evaluated at a random point $\hat{\theta}$ has the following representation for each n,

$$\mathbb{G}_n(\hat{\theta}) - \mathbb{G}_n(\theta_0) = O_p\left(\|\hat{\theta} - \theta_0\|^{\frac{1}{2}-\gamma}\right) + R_n(\hat{\theta}) - R_n(\theta_0). \tag{12}$$

The remainder here is uniformly bounded since $|R_n(\hat{\theta}) - R_n(\theta_0)| \leq 2\sup_{0\leq t\leq 1}|R_n(t)|$. Hence, for any $\sqrt{n}$-consistent estimator $\hat{\theta}$ and any small $\gamma > 0$, we get

$$\mathbb{G}_n(\hat{\theta}) - \mathbb{G}_n(\theta_0) = O_p\left(\frac{1}{n^{1/4-\gamma}}\right). \tag{13}$$

This results in a nearly $O_p\left(\frac{1}{n^{3/4}}\right)$ term in the stochastic expansion of $\hat{\theta}$, see equation (5). □

Theorem 1 in Section 4 provides an analog of the KMT theorem that accommodates general quantile models, and Corollary 1 shows that the sample moments can be approximated by a Hölder continuous process such that the remainder rate of nearly $O_p\left(\frac{1}{n^{3/4}}\right)$ is still valid.

4 Higher-order expansions and equivalence of quantile estimators

In the previous section, we illustrated the key ideas behind the stochastic expansion of a simple quantile estimator. In this section, we obtain stochastic (asymptotically linear) expansions of general conditional quantile estimators and provide higher-order equivalence results.

4.1 Coupling of sample moments with a Hölder continuous process

The quantile sample moment conditions are discontinuous functions of the parameter θ, which invalidates standard arguments for deriving higher-order properties. As we have seen in the previous section in the case of the sample quantile estimator, the sample moment conditions can be approximated by a Brownian bridge process, which has continuous trajectories and admits a strong version of stochastic equicontinuity that we call ℓ_1 Hölder continuity. Namely, the expectation of the ℓ_1 norm of the increments of the process is bounded by a

polynomial function of the increments. Here we extend this one-dimensional result to derive a coupling of the quantile sample moments with a Hölder continuous process.

To develop intuition for our main arguments, consider the following example.

Example 4 (QR with binary regressor)**.** *Consider a classical QR model with one exogenous binary regressor and an intercept. In this case, the instruments are equal to the regressors, $Z_i = W_i$, where $W_i = (W_{1i}, W_{2i})' \in \{(1,0)', (1,1)'\}$, and the moment functions are $(1\{Y_i \leq W_i'\theta\} - \tau)W_{1i}$ and $(1\{Y_i \leq W_i'\theta\} - \tau)W_{2i}$. The corresponding empirical processes are indexed by the 2-dimensional parameter $\theta \in \mathbb{R}^2$, and hence the standard univariate coupling results like the KMT theorem are not applicable. Our theoretical argument is based on the following observation:*

$$1\{Y_i \leq W_i'\theta\}W_{1i} = 1\{Y_i \leq y_1, W_i = (1,0)\} + 1\{Y_i \leq y_2, W_i = (1,1)\}, \tag{14}$$

$$1\{Y_i \leq W_i'\theta\}W_{2i} = 1\{Y_i \leq y_2, W_i = (1,1)\}, \tag{15}$$

where $y_1 = (1,0)\theta$ and $y_2 = (1,1)\theta$ are scalar parameters. This observation allows us to separate the empirical process indexed by θ into processes with 1-dimensional parameters to which univariate coupling results can be applied. □

To derive a coupling of the quantile sample moments with a Hölder continuous process, we develop conditions under which the dimension reduction technique illustrated in Example 4 can be extended to general conditional quantile models. Specifically, we will show that one can separate the empirical process corresponding to the quantile moment conditions, which is indexed by the k-dimensional parameter θ, into weighted conditional empirical processes that take the following form:

$$\mathbb{Z}_n(y,a) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i 1\{Y_i \leq y, A_i = a\} - \mathbb{E} Z_i 1\{Y_i \leq y, A_i = a\} \right), \text{ for } a \in \{0,1\}, \tag{16}$$

where A_i is a Bernoulli random variable. The next theorem provides a coupling result for $\mathbb{Z}_n(y,a)$.

Theorem 1. *Suppose that $Z \in \mathbb{R}$ and $|Z| < m < \infty$ a.s.. Suppose further that Y has density conditional on Z and A that is bounded by $\bar{f}$. Then, for any $a \in \{0,1\}$, the process $\mathbb{Z}_n(y,a)$ can be a.s. represented as*

$$\mathbb{Z}_n(y,a) = \mathbb{Z}(y,a) + R_n(y,a), \tag{17}$$

where $\mathbb{Z}(y,a)$ is a zero mean process with a.s. Hölder continuous paths that has increments

with bounds

$$\limsup_{r\to 0^+} \sup_{y_1,y_2\in\mathbb{R},|y_2-y_1|<r} \frac{|\mathbb{Z}(y_2,a)-\mathbb{Z}(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2-y_1|)}} \leq 4\sqrt{2}m, \ \ a.s. \tag{18}$$

$$\mathbb{E}\sup_{y_1\neq y_2} \frac{|\mathbb{Z}(y_2,a)-\mathbb{Z}(y_1,a)|}{|y_2-y_1|^{\frac{1}{2}-\gamma}} \leq m\tilde{c}_\gamma \bar{f}^{\frac{1}{2}-\gamma}, \quad \textit{for any } \gamma\in(0,1/2), \tag{19}$$

where $\Psi(x) \triangleq x\log(1/x)$ *and* $R_n(y,a)$ *is such that* $\mathbb{E}(R_n(y,a))=0$ *for all* $y\in\mathbb{R}$ *and*

$$\mathbb{E}\sup_{y\in\mathbb{R}_+}|R_n(y,a)| \leq m\tilde{c}_1\frac{\log n+\tilde{c}_0}{\sqrt{n}}, \tag{20}$$

with constants $\tilde{c}_1$*,* $\tilde{c}_0$*, and* $\tilde{c}_\gamma$ *that do not depend on* n*,* x*, or the distribution of* (Y,Z,A)*.*

Proof. See Appendix A. □

Equation (17) corresponds to the a.s. representation implied by the KMT theorem. The moment bounds on the remainder term could be replaced by the corresponding tail probability bounds, as in the KMT theorem; however, we found it more convenient to work with moment bounds for deriving stochastic expansions.

Equation (19) generalizes Example 3 to the case of a stochastically weighted CDF. [16] This equation shows the relationship between the bound on the conditional density, $\bar{f}$, and the modulus of continuity of the quantile sample moments. Whether the modulus of continuity $r^{\frac{1}{2}-\gamma}$ in Equation (19) can be replaced with $r^{\frac{1}{2}}$ up to a log term is an open question.[17]

Remark 1 (**Unbounded support**)**.** *In the proof, we treat the support of the instruments as fixed. This assumption is made to simplify the exposition. We can allow the diameter of the support,* m*, to grow with sample size at a slow rate. In particular, if* Z *is a sub-Gaussian or a sub-exponential random variable, the effective rate of growth of* m *will be* $\sqrt{\log n}$ *or* $\log n$ *correspondingly. We leave this extension for future work.* □

Under the following sufficient conditions, Theorem 1 implies a coupling result for the quantile sample moments with nearly-optimal remainder rates, see Corollary 1.

Assumption 2 (Regressors and instruments)**.**

1. *There exists* $m<\infty$ *such that* $\|Z\|<m$ *and* $\|W\|<m$ *a.s.*

[16] Because of the stochastic Z_i assumption, our results are not suitable for studying classical QR with fixed designs. We conjecture that one can develop analogous results based on strong approximations of (deterministically) weighted empirical processes (e.g., Csörgő et al., 1986).

[17] The recent work by Fischer and Nappo (2009) suggests that it may be possible to achieve $r^{\frac{1}{2}}$ up to a log term.

2. $\|W_i\| > 0$ *a.s. and* $\tilde{W}_i \triangleq W_i/\|W_i\| \in \{\tilde{w}_1, \ldots, \tilde{w}_s\} \subset \mathbb{R}^k$ *where* $s \ll n$.

Bounded support of the instruments is required for applying Theorem 1. Assumption 2.2 is a restriction on the support of the regressors W_i, which enables extending Theorem 1 to general conditional quantile estimators. In particular, this assumption allows us to reduce the k-dimensional empirical process to s univariate processes to which Theorem 1 can be applied. Such a dimension reduction is sufficient for obtaining explicit remainder rates for general conditional quantile estimators that are nearly-optimal in the sense that they match, up to a logarithmic factor, the classical Bahadur-Kiefer remainder rate in the univariate order statistics case. In Appendix E, we consider an alternative to Assumption 2.2, which allows for continuous directions $\tilde{W}_i$, see Remark 2 below.

We impose the following additional restrictions on the conditional density.

Assumption 3 (Conditional density)**.**

1. *The conditional density of* Y_i *given* (W_i, Z_i), $f_Y(y|W, Z)$, *exists and is a.s. uniformly bounded in* y *on* $supp(Y)$ *by* $\bar{f}$.

2. $f_Y(y|W, Z)$ *is a.s. twice continuously differentiable on* $supp(Y)$.

The uniform bound in Assumption 3.1 is essential for applying the coupling in Theorem 1. This assumption is standard in the literature (e.g., Chernozhukov and Hansen, 2006, Assumption R3). Under Assumption 3.2, the Jacobian of the population moment functions $G(\cdot) \triangleq \partial_\theta g(\cdot)$ is twice continuously differentiable (see Lemma B.1); we will use this property for second-order Taylor expansions of the population moments. Such smoothness assumptions are standard in the literature on higher-order properties of moment-based estimators (e.g., Rilstone et al., 1996; Newey and Smith, 2004).

To state the coupling result for the quantile sample moments, it is useful to introduce the following notation. For $\theta \in \Theta$, we define

$$g^\circ(\theta) \triangleq \mathbb{E}1\{Y \leq W'\theta\}Z, \tag{21}$$

$$B_n^\circ(\theta) \triangleq \sqrt{n}(\mathbb{E}_n 1\{Y \leq W'\theta\}Z - g^\circ(\theta)), \tag{22}$$

$$B_n(\theta) \triangleq B_n^\circ(\theta) - B_n^\circ(\theta_0). \tag{23}$$

Corollary 1. *Suppose that Assumptions 2 and 3.1 hold. Then*

$$B_n(\theta) = B(\theta) + R_n(\theta), \tag{24}$$

where $B(\theta)$ *is a* k*-dimensional* $\frac{1}{2} - \gamma$ *Hölder continuous process with expected* ℓ_1 *modulus bounded by* $ksm\tilde{c}_\gamma(m\bar{f})^{\frac{1}{2}-\gamma}$ *and* $\mathbb{E}\sup_{\theta\in\Theta}\|R_n(\theta)\|_1$ *is bounded by* $ksm\tilde{c}_1\frac{\log n+\tilde{c}_0}{\sqrt{n}}$.

Proof. Define $\tilde{Y} \triangleq \frac{Y}{\|W\|}$ and $A^q \triangleq 1\{\tilde{W} = \tilde{w}_q\}$ for $q = 1, \ldots, s$. Consider the empirical process

$$B_n^q(y) \triangleq \mathbb{G}_n(Z1\{\tilde{Y} \leq y, A^q = 1\}), \tag{25}$$

where $Z \in \mathbb{R}^k$ is the vector of instruments and $y \in \mathbb{R}$. By Assumption 2.2, the events $\{\tilde{W} = \tilde{w}_q\}$, $q = 1, \ldots, s$, form a partition of the probability space, so for any $\theta \in \Theta$,

$$B_n(\theta) = \mathbb{G}_n(1\{\tilde{Y} \leq \tilde{W}'\theta\}Z) = \sum_{q=1}^{s} B_n^q\left(\tilde{w}_q'\theta\right). \tag{26}$$

By Assumption 3.1, $f_Y(y|W, Z) \leq \bar{f}$ for all y. By Assumption 2.1, it follows that $f_{\tilde{Y}}(y|W, Z) \leq m\bar{f}$. Finally, for each $q = 1, \ldots, s$, apply (19) from Theorem 1 and Lemma A.4 componentwise to $B_n^q(\cdot)$ to get the desired result. □

Corollary 1 shows that one can explicitly construct a smooth approximation of the sample moments process with a uniform error bound. This coupling is useful for deriving higher-order expansions as it allows us to relate the non-smooth sample moment process to a smooth process, which can be studied using existing approaches. To our knowledge, Corollary 1 provides the first coupling of the general ($k \geq 1$) quantile sample moment process with nearly-optimal remainder rate (see Rio, 1994, for optimal rates of strong approximation for generic empirical processes).[18]

The coupling in Corollary 1 may be of independent interest. As in Theorem 1, Corollary 1 gives explicit moment bounds that hold for any finite n. Results of this nature are particularly useful for establishing the uniform (with respect to classes of DGPs) validity of inference procedures. For example, coupling techniques have been used for inference under shape restrictions (Chernozhukov et al., 2020).

4.2 Bahadur-Kiefer representations of general quantile estimators

Consider the infeasible linear and unbiased estimator

$$\hat{\theta}_1 \triangleq \theta_0 - G^{-1}(\theta_0)\frac{1}{n}\sum_{i=1}^{n} Z_i\left(1\{Y_i \leq W_i'\theta_0\} - \tau\right). \tag{27}$$

The linear structure of this estimator makes it easy to study the quality of the asymptotic normal approximation using the corresponding results on explicit CLT precision bounds for

[18] We thank Andres Santos for pointing out this reference to us.

multivariate sample means. Under some regularity conditions, the estimator $\hat{\theta}_1$ is first-order equivalent to most conditional quantile estimators (e.g., Koenker and Bassett, 1978; Chernozhukov and Hansen, 2006; Kaplan and Sun, 2017; Kaido and Wüthrich, 2019). However, it is important to have a bound on the higher-order terms to study the precision of the normal approximation and bootstrap procedures. Rilstone et al. (1996) and Newey and Smith (2004) provide results for smooth objective functions suggesting the higher-order properties of first-order equivalent estimators may be very different.[19] We focus on estimators with non-smooth objective functions such that these results are not applicable in our context (see Section 3.2).

For conditional quantile models, explicit remainder bounds in asymptotic linear representations are only available for classical QR (e.g., Zhou and Portnoy, 1996; Knight, 2002) and series QR (Belloni et al., 2019). As explained in Section 2.2, these results cannot directly be extended to general conditional quantile models such as IVQR with endogenous covariates. In what follows, we use Theorem 1 and Corollary 1 to obtain such bounds for a much broader class of estimators.

Consider an infeasible single Newton step correction corresponding to the minimization of $\|\hat{g}(\theta)\|_2$, $T(\theta) \triangleq \theta - G^{-1}\hat{g}(\theta)$. The next theorem provides a Bahadur-Kiefer expansion with an explicit bound on the remainder term for any estimator of θ_0 after applying a single Newton step.[20]

Theorem 2. *Suppose that Assumptions 1–3 hold. Suppose further that $\hat{\theta} = \theta_0 + R_n$. Then*

$$\hat{\theta}_{1-step} \triangleq T(\hat{\theta}) = \hat{\theta}_1 + O_p\left(\frac{sm^{\frac{3}{2}} R_n^{\frac{1}{2}-\gamma}}{n^{1/2}}\right) + O_p(\|R_n\|^2). \tag{28}$$

Proof. By Lemma B.2 , $\hat{\theta}$ satisfies

$$\begin{aligned} &G(\theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)'\partial_\theta G(\theta_0)(\hat{\theta} - \theta_0) \\ &= \hat{g}(\hat{\theta}) - \frac{1}{\sqrt{n}} B_n^\circ(\theta_0) - \tau(\mathbb{E}Z - \mathbb{E}_n Z) - \frac{1}{\sqrt{n}} B_n(\hat{\theta}) + O_p\left(\|R_n\|^3\right). \end{aligned} \tag{29}$$

[19] Kaplan and Sun (2017) derive higher-order properties of smoothed IVQR estimators.

[20] Arcones and Mason (1997) provide a second-order distributional analysis of 1-step estimators based on generic estimating equations. However, their framework cannot directly accommodate general conditional quantile models. We leave the exact distributional analysis of the Bahadur-Kiefer remainder in our setup for future research.

By Assumption 2, we have

$$\hat{\theta} - G^{-1}(\theta_0)\hat{g}(\hat{\theta}) = \hat{\theta}_1 - G^{-1}(\theta_0)\frac{1}{\sqrt{n}}B_n(\hat{\theta}) + O_p\left(\|R_n\|^2\right).$$

By Lemma B.5, which uses Corollary 1, $\frac{1}{\sqrt{n}}B_n(\hat{\theta}) = O_p\left(\frac{sm^{\frac{3}{2}}R_n^{\frac{1}{2}-\gamma}}{n^{1/2}}\right)$. This result completes the proof. □

In particular, Theorem 2 implies that $\hat{\theta}_{1-step}$ has an asymptotically linear representation with nearly $O_p\left(\frac{1}{n^{3/4}}\right)$ rate if $R_n = o_p\left(\frac{1}{n^{1/4}}\right)$. Moreover, if $R_n = O_p(\frac{1}{\sqrt{n}})$, then for the two step estimator $\hat{\theta}_{2-step} \triangleq T(T(\hat{\theta}))$, we get the nearly conventional order of the remainder, $O_p(\frac{1}{n^{3/4}})$, as in the sample quantile case (see p.122 in Jurečková et al., 2012). This implication for the 2-step estimator is analogous to the results in Robinson (1988) for smooth estimators.[21]

Theorem 2 also complements Zhu (2019), who established first-order equivalence of k-step estimators and GMM estimators for the IVQR model, by providing an explicit bound on the remainder term.

