Journal of Econometrics 188 (2015) 250–263

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Extremum estimation and numerical derivatives

Han Hong [a], Aprajit Mahajan [b], Denis Nekipelov [c,d,*]

[a] Department of Economics, Stanford University, Stanford, CA 94305, United States
[b] Department of Economics, UCLA, Los Angeles, 90095, United States
[c] Department of Economics, University of Virginia, Charlottesville, VA 22904, United States
[d] Department of Computer Science, University of Virginia, Charlottesville, VA 22904, United States

ARTICLE INFO

Article history:
Received 10 July 2012
Received in revised form
20 November 2013
Accepted 5 May 2014
Available online 25 March 2015

JEL classification:
C14
C52

Keywords:
Numerical derivative
Entropy condition
Stochastic equicontinuity

ABSTRACT

Finite-difference approximations are widely used in empirical work to evaluate derivatives of estimated functions. For instance, many standard optimization routines rely on finite-difference formulas for gradient calculations and estimating standard errors. However, the effect of such approximations on the statistical properties of the resulting estimators has only been studied in a few special cases. This paper investigates the impact of commonly used finite-difference methods on the large sample properties of the resulting estimators. We find that first, one needs to adjust the step size as a function of the sample size. Second, higher-order finite difference formulas reduce the asymptotic bias analogous to higher order kernels. Third, we provide weak sufficient conditions for uniform consistency of the finite-difference approximations for gradients and directional derivatives. Fourth, we analyze numerical gradient-based extremum estimators and find that the asymptotic distribution of the resulting estimators may depend on the sequence of step sizes. We state conditions under which the numerical derivative based extremum estimator is consistent and asymptotically normal. Fifth, we generalize our results to semiparametric estimation problems. Finally, we demonstrate that our results apply to a range of nonstandard estimation procedures.

1. Introduction

When the analytical gradient of the objective function used in an extremum estimation problem is not available, numerical optimization routines often use finite-difference approximations to the analytical gradient instead. This involves the choice of a step size parameter. The approximation algorithm in the optimization routine introduces statistical noise but typically this noise is not accounted for while performing inference. In this paper, we provide weak conditions for the consistency of numerical derivative estimates and demonstrate that the use of finite approximation can affect both the rate of convergence and the asymptotic distribution of the resulting estimator. This result has important implications for the practical use of numerical optimization routines. In particular, the choice of the numerical step size should depend on the sample size. Further, in some situations the asymptotic distribution may depend on the particular step size sequence chosen.

We consider a general framework where the objective function is computed from an i.i.d. sample and the numerical gradient-based optimization routine uses finite-difference formulas to approximate the gradient of the objective function (which can depend on finite or infinite-dimensional unknown parameters). Our framework applies generally in empirical work which includes, for example, search models that use simulated method of moments. Aspects of this problem have received some attention in the previous literature. Pakes and Pollard (1989), Newey and McFadden (1994) and Murphy and Van der Vaart (2000) provided sufficient conditions for numerical derivative approximations to consistently estimate the asymptotic variance in a parametric model. The properties of numerical derivatives have predominantly been investigated only for smooth models. For instance, Anderssen and Bloomfield (1974) analyzed derivative computations for functions that are approximated by polynomial interpolation. L'Ecuyer and Perron (1994) considered asymptotic properties of numerical derivatives for the class of general smooth regression models, while Andrews (1997) developed a stopping-rule for approximate global optimization for GMM objective functions. However, to the best of our knowledge understanding the impact of numerical differentiation on the statistical properties

* Correspondence to: 237 Monroe Hall Charlottesville, VA 22904-4182, United States.
E-mail address: denis@virginia.edu (D. Nekipelov).

http://dx.doi.org/10.1016/j.jeconom.2014.05.019
0304-4076/

of general extremum estimators is still an open question. In this paper, we analyze local optimization algorithms that are less general than global optimizers and that are only consistent when local and global identification conditions coincide, but we take into account the numerical errors involved in computing the local optimizer.

In an important recent paper, Kristensen and Salanié (2010, 2013) considered a high level framework for bias reduction techniques and higher order improvement properties in estimators that are based on approximations of objective functions. This high level framework includes many approximate estimators, including the numerical differentiation based optimization methods that we study as special cases. Our results complement their paper and enrich their high level framework. We focus on first order asymptotics. While beyond the scope of this paper, it will be interesting in future work to investigate the possibility of higher order improvements along the lines of Kristensen and Salanié (2010, 2013).

Our results include fairly weak sufficient conditions required for consistency, rates of convergence and the asymptotic distribution for several classes of numerically computed extremum estimators. We find that the choice of the step size for consistency and the limiting distribution depend on the interplay between the order of the numerical differentiation and the properties of the sample objective function. Specifically, we find that if the sample objective function is very smooth, then the step size for numerical differentiation can be chosen to approach zero at an arbitrarily fast rate. For a non-differentiable objective function, however, the step size should not converge to zero too rapidly as the sample increases.

More generally, we make no normative claims about the use of finite-difference based approximations but rather demonstrate that commonly used procedures based on numerical derivatives may have profound implications for the large sample properties of the resultant estimators, an issue that has heretofore been relatively neglected. As a result, we do not consider the issue of "correcting" the behavior of the numerical gradient-based estimation procedures, for instance, by smoothing the objective function. But rather we focus on the specific smoothing implied by commonly used numerical optimization routines that are standard in empirical work. We conjecture that the behavior for the step size should resemble the choice of a bandwidth in a smoothed case.

The paper is organized as follows. Section 2 analyzes uniformly consistent numerical estimation in the context of the Jacobian of a moment model. Section 3 studies the impact of numerical derivative based optimization method on the asymptotic properties of the resulting parametric extremum estimators. Section 4 presents Monte Carlo simulation evidence. Finally Section 5 concludes.

2. Consistent derivative estimation

2.1. Numerical differentiation using finite differences

Finite difference methods (e.g. Judd (1998)) are often used for the numerical approximation of derivatives. To illustrate the implementation of the finite-difference formula for a univariate function $g(y)$, we can use a step size ϵ and construct a one-sided derivative estimate $\hat{g}'(y) = \frac{g(y+\epsilon)-g(y)}{\epsilon}$, or a two-sided derivative estimate $\hat{g}'(y) = \frac{g(y+\epsilon)-g(y-\epsilon)}{2\epsilon}$. More generally, the kth derivative of $g(y)$ for a d-dimensional y, where $k = \sum_{j=1}^{d} k_j$, can be estimated by a linear operator, denoted by $L_{k,p}^{\epsilon} g(y)$, that makes use of a pth order two-sided formula:

$$L_{k,p}^{\epsilon} g(y) = \frac{1}{\epsilon^k} \sum_{l_1=-p}^{p} \cdots \sum_{l_D=-p}^{p} c_{l_1 \ldots l_D} g\left(y + \sum_{j=1}^{d} l_j \epsilon e_j\right).$$

In the above e_j are vectors of the same dimensionality as argument x with one entry equal to one and other entries equal to zero. The usual two sided derivative formula refers to the case when $p = 1$. When $p \geq 1$, these are called higher order finite differences. For a given p, when the weights $c_{l_1,\ldots,l_d}$ are chosen appropriately, the error in approximating $\frac{\partial^k g(y)}{\partial y_1^{k_1} \ldots \partial y_d^{k_d}}$ with $L_{k,p}^{\epsilon} g(y)$ will be small:

$$L_{k,p}^{\epsilon} g(y) - \frac{\partial^k g(y)}{\partial y_1^{k_1} \ldots \partial y_d^{k_d}} = O(\epsilon^{2p+1-k}).$$

To obtain the coefficients $c_{l_1,\ldots,l_d}$ for the finite-difference approximation we need to evaluate the order of approximation error for a derivative of interest to guarantee that the formula with r terms delivers a precise expression for the first derivative of all polynomials of degree less than or equal to r. For instance, consider for the case where $d = 1$ and $r = 2p$, the following Taylor expansion:

$$\begin{aligned} L_{k,p}^{\epsilon} g(y) &= \frac{1}{\epsilon^k} \sum_{l=-p}^{p} c_l \left[\sum_{i=0}^{r} \frac{g^{(i)}(y)}{i!} (l\epsilon)^i + O\left(\epsilon^{r+1}\right) \right] \\ &= \sum_{i=0}^{r} g^{(i)}(y) \frac{\epsilon^i}{\epsilon^k} \sum_{l=-p}^{p} \frac{c_l l^i}{i!} + O\left(\epsilon^{r+1-k}\right). \end{aligned}$$

The coefficients c_l are therefore determined by a system of equations where $\delta_{i,k}$ is the Kronecker symbol that equals 1 if $i = k$ and is zero otherwise:

$$\sum_{l=-p}^{p} c_l l^i = i! \delta_{i,k}, \quad \text{for } i = 0, \ldots, r.$$

We are mostly concerned with first derivatives where $k = 1$. In this case we use $L_{1,p}^{\epsilon,y_j}$ to highlight the element of y to which the linear operator applies.