Remark 2 (**Continuous directions** $W_i/\|W_i\|$)**.** *v Assumption 2.2 is a transparent sufficient condition allowing us to derive Bahadur-Kiefer expansions and prove equivalence results in the following subsections. However, this condition may not be plausible in some applications. Assumption 2.2 only enters the theoretical arguments through Lemma B.5. In Appendix E, we provide an alternative to Lemma B.5, which allows for continuously distributed directions $W_i/\|W_i\|$. The same remark applies to all results in Sections 4.3 and 4.4. (The results on the bias in Section 5 do not rely on Assumption 2.2.)* □

4.3 Higher-order equivalence results for quantile estimators

In stark contrast with the smooth extremum estimators (e.g., Andrews, 2002a,b), additional Newton steps may not result in convergence of the k-step estimator to the exact minimizer of the GMM objective function. The reason is that the Newton steps are not guaranteed to reduce the objective function $\|\hat{g}(\hat{\theta})\|$ below $O_p\left(\frac{sm^{\frac{3}{2}}}{n^{3/4-\gamma}}\right)$ since this function is non-smooth. The following theorem gives a general result on the norm of the objective for a generic estimator $\hat{\theta}$.

Theorem 3. *Suppose that Assumptions 1–3 hold. Any estimator of the form $\hat{\theta} = \hat{\theta}_1 + R_n$,*

[21] The same implication would hold for k-step estimators with G^{-1} replaced by a consistent estimator with sufficiently fast convergence rate. See discussion of such estimators in Section 5.2.

where $\|R_n\| = o_p\left(\frac{1}{\sqrt{n}}\right)$, *satisfies*

$$\hat{g}(\hat{\theta}) = G(\theta_0)R_n + O_p\left(\frac{sm^{\frac{3}{2}}}{n^{3/4-\gamma}}\right). \tag{30}$$

for any small $\gamma > 0$.

Proof. Since $\|R_n\| = o_p\left(\frac{1}{\sqrt{n}}\right)$, $\hat{\theta} = \theta_0 + O_p\left(\frac{1}{\sqrt{n}}\right)$. As in the proof of Theorem 2, we have

$$R_n = \hat{\theta} - \hat{\theta}_1 = G^{-1}(\theta_0)\hat{g}(\hat{\theta}) - G^{-1}(\theta_0)\frac{1}{\sqrt{n}}B_n(\hat{\theta}) + O_p\left(\frac{1}{n}\right). \tag{31}$$

By Lemma B.5, which uses Corollary 1, $\frac{1}{\sqrt{n}}B_n(\hat{\theta}) = O_p\left(\frac{sm^{\frac{3}{2}}\|\hat{\theta}-\theta_0\|^{\frac{1}{2}-\gamma}}{n^{1/2}}\right) = O_p\left(\frac{sm^{\frac{3}{2}}}{n^{3/4-\gamma/2}}\right)$ for any small $\gamma > 0$. Hence equation (31) becomes

$$\hat{g}(\hat{\theta}) = O_p\left(\frac{sm^{\frac{3}{2}}}{n^{3/4-\gamma/2}}\right) + O_p\left(\frac{1}{n}\right) + G(\theta_0)R_n. \tag{32}$$

The term $O_p\left(\frac{1}{n}\right)$ is negligible when compared to the first term, which completes the proof. □

The previous theorem suggests that the order of magnitude of the sample moments evaluated at the estimated value $\hat{\theta}$ additively depends on the remainder of the Bahadur-Kiefer expansion. The following statement provides the converse result: estimators $\hat{\theta}^*$, $\hat{\theta}$, for which the values $\hat{g}(\hat{\theta}^*)$ and $\hat{g}(\hat{\theta})$ are close to each other, are equivalent up the order of magnitude of $\hat{g}(\hat{\theta}^*) - \hat{g}(\hat{\theta})$.

Theorem 4. *Suppose that Assumptions 1–3 hold. Consider any pair of asymptotically linear estimators* $\hat{\theta} = \hat{\theta}_1 + o_p(\frac{1}{\sqrt{n}})$ *and* $\hat{\theta}^* = \hat{\theta}_1 + o_p(\frac{1}{\sqrt{n}})$. *If* $\hat{g}(\hat{\theta}) - \hat{g}(\hat{\theta}^*) = O_p\left(\frac{|R_n|}{\sqrt{n}}\right)$ *for some bounded sequence* R_n, *then the following is true*

$$\hat{\theta} - \hat{\theta}^* = O_p\left(\frac{|R_n|}{\sqrt{n}}\right) + O_p\left(\frac{m^3 s^2}{n^{1-\gamma}}\right). \tag{33}$$

Proof. By Lemma B.2, both estimators satisfy representation (29). Then, for any small $\gamma > 0$,

$$\hat{\theta} - \hat{\theta}^* = -G^{-1}(\theta_0)\frac{B_n(\hat{\theta}) - B_n(\hat{\theta}^*)}{\sqrt{n}} + G^{-1}(\theta_0)(\hat{g}(\hat{\theta}) - \hat{g}(\hat{\theta}^*)) + O_p\left(\frac{s^2 m^3}{n^{5/4-\gamma}}\right). \tag{34}$$

The the result follows immediately from Lemma B.6. □

One direct implication of Theorems 3 and 4 is that any $\sqrt{n}$-consistent estimators of θ_0 become equivalent up to nearly $O_p\left(\frac{1}{n^{3/4}}\right)$ after a single Newton step correction. This result is reminiscent of the results obtained by Andrews (2002a) for smooth extremum estimators, although the non-differentiability of the objective function leads to equivalence of order nearly $\frac{1}{n^{3/4}}$ instead of $\frac{1}{n^{3/2}}$.

4.4 Stochastic expansion of exact estimators

The previous results hold for generic estimators. Theorem 3 shows the connection between the remainder in the Bahadur-Kiefer expansion and the sample moments evaluated at the estimator. It is therefore useful to study estimators that precisely minimize a finite-dimensional ℓ_p norm of the sample moments

$$\hat{\theta}_{\ell_p} = \text{argmin}_{\theta \in \Theta} ||\hat{g}(\theta)||_p. \tag{35}$$

This class of exact estimators includes GMM, which corresponds to $||\cdot||_2$ norm as in Chen and Lee (2018) for just-identified models, and the estimator proposed by Zhu (2019), which corresponds to $||\cdot||_\infty$. In the Monte Carlo section of this paper we consider $||\cdot||_1$ for computational convenience; see Appendix D.[22] The classical QR estimator implemented using exact LP algorithms is another leading example to which our results apply. Indeed, exact LP estimators yield exact zeros of the subgradient, which differs from sample moment functions by at most k-terms, or $O_p(\frac{k}{n})$.

The next theorem provides a bound on the minimal norm of the sample moments for the exact estimator and an explicit stochastic expansion up to order nearly $\frac{1}{n^{5/4}}$. We use the following notation,

$$G \triangleq G(\theta_0), \tag{36}$$

$$\partial_\theta G_j \triangleq \frac{\partial^2 g_j(\theta_0)}{\partial\theta\partial\theta'}. \tag{37}$$

For any $x \in \mathbb{R}^k$, we will use $x'\partial_\theta G x$ to denote a vector with components $x'\partial_\theta G_j x$ for $j = 1, \ldots, k$.

Theorem 5. *Suppose that Assumptions 1–3 hold. Consider $\hat{\theta}_{\ell_p}$ obtained from program (35)*

[22] The cases $||\cdot||_1$ and $||\cdot||_\infty$ have computationally convenient MILP representations, while the MILP formulation for $||\cdot||_2$ proposed in Chen and Lee (2018) has many more decision variables.

for some $p \in [1, \infty]$. *Then*

$$\hat{\theta}_{\ell_p} - G^{-1}\hat{g}(\hat{\theta}_{\ell_p}) = \hat{\theta}_1 - G^{-1}\left[\frac{B_n(\hat{\theta}_{\ell_p})}{\sqrt{n}} + \frac{1}{2}(\hat{\theta}_1 - \theta_0)'\partial_\theta G(\hat{\theta}_1 - \theta_0)\right] + R_n, \tag{38}$$

where $\hat{g}(\hat{\theta}_{\ell_p}) = O_p\left(\frac{m^3 s^2}{n^{1-\gamma}}\right)$, $B_n(\hat{\theta}_{\ell_p}) = O_p\left(\frac{sm^{3/2}}{n^{1/4-\gamma}}\right)$, *and* $R_n = O_p\left(\frac{sm^{3/2}}{n^{5/4-\gamma}}\right)$, *for any small* $\gamma > 0$.

Proof. By Lemma B.2 applied to $\hat{\theta}_{\ell_p}$,

$$\hat{\theta}_{\ell_p} - G^{-1}\hat{g}(\hat{\theta}_{\ell_p}) = \hat{\theta}_1 - G^{-1}\left[\frac{B_n(\hat{\theta}_{\ell_p})}{\sqrt{n}} + \frac{1}{2}(\hat{\theta}_{\ell_p} - \theta_0)'\partial_\theta G(\hat{\theta}_{\ell_p} - \theta_0)\right] + O_p(\|\hat{\theta}_{\ell_p} - \theta_0\|^3). \tag{39}$$

By Lemma B.4, $\hat{\theta}_{\ell_p} - \theta_0 = O_p\left(\frac{1}{\sqrt{n}}\right)$. The result $B_n(\hat{\theta}_{\ell_p}) = O_p\left(\frac{sm^{3/2}}{n^{1/4-\gamma}}\right)$ follows from Lemma B.5. Under Assumptions 1–3, Lemma B.7, which is based on the Hölder continuity of B_n, yields

$$\hat{g}(\hat{\theta}_{\ell_p}) = O_p\left(\frac{m^3 s^2}{n^{1-\gamma}}\right). \tag{40}$$

To complete the proof, notice that

$$(\hat{\theta}_{\ell_p} - \theta_0)'\partial_\theta G(\hat{\theta}_{\ell_p} - \theta_0) = (\hat{\theta}_1 - \theta_0)'\partial_\theta G(\hat{\theta}_1 - \theta_0) + O_p\left(\frac{sm^{3/2}}{n^{5/4-\gamma}}\right). \tag{41}$$

□

Theorem 5 implies that any two estimators $\hat{\theta}_{\ell_p}$ and $\hat{\theta}_{\ell_{p'}}$ yield sample moments that differ by at most $O_p\left(\frac{m^3 s^2}{n^{1-\gamma}}\right)$. Then Theorem 4 implies the following corollary.

Corollary 2. *Under Assumptions 1–3, the difference between any two solutions to* (35), *possibly corresponding to different norms, is at most* $O_p\left(\frac{m^3 s^2}{n^{1-\gamma}}\right)$ *for any small* $\gamma > 0$.

This corollary generalizes Example 2 to general conditional quantile models. In particular, Corollary 2 implies the equivalence of the exact MIP estimators proposed by Chen and Lee (2018) and Zhu (2019) for just identified models. Pouliot (2019) also proposed a dual MIP approach based on the inverse quantile regression estimator of Chernozhukov and Hansen (2006). His MIP formulation has sample moment equations constraints, which generally cannot be satisfied exactly. Therefore, the asymptotic rate of the tolerance level on the sample moment constraints would determine the order of equivalence of the dual estimator and $\hat{\theta}_{\ell_p}$.

5 Higher-order bias of exact estimators

5.1 Bias formula

Here we provide the bias formula for $\hat{\theta}_{\ell_p}$ that is a *corner solution* to program (35). We say that a solution is a corner solution if it is a corner of the polygon corresponding to the closure of the argmin set of (35). We focus on corner solutions because such solutions are the default output of exact MILP solvers. In particular, our bias formula also applies to classical QR implemented using exact linear programming algorithms. Importantly, the proof of the bias formula only relies on stochastic equicontinuity of the sample moment empirical process and does not require the stronger Hölder property in Theorem 1 and Corollary 1. As a result, the formula remains valid even if Assumption 2.2 (or the alternative conditions in Appendix E) is violated.

Note that, under the maintained assumptions, the true parameter θ_0 has an equivalent alternative definition as a solution to

$$g^*(-\theta_0) \triangleq \mathbb{E}[(1\{-Y \leq W'(-\theta_0)\} - (1-\tau))Z] = 0. \tag{42}$$

The sample analog of (42), $\hat{g}^*(-\theta)$, is different from $\hat{g}(\theta)$, and thus would typically deliver a different corner solution, which we refer to as $\hat{\theta}^*_{\ell_p}$. To resolve this apparently arbitrary choice of the corner solutions, Theorem 6 below is written in a symmetric form. Namely, the choice of the sample moment conditions, $\hat{g}(\theta)$ or $\hat{g}^*(-\theta)$, does not affect the resulting bias formula (50), which is symmetric with respect to the permutation of $(Y, \theta, \tau, \hat{\theta}_{\ell_p})$ and $(-Y, -\theta, 1-\tau, -\hat{\theta}^*_{\ell_p})$.

We use the following notation

$$\varepsilon_i \triangleq Y_i - W_i'\theta_0, \tag{43}$$

$$f_\varepsilon(0|W,Z) \triangleq f_Y(y - W'\theta_0|W,Z), \tag{44}$$

$$\kappa_1(\tau) \triangleq \left(\tau - \frac{1}{2}\right)\mathbb{E}f_\varepsilon(0|W,Z)ZW'G^{-1}Z, \tag{45}$$

$$\kappa_2(\tau) \triangleq n\mathbb{E}\frac{\hat{g}(\hat{\theta}_{\ell_p}) + \hat{g}^*(-\hat{\theta}_{\ell_p})}{2} = \frac{n}{2}\mathbb{E}Z_i 1\{\hat{\varepsilon}_i = 0\}, \tag{46}$$

$$\hat{\varepsilon}_i \triangleq Y_i - W_i'\hat{\theta}_{\ell_p}, \tag{47}$$

$$\Omega \triangleq Var[Z(1\{Y \leq W'\theta_0\} - \tau)], \tag{48}$$

$$Q_j \triangleq vec\left[(G^{-1})'\partial_\theta G_j G^{-1}\right],\ Q \text{ is a matrix with columns } Q_j, \tag{49}$$

where the operator $vec(\cdot)$ denotes the standard matrix vectorization.

Theorem 6. *Consider any corner solution $\hat{\theta}_{\ell_p}$ of program (35). Suppose that Assumptions 1, 2.1, and 3 hold. Then*

$$\mathbb{E}\hat{\theta}_{\ell_p} - \theta_0 = G^{-1}\left[\mathbb{E}\hat{g}(\hat{\theta}_{\ell_p}) - \frac{\kappa_1(\tau)}{n} - \frac{\kappa_2(\tau)}{n} - \frac{1}{2n}Q'vec(\Omega)\right] + o\left(\frac{1}{n}\right). \tag{50}$$

Proof. Following the proof of Theorem 5, Lemma B.2 implies

$$\hat{\theta}_{\ell_p} - G^{-1}\hat{g}(\hat{\theta}_{\ell_p}) = \hat{\theta}_1 - G^{-1}\left[\frac{B_n(\hat{\theta}_{\ell_p})}{\sqrt{n}} + \frac{1}{2}(\hat{\theta}_1 - \theta_0)'\partial_\theta G(\hat{\theta}_1 - \theta_0)\right] + R_n, \tag{51}$$

where $R_n = o_p\left(\frac{1}{n}\right)$ is implied by Lemma B.4. Lemma C.1, which relies on a novel non-smooth calculus argument, implies

$$\frac{1}{\sqrt{n}}\mathbb{E}B_n(\hat{\theta}_{\ell_p}) = \frac{1}{n}\kappa_1(\tau) + \frac{1}{n}\kappa_2(\tau) + o\left(\frac{1}{n}\right). \tag{52}$$

For correctly specified models, $\mathbb{E}\hat{\theta}_1 = \theta_0$ and

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_1 - \theta_0)'\partial_\theta G_j(\hat{\theta}_1 - \theta_0) &= \mathbb{E}\hat{g}'(\theta_0)(G^{-1})'\partial_\theta G_j G^{-1}\hat{g}(\theta_0) && (53)\\
&= vec((G^{-1})'\partial_\theta G_j G^{-1})'vec(Var(m(Y_i, W_i, Z_i, \theta_0))) && (54)\\
&= \frac{1}{n}Q_j' vec(\Omega). && (55)
\end{aligned}$$

The statement of the theorem follows immediately. □

The key technical ingredient in the proof of Theorem 6 is Lemma C.1, which is based on a directional Taylor expansion of the conditional PDF. Specifically, conditioning on the estimator allows us to compute a Taylor expansion of the discontinuous sample moments under the expectation operator.

The formula (50) has four components:

1. $\mathbb{E}G^{-1}\hat{g}(\hat{\theta}_{\ell_p})$, the bias from non-zero sample moments;
2. $G^{-1}\frac{\kappa_1(\tau)}{n}$, the bias from the covariance of linear influence of a single observation on $\hat{\theta}_{\ell_p}$ and the sample moments;
3. $G^{-1}\frac{\kappa_2(\tau)}{n}$, the bias from non-zero probability of $\hat{\varepsilon}_i = 0$;
4. $G^{-1}\frac{Q'vec(\Omega)}{2n}$, the bias from non-uniform conditional distribution of Y given (W, Z).

Component (1) is equal to zero when it is possible to find exact zeros to the sample moments. That is generally only possible in the univariate case when τn is an integer. Component (2)

is generally equal to zero only at the median ($\tau = \frac{1}{2}$) and is linear in τ. Component (3) is always present for corner solutions. Component (4) is typically present in most non-linear estimators with non-zero Hessian of the score function (see, for example, Rilstone et al., 1996). For example, in the simple location model used in the simulations in Section 5.3, this component captures deviations from the uniform distribution in the error term.

It is instructive to study those four components of the asymptotic bias formula in the uniform sample quantile case, where exact formulas are available.

Example 5 (Example 2 cont.)**.** *Suppose we are interested in estimating the τ-quantile of a uniformly distributed outcome variable Y. This is a special case of the general framework with $W = Z = 1$, $f_Y(y) = 1\{0 \le y \le 1\}$. We start by discussing the four components in this special case. There are two ways of defining an estimator: as a minimizer of $|\hat{g}(\theta)|$ or as a minimizer of $|\hat{g}^*(-\theta)|$, where*

$$\hat{g}(\theta) = \mathbb{E}_n(1\{Y \le \theta\} - \tau), \tag{56}$$

$$\hat{g}^*(-\theta) = \mathbb{E}_n 1\{-Y \le -\theta\} - (1 - \tau). \tag{57}$$

The derivatives of the population moment conditions $g(\theta) = g^*(-\theta) = 0$ are $G = 1$, $\partial_\theta G = 0$ and $G^* \triangleq \partial_\theta g^*(-\theta) = -1$, $\partial_\theta G^* = 0$, respectively. In either case, the closure of the argmin set will be $[Y_{(k)}, Y_{(k+1)}]$, where $k \triangleq \lfloor \tau n \rfloor$. If the fractional part $\{\tau n\} \triangleq \tau n - \lfloor \tau n \rfloor \le \frac{1}{2}$, a minimizer of $|\hat{g}(\theta)|$ ($|\hat{g}^*(-\theta)|$) is the order statistic $Y_{(k)}$ ($Y_{(k+1)}$, respectively); if $\{\tau n\} \ge \frac{1}{2}$, a minimizer of $|\hat{g}(\theta)|$ ($|\hat{g}^*(-\theta)|$) is $Y_{(k+1)}$ ($Y_{(k)}$, respectively)— on the real line $\mathbb{R}^1$, all norms $\|\cdot\|_p$, $p \in [1, \infty]$, are just the absolute value. Let us now compute the bias of the corner solutions $Y_{(k)}$ and $Y_{(k+1)}$ using formula (50) corresponding to either $\hat{g}(\theta)$ or $\hat{g}^*(-\theta)$, depending on the value of $\{\tau n\}$.

Suppose $\{\tau n\} \le \frac{1}{2}$. Then Component (1) takes forms

$$G^{-1}\hat{g}(Y_{(k)}) = \frac{k - \tau n}{n}, \tag{58}$$

$$(G^*)^{-1}\hat{g}^*(-Y_{(k+1)}) = (G^*)^{-1}\frac{n - k - (1-\tau)n}{n} = \frac{k - \tau n}{n}. \tag{59}$$

Component (2) is equal to $\frac{1}{n}(\tau - \frac{1}{2})$ for both estimators. Component (3) is

$$G^{-1}\frac{\hat{g}(Y_{(k)}) + \hat{g}^*(-Y_{(k)})}{2} = \frac{1}{2n}, \tag{60}$$

$$(G^*)^{-1}\frac{\hat{g}(Y_{(k+1)}) + \hat{g}^*(-Y_{(k+1)})}{2} = -\frac{1}{2n}. \tag{61}$$

Finally, since $\partial_\theta G = \partial_\theta G^* = 0$, Component (4) is equal to zero for both estimators. Overall,

formula (50) yields asymptotic bias expansions

$$\mathbb{E}Y_{(k)} - \tau = \frac{k - \tau n}{n} + \frac{1}{n}\left(\frac{1}{2} - \tau\right) - \frac{1}{2n} + o\left(\frac{1}{n}\right) = -\frac{\{\tau n\}}{n} - \frac{\tau}{n} + o\left(\frac{1}{n}\right), \tag{62}$$

$$\mathbb{E}Y_{(k+1)} - \tau = \frac{k - \tau n}{n} + \frac{1}{n}\left(\frac{1}{2} - \tau\right) + \frac{1}{2n} + o\left(\frac{1}{n}\right) = -\frac{\{\tau n\}}{n} + \frac{1-\tau}{n} + o\left(\frac{1}{n}\right). \tag{63}$$

The case $\{\tau n\} \geq \frac{1}{2}$ is analogous and results in the same formulas for asymptotic bias.

The exact bias formulas are given by (e.g., Ahsanullah et al., 2013)

$$\mathbb{E}Y_{(k)} - \tau = \frac{k}{n+1} - \tau = -\frac{\{\tau n\}}{n+1} - \frac{\tau}{n+1}, \tag{64}$$

$$\mathbb{E}Y_{(k+1)} - \tau = \frac{k+1}{n+1} - \tau = -\frac{\{\tau n\}}{n+1} + \frac{1-\tau}{n+1}. \tag{65}$$

Comparing these formulas with the asymptotic formulas (62) and (63), we see that they indeed coincide up to $O\left(\frac{1}{n^2}\right)$. Figure 2 illustrates the exact and the asymptotic bias formula (scaled by n) for $n = 10$. □

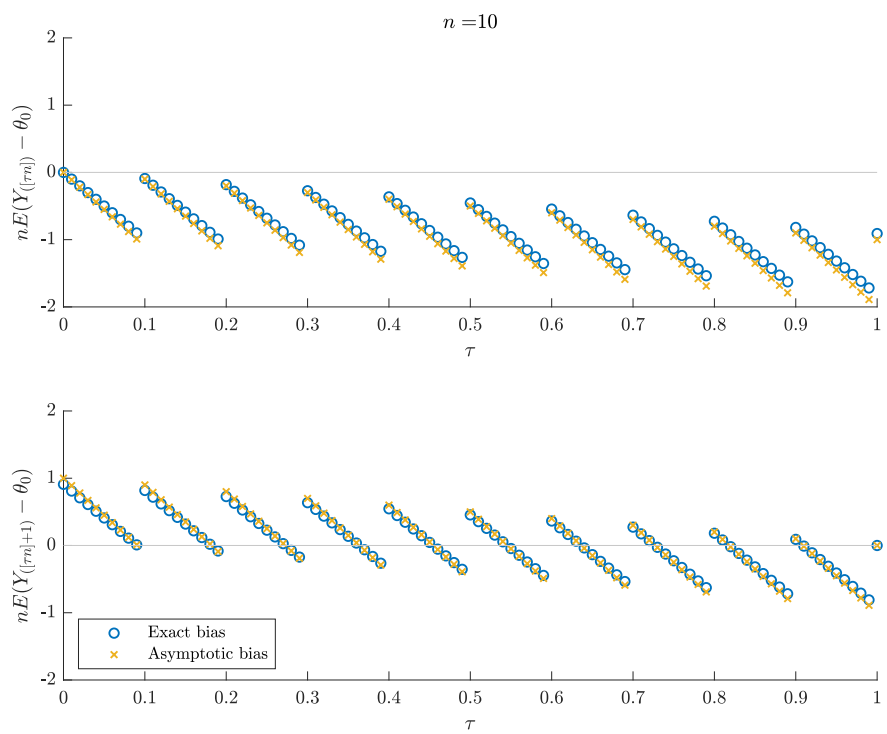

Figure 2: Exact and asymptotic bias (scaled by n) for $Y_{(\lfloor \tau n \rfloor)}$ and $Y_{(\lfloor \tau n+1 \rfloor)}$, where $Y \sim \text{Uniform}(0,1)$, $n = 10$.

There are very few results about the higher order bias of QR. Portnoy (2012) derived a near root-n Gaussian approximation for the QR process, which implies near $\frac{1}{n}$ order of the

bias. However, he did not provide explicit bias formulas. We are not aware of any results on the higher order bias of IVQR estimators. We contribute to this literature by providing the first complete formula for the bias up to $O\left(\frac{1}{n}\right)$ in for both QR and IVQR estimators.

We would like to mention an interesting approach to higher-order bias analysis of non-smooth estimators that is based on the generalized functions heuristic (e.g., Phillips, 1991). In recent work, Lee et al. (2017, 2018) derived an asymptotic bias formula for classical QR and IVQR under the assumption that the sample moments are equal to zero. Example 1 shows that this assumption is violated even in simple cases. Moreover, the generalized function approach neglects the multiplicity of solutions (cf. Example 2). As a result, the bias formulas in Lee et al. (2017, 2018) neglect Components (1) and (3) of the bias formula (50) in Theorem 6. Both the analytical Example 5 as well as the Monte Carlo evidence presented in Section 5.3 suggest that these two components contribute substantially to the higher-order bias.

Finally, our results imply that exact quantile estimators based on non-smooth moment conditions exhibit a lower bias than those based on smoothed moment condition. One can evaluate the order of bias in smoothed estimators using the higher-order MSE. Kaplan and Sun (2017) show that the higher-order MSE of IVQR estimators can be reduced using a smoothed estimating equations approach. In particular, they can guarantee that the bias is $O(n^{-\alpha})$ for some $\frac{1}{2} < \alpha < 1$, where α depends on the smoothness of the PDF of Y given (W, Z) (Section 5 in Kaplan and Sun, 2017). By contrast, Theorem 6 shows that the 1-step corrected exact non-smooth estimator has bias of order $\frac{1}{n}$, which is substantially lower than $O(n^{-\alpha})$ for any $\frac{1}{2} < \alpha < 1$. In Section 5.2, we propose a bias correction that reduces the bias of exact estimators even further.

5.2 Feasible bias correction

The theoretical bias formula (50) has a feasible counterpart that can be used for bias correction of an exact estimator $\hat{\theta}_{\ell_p}$.

Bias Components (1) and (3) depend on the realization of $\hat{\theta}_{\ell_p}$ and, thus, to our knowledge, cannot be consistently estimated using existing methods. Therefore, we consider the following symmetric 1-step correction of $\hat{\theta}_{\ell_p}$,

$$\hat{\theta}_{sym} = \hat{\theta}_{\ell_p} - G^{-1}\frac{\left(\hat{g}(\hat{\theta}_{\ell_p}) - \hat{g}^*(-\hat{\theta}_{\ell_p})\right)}{2}. \tag{66}$$

The 1-step estimator $\hat{\theta}_{sym}$ is first-order equivalent to $\hat{\theta}_{\ell_p}$ (see Lemma B.4 in the Appendix).[23]

[23] We emphasize that the symmetric one 1-step estimator is different from standard 1-step estimators as

Our key insight is that $\hat{\theta}_{sym}$ admits a concise bias formula that can be consistently estimated using sample analogs,

$$\mathbb{E}\hat{\theta}_{sym} - \theta_0 = -\frac{1}{n}G^{-1}\left[\kappa_1(\tau) + \frac{1}{2}Q'vec(\Omega)\right] + o\left(\frac{1}{n}\right). \tag{67}$$

The corresponding bias-corrected estimator is then defined as follows,

$$\hat{\theta}_{bc} = \hat{\theta}_{sym} + \frac{1}{n}G^{-1}\left[\kappa_1(\tau) + \frac{1}{2}Q'vec(\Omega)\right], \tag{68}$$

Estimators of the Jacobian G are readily available (e.g., Powell, 1986; Chernozhukov and Hansen, 2006; Angrist et al., 2006; Kato, 2012; Hong et al., 2015). The variance of the sample moments Ω can be estimated using its sample analog. The parameter $\kappa_1(\tau)$ has the same structure as G and can be estimated accordingly. The remaining parameters $\partial_\theta G_j$ take the form of weighted-average derivatives. Consistent kernel estimators of parameters with such structure were considered, for example, in Powell et al. (1989). In sum, feasible bias correction can proceed based on well-established plug-in estimators. In particular, this approach does not rely on any resampling or simulation-based methods. This is important in practice as exact estimators can be computationally quite expensive.

5.3 Monte Carlo validation of the bias formula

To validate our bias formula, we perform Monte Carlo simulations based on the following simple location model:

$$Y = W'\beta + F^{-1}(U), \quad U \mid W \sim \text{Uniform}(0,1), \tag{69}$$

where $W = (1, X)'$, $\beta = (0, 1)'$, and $X \sim \text{Uniform}(0, 1)$. In this model, the conditional quantile of Y given W is

$$Q_Y(\tau|W) = W'\theta_0(\tau), \tag{70}$$

where $\theta_0(\tau) = (F^{-1}(\tau), 1)$. We use the location model (69) because there are explicit formulas for all the bias components, allowing us to avoid any approximation errors. We consider three different choices for F:

DGP1 (Uniform)	$F(y) = \int_{-\infty}^{y} 1\{t \in [0,1]\}dt$
DGP2 (Triangular)	$F(y) = \int_{-\infty}^{y} 2t1\{t \in [0,1]\}dt$
DGP3 (Cauchy)	$F(y) = \int_{-\infty}^{y} \frac{1}{\pi(1+(4t)^2)}dt$

considered, for instance, in Zhu (2019).

The asymptotic bias formula in the case of DGP1 is expected to be precise even for small n since all the higher-order derivatives of $f_Y(y|W,Z)$ are zero. By construction, $Q = 0$ such that Component (4) of the bias formula (50) is zero. DGP2 has non-zero first derivative of $f_Y(y|W,Z)$. Unlike DGP1, this DGP has $Q \neq 0$, which allows us to assess the contribution of Component (4). At the same time, the second and higher-order derivatives of $f_Y(y|W,Z)$ are equal to zero, such that the population moments are quadratic in θ. As a result, the asymptotic formula then fully captures the shape of the population moments. In addition, $f_Y(y|W,Z)$ is asymmetric around the median. This feature is useful to illustrate the effects of small $f_Y(y|W,Z)$ (and thus large G^{-1}) on the bias. Finally, DGP3 is more complex and features all the four components of the bias. The bias correction under DGP3 is less precise than under DGP1 and DGP2 since the influence of the neglected higher-order terms of the population moments is $o\left(\frac{1}{n}\right)$. The remaining bias can potentially be quite substantial for tail quantities and for small n.

We use the MILP formulations defined in Appendix D to compute $\hat{\theta}_{\ell_1}$. We compare the bias of $\hat{\theta}_{\ell_1} = (\hat{\theta}_{\ell_1,1}, \hat{\theta}_{\ell_1,2})'$ and $\hat{\theta}_{bc} = (\hat{\theta}_{bc,1}, \hat{\theta}_{bc,2})'$ computed using formula (68). We use the sample size $n = 50$ and perform 20,000 Monte Carlo simulations for a grid of values for τ. The results of the experiments are summarized in Figure 3.

To evaluate precision of the Monte Carlo integration, we compute MCSE, standard errors based on CLT with sample size 20,000. We report $3 \times$ MCSE to account for the joint testing of 18 hypotheses. Despite its approximate nature, the asymptotic bias correction formula systematically reduces the bias across all the designs and quantile levels. In the case of DGP1, the bias correction leaves a remainder bias that is statistically indistinguishable from zero, given the number of simulations. This result is consistent with the exact comparison presented in Figure 2. The bias correction for DGP2 is very precise at the right tail of the distribution, since the asymptotic formula precisely captures the population moments. The remaining bias in the left tail is still significant.

For DGP3, the bias reduction is even more noticeable than in the previous two designs. However, for $\theta_{0,1}$, the remaining bias is still significant for most $\tau \neq 0.5$; for $\theta_{0,2}$, the remaining bias is statistically indistinguishable from zero at all but two quantiles. The comparison with DGP2 suggests that the large size of the remainder bias is partially due to the fact the asymptotic formula only captures the quadratic terms of the population moments, and partially due to the small density $f_Y(y|W,Z)$ away from the median.

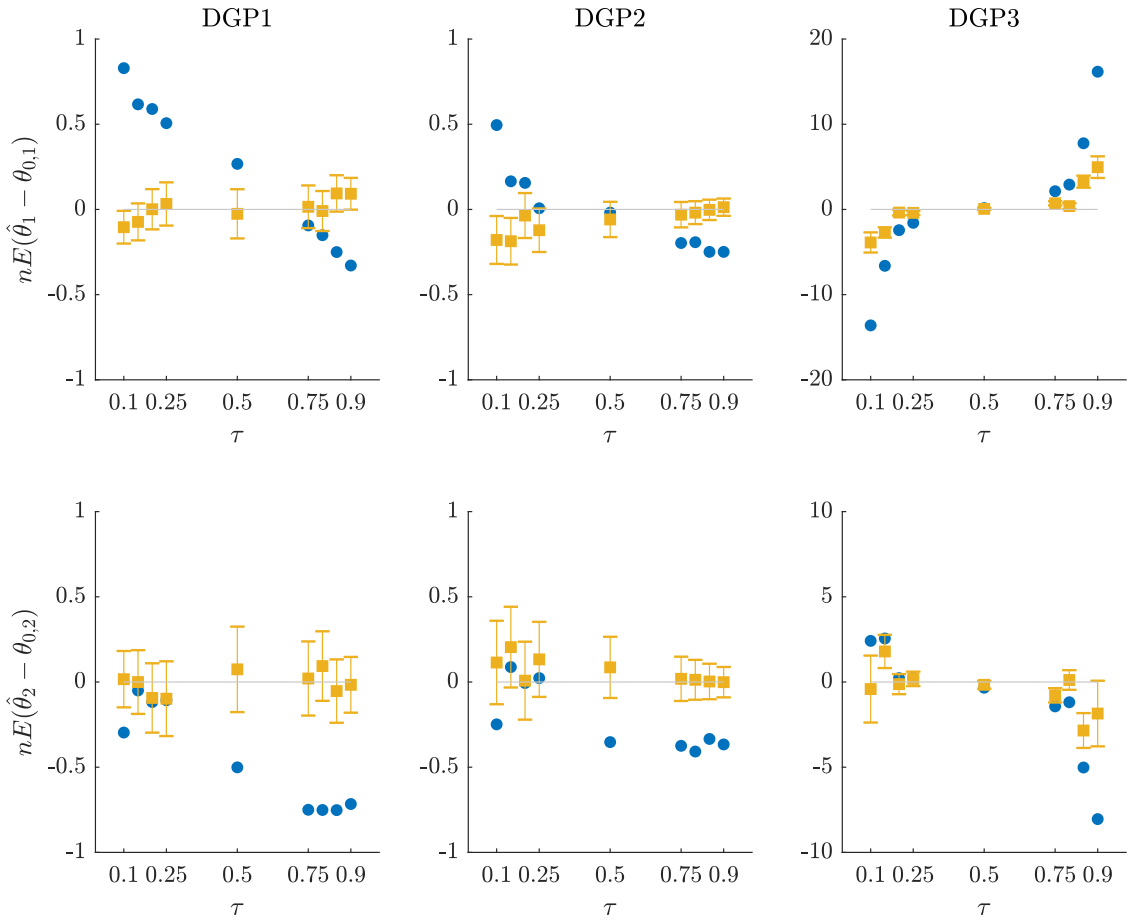

Figure 3: Bias of $\hat{\theta}_{\ell_p}$ (blue circle) and $\hat{\theta}_{bc}$ (gold squares), both scaled by n with $n = 50$ for DGP1-3. We report results for quantile levels $\tau \in \{0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 0.8, 0.85, 0.9\}$. Based on Monte Carlo simulations with 20,000 repetitions. The error bands correspond to 3 $\times$ MCSE to account for the joint testing of 18 hypotheses. Note that the scale of the y-axis for DGP3 is different.