The usual two sided formula corresponds to $p = 1$, $c_{-1} = -1/2$, $c_0 = 0$ and $c_1 = 1/2$. For second order first derivatives where $p = 2$ and $k = 1$, $c_1 = 1/12$, $c_{-1} = -1/12$, $c_2 = -2/3$, $c_{-2} = 2/3$, $c_0 = 0$. In addition to the central numerical derivative, left and right numerical derivatives can also be defined analogously. Since they generally have larger approximation errors than central numerical derivatives, we will generally focus on central derivatives.

In general the step size ϵ can be chosen differently for different elements of the argument vector. It might also be possible to adapt the equal distance grid to a variable distance grid of the form $L_{k,p}^{\epsilon} g(y) = \frac{1}{\epsilon^k} \sum_{l=-p}^{p} c_l g(y + t_l \epsilon)$ for a scalar y, where t_l can be different from l. In addition both the step size and the grid distance can be made data-dependent. We leave this for future research.

Finally, for most of the statistical analysis in the rest of the paper we assume away machine imprecision. Machine precision imposes a lower bound on the step size in conjunction with the statistical lower bound.

2.2. Weak sufficient conditions for consistent Jacobian estimation

In this section we provide conditions on the step size for consistent derivative estimation, as required for instance, in variance estimation. These conditions are much weaker than those previously established in the literature. In particular, as long as a local uniformity condition holds, there is no interaction between the step size choice and the statistical uncertainty in parameter estimation. We focus on the unconditional parametric case in this section to best convey intuition, and develop the conditional semiparametric cases in the Appendix.

Consider a parametric unconditional moment model defined by the sample and population moment conditions: $\hat{g}(\theta) = \frac{1}{n} \sum_{i=1}^{n}$

$g(Y_i, \theta)$ and $g(\theta) = Eg(Y_i, \theta)$ where $g(\theta) = 0$ if and only if $\theta = \theta_0$, which lies in the interior of the parameter space Θ. The goal is to estimate $G(\theta_0) = \frac{\partial g(\theta_0)}{\partial \theta}$ using $L_{1,p}^{\epsilon_n} \hat{g}(\hat{\theta}) = \left(L_{1,p}^{\epsilon_n, \hat{\theta}_j} \hat{g}(\hat{\theta}), j = 1, \ldots, d\right)$, where $\hat{\theta}$ is typically a $\sqrt{n}$ consistent estimator of θ_0. The main intuition of this section can be understood by focusing on the case when both $g(Y_i, \theta)$ and θ are scalars.

In the following, we decompose the error of approximating $G(\theta_0)$ with $L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right)$ into three components: $L_{1,p}^{\epsilon_n} \hat{g}(\hat{\theta}) - G(\theta_0) = \hat{G}_1(\hat{\theta}) + G_2(\hat{\theta}) + G_3(\hat{\theta})$, where

$$\hat{G}_1(\hat{\theta}) = L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right) - L_{1,p}^{\epsilon_n} g\left(\hat{\theta}\right), \tag{2.1}$$

and

$$G_2\left(\hat{\theta}\right) = L_{1,p}^{\epsilon_n} g\left(\hat{\theta}\right) - G\left(\hat{\theta}\right), \qquad G_3\left(\hat{\theta}\right) = G\left(\hat{\theta}\right) - G\left(\theta_0\right).$$

In the above, $G_3\left(\hat{\theta}\right)$ represents the estimation error induced by the difference between $\hat{\theta}$ and θ_0. $G_2\left(\hat{\theta}\right)$ represents the bias induced by replacing analytic derivatives with numerical differentiation. Finally, $G_1\left(\hat{\theta}\right)$ represents an "empirical process" term that controls the sampling variation induced by estimating the population moment condition with its empirical analog. We discuss how to control each of these three terms in turn. Notice first that the sample size dependent step size ϵ_n does not play a role in $G_3(\hat{\theta})$. The bias term $G_2\left(\hat{\theta}\right)$ can be controlled if the bias reduction is uniformly small in a neighborhood of θ_0.