6 Directions for future research

We would like to conclude the paper by outlining some promising avenues for future research. The seminal contributions by Robinson (1988), Rilstone et al. (1996), Andrews (2002b), and Newey and Smith (2004), among many others, have stimulated an extensive literature on small sample performance of GMM estimators with smooth sample moments. We hope that our paper will, in turn, generate interest in the small sample analysis of models with discontinuous estimating equations, which include (but are not limited to) conditional quantile models. In particular, let us outline three directions for further research.

First, an interesting extension of our results would be to the case of overidentified models

that fully utilize the information in the conditional moment restrictions implied by conditional quantile models.

Second, it would be useful to extend our higher-order expansions and bias correction formulas to panel and time series data; see for instance Anatolyev (2005) for higher-order results in the case of smooth sample moments.

Finally, our Bahadur-Kiefer expansions (Theorems 2 and 5) can be used to study and compare the higher-order properties of competing resampling procedures such as the nonparametric bootstrap, the wild bootstrap, and the pivotal bootstrap. Furthermore, we conjecture that the symmetric 1-step correction may enable convenient resampling-based bias correction procedures such as the jackknife.

Acknowledgements

We are grateful to Victor Chernozhukov, Zheng Fang, Dalia Ghanem, Jiaying Gu, Nail Kashaev, Roger Koenker, Vladimir Koltchinskii, Simon Lee, Blaise Melly, Roger Moon, Hashem Pesaran, Joris Pinkse, Wolfgang Polonik, Stephen Portnoy, Geert Ridder, Andres Santos, Yuanyuan Wan, and seminar participants at UC Berkeley, UC Davis, University of Toronto, and USC for valuable comments. All errors and omissions are our own.

References

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental variable estimates of the effect of subsidized training on the quantile of trainee earnings," *Econometrica*, 70, pp. 91–117.

Adrian, T., N. Boyarchenko, and D. Giannone (2019): "Vulnerable Growth," *American Economic Review*, 109, 1263–89.

Adrian, T. and M. K. Brunnermeier (2016): "CoVaR," *American Economic Review*, 106, 1705–41.

Ahsanullah, M., V. B. Nevzorov, and M. Shakil (2013): *An introduction to order statistics*, vol. 8, Springer.

Amemiya, T. (1982): "Two stage least absolute deviations estimators," *Econometrica: Journal of the Econometric Society*, 689–711.

Anatolyev, S. (2005): "GMM, GEL, serial correlation, and asymptotic bias," *Econometrica*, 73, 983–1002.

Andrews, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica: Journal of the Econometric Society*, 817–858.

——— (2002a): "Equivalence of the higher order asymptotic efficiency of k-step and extremum statistics," *Econometric Theory*, 18, 1040–1085.

——— (2002b): "Higher-order improvements of a computationally attractive k-step bootstrap for extremum estimators," *Econometrica*, 70, 119–162.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile regression under misspecification, with an application to the US wage structure," *Econometrica*, 74, 539–563.

Arcones, M. A. and D. M. Mason (1997): "A general approach to Bahadur-Kiefer representations for M-estimators," *Mathematical Methods of Statistics*, 6, 267–292.

Bahadur, R. R. (1966): "A note on quantiles in large samples," *The Annals of Mathematical Statistics*, 37, 577–580.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2019): "Conditional quantile processes based on series or many regressors," *Journal of Econometrics*, 213, 4 – 29, annals: In Honor of Roger Koenker.

Berthet, P., D. M. Mason, et al. (2006): "Revisiting two strong approximation results of Dudley and Philipp," in *High dimensional probability*, Institute of Mathematical Statistics, 155–172.

Bickel, P. J. (1975): "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association*, 70, 428–434.

Chen, L.-A. and S. Portnoy (1996): "Two-stage regression quantiles and two-stage trimmed least squares estimators for structural equation models," *Communications in Statistics-theory and Methods*, 25, 1005–1032.

Chen, L.-Y. and S. Lee (2018): "Exact computation of GMM estimators for instrumental variable quantile regression models," *Journal of Applied Econometrics*, 33, 553–567.

Chernozhukov, V. and I. Fernández-Val (2011): "Inference for extremal conditional quantile models, with an application to market and birthweight risks," *The Review of Economic Studies*, 78, 559–589.

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on counterfactual distributions," *Econometrica*, 81, 2205–2268.

Chernozhukov, V. and C. Hansen (2005): "An IV model of quantile treatment effects," *Econometrica*, 73, 245–261.

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.

——— (2008): "Instrumental variable quantile regression: A robust inference approach," *Journal of Econometrics*, 142, 379–398.

——— (2013): "Quantile Models with Endogeneity," *Annual Review of Economics*, 5, pp.

57–81.

Chernozhukov, V., C. Hansen, and K. Wüthrich (2017): "Instrumental Variable Quantile Regression," in *Handbook of Quantile Regression*, ed. by V. Chernozhukov, X. He, R. Koenker, and L. Peng, CRC Chapman-Hall, pp. 119–143.

Chernozhukov, V. and H. Hong (2003): "An MCMC approach to classical estimation," *Journal of Econometrics*, 115, pp. 293–346.

Chernozhukov, V., W. Newey, and A. Santos (2020): "Constrained Conditional Moment Restriction Models," Working Paper.

Csörgő, M., S. Csörgő, L. Horváth, and D. M. Mason (1986): "Weighted empirical and quantile processes," *The Annals of Probability*, 31–85.

Dhaene, G. and K. Jochmans (2015): "Split-panel jackknife estimation of fixed-effect models," *The Review of Economic Studies*, 82, 991–1030.

Dudley, R. M. (2014): *Uniform central limit theorems*, vol. 142, Cambridge university press.

Elsner, J. B., J. P. Kossin, and T. H. Jagger (2008): "The increasing intensity of the strongest tropical cyclones," *Nature*, 455, 92–95.

Fernandes, M., E. Guerre, and E. Horta (2021): "Smoothing quantile regressions," *Journal of Business & Economic Statistics*, 39, 338–357.

Fischer, M. and G. Nappo (2009): "On the moments of the modulus of continuity of Itô processes," *Stochastic Analysis and Applications*, 28, 103–122.

Foresi, S. and F. Peracchi (1995): "The conditional distribution of excess returns: An empirical analysis," *Journal of the American Statistical Association*, 90, 451–466.

Frölich, M. and B. Melly (2013): "Unconditional Quantile Treatment Effects Under Endogeneity," *Journal of Business & Economic Statistics*, 31, pp. 346–357.

Hahn, J. and W. Newey (2004): "Jackknife and analytical bias reduction for nonlinear panel models," *Econometrica*, 72, 1295–1319.

Hong, H., A. Mahajan, and D. Nekipelov (2015): "Extremum estimation and numerical derivatives," *Journal of Econometrics*, 188, 250–263.

Horowitz, J. L. (1998): "Bootstrap methods for median regression models," *Econometrica*, 1327–1351.

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

Jurečková, J. and P. K. Sen (1987): "A Second-Order Asymptotic Distributional Representation of M-Estimators with Discontinuous Score Functions," *Ann. Probab.*, 15, 814–823.

Jurečková, J., P. K. Sen, and J. Picek (2012): *Methodology in robust and nonpara-*

metric statistics, CRC Press.

Kaido, H. and K. Wüthrich (2019): "Decentralization Estimators for Instrumental Variable Quantile Regression Models," *arXiv preprint arXiv:1812.10925*.

Kallenberg, O. (2006): *Foundations of modern probability*, Springer Science & Business Media.

Kaplan, D. M. and Y. Sun (2017): "Smoothed estimating equations for instrumental variables quantile regression," *Econometric Theory*, 33, 105–157.

Kato, K. (2012): "Asymptotic normality of Powell's kernel estimator," *Annals of the Institute of Statistical Mathematics*, 64, 255–273.

Kiefer, J. (1967): "On Bahadur's Representation of Sample Quantiles," *The Annals of Mathematical Statistics*, 38, 1323–1342.

Knight, K. (2002): "Comparing conditional quantile estimators: first and second order considerations," Working Paper.

——— (2003): "On the Second Order Behaviour of the Bootstrap of L_1 Regression Estimators," *Journal of The Iranian Statistical Society*, 2, 21–42.

Koenker (2005): *Quantile Regression*, Cambridge University Press.

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Koltchinskii, V. I. (1994): "Komlós-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions," *Journal of Theoretical Probability*, 7, 73–118.

Komlós, J., P. Major, and G. Tusnády (1975): "An approximation of partial sums of independent RV'-s, and the sample DF. I," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32, 111–131.

——— (1976): "An approximation of partial sums of independent RV's, and the sample DF. II," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34, 33–58.

Lazarus, E., D. J. Lewis, and J. H. Stock (2021): "The Size-Power Tradeoff in HAR Inference," *forthcoming at Econometrica*, 1–60.

Lee, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141, pp. 1131–1158.

Lee, T.-H., A. Ullah, and H. Wang (2017): "The Second-order Bias and MSE of Quantile Estimators," *Unpublished manuscript*.

——— (2018): "The second-order bias of quantile estimators," *Economics Letters*, 173, 143–147.

Ma, L. and R. Koenker (2006): "Quantile regression methods for recursive structural equation models," *Journal of Econometrics*, 134, pp. 471 – 506.

Massart, P. (1989): "Strong approximation for multivariate empirical and related pro-

cesses, via KMT constructions," *The Annals of probability*, 266–291.

Melly, B. and K. Wüthrich (2017): "Local quantile treatment effects," in *Handbook of Quantile Regression*, ed. by V. Chernozhukov, X. He, R. Koenker, and L. Peng, CRC Chapman-Hall, 145–164.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.

Newey, K. and D. McFadden (1994): "Large sample estimation and hypothesis," *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, 2112–2245.

Newey, W. K. and R. J. Smith (2004): "Higher order properties of GMM and generalized empirical likelihood estimators," *Econometrica*, 72, 219–255.

Phillips, P. C. B. (1991): "A shortcut to LAD estimator asymptotics," *Econometric theory*, 450–463.

Pollard, D. (2002): *A user's guide to measure theoretic probability*, vol. 8, Cambridge University Press.

Portnoy, S. (2012): "Nearly root-n approximation for regression quantile processes," *The Annals of Statistics*, 40, 1714–1736.

Pouliot, G. A. (2019): "Instrumental Variables Quantile Regression with Multivariate Endogenous Variable," *Unpublished Working Paper*.

Powell, J. L. (1983): "The asymptotic normality of two-stage least absolute deviations estimators," *Econometrica: Journal of the Econometric Society*, 1569–1575.

——— (1986): "Censored regression quantiles," *Journal of econometrics*, 32, 143–155.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

Rilstone, P., V. Srivastava, and A. Ullah (1996): "The second-order bias and mean squared error of nonlinear estimators," *Journal of Econometrics*, 75, 369 – 395.

Rio, E. (1994): "Local invariance principles and their application to density estimation," *Probability Theory and Related Fields*, 98, 21–45.

Robinson, P. M. (1988): "The stochastic difference between econometric statistics," *Econometrica: Journal of the Econometric Society*, 531–548.

Ronchetti, E. and R. Sabolová (2016): "Saddlepoint tests for quantile regression," *Canadian Journal of Statistics*, 44, 271–299.

Rothe, C. and D. Wied (2013): "Misspecification testing in a class of conditional distributional models," *Journal of the American Statistical Association*, 108, 314–324.

Rothenberg, T. J. (1984): "Approximating the distributions of econometric estimators and test statistics," *Handbook of econometrics*, 2, 881–935.

Schilling, R. L. and L. Partzsch (2014): *Brownian motion: an introduction to stochas-*

tic processes, Walter de Gruyter GmbH & Co KG.

Shorack, G. R. (2017): *Probability for statisticians*, Springer.

Shorack, G. R. and J. A. Wellner (2009): *Empirical processes with applications to statistics*, SIAM.

Sun, Y., P. C. B. Phillips, and S. Jin (2008): "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing," *Econometrica*, 76, 175–194.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Science & Business Media.

Wüthrich, K. (2019): "A closed-form estimator for quantile treatment effects with endogeneity," *Journal of Econometrics*, 210, 219 – 235.

Wüthrich, K. (2020): "A Comparison of Two Quantile Models With Endogeneity," *Journal of Business & Economic Statistics*, 38, 443–456.

Zhou, K. Q. and S. L. Portnoy (1996): "Direct use of regression quantiles to construct confidence sets in linear models," *The Annals of Statistics*, 24, 287–306.

Zhu, Y. (2019): "Learning non-smooth models: instrumental variable quantile regressions and related problems," *arXiv preprint arXiv:1805.06855*.

Online appendix

A Coupling of sample moments, proofs

A.1 Proof of Theorem 1

The proof consists of four steps below which depend on auxiliary lemmas in Section A.2.

Step 1 (bound on the approximation error). First, let us split Z into a sum of a positive and negative parts $Z = Z^+ - Z^- \triangleq \max(Z, 0) - \min(-Z, 0)$. We will consider the positive part Z^+ only, since the argument for the negative part is analogous.

Denote

$$\mathcal{Z}_0 \triangleq [0, m], \tag{71}$$

$$\mathcal{Z}_{j,\ell} \triangleq \left(\frac{jm}{2^\ell}, \frac{(j+1)m}{2^\ell}\right], \quad j \in \left\{0, \ldots, 2^\ell - 1\right\}, \ \ell \geq 0. \tag{72}$$

Hence $\mathcal{Z}_{j,\ell}$ is a subinterval of $[0, m]$ with length $m2^{-\ell}$ and midpoint $m(j+\frac{1}{2})/2^\ell$. The intervals are nested such that the union of $\mathcal{Z}_{2j,\ell+1}$ and $\mathcal{Z}_{2j+1,\ell+1}$ constitutes $\mathcal{Z}_{j,\ell}$[24]. Denote the union of even numbered subintervals at level $\ell > 0$ as $\mathcal{Z}_\ell$, i.e.

$$\mathcal{Z}_\ell \triangleq \cup_{j=0}^{2^{\ell-1}-1} \mathcal{Z}_{2j,\ell}. \tag{73}$$

Using these intervals, we can define an approximating sequence of simple random variables

$$Z^{+,\bar{\ell}} \triangleq \left(1 - \sum_{\ell=1}^{\bar{\ell}} \frac{1}{2^\ell} 1\left\{Z \in \mathcal{Z}_\ell\right\}\right) m. \tag{74}$$

By construction, $|Z^+ - Z^{+,\bar{\ell}}| \leq m/2^{\bar{\ell}}$.

Hence, the empirical process $\mathbb{Z}_n^+(y, a) \triangleq \mathbb{G}_n 1\{Y \leq y, A = a\} Z^+$ can be written as

$$m^{-1}\mathbb{Z}_n^+(y, a) = \mathbb{G}_n 1\{Y \leq y, A = a, Z \in \mathcal{Z}_0\} - \sum_{\ell=1}^{\infty} 2^{-\ell} \mathbb{G}_n 1\{Y \leq y, A = a, Z \in \mathcal{Z}_\ell\} \tag{75}$$

$$\triangleq \mathbb{Z}_n^{+,0}(y, a) - \sum_{\ell=1}^{\infty} 2^{-\ell} \mathbb{Z}_n^{+,\ell}(y, a), \tag{76}$$

where the last line defines $\mathbb{Z}_n^{+,\ell}$, for $\ell \geq 0$.

By Lemmas A.2 and A.3, for each $\ell \geq 0$, there exists an approximating Brownian bridge

[24] This is the *Hungarian construction* (see, for example, p.252 in Pollard, 2002).

$\mathbb{Z}^{+,\ell}(y,a)$, such that the remainder

$$R_n^{+,\ell}(y,a) \triangleq \mathbb{Z}_n^{+,\ell}(y,a) - \mathbb{Z}^{+,\ell}(y,a) \tag{77}$$

satisfies the bound $\mathbb{E}\sup_{y\in\mathbb{R}} |R_n^{+,\ell}(y,a)| \leq \tilde{c}_1 \frac{\log n + \tilde{c}_0}{\sqrt{n}}$.

Now, with a convenient abuse of notation, define

$$\mathbb{Z}_n^{+,\bar{\ell}}(y,a) \triangleq \mathbb{Z}_n^{+,0}(y,a) - \sum_{\ell=1}^{\bar{\ell}} 2^{-\ell} \mathbb{Z}_n^{+,\ell}(y,a), \tag{78}$$

$$\mathbb{Z}^{+,\bar{\ell}}(y,a) \triangleq \mathbb{Z}^{+,0}(y,a) - \sum_{\ell=1}^{\bar{\ell}} 2^{-\ell} \mathbb{Z}^{+,\ell}(y,a). \tag{79}$$

The total remainder of the approximation of $\mathbb{Z}_n^{+,\bar{\ell}}$ by $\mathbb{Z}^{+,\bar{\ell}}$ is then

$$R_n^{+,\bar{\ell}}(y,a) \triangleq \mathbb{Z}_n^{+,\bar{\ell}}(y,a) - \mathbb{Z}^{+,\bar{\ell}}(y,a) = R_n^{+,\ell}(y,a) - \sum_{\ell=1}^{\bar{\ell}} 2^{-\ell} R_n^{+,\ell}(y,a). \tag{80}$$

By the triangle inequality,

$$\mathbb{E}\sup_{y\in\mathbb{R}} |R_n^{+,\bar{\ell}}(y,a)| \leq \sum_{\ell=0}^{\bar{\ell}} 2^{-\ell} \mathbb{E}\sup_{y\in\mathbb{R}} |R_n^{+,\ell}(y,a)| \leq \left(\sum_{\ell=0}^{\bar{\ell}} 2^{-\ell}\right) \tilde{c}_1 \frac{\log n + \tilde{c}_0}{\sqrt{n}}. \tag{81}$$

Note that the bound $\sum_{\ell=0}^{\bar{\ell}} 2^{-\ell} \leq 2$ does not depend on $\bar{\ell}$. Then we can define an approximating process

$$\mathbb{Z}^{+}(y,a) \triangleq m\, \mathbb{Z}^{+,\bar{\ell}}(y,a), \tag{82}$$

so that

$$\mathbb{Z}_n^{+}(y,a) = \mathbb{Z}^{+}(y,a) + R_n^{+}(y,a), \tag{83}$$

where

$$\mathbb{E}\sup_{y\in\mathbb{R}} |R_n^{+}(y,a)| \leq \left(\sum_{\ell=0}^{\bar{\ell}} 2^{-\ell}\right) \tilde{c}_1 \frac{\log n + \tilde{c}_0}{\sqrt{n}} m + m\mathbb{E} \sum_{\ell=\bar{\ell}+1}^{\infty} 2^{-\ell} \sup_{y\in\mathbb{R}} |\mathbb{Z}_n^{+,\ell}(y,a)| \tag{84}$$

$$\leq \left(\sum_{\ell=0}^{\bar{\ell}} 2^{-\ell}\right) \tilde{c}_1 \frac{\log n + \tilde{c}_0}{\sqrt{n}} m + m \sum_{\ell=\bar{\ell}+1}^{\infty} 2^{-\ell} \mathbb{E}\sup_{y\in\mathbb{R}} |\mathbb{Z}_n^{+,\ell}(y,a)|. \tag{85}$$

By the triangle inequality,

$$\mathbb{E}\sup_{y\in\mathbb{R}}|\mathbb{Z}_n^{+,\ell}(y,a)| \le \mathbb{E}\sup_{y\in\mathbb{R}}|\mathbb{Z}^{+,\ell}(y,a)| + \mathbb{E}\sup_{y\in\mathbb{R}}|R_n^{+,\ell}(y,a)|. \tag{86}$$

By the KMT theorem, the first term on the right is bounded by the expectation of a supremum of a standard Brownian bridge over $[0,1]$, denoted as constant c_B. The constant c_B is finite by the definition of a Brownian bridge and the tail bounds for Brownian motions (Shorack and Wellner, 2009, p.34). The second term is bounded by $\tilde{c}_1\frac{\log n+\tilde{c}_0}{\sqrt{n}}$. Note that both bounds do not depend on ℓ, and hence

$$\mathbb{E}\sup_{y\in\mathbb{R}}|R_n^{+}(y,a)| \le 2\tilde{c}_1\frac{\log n+\tilde{c}_0}{\sqrt{n}}m + m2^{-(\bar{\ell}-1)}c_B. \tag{87}$$

We can take $\bar{\ell} = 1+\log_2(\sqrt{n})$ so that

$$\mathbb{E}\sup_{y\in\mathbb{R}}|R_n^{+}(y,a)| \le 2\tilde{c}_1\frac{\log n+\tilde{c}_0+c_B/(2\tilde{c}_1)}{\sqrt{n}}m = c_1\frac{\log n+c_0}{\sqrt{n}}m, \tag{88}$$

for appropriately defined constants c_0, c_1.

The bound on the remainder of the analogous approximation for $\mathbb{Z}^-$ is the same. To complete the argument, take $\mathbb{Z}(y,a) \triangleq \mathbb{Z}^+(y,a) - \mathbb{Z}^-(y,a)$ and apply the triangle inequality to bound the total remainder.

Step 2 ($\mathbb{Z}^{+,\ell}(\cdot,a)$ is a.s. 1/2-Hölder, up to a log term). Denote $\psi(y) = P\{Y\le y, Z\in\mathcal{Z}_\ell, A=a\}$. Since $\mathbb{Z}^{+,\ell}(y,a) = B^{(\ell)}(y) - yB^{(\ell)}(1)$ for some Brownian motion $B^{(\ell)}$, we have

$$\mathbb{Z}^{+,\ell}(y_2,a) - \mathbb{Z}^{+,\ell}(y_1,a) = B^{(\ell)}(\psi(y_2)) - B^{(\ell)}(\psi(y_1)) - (\psi(y_2)-\psi(y_1))B^{(\ell)}(1). \tag{89}$$

Therefore,

$$\frac{|\mathbb{Z}^{+,\ell}(y_2,a)-\mathbb{Z}^{+,\ell}(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2-y_1|)}} \le \frac{|B^{(\ell)}(\psi(y_2))-B^{(\ell)}(\psi(y_1))|}{\sqrt{\Psi(\bar{f}|y_2-y_1|)}} + \frac{|\psi(y_2)-\psi(y_1)|\cdot|B^{(\ell)}|}{\sqrt{\Psi(\bar{f}|y_2-y_1|)}} \tag{90}$$

$$= \underbrace{\frac{|B^{(\ell)}(\psi(y_2))-B^{(\ell)}(\psi(y_1))|}{\sqrt{2\Psi(|\psi(y_2)-\psi(y_1)|)}}}_{(i)}\underbrace{\sqrt{2\cdot\frac{\Psi(|\psi(y_2)-\psi(y_1)|)}{\Psi(\bar{f}|y_2-y_1|)}}}_{(ii)} \tag{91}$$

$$+\underbrace{\sqrt{|\psi(y_2)-\psi(y_1)|}\cdot\sqrt{\frac{|\psi(y_2)-\psi(y_1)|}{\bar{f}|y_2-y_1|\log(1/\bar{f}|y_2-y_1|)}}\cdot|B^{\ell}(1)|}_{(iii)}. \tag{92}$$

The goal of this step is to derive an a.s. bound of $\limsup\sup \triangleq \limsup_{r\to 0+}\sup_{0<|y_2-y_1|\le r}$ of the left-hand side.

By Lévy's modulus of continuity theorem (Theorem 10.6 in Schilling and Partzsch, 2014), we have

$$\limsup_{r\to 0+} \sup_{x_1,x_2\in[0,1],|x_2-x_1|\le r} \frac{|B^\ell(x_2)-B^\ell(x_1)|}{\sqrt{2\Psi(r)}} = 1 \quad \text{a.s.}, \tag{93}$$

so, since $\psi(\cdot)\in[0,1]$, the lim sup sup of term (i) is a.s. bounded by 1 too.

In the second term (ii), we have, for $|y_2-y_1|$ small enough,

$$0 < \frac{\Psi(|\psi(y_2)-\psi(y_1)|)}{\Psi(\bar{f}|y_2-y_1|)} \le 1, \tag{94}$$

since $\Psi(x)$ is positive and strictly increasing in $(0,e^{-1})$ (see Lemma 10.4 in Schilling and Partzsch, 2014). Hence lim sup sup of the term (ii) is bounded by $\sqrt{2}$ a.s..

Finally, in the term (iii),

$$\sqrt{|\psi(y_2)-\psi(y_1)|} \le \sqrt{\bar{f}|y_2-y_1|}, \tag{95}$$

and therefore

$$\limsup_{r\to 0+} \sup_{0<y_2-y_1\le r} \sqrt{|\psi(y_2)-\psi(y_1)|} = 0 \quad \text{a.s.} \tag{96}$$

Since

$$\sqrt{\frac{|\psi(y_2)-\psi(y_1)|}{\bar{f}|y_2-y_1|\log(1/\bar{f}|y_2-y_1|)}} \le \sqrt{\frac{|\psi(y_2)-\psi(y_1)|}{\bar{f}|y_2-y_1|}} \le 1, \tag{97}$$

and $\limsup_{r\to 0+}\sup_{0<|y_2-y_1|\le r}|B^\ell(1)| = |B^\ell(1)|$ a.s., we have

$$\limsup_{r\to 0+} \sup_{0<|y_2-y_1|\le r} (iii) = 0 \quad \text{a.s.} \tag{98}$$

Combining the bounds for lim sup sup of terms (i), (ii) and (iii), we obtain

$$\limsup_{r\to 0+} \sup_{0<|y_2-y_1|\le r} \frac{|\mathbb{Z}^{+,\ell}(y_2,a)-\mathbb{Z}^{+,\ell}(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2-y_1|)}} \le \sqrt{2} \quad \text{a.s.} \tag{99}$$

Step 3 ($\mathbb{Z}(\cdot,a)$ is a.s. $1/2$-Hölder, up to a log term). By the definition of $\mathbb{Z}^+(y,a)$ in Equation (82),

$$m^{-1}(\mathbb{Z}^+(y_2,a)-\mathbb{Z}^+(y_1,a)) = \mathbb{Z}^{+,0}(y_2,a)-\mathbb{Z}^{+,0}(y_1,a) - \sum_{\ell=1}^{\bar{\ell}} \frac{1}{2^\ell}\cdot(\mathbb{Z}^{+,\ell}(y_2,a)-\mathbb{Z}^{+,\ell}(y_1,a)). \tag{100}$$

Dividing by $\sqrt{\Psi(\bar{f}|y_2 - y_1|)}$ and using the triangular inequality, we obtain

$$m^{-1}\frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2 - y_1|)}} \leq \frac{|\mathbb{Z}^{+,0}(y_2,a) - \mathbb{Z}^{+,0}(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2 - y_1|)}} + \sum_{\ell=1}^{\bar{\ell}} \frac{1}{2^\ell} \cdot \frac{|\mathbb{Z}^{+,\ell}(y_2,a) - \mathbb{Z}^{+,\ell}(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2 - y_1|)}}. \tag{101}$$

Taking $\limsup\sup$ of both sides and multiplying by m yields

$$\limsup_{r\to 0+} \sup_{0<y_2-y_1\leq r} \frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2 - y_1|)}} \leq m(\sqrt{2} + \sqrt{2}) = 2m\sqrt{2}. \tag{102}$$

The case of $\mathbb{Z}^-$ is analogous and yields the overall bound

$$\limsup_{r\to 0+} \sup_{0<|y_2-y_1|\leq r} \frac{|\mathbb{Z}(y_2,a) - \mathbb{Z}(y_1,a)|}{\sqrt{\Psi(\bar{f}|y_2 - y_1|)}} \leq 4m\sqrt{2}. \tag{103}$$

Step 4 ($\mathbb{Z}$ is $(1/2-\gamma)$-Hölder in expectation). First, note that, for any fixed $\gamma \in (0, \frac{1}{2})$,

$$\begin{aligned}
&\mathbb{E} \sup_{y_1\neq y_2} \frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{|y_2 - y_1|^{\frac{1}{2}-\gamma}} \\
&\leq \mathbb{E} \sup_{\substack{y_1\neq y_2 \\ \psi(y_1)\neq\psi(y_2)}} \frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{|\psi(y_2) - \psi(y_1)|^{\frac{1}{2}-\gamma}} \left(\frac{|\psi(y_2) - \psi(y_1)|}{|y_2 - y_1|}\right)^{\frac{1}{2}-\gamma} + \mathbb{E} \sup_{\substack{y_1\neq y_2 \\ \psi(y_1)=\psi(y_2)}} \frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{|y_2 - y_1|^{\frac{1}{2}-\gamma}} \\
&\leq \mathbb{E} \sup_{\psi(y_1)\neq\psi(y_2)} \frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{|\psi(y_2) - \psi(y_1)|^{\frac{1}{2}-\gamma}} \cdot \bar{f}^{\frac{1}{2}-\gamma}.
\end{aligned} \tag{104}$$

By the definition of $\mathbb{Z}^+(y,a)$ in Equation (82),

$$m^{-1}(\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)) = \mathbb{Z}^{+,0}(y_2,a) - \mathbb{Z}^{+,0}(y_1,a) - \sum_{\ell=1}^{\bar{\ell}} \frac{1}{2^\ell} \cdot (\mathbb{Z}^{+,\ell}(y_2,a) - \mathbb{Z}^{+,\ell}(y_1,a)).$$

Dividing by $|\psi(y_2) - \psi(y_1)|^{\frac{1}{2}-\gamma}$ and using the triangular inequality, we obtain

$$m^{-1}\frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{|\psi(y_2) - \psi(y_1)|^{\frac{1}{2}-\gamma}} \leq \frac{|\mathbb{Z}^{+,0}(y_2,a) - \mathbb{Z}^{+,0}(y_1,a)|}{|\psi(y_2) - \psi(y_1)|^{\frac{1}{2}-\gamma}} + \sum_{\ell=1}^{\infty} \frac{1}{2^\ell} \cdot \frac{|\mathbb{Z}^{+,\ell}(y_2,a) - \mathbb{Z}^{+,\ell}(y_1,a)|}{|\psi(y_2) - \psi(y_1)|^{\frac{1}{2}-\gamma}}. \tag{105}$$

Taking $\mathbb{E}\sup_{\psi(y_1)\neq\psi(y_2)}$ of both sides and multiplying by m yields

$$\mathbb{E} \sup_{\psi(y_1)\neq\psi(y_2)} \frac{|\mathbb{Z}^+(y_2,a) - \mathbb{Z}^+(y_1,a)|}{|\psi(y_2) - \psi(y_1)|^{\frac{1}{2}-\gamma}} \leq 2C_\gamma m < \infty, \tag{106}$$

where the bound holds by Lemma A.5, since $\mathbb{Z}^{+,\ell}$, $\ell \geq 0$, are Brownian bridges.

Combining with (104) yields

$$\mathbb{E} \sup_{y_1 \neq y_2} \frac{|\mathbb{Z}^+(y_2, a) - \mathbb{Z}^+(y_1, a)|}{|y_2 - y_1|^{\frac{1}{2}-\gamma}} \leq 2C_\gamma m \bar{f}^{\frac{1}{2}-\gamma}, \tag{107}$$

which is the definition of $(1/2 - \gamma)$-Hölder continuity of $\mathbb{Z}^+$ in expectation.

The case of $\mathbb{Z}^-$ is analogous and yields the overall bound

$$\mathbb{E} \sup_{y_1 \neq y_2} \frac{|\mathbb{Z}(y_2, a) - \mathbb{Z}(y_1, a)|}{|y_2 - y_1|^{\frac{1}{2}-\gamma}} \leq 4C_\gamma m \bar{f}^{\frac{1}{2}-\gamma}. \tag{108}$$

□

A.2 Auxiliary Lemmas

Lemma A.1. *Let Y be a r.v. with CDF $F(y)$. Then there exist a uniformly distributed r.v. V such that $F^{-1}(V) = Y$ a.s.*

Proof. This result follows immediately from Proposition 3.2 in Shorack (2017). □

Consider a simple r.v. $A \in \{0, 1\}$ defined on the same probability space with Y. Let $\mathbb{G}_n(y, a) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n (1\{Y_i \leq y, A_i = a\} - P\{Y_i \leq y, A_i = a\})$ for $a \in \{0, 1\}$.

Lemma A.2. *For all $a \in \{0, 1\}$, there exist a tight Brownian Bridge $\{B^\circ(t) : 0 \leq t \leq 1\}$ such that, for all $x \geq 0$,*

$$P\left\{\sup_{y \in \mathbb{R}} |\mathbb{G}_n(y, a) - B^\circ(P\{Y \leq y, A = a\})| \geq c_1 \frac{x + \log n}{\sqrt{n}}\right\} \leq c_0 \exp(-x), \tag{109}$$

where the constants c_1 and c_0 do not depend on n, a, and x .

Proof. First, consider the case $a = 1$. Note that for any CDF F and its left-continuous inverse F^{-1} the following holds: for any $p \in [0, 1]$ and $x \in \mathbb{R}$,

$$F^{-1}(p) \leq x \text{ if and only if } p \leq F(x).$$

Let V be the uniform r.v. such that $F^{-1}(V) = Y$ a.s. as in Lemma A.1. Then

$$\{Y \leq y, A = 1\} = \{F^{-1}(V) \leq y, A = 1\} = \{V \leq F^{-1}(y), A = 1\} = \{V + 2(1 - A) \leq F(y)\}.$$

Now let $\tilde{F}$ be the CDF of $V + 2(1 - A)$ and let U be the uniform r.v. such that $\tilde{F}^{-1}(U) =$

$V + 2(1-A)$ a.s. as in Lemma A.1. Then

$$\{V + 2(1-A) \le F(y)\} = \{\tilde{F}^{-1}(U) \le F(y)\} = \{U \le \tilde{F}(F(y))\}$$

By the KMT theorem, there exist a tight Brownian bridge $B^\circ(u)$, such that

$$R_n \triangleq \sup_{0\le u\le 1} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1\{U_i \le u\} - P\{U_i \le u\}) - B^\circ(u) \right| \tag{110}$$

is uniformly tight and satisfies, for any $x > 0$,

$$P\left\{ R_n \ge c_1 \frac{x + \log n}{\sqrt{n}} \right\} \le c_0 \exp(-x), \tag{111}$$

where constants c_1 and c_0 depend neither on x nor n.