Throughout the paper we maintain the following suitable measurability requirement, which we will not refer to explicitly for the sake of brevity.

Assumption 1. The parameter space $\Theta \subset \mathbb{R}^p$ has a compact cover. For each n, there exists a countable subset $T_n \subset \Theta$ such that

$$P\left(\sup_{\theta \in \Theta} \inf_{\theta' \in T_n} \|\hat{g}(\theta) - \hat{g}\left(\theta'\right)\|^2 > 0\right) = 0.$$

This condition states that the values of the moment function on the parameter space Θ can be approximated arbitrarily well (with probability one) by its values on a countable subset of Θ. This condition is satisfied when the moment condition is right continuous, and thus allows for discontinuous moment functions. More precisely, Assumption 1 is a sufficient condition for the moment function to be *image admissible Suslin*. As discussed in Dudley (1999) and Kosorok (2008) this property will be required to establish the functional uniform law of large numbers needed for consistency.

Assumption 2. A mean value expansion of order $2p+1$ applies to the limiting function $g(\theta)$ uniformly in a neighborhood $\mathcal{N}(\theta_0)$ of θ_0. For all sufficiently small $|\epsilon|$ and $r = 2p+1$,

$$\sup_{\theta \in \mathcal{N}(\theta_0)} \left| g(\theta + \epsilon) - \sum_{l=0}^{r} \frac{\epsilon^l}{l!} g^{(l)}(\theta) \right| = O\left(|\epsilon|^{r+1}\right).$$

An immediate consequence of this assumption is that $G_2\left(\hat{\theta}\right) = O\left(\epsilon_n^{2p+1}\right)$ when $\hat{\theta}$ is consistent for θ_0. We are left with $\hat{G}_1\left(\hat{\theta}\right)$. The weakest possible condition to control $\hat{G}_1\left(\hat{\theta}\right)$ that covers all the models that we are interested in seems to come from a convergence rate result in Pollard (1984).

Assumption 3. For each $j = 1, \ldots, d$, consider functions $g(y, \theta)$ contained in class $\mathcal{F}_j = \left\{g(\cdot, \theta + e_j \epsilon),\ \theta \in \Theta\right\}$ for $\epsilon > 0$. Assume

(i) All $g \in \mathcal{F}_j$ are globally bounded such that $\|F\| = \sup_{\theta \in \Theta^\epsilon} |g(Y_i, \theta)| < C_1 \ll \infty$, where $\Theta^\epsilon = \{\theta : \inf_{\theta' \in \Theta} d\left(\theta, \theta'\right) \leq C_2 \epsilon\}$ for $d\left(\theta, \theta'\right)$ being the Euclidean distance between θ and θ'.

(ii) The moment function is Lipschitz-continuous in **mean square** in some neighborhood of θ_0. That is for sufficiently small $\epsilon > 0$ there exists a constant C_3 such that for each $j = 1, \ldots, p$

$$\sup_{\theta \in \mathcal{N}(\theta_0)} E\left[\left(g\left(Y_i, \theta + \epsilon e_j\right) - g\left(Y_i, \theta - \epsilon e_j\right)\right)^2\right] \leq C_3 \epsilon.$$

(iii) The graphs (defined on p. 27 of Pollard (1984)) of functions from $\mathcal{F}_j$ form a polynomial class of sets for any $\epsilon \to 0$ and for all j, in the sense of Definition 13 of Pollard (1984) (p. 17).

Many functions used in applications fall in this category. These include moment conditions that are Lipschitz continuous in parameters, and discontinuous moment conditions that involve the indicator function. In certain simulated method of moment models, for example those considered by Pakes and Pollard (1989), an example of the moment function $g(y_i, \theta)$ takes a form that is similar to

$$g(y_i, \theta) = \frac{1}{S} \sum_{s=1}^{S} z_i \left(w_i - 1\left(x_i' \theta - \epsilon_{is} \geq 0\right)\right),$$

where z_i is a set of bounded instrumental variables, ϵ_{is} is the simulated error term, and S is the finite number of simulations for each observation. In this particular simulated method of moment problem, Assumption 3(iii) holds because of Lemma 28 in Pollard (1984). In particular, Pakes and Pollard (1989) (p. 1043) advocates the use of numerical differentiation for gradient estimation. We will show that their conditions are too strong. Our results do not yet cover simulation estimators in which the number of simulations for each observation increases without bound, such as the maximum simulated likelihood estimator.