Therefore,

$$\begin{aligned} R_n^Y &\triangleq \sup_{y\in\mathbb{R}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1\{Y_i \le y, A_i = 1\} - P\{Y_i \le y, A_i = 1\}) - B^\circ(P\{Y_i \le y, A_i = 1\}) \right| \\ &= \sup_{y\in\mathbb{R}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1\{U_i \le \tilde{F}(F(y))\} - P\{U_i \le \tilde{F}(F(y))\}) - B^\circ(P\{U_i \le \tilde{F}(F(y))\}) \right| \\ &\le \sup_{0\le u\le 1} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1\{U_i \le u\} - P\{U_i \le u\}) - B^\circ(u) \right| = R_n \end{aligned}$$

This establishes that $R_n^Y \le R_n$ a.s., which means that R_n^Y also satisfies the tail bound (111).

The case $a = 0$ is analogous since $A = 0$ is equivalent to $A^c = 1$, where $A^c = 1 - A$. This completes the proof.

□

Lemma A.3. *Let a sequence of r.v.* $R_n \ge 0$ *satisfy*

$$P\left\{ R_n \ge c_1 \frac{x + \log n}{\sqrt{n}} \right\} \le c_0 \exp(-x) \text{ for all } x \ge 0. \tag{112}$$

Then

$$\mathbb{E} R_n \le \tilde{c}_1 \frac{\log n + \tilde{c}_0}{\sqrt{n}}. \tag{113}$$

Proof. Define $\xi_n \triangleq \frac{\sqrt{n}}{c_1} R_n$ and consider the following decomposition

$$\mathbb{E}\xi_n = \mathbb{E}\xi_n 1\{\xi_n \leq \log n\} + \mathbb{E}\xi_n 1\{\xi_n > \log n\}.$$

The first term is bounded by $\log n$. The second term can be bounded as follows:

$$\begin{aligned}\mathbb{E}\xi_n 1\{\xi_n > \log n\} &= \mathbb{E}(\xi_n - \log n)1\{\xi_n - \log n > 0\} + \log n \cdot P(\xi_n - \log n > 0) \\ &\leq \mathbb{E}\zeta_n + c_0 \log n,\end{aligned}$$

where $\zeta_n = (\xi_n - \log n)1\{\xi_n - \log n > 0\}$.

By assumption of the lemma, for $x \geq 0$,

$$P\left\{\zeta_n \geq x\right\} = P\left\{R_n \geq c_1 \frac{x + \log n}{\sqrt{n}}\right\} \leq c_0 \exp(-x) \text{ for all } x \geq 0,$$

Therefore, by the Fubini theorem (e.g., Kallenberg, 2006, Lemma 3.4), $\mathbb{E}\zeta_n \leq c_0$. Overall, we obtain

$$\mathbb{E}\xi_n \leq \log n + c_0 + c_0 \log n = c_0 + (1 + c_0) \log n \tag{114}$$

or, substituting the expression for ξ_n,

$$\mathbb{E}R_n \leq \frac{c_1(1 + c_0) \log n + c_1 c_0}{\sqrt{n}}. \tag{115}$$

Taking $\tilde{c}_1 = c_1(1 + c_0)$ and $\tilde{c}_0 = c_0/(1 + c_0)$ completes the proof. □

Lemma A.4. *If a sequence of stochastic processes $\theta \mapsto X_n(\theta)$ defined on $\Theta \subset \mathbb{R}^k$ satisfies the bound*

$$\mathbb{E} \sup_{\theta' \neq \theta} \frac{|X_n(\theta') - X_n(\theta)|}{||\theta' - \theta||^{\frac{1}{2} - \gamma}} \leq C_\gamma < \infty \tag{116}$$

for some constant C_γ that does not depend on n, then for any random sequences θ'_n, θ_n we have

$$X_n(\theta'_n) - X_n(\theta_n) = O_p\left(||\theta'_n - \theta_n||^{\frac{1}{2} - \gamma}\right). \tag{117}$$

Proof. For $M > 0$, Markov's inequality implies

$$P\left(|X_n(\theta'_n) - X_n(\theta_n)| > M||\theta'_n - \theta_n||^{\frac{1}{2} - \gamma}\right) \leq P\left(\sup_{\theta' \neq \theta} \frac{|X_n(\theta') - X_n(\theta)|}{||\theta' - \theta||^{\frac{1}{2} - \gamma}} > M\right) \tag{118}$$

$$\leq M^{-1} \mathbb{E} \sup_{\theta' \neq \theta} \frac{|X_n(\theta') - X_n(\theta)|}{||\theta' - \theta||^{\frac{1}{2} - \gamma}} \leq M^{-1} C_\gamma \tag{119}$$

Therefore, the left-hand side can be made arbitrarily small (for all n) by choosing large enough M. The statement of the theorem follows. □

Lemma A.5. *Any sequence of Brownian bridges $t \mapsto BB_n(t)$ defined on $\Theta = [0,1]$ satisfies the bound* (116).

Proof. Since $BB_n(t) = BM_n(t) - tBM_n(1)$ for some Brownian motion $BM_n(\cdot)$, we have

$$BB_n(t') - BB_n(t) = BM_n(t') - BM_n(t) - (t'-t)BM_n(1). \tag{120}$$

For the first term in (120), we have

$$\mathbb{E}\sup_{t'\neq t}\frac{|BM_n(t') - BM_n(t)|}{|t'-t|^{\frac{1}{2}-\gamma}} \leq \mathbb{E}\sup_{0<|t'-t|<1}\frac{|BM_n(t') - BM_n(t)|}{|t'-t|^{\frac{1}{2}-\gamma}} + \mathbb{E}|BM_n(1) - BM_n(0)| < \infty, \tag{121}$$

where the last inequality holds since the first term on the right is finite by Theorem 10.1 and Corollary 10.2 in Schilling and Partzsch (2014).

For the second term in (120), we have

$$\mathbb{E}\sup_{t'\neq t}\frac{|(t'-t)BM_n(1)|}{|t'-t|^{\frac{1}{2}-\gamma}} = \mathbb{E}\sup_{t'\neq t}|t'-t|^{\frac{1}{2}-\gamma}|BM_n(1)| \leq \mathbb{E}|BM_n(1)| < \infty. \tag{122}$$

Combining (121) and (122) yields

$$\mathbb{E}\sup_{t'\neq t}\frac{|BB_n(t') - BB_n(t)|}{|t'-t|^{\frac{1}{2}-\gamma}} \leq \mathbb{E}\sup_{t'\neq t}\frac{|BM_n(t') - BM_n(t)|}{|t'-t|^{\frac{1}{2}-\gamma}} + \mathbb{E}\sup_{t'\neq t}\frac{|(t'-t)BM_n(1)|}{|t'-t|^{\frac{1}{2}-\gamma}} < \infty. \tag{123}$$

The conclusion follows since no bounds in this proof depend on n. □

B Bahadur-Kiefer representation, proofs

Lemma B.1. *Under Assumptions 2.1 and 3.2, $g(\theta)$ is three times continuously differentiable.*

Proof. By definition, $g(\theta) = \mathbb{E}(1\{Y \leq W'\theta\} - \tau)Z = \mathbb{E}(\mathbb{E}(F_Y(W'\theta|W, Z) - \tau)Z)$. The result then follows from the dominated convergence theorem. □

Lemma B.2. *Suppose Assumptions 2.1 and 3.2 hold. Then for any estimator $\hat{\theta}$, we have a representation*

$$\begin{aligned}\hat{g}(\hat{\theta}) = &\frac{1}{\sqrt{n}}B_n^\circ(\theta_0) + \tau(\mathbb{E}Z - \mathbb{E}_n Z) + \frac{1}{\sqrt{n}}B_n(\hat{\theta}) \\ &+ g(\theta_0) + G(\theta_0)(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)'\partial_\theta G(\theta_0)(\hat{\theta} - \theta_0) + O_p\left(\|\hat{\theta} - \theta_0\|^3\right). \end{aligned} \tag{124}$$

Proof. By definition

$$\hat{g}(\hat{\theta}) = \mathbb{E}_n 1\{Y \leq W'\hat{\theta}\}Z - \tau\mathbb{E}_n Z, \tag{125}$$

$$= \frac{1}{\sqrt{n}}B_n^\circ(\hat{\theta}) + g^\circ(\hat{\theta}) - \tau\mathbb{E}_n Z \tag{126}$$

$$= \frac{1}{\sqrt{n}}B_n^\circ(\theta_0) + \frac{1}{\sqrt{n}}B_n(\hat{\theta}) + \tau(\mathbb{E}Z - \mathbb{E}_n Z) + g(\hat{\theta}) \tag{127}$$

By Lemma B.1, $g(\cdot)$ is three times continuously differentiable. The Taylor theorem implies that there exist a neighborhood of θ_0 such that for any θ in the neighborhood,

$$g(\theta) = g(\theta_0) + G(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)'\partial_\theta G(\theta_0)(\theta - \theta_0)) + R(\theta), \tag{128}$$

where $R(\theta) = O\left(\|\theta - \theta_0\|^3\right)$. Then (124) follows immediately. □

Lemma B.3. *Under Assumptions 1.1, 2.1, and 3.1, any estimator $\hat{\theta}_{\ell_p}$ that minimizes $\theta \mapsto \|\hat{g}(\theta)\|_p$ satisfies $\hat{\theta}_{\ell_p} \xrightarrow{p} \theta_0$.*

Proof. By Assumption 1.1,

$$\arg\min_{\theta\in\Theta} \|g(\theta)\|_p = \theta_0. \tag{129}$$

Assumptions 2.1 and 3.1 imply that the function class $\theta \mapsto Z_i(1\{Y_i \leq W_i'\theta\} - \tau)$ is Donsker and thus Glivenko-Cantelli, and hence[25]

$$\sup_{\theta\in\Theta} |\hat{g}(\theta) - g(\theta)| = \sup_{\theta\in\Theta} |(\mathbb{E}_n - \mathbb{E})Z_i(1\{Y_i \leq W'\theta\} - \tau)| \xrightarrow{a.s.} 0. \tag{130}$$

[25] See Chernozhukov and Hansen (2006, Lemma B.2) for a more detailed discussion.

By the argmin theorem (Theorem 2.1 in Newey and McFadden, 1994), applied to $Q_n(\theta) = \|\hat{g}(\theta)\|_p$, we get $\hat{\theta}_{\ell_p} \xrightarrow{p} \theta_0$. □

Lemma B.4. *Under Assumptions 1, 2.1, and 3, for any estimator $\hat{\theta}_{\ell_p}$ that minimizes $\|\hat{g}(\theta)\|_p$, we have*

$$\hat{\theta}_{\ell_p} = \hat{\theta}_1 + o_p\left(\frac{1}{\sqrt{n}}\right), \tag{131}$$

$$\|\hat{g}(\hat{\theta}_{\ell_p})\|_p = o_p\left(\frac{1}{\sqrt{n}}\right), \tag{132}$$

where $\hat{\theta}_1$ is introduced in equation (27).

Proof. The proof proceeds in four steps.

Step 1. Notice that under the assumptions of the lemma, the empirical process $B_n^\circ(\theta)$ is Donsker and stochastically equicontinuous (see Chernozhukov and Hansen, 2006, Lemma B.2).

Step 2. By definition, $\hat{\theta}_1$ can be written as

$$\hat{\theta}_1 = \theta_0 - G^{-1}\left[\tau(\mathbb{E}Z - \mathbb{E}_n Z) + \frac{1}{\sqrt{n}} B_n^\circ(\theta_0)\right] = \theta_0 + O_p\left(\frac{1}{\sqrt{n}}\right), \tag{133}$$

where θ_0 and $G^{-1} \triangleq \partial g(\theta_0)$ are well-defined by Assumption 1 and CLT holds as an implication of Assumption 2.1.

By Lemma B.2 and Step 1,

$$\begin{aligned}\hat{g}(\hat{\theta}_1) &= \frac{1}{\sqrt{n}} B_n^\circ(\theta_0) + \tau(\mathbb{E}Z - \mathbb{E}_n Z) + \frac{1}{\sqrt{n}} B_n(\hat{\theta}_1) \\ &\quad + g(\theta_0) + G(\theta_0)(\hat{\theta}_1 - \theta_0) + \frac{1}{2}(\hat{\theta}_1 - \theta_0)'\partial_\theta G(\theta_0)(\hat{\theta}_1 - \theta_0) + O_p\left(n^{-\frac{3}{2}}\right) \qquad (134) \\ &= \frac{1}{\sqrt{n}} B_n(\hat{\theta}_1) + O_p\left(\frac{1}{n}\right) = O_p\left(n^{-\frac{1}{2}}\right). \qquad (135)\end{aligned}$$

Since $\hat{\theta}_{\ell_p}$ is defined as the estimator with the minimal norm,

$$\|\hat{g}(\hat{\theta}_{\ell_p})\|_p \leq \|\hat{g}(\hat{\theta}_1)\|_p = O_p\left(n^{-\frac{1}{2}}\right). \tag{136}$$

Step 3. By Lemma B.2 and Steps 1 and 2, $\hat{\theta}_{\ell_p}$ satisfies

$$\begin{aligned}
&G(\theta_0)(\hat{\theta}_{\ell_p} - \theta_0) + \frac{1}{2}(\hat{\theta}_{\ell_p} - \theta_0)'\partial_\theta G(\theta_0)(\hat{\theta}_{\ell_p} - \theta_0) \\
&= \hat{g}(\hat{\theta}_{\ell_p}) - \frac{1}{\sqrt{n}} B_n^\circ(\theta_0) - \tau(\mathbb{E}Z - \mathbb{E}_n Z) - \frac{1}{\sqrt{n}} B_n(\hat{\theta}_{\ell_p}) + O_p\left(\|\hat{\theta}_{\ell_p} - \theta_0\|^3\right) \qquad (137) \\
&= O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\|\hat{\theta}_{\ell_p} - \theta_0\|^3\right). \qquad (138)
\end{aligned}$$

By Lemma B.3, $\|\hat{\theta}_{\ell_p} - \theta_0\| \xrightarrow{p} 0$. By Assumption 1.2,

$$\hat{\theta}_{\ell_p} - \theta_0 + O_p\left(\|\hat{\theta}_{\ell_p} - \theta_0\|^2\right) = O_p\left(\frac{1}{\sqrt{n}}\right), \qquad (139)$$

$$O_p\left(\|\hat{\theta}_{\ell_p} - \theta_0\|\right) = O_p\left(\frac{1}{\sqrt{n}}\right), \qquad (140)$$

which implies $\hat{\theta}_{\ell_p} = \theta_0 + O_p\left(\frac{1}{\sqrt{n}}\right)$. So consistency of $\hat{\theta}_{\ell_p}$ implies by Step 1 that

$$B_n(\hat{\theta}_{\ell_p}) = B_n^\circ(\hat{\theta}_{\ell_p}) - B_n^\circ(\theta_0) = o_p(1). \qquad (141)$$

Step 4. Consider $\hat{\theta}_2 \triangleq \hat{\theta}_1 - G^{-1}\frac{B_n(\hat{\theta}_1)}{\sqrt{n}} = \hat{\theta}_1 + o_p\left(\frac{1}{\sqrt{n}}\right)$. Then by the stochastic equicontinuity of B_n and Lemma B.2,

$$\hat{g}(\hat{\theta}_2) = \frac{B_n(\hat{\theta}_2) - B_n(\hat{\theta}_1)}{\sqrt{n}} + o_p\left(\frac{1}{\sqrt{n}}\right) = o_p\left(\frac{1}{\sqrt{n}}\right). \qquad (142)$$

This implies

$$\|\hat{g}(\hat{\theta}_{\ell_p})\|_p \leq \|\hat{g}(\hat{\theta}_2)\|_p = o_p\left(n^{-\frac{1}{2}}\right). \qquad (143)$$

The remaining result then follows from (138).

□

Lemma B.5. *Suppose that Assumptions 2 and 3.1 hold.*[26] *For any pair of estimators* $\hat{\theta}, \hat{\theta}^* \in \Theta$*, the following property holds*

$$B_n(\hat{\theta}) - B_n(\hat{\theta}^*) = O_p\left(sm(m\bar{f}\|\hat{\theta} - \hat{\theta}^*\|)^{\frac{1}{2}-\gamma}\right) + O_p\left(\frac{sm\log n}{\sqrt{n}}\right). \qquad (144)$$

Proof. In view of Corollary 1, the proof is analogous to the argument in Example 3. □

[26] See Appendix E for an alternative to this lemma that does not rely on Assumption 2.

Lemma B.6. *Suppose Assumptions 1.2, 2 and 3.1 hold. Then for any estimators $\hat{\theta}$ and $\hat{\theta}^*$ that satisfy*

$$\hat{\theta} - \hat{\theta}^* = -G^{-1}(\theta_0)\frac{B_n(\hat{\theta}) - B_n(\hat{\theta}^*)}{\sqrt{n}} + \frac{R_n}{\sqrt{n}}, \tag{145}$$

for some sequence of random vectors $R_n = O_p(1)$, we have, for any small positive γ,

$$\hat{\theta} - \hat{\theta}^* = O_p\left(\frac{m^3 s^2 \bar{f}}{n^{1-\gamma}}\right) + O_p\left(\frac{\|R_n\|}{\sqrt{n}}\right). \tag{146}$$

Proof. By definition

$$B_n(\hat{\theta}) - B_n(\hat{\theta}^*) = B_n^\circ(\hat{\theta}) - B_n^\circ(\hat{\theta}^*). \tag{147}$$

By Lemma B.5, for any $\gamma \in (0, \frac{1}{2})$,

$$B_n(\hat{\theta}) - B_n(\hat{\theta}^*) = O_p\left(sm(m\bar{f}\|\hat{\theta} - \hat{\theta}^*\|)^{\frac{1}{2}-\gamma}\right) + O_p\left(sm\frac{\log n}{\sqrt{n}}\right). \tag{148}$$

By equation (145), we get

$$\hat{\theta} - \hat{\theta}^* = O_p\left(\frac{sm(m\bar{f}\|\hat{\theta} - \hat{\theta}^*\|)^{\frac{1}{2}-\gamma}}{\sqrt{n}}\right) + \frac{\zeta_n}{\sqrt{n}}. \tag{149}$$

where $\zeta_n = O_p\left(sm\frac{\log n}{\sqrt{n}}\right) + R_n$. There are two possibilities. If ζ_n converges to zero slower than $O_p\left(sm(m\bar{f}\|\hat{\theta} - \hat{\theta}^*\|)^{\frac{1}{2}-\gamma}\right)$, then, according to the big-O notation,

$$\hat{\theta} - \hat{\theta}^* = O_p\left(\frac{\zeta_n}{\sqrt{n}}\right), \tag{150}$$

which implies the result of the lemma since $\frac{\log n}{n} = o(n^{-(1-\gamma)})$. The other possibility is $\zeta_n = o_p\left(sm\sqrt{\bar{f}\|\hat{\theta} - \hat{\theta}^*\|}\right)$. Then we can ignore ζ_n, i.e.

$$\|\hat{\theta} - \hat{\theta}^*\| = \frac{sm(m\bar{f}\|\hat{\theta} - \hat{\theta}^*\|)^{\frac{1}{2}-\gamma}}{\sqrt{n}} O_p(1), \tag{151}$$

which implies

$$\|\hat{\theta} - \hat{\theta}^*\| = \left(O_p\left(\frac{m^{\frac{3}{2}-\gamma} s \bar{f}^{\frac{1}{2}-\gamma}}{\sqrt{n}}\right)\right)^{\frac{1}{1/2+\gamma}} = O_p\left(m^{\frac{3-2\gamma}{1+2\gamma}} s^{\frac{2}{1+2\gamma}} \bar{f}^{\frac{1-2\gamma}{1+2\gamma}} n^{-\frac{1}{1+2\gamma}}\right). \tag{152}$$

This again implies equation (150) and completes the proof. □

Lemma B.7. *Under Assumptions 1–2, for any estimator $\hat{\theta}$ that minimizes $\|\hat{g}(\theta)\|_p$, we have, for all small enough $\gamma > 0$,*

$$\hat{g}(\hat{\theta}_{\ell_p}) = O_p\left(\frac{m^3 s^2}{n^{1-\gamma}}\right). \tag{153}$$

Proof. The proof proceeds in three steps. Throughout the proof we will treat $\bar{f}$ as a constant to simplify algebra.

Step 1. By Lemmas B.5 and B.4,

$$B_n(\hat{\theta}_{\ell_p}) = O_p\left(sm(mn^{-1/2})^{\frac{1}{2}-\gamma}\right) + O_p\left(\frac{sm\log n}{\sqrt{n}}\right) = O_p\left(sm^{\frac{3}{2}}\frac{1}{n^{\frac{1-2\gamma}{4}}}\right). \tag{154}$$

Step 2. Consider the following estimator

$$\hat{\theta}_2 \triangleq \hat{\theta}_1 - \frac{G^{-1}B_n(\hat{\theta}_{\ell_p})}{\sqrt{n}} \tag{155}$$

By Lemma B.2 , we get

$$\hat{g}(\hat{\theta}_2) = \frac{1}{\sqrt{n}}B_n^{\circ}(\theta_0) + (\tau\mathbb{E}Z - \tau\mathbb{E}_n Z) + \frac{1}{\sqrt{n}}B_n(\hat{\theta}_2) \tag{156}$$

$$+ g(\theta_0) + G(\theta_0)(\hat{\theta}_2 - \theta_0) + (\hat{\theta}_2 - \theta_0)'\frac{\partial G(\theta_0)}{\partial\theta}(\hat{\theta}_2 - \theta_0) + O_p\left(\frac{1}{n^{3/2}}\right), \tag{157}$$

Then, by definition of $\hat{\theta}_2$,

$$\hat{g}(\hat{\theta}_2) = \frac{B_n(\hat{\theta}_2) - B_n(\hat{\theta}_{\ell_p})}{\sqrt{n}} + (\hat{\theta}_2 - \theta_0)'\frac{\partial G(\theta_0)}{\partial\theta}(\hat{\theta}_2 - \theta_0) + O_p\left(\frac{1}{n^{3/2}}\right). \tag{158}$$

Also, by definition,

$$B_n(\hat{\theta}_2) - B_n(\hat{\theta}_{\ell_p}) = B_n^{\circ}(\hat{\theta}_2) - B_n^{\circ}(\hat{\theta}_{\ell_p}). \tag{159}$$

By Lemma B.5,

$$B_n(\hat{\theta}_2) - B_n(\hat{\theta}_{\ell_p}) = O_p\left(sm^{\frac{3}{2}}\|\hat{\theta}_2 - \hat{\theta}_{\ell_p}\|^{\frac{1}{2}-\gamma}\right) + O_p\left(sm\frac{\log n}{\sqrt{n}}\right). \tag{160}$$

Then (158) becomes

$$\hat{g}(\hat{\theta}_2) = O_p\left(\frac{sm^{\frac{3}{2}}\|\hat{\theta}_2 - \hat{\theta}_{\ell_p}\|^{\frac{1}{2}-\gamma}}{\sqrt{n}}\right) + O_p\left(sm\frac{\log n}{n}\right). \tag{161}$$

By Lemma B.2 applied to $\hat{\theta}_{\ell_p}$ and the definition of $\hat{\theta}_1$,

$$\hat{\theta}_{\ell_p} = \hat{\theta}_1 + G^{-1}\hat{g}(\hat{\theta}_{\ell_p}) - \frac{G^{-1}B_n(\hat{\theta}_{\ell_p})}{\sqrt{n}} - G^{-1}(\hat{\theta}_{\ell_p} - \theta_0)'\frac{\partial G(\theta_0)}{\partial \theta}(\hat{\theta}_{\ell_p} - \theta_0) + O_p\left(\frac{1}{n^{3/2}}\right). \quad (162)$$

So by (157) and the definition of $\hat{\theta}_2$, we get

$$\hat{\theta}_{\ell_p} - \hat{\theta}_2 = G^{-1}\hat{g}(\hat{\theta}_{\ell_p}) + O_p(n^{-1}). \quad (163)$$

So (161) becomes

$$\hat{g}(\hat{\theta}_2) = O_p\left(\frac{sm^{\frac{3}{2}}}{\sqrt{n}}\|\hat{g}(\hat{\theta}_{\ell_p})\|^{\frac{1}{2}-\gamma}\right) + O_p\left(sm\frac{\log n}{n}\right). \quad (164)$$

Step 3. From (164), we obtain

$$\|\hat{g}(\hat{\theta}_{\ell_p})\|_p \leq \|\hat{g}(\hat{\theta}_2)\|_p = O_p\left(\frac{sm^{\frac{3}{2}}}{\sqrt{n}}\|\hat{g}(\hat{\theta}_{\ell_p})\|^{\frac{1}{2}-\gamma}\right) + O_p\left(sm\frac{\log n}{n}\right). \quad (165)$$

On the right hand side of this inequality, suppose that the first term dominates the second term. Then $\|\hat{g}(\hat{\theta}_{\ell_p})\| = O_p\left(sm\frac{\log n}{n}\right)$, which has order no larger than $O\left(\frac{m^3s^2}{n^{1-\gamma}}\right)$, from which the statement of the lemma follows.

Otherwise, if the second term dominates the first term, we have

$$\|\hat{g}(\hat{\theta}_{\ell_p})\|^{\frac{1}{2}+\gamma} \leq O_p\left(\frac{sm^{\frac{3}{2}}}{\sqrt{n}}\right) \quad (166)$$

or, after exponentiation,

$$\|\hat{g}(\hat{\theta}_{\ell_p})\| \leq O_p\left(\frac{s^{\frac{2}{1+2\gamma}}m^{\frac{3}{1+2\gamma}}}{n^{\frac{1}{1+2\gamma}}}\right) \leq O_p\left(\frac{s^2m^3}{n^{1-\gamma}}\right), \quad (167)$$

which implies the statement of the lemma.

□

C Higher-order bias, proofs

Lemma C.1. *Consider a corner solution $\hat{\theta} = \hat{\theta}_{\ell_p}$. Under Assumptions 1, 2.1, and 3, we have*

$$\mathbb{E}\frac{1}{\sqrt{n}}B_n(\hat{\theta}) = \frac{1}{n}\kappa(\tau) + o\left(\frac{1}{n}\right) \tag{168}$$

where

$$\kappa(\tau) \triangleq \mathbb{E}\left(\tau - \frac{1}{2}\right) f_\varepsilon(0|W,Z)ZW'G^{-1}Z + n\mathbb{E}\left(\frac{\hat{g}(\hat{\theta}) + \hat{g}^*(-\hat{\theta})}{2}\right). \tag{169}$$

Proof. The proof proceeds in five steps.

Step 1. Note that

$$\frac{1}{\sqrt{n}}\mathbb{E}B_n(\hat{\theta}) = \frac{1}{\sqrt{n}}\mathbb{E}\left(B_n^\circ(\hat{\theta}) - B_n^\circ(\theta_0)\right) \tag{170}$$

$$= \mathbb{E}\left(1\{Y \le W'\hat{\theta}\}Z\right) - \mathbb{E}g^\circ(\hat{\theta}). \tag{171}$$

Step 2. Define $\hat{\varepsilon}_i \triangleq Y_i - W_i'\hat{\theta}$ and split the first term in the equation above as follows:

$$\mathbb{E}1\{Y_i \le W_i'\hat{\theta}\}Z_i = \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i + \mathbb{E}1\{\hat{\varepsilon}_i < 0\}Z_i. \tag{172}$$

Lemma B.4 implies

$$\hat{\theta} = \theta_0 - \frac{1}{n}G^{-1}\sum_{i=1}^{n}\left(1\{Y_i \le W_i'\theta_0\} - \tau\right)Z_i + R_n, \tag{173}$$

where $R_n = o_p(n^{-1/2})$. We can use this structure to isolate an influence of observation i, $\frac{1}{n}\lambda_i \triangleq -\frac{1}{n}W_i'G^{-1}Z_i(1\{Y_i \le W_i'\theta_0\} - \tau)$. The indicator $1\{\hat{\varepsilon}_i < 0\}$ can be rewritten as

$$1\left\{Y_i < W_i'\hat{\theta}_{-i} + \frac{1}{n}\lambda_i\right\}, \tag{174}$$

where

$$\hat{\theta}_{-i} \triangleq \theta_0 - \frac{1}{n}G^{-1}\sum_{j=1, j\neq i}^{n}\left(1\{Y_j \le W_j'\theta_0\} - \tau\right)Z_i \tag{175}$$

is equal to $\hat{\theta}$ without the linear influence of the observation i. Then, using Taylor's theorem, the term $\mathbb{E}Z_iP(Y_i < W_i'\hat{\theta}_{-i} + \frac{1}{n}\lambda_i|1\{Y_i \le W_i'\theta_0\}, Z_i, W_i)$ can be represented as

$$\mathbb{E}Z_iP(Y_i < W_i'\hat{\theta}_{-i}|Z_i, W_i) + \mathbb{E}\frac{1}{n}Z_i\lambda_i f_Y(W_i'\hat{\theta}_{-i}|\hat{\theta}_{-i}, \lambda_i, Z_i, W_i) + O\left(\frac{1}{n^2}\right). \tag{176}$$

By Assumption 3, $f_Y(y|W_i, Z_i)$ is uniformly bounded and

$$\mathbb{E}\phi(W_i, Z_i) f_Y(y|\hat{\theta}_{-i}, \lambda_i, W_i, Z_i) = \mathbb{E}\phi(W_i, Z_i) f_Y(y|W_i, Z_i) \leq \mathbb{E}\phi(W_i, Z_i)\bar{f} \tag{177}$$

for any non-negative measurable function $\phi(W_i, Z_i)$. The same is true for the derivative of the density. So $P(f_Y(y|\hat{\theta}_{-i}, \lambda_i, W_i, Z_i) = \infty) = 0$ and $P(\partial f_Y(y|\hat{\theta}_{-i}, \lambda_i, W_i, Z_i) = \infty) = 0$, which justifies the Taylor expansion above. By a.s. smoothness of $f_Y(y|\hat{\theta}_{-i}, \lambda_i, Z_i, W_i)$ and equation (173),

$$\mathbb{E}Z_i\lambda_i f_Y(W_i'\hat{\theta}_{-i}|\hat{\theta}_{-i}, W_i, \lambda_i, Z_i) = \mathbb{E}Z_i\lambda_i f_Y(W_i'\theta_0|W_i, Z_i, 1\{Y_i \leq W_i'\theta_0\}) + O\left(\frac{1}{\sqrt{n}}\right). \tag{178}$$

Notice that the side of the density in the Taylor expansion depends on the direction of the deviation $\lambda_i^\circ \triangleq -\frac{1}{n}W_i'G^{-1}Z_i 1\{Y_i \leq W_i'\theta_0\}$,

$$\mathbb{E}Z_i\lambda_i^\circ f_Y(W_i'\theta_0|\lambda_i^\circ, Z_i, W_i) = \tag{179}$$

$$= \mathbb{E}\left(-Z_iW_i'G^{-1}Z_i 1\{Y_i \leq W_i'\theta_0\} \lim_{t\to 0} \frac{1\{W_i'\theta_0 - t1\{\lambda_i^\circ \leq 0\} < Y_i \leq W_i'\theta_0 + t1\{\lambda_i^\circ > 0\}\}}{t}\right) \tag{180}$$

$$= -\lim_{t\to 0}\mathbb{E}Z_iW_i'G^{-1}Z_i 1\{Y_i \leq W_i'\theta_0\}\frac{1\{W_i'\theta_0 - t1\{\lambda_i^\circ \leq 0\} < Y_i \leq W_i'\theta_0 + t1\{\lambda_i^\circ > 0\}\}}{t} \tag{181}$$

$$= -\mathbb{E}Z_iW_i'G^{-1}Z_i 1\{-W_i'G^{-1}Z_i < 0\}\lim_{t\to 0} 1\{\varepsilon_i \leq 0\}\frac{1\{-t < \varepsilon_i \leq 0\}}{t} \tag{182}$$

$$- \mathbb{E}Z_iW_i'G^{-1}Z_i 1\{-W_i'G^{-1}Z_i > 0\}\lim_{t\to 0} 1\{\varepsilon_i \leq 0\}\frac{1\{0 < \varepsilon_i \leq t\}}{t} \tag{183}$$

$$= -\mathbb{E}Z_iW_i'G^{-1}Z_i 1\{W_i'G^{-1}Z_i > 0\} f_\varepsilon(0|Z_i, W_i). \tag{184}$$

So (172) becomes

$$\mathbb{E}Z_iP(Y_i < W_i'\hat{\theta}_{-i}|Z_i, W_i) + \frac{1}{n}\mathbb{E}Z_iW_i'G^{-1}Z_i(\tau - 1\{W_i'G^{-1}Z_i > 0\}) f_\varepsilon(0|W_i, Z_i)$$
$$+ \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i + O\left(\frac{1}{n^{3/2}}\right). \tag{185}$$

Step 3. Now consider the second term $\mathbb{E}g^\circ(\hat{\theta})$. Let $(Y_{n+1}, W_{n+1}, Z_{n+1})$ be a copy of (Y, W, Z), which is independent of the sample $\{Y_i, W_i, Z_i\}_{i=1}^n$. Also define $\lambda_{n+1,i} = -\frac{1}{n}W_{n+1}'G^{-1}Z_i(1\{Y_i \leq W_i'\theta_0\} - \tau)$, which satisfies $\mathbb{E}\lambda_{n+1,i} = 0$. Then

$$\mathbb{E}g^\circ(\hat{\theta}) = \mathbb{E}1\{Y_{n+1} \leq W_{n+1}'\hat{\theta}\}Z_{n+1} \tag{186}$$

$$= \mathbb{E}P\{Y_{n+1} \leq W_{n+1}'\hat{\theta}_{-i} - \frac{1}{n}\lambda_{n+1,i}|W_{n+1}, Z_{n+1}\}Z_{n+1} \tag{187}$$

$$= \mathbb{E}P\{Y_{n+1} < W_{n+1}'\hat{\theta}_{-i}|W_{n+1}, Z_{n+1}\}Z_{n+1} + O\left(\frac{1}{n^{3/2}}\right). \tag{188}$$

Combining this equality with (185) yields

$$\mathbb{E}\left(1\{Y \leq W'\hat{\theta}\}Z\right) - \mathbb{E}g(\hat{\theta}) \tag{189}$$

$$= \mathbb{E}Z_i P\{Y_i < W_i'\hat{\theta}_{-i}|W_i, Z_i\} - \mathbb{E}Z_{n+1}P\{Y_{n+1} < W_{n+1}'\hat{\theta}_{-i}|W_{n+1}, Z_{n+1}\} \tag{190}$$

$$+ \frac{1}{n}\mathbb{E}Z_i W_i' G^{-1} Z_i(\tau - 1\{W_i'G^{-1}Z_i > 0\}) f_\varepsilon(0|W_i, Z_i) + \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i + O\left(\frac{1}{n^{3/2}}\right). \tag{191}$$

Step 4. By expansion (173)

$$\hat{\theta}_{-i} = \theta_0 - \frac{1}{n}G^{-1}\sum_{j=1, j\neq i}^{n}\left(1\{\varepsilon_j \leq 0\} - \tau\right)Z_j + R_n. \tag{192}$$

Define

$$\zeta_{-i,n} \triangleq \frac{n-1}{n}\frac{1}{n-1}G^{-1}\sum_{j=1, j\neq i}^{n}\left(1\{\varepsilon_j \leq 0\} - \tau\right)Z_j, \tag{193}$$

so that $\zeta_{-i,n}$ is a zero mean r.v. that is independent of Y_i. Therefore,

$$\mathbb{E}Z_i P\{Y_i < W_i'\hat{\theta}_{-i}|W_i, Z_i\} = \mathbb{E}Z_i P(Y_i - W_i'\zeta_{-i,n} < W_i'W_i'R_n|W_i, Z_i) \tag{194}$$

$$= \mathbb{E}Z_i P(\xi_i < W_i'R_n|W_i, Z_i), \tag{195}$$

where $\xi_i \triangleq Y_i - W_i'\zeta_{-i,n}$ is a r.v. with PDF conditional on (W_i, Z_i) by Assumption 3.1 and $\zeta_{-i,n}$ is independent of Y_i.

Apply the Taylor theorem to obtain

$$\mathbb{E}Z_i P\{Y_i < W_i'\hat{\theta}_{-i}|W_i, Z_i, W_i'R_n\} \tag{196}$$

$$= \mathbb{E}Z_i P(\xi_i < W_i'\theta_0|W_i, Z_i) + \mathbb{E}W_i'R_n f_{\xi_i}(W_i'\theta_0|W_i, Z_i, R_n) \tag{197}$$

$$+ \frac{1}{2}\mathbb{E}R_n'W_i\partial f_{\xi_i}(W_i'\theta_0|W_i, Z_i, R_n)W_i'R_n + \mathbb{E}O_p\left(\|R_n\|^3\right). \tag{198}$$

By the Bahadur expansion (173), $R_n = O_p(\frac{1}{\sqrt{n}})$. Hence, (198) becomes

$$\mathbb{E}Z_i P(Y_i - W_i'\zeta_{-i,n} < W_i'\theta_0|W_i, Z_i) + \mathbb{E}W_i'R_n f_{\varepsilon_i}(0|W_i, Z_i, R_n) + o\left(\frac{1}{n}\right). \tag{199}$$

Similarly,

$$\mathbb{E}Z_{n+1}P\{Y_{n+1} < W'_{n+1}\hat{\theta}_{-i,n}|W_{n+1}, Z_{n+1}\} \tag{200}$$
$$= \mathbb{E}Z_{n+1}P(Y_{n+1} - W'_{n+1}\zeta_{-i} < W'_{n+1}\theta_0|W_{n+1}, Z_{n+1}) \tag{201}$$
$$+ \mathbb{E}W'_{n+1}R_n f_{\varepsilon_{n+1}}(0|W_{n+1}, Z_{n+1}, R_n) + o\left(\frac{1}{n}\right) \tag{202}$$
$$= \mathbb{E}Z_i P(Y_i - W'_i\zeta_{-i,n} < W'_i\theta_0|W_i, Z_i) + \mathbb{E}Z_{i,n}W'_i(\mathbb{E}R_n)f_{\varepsilon_i}(0|W_i, Z_i) + o\left(\frac{1}{n}\right). \tag{203}$$

Note that

$$\lim_{t\to 0}\mathbb{E}Z_iW'_i(R_n - \mathbb{E}R_n)(f_{\varepsilon_i}(0|W_i, Z_i, R_n) - f_{\varepsilon_i}(0|W_i, Z_i)) \tag{204}$$
$$= \mathbb{E}Z_iW'_i(R_n - \mathbb{E}R_n)\lim_{t\to 0}\frac{1}{t}(\mathbb{E}(1\{0 < \varepsilon_i \le t\}|W_i, Z_i, R_n) - \mathbb{E}(1\{0 < \varepsilon_i \le t\}|W_i, Z_i) \tag{205}$$
$$= \lim_{t\to 0}\mathbb{E}Z_iW'_i(R_n - \mathbb{E}R_n)\frac{1}{t}(1\{0 < \varepsilon_i \le t\} - 1\{0 < \varepsilon_i \le t\}) = 0. \tag{206}$$

To summarize, (191) becomes

$$\mathbb{E}\left(1\{Y \le W'\hat{\theta}\}Z\right) - \mathbb{E}g^\circ(\hat{\theta}) \tag{207}$$
$$= \frac{1}{n}\mathbb{E}Z_iW'_iG^{-1}Z_i(\tau - 1\{W'_iG^{-1}Z_i > 0\})f_\varepsilon(0|W_i, Z_i) + \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i + o\left(\frac{1}{n}\right). \tag{208}$$

Step 5. Formula (208) can be rewritten as

$$\mathbb{E}\left(1\{Y \le W'\hat{\theta}\}Z\right) - \mathbb{E}g^\circ(\hat{\theta}) \tag{209}$$
$$= \frac{1}{n}\mathbb{E}Z_iW'_iG^{-1}Z_i(\tau - 1\{W'_iG^{-1}Z_i > 0\})f_\varepsilon(0|W_i, Z_i) + \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i + o\left(\frac{1}{n}\right) \tag{210}$$
$$= \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i - \mathbb{E}\left(1\{Y_i \ge W'_i\hat{\theta}\}Z_i\right) + \mathbb{E}\left(1\{Y_{n+1} \ge W'_{n+1}\hat{\theta}\}Z_{n+1}\right) \tag{211}$$
$$= \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i - \mathbb{E}\left(1\{-Y_i \le W'_i(-\hat{\theta})\} - (1-\tau)\right)Z_i + \mathbb{E}\left(1\{-Y_{n+1} \le W'_{n+1}(-\hat{\theta})\} - (1-\tau)\right)Z_{n+1} \tag{212}$$
$$= \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i - \left[\frac{1}{n}\mathbb{E}Z_iW'_iG^{-1}Z_i(1 - \tau - 1\{W'_iG^{-1}Z_i > 0\})f_\varepsilon(0|W_i, Z_i) + \mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i\right] + o\left(\frac{1}{n}\right). \tag{213}$$

This implies

$$\frac{1}{2}\mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i = \frac{1}{n}\mathbb{E}Z_iW'_iG^{-1}Z_i(1\{W'_iG^{-1}Z_i > 0\} - 1/2)f_\varepsilon(0|W_i, Z_i) + o\left(\frac{1}{n}\right) \tag{214}$$

and

$$\mathbb{E}1\{\hat{\varepsilon}_i = 0\}Z_i = \mathbb{E}\left(\frac{\hat{g}(\hat{\theta}) + \hat{g}^*(-\hat{\theta})}{2}\right). \tag{215}$$

Hence, equation (208) can be rewritten as

$$\mathbb{E}\left(1\{Y \leq W'\hat{\theta}\}Z\right) - \mathbb{E}g^{\circ}(\hat{\theta}) \tag{216}$$

$$= \frac{1}{n}\mathbb{E}Z_i W_i' G^{-1} Z_i \left(\tau - \frac{1}{2}\right) f_{\varepsilon}(0|W_i, Z_i) + \mathbb{E}\left(\frac{\hat{g}(\hat{\theta}) + \hat{g}^*(-\hat{\theta})}{2}\right) + o\left(\frac{1}{n}\right). \tag{217}$$

□

D MILP implementation

Consider the following exact estimator

$$\hat{\theta}_{\ell_1} = \text{argmin}_{\theta \in \Theta} ||\hat{g}(\theta)||_1. \tag{218}$$

The underlying optimization problem can be equivalently reformulated as a mixed integer linear program (MILP) with special ordered set (SOS) constraints,

$$\begin{aligned}
&\min_{e,\theta,r,s,t} \boldsymbol{\iota}' t \\
&\text{s.t.} \\
&\varepsilon_i = r_i - s_i = Y_i - W_i'\theta,\ i = 1, \ldots, n, \\
&(r_i, e_i) \in SOS_1,\ i = 1, \ldots, n, \\
&(s_i, 1 - e_i) \in SOS_1,\ i = 1, \ldots, n, \\
&r_i \geq 0, s_i \geq 0,\ i = 1, \ldots, n, \\
&e_i \in \{0, 1\}, \\
&-t_l \leq Z_l'(e - \tau\boldsymbol{\iota}) \leq t_l,\ l = 1, \ldots, d.
\end{aligned}$$

where Z_l is an $n \times 1$ vector of realizations of instrument l. All constraints except the last one coincide with the ones derived by Chen and Lee (2018) in Appendix C.1. The last constraint ensures that the objective function is the ℓ_1 norm of the just identifying moment conditions.

We also considered the big-M formulation while performing the Monte Carlo analyses. The big-M formulation has certain computational advantages, although the arbitrary choice of tuning parameters may result in sub-optimal solutions. This problem is more prominent for tail quantiles. Consistent with our theory, the choice of tuning parameters in the big-M formulation may affect the asymptotic bias. We prefer the above SOS formulation because it does not depend on tuning parameters as the big-M MILP/MIQP formulations considered

in Chen and Lee (2018) and Zhu (2019).[27]

E An alternative to Lemma B.5

The key ingredient for deriving the Bahadur-Kiefer expansions is Lemma B.5. This lemma follows from our novel coupling of the sample moments with a Hölder continuous process and relies on the support restriction in Assumption 2.2. Here we outline an alternative to Lemma B.5, which does not rely on Assumption 2.2.[28] For simplicity, we treat m and $\bar{f}$ as fixed and omit them from the remainder bounds.

We impose the following assumptions.

Assumption 4 (Regressors)**.** *The distribution of $\tilde{W} \triangleq W/\|W\|$ admits a density.*

Assumption 2.2 corresponds to the case where the directions of W have finite support. Assumption 4 focuses on the complementary case where the directions have continuous support and distribution. Note that we can combine the two cases via an appropriate partitioning of the probability space by whether $\tilde{W}$ satisfies Assumption 2.2 or Assumption 4.

Lemma B.5 ultimately relies on the approximate Hölder continuity of the sample moment process. This property can be formalized in terms of tightness of a certain empirical process. Namely, given a sequence $r_n \to \infty$, define the following empirical process indexed by $\theta \in \Theta$, $\|h\| = 1$, $h \in \mathbb{R}^k$, and $s \in \left[-\frac{c}{r_n}, \frac{c}{r_n}\right]$ for some constants $c > 0$ and $0 < \gamma < \frac{1}{2}$,

$$\Xi_n^{\gamma, r_n}(\theta, h, s) \triangleq \mathbb{G}_n r_n^{1/2-\gamma} Z \left(1\left\{\frac{\tilde{Y} - \tilde{W}'\theta}{\tilde{W}'h} \leq s\right\} - 1\left\{\frac{\tilde{Y} - \tilde{W}'\theta}{\tilde{W}'h} \leq 0\right\}\right) 1\{\tilde{W}'h > 0\}. \quad (219)$$

As in the case of Assumption 2.2, $\gamma < 0$ yields an unbounded pointwise covariance $\Xi_n^{\gamma, r_n}(\theta, h, s)$ as $n \to \infty$.

Notice that for any fixed θ and h, we can use the conditional CDF transform $F_{\theta,h}(s|Z)$ corresponding to the random variable $\frac{\tilde{Y} - \tilde{W}'\theta}{\tilde{W}'h}$ to get a representation

$$Z1\left\{\frac{\tilde{Y} - \tilde{W}'\theta}{\tilde{W}'h} \leq s, \tilde{W}'h > 0\right\} = Z1\left\{U_{\theta,h} \leq u, A_h = 1\right\}, \quad (220)$$

[27] These papers pick the value of the tuning parameter M as a solution to a linear program that in turn depends on the choice of an arbitrary box around a linear IV estimate. This is problematic if there is a lot of heterogeneity in the coefficients across quantiles. Moreover, in the linear model with heavy tailed residuals, the linear IV estimator is not consistent.

[28] We thank Wolfgang Polonik and Zheng Fang for suggesting to explore VC-class arguments to bound the modulus of continuity.

where $U_{\theta,h}$ has the uniform $[0,1]$ distribution conditional on Z, $u = F_{\theta,h}(s|Z)$ and $A_h = 1\{\tilde{W}'h > 0\}$. Then Theorem 1 would imply that $\sup_s \|\Xi_n^{\gamma,r_n}(\theta,h,s)\|$ is tight for each θ, h. Unfortunately, this pointwise result is not strong enough for proving Lemma B.5 under Assumption 4. However, one can obtain an analogous result under the following alternative assumption.

Assumption 5 (Asymptotic tightness)**.** *The empirical process* (219) *is asymptotically tight.*

Assumption 5 could be verified using results on the modulus of continuity of empirical processes indexed by VC classes of functions; see, for example, Chapter 3.2 in van der Vaart and Wellner (1996).

Lemma E.1. *Suppose Assumptions 4 and 5 hold. For any pair of estimators $\hat{\theta}$ and $\hat{\theta}^* \in \mathbb{R}^d$ such that $\hat{\theta}^* = \hat{\theta} + O_p\left(\frac{1}{r_n}\right)$ for some increasing sequence r_n, the following property holds for any $0 < \gamma < \frac{1}{2}$*

$$B_n(\hat{\theta}) - B_n(\hat{\theta}^*) = O_p\left(\|\hat{\theta} - \hat{\theta}^*\|^{\frac{1}{2}-\gamma}\right) + O_p\left(\frac{1}{\sqrt{n}}\right). \tag{221}$$

Proof. The proof proceeds in two steps.

Step 1. Recall that

$$B_n^\circ(\theta) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n 1\{Y_i \le W_i'\theta\} - \mathbb{E}1\{Y_i \le W_i'\theta\}. \tag{222}$$

Notice that

$$\begin{aligned} 1\{Y \le W'\theta^*\} =& 1\{Y \le W'\theta^*, \tilde{W}'h > 0\} + 1\{Y \le W'\theta^*, \tilde{W}'h < 0\} \\ &+ 1\{Y \le W'\theta^*, \tilde{W}'h = 0\}. \end{aligned} \tag{223}$$

By Assumption 4, $P\{Y \le W'\theta^*, \tilde{W}'h = 0\} \le P\{\tilde{W}'h = 0\} = 0$ for all $\|h\| = 1$. Then

$$\mathbb{G}_n\{Y \le W'\theta^*, \tilde{W}'h = 0\} = \frac{1}{\sqrt{n}} \sum_{i=1}^n (1\{Y \le W'\theta^*, \tilde{W}'h = 0\} - 0) = O_p\left(\frac{k}{\sqrt{n}}\right), \tag{224}$$

where the last equality follows from the fact that $\sum_{i=1}^n 1\{\tilde{W}_i'h = 0\} \le k$ a.s. since $\tilde{W}$ has a density. Indeed, any k independent observations $\tilde{W}_i$ would be a.s. linearly independent. Hence any non-zero vector h can be a solution to at most $k-1$ equations $\tilde{W}_i'h = 0$ a.s.

The other two components in equation (223) can be rewritten as

$$1\{Y \le W'\theta^*, \tilde{W}'h > 0\} = 1\left\{\frac{\tilde{Y} - \tilde{W}'\theta}{\tilde{W}'h} \le s, \tilde{W}'h > 0\right\} \tag{225}$$

and

$$1\{Y \leq W'\theta^*, \tilde{W}'h < 0\} = 1\left\{\frac{\tilde{Y} - \tilde{W}'\theta}{\tilde{W}'h} \geq s, \tilde{W}'h < 0\right\} \tag{226}$$

$$= 1\left\{\frac{\tilde{Y} - \tilde{W}'\theta}{\tilde{W}'(-h)} \leq -s, \tilde{W}'(-h) > 0\right\}, \tag{227}$$

where $h \triangleq \frac{\theta^* - \theta}{s}$ and $s \triangleq \|\theta^* - \theta\|$.

Step 2. Since $\hat{\theta}^* - \hat{\theta} = O_p\left(\frac{1}{r_n}\right)$, we can select a constant c such that $\|\hat{\theta}^* - \hat{\theta}\| \leq \frac{c}{r_n}$ with arbitrarily high probability. Then by definition of $B_n(\theta)$ and (223), with probability approaching 1 we get

$$B_n(\hat{\theta}^*) - B_n(\hat{\theta}) = \frac{1}{r_n^{\frac{1}{2}-\gamma}}\left(\Xi_n^{\gamma,r_n}\left(\hat{\theta}, \hat{h}, \hat{s}\right) + \Xi_n^{\gamma,r_n}\left(\hat{\theta}, -\hat{h}, -\hat{s}\right)\right) + O_p\left(\frac{2k}{\sqrt{n}}\right), \tag{228}$$

where $\hat{h} \triangleq \frac{\hat{\theta} - \hat{\theta}^*}{\|\hat{\theta}^* - \hat{\theta}\|}$ and $\hat{s} \triangleq \|\hat{\theta}^* - \hat{\theta}\|$. By Assumption 5, $\Xi_n^{\gamma,r_n}\left(\hat{\theta}, \hat{h}, \hat{s}\right) = O_p(1)$ and $\Xi_n^{\gamma,r_n}\left(\hat{\theta}, -\hat{h}, -\hat{s}\right) = O_p(1)$. To complete the proof, notice that $O_p\left(\frac{1}{r_n}\right) = O_p\left(\|\hat{\theta} - \hat{\theta}^*\|\right)$. □