By Lemmas 25 and 36 of Pollard (1984), Assumption 3 defines the Euclidean property of a class of functions, and implies that there exist universal constants $A > 0$ and $V > 0$ such that for any $\mathcal{F}_n \subset \mathcal{F}$ with envelope function $\|F_n\|$,

$$\sup_{\mathcal{Q}} N_1(\epsilon \mathcal{Q} F_n, \mathcal{Q}, \mathcal{F}_n) \leq A \epsilon^{-V},$$

$$\sup_{\mathcal{Q}} N_2\left(\epsilon \left(\mathcal{Q} F_n^2\right)^{1/2}, \mathcal{Q}, \mathcal{F}_n\right) \leq A \epsilon^{-V},$$

where $N_1(\cdot)$ and $N_2(\cdot)$ are covering numbers defined in Pollard (1984) (p. 25 and p. 31) for probability measures $\mathcal{Q}$.

Lemma 1. *Under Assumption 3, if $n\epsilon_n / \log n \to \infty$, then for δ small enough,*

$$\sup_{d(\theta, \theta_0) \leq \delta} \|L_{1,p}^{\epsilon_n} \hat{g}(\theta) - L_{1,p}^{\epsilon_n} g(\theta)\| = o_p(1).$$

Consequently, Assumption 3 implies that $\hat{G}_1\left(\hat{\theta}\right) = o_p(1)$ if $d\left(\hat{\theta}, \theta_0\right) = o_p(1)$.

Proof. The argument follows directly from Theorem 2.37 in Pollard (1984) by verifying its conditions. For each n and each ϵ_n, consider the class of functions $\mathcal{F}_n = \{\epsilon_n L_{1,p}^{\epsilon_n} g(\cdot, \theta), \theta \in \mathcal{N}(\theta_0)\}$, with envelope function F, such that $PF \leq C$. Then we can write

$$\sup_{d(\theta, \theta_0) \leq o(1)} \epsilon_n \|L_{1,p}^{\epsilon_n} \hat{g}(\theta) - L_{1,p}^{\epsilon_n} g(\theta)\| \leq \sup_{f \in \mathcal{F}_n} |P_n f - Pf|.$$

For each $f \in \mathcal{F}_n$, note that $Ef^2 = E\left(\epsilon_n L_{1,p}^{\epsilon_n} g(\cdot,\theta)\right)^2 = O(\epsilon_n)$ because of Assumption 3(ii).[1] The lemma then follows immediately by taking $\alpha_n = 1$ and $\delta_n^2 = \epsilon_n$ in Theorem 2.37 of Pollard (1984). □

Theorem 1. *Suppose Assumptions 2 and 3 hold and* $\epsilon_n \to 0$, $n\epsilon_n/\log n \to \infty$ *and* $d\left(\hat{\theta},\theta_0\right) = o_p(1)$. *Then,* $L_{1,p}^{\epsilon_n}\hat{g}\left(\hat{\theta}\right) \xrightarrow{p} G(\theta_0)$.

In most situations $d\left(\hat{\theta},\theta_0\right) = O_p\left(n^{-\gamma}\right)$ for some $\gamma > 0$. Typically $\gamma = 1/2$. One might hope to further weaken the requirement of the $\log n$ term when uniformity is only confined to a shrinking neighborhood of size $n^{-\gamma}$. However, this is not possible unless the moment function satisfies additional smoothness conditions.

The result of Theorem 1 can be improved to allow for ϵ_n to approach zero at a rate even faster than $\log n/n$ if we are willing to impose the following stronger assumption, which holds for smoother Hölder-continuous functions. In the following E^* and E_P^* stand for outer expectation and outer expectation under measure P.

Assumption 4. In addition to Assumption 3, for all sufficiently small ϵ and all $\theta \in \Theta$, if we define $\mathbb{G}_n(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (g(Z_i,\theta) - g(\theta))$, then

$$E^* \sup_{\theta',\theta\in\mathcal{N}(\theta_0),d(\theta,\theta')\le\delta} |\mathbb{G}_n(\theta') - \mathbb{G}_n(\theta)| \lesssim \phi_n(\delta),$$

for functions $\phi_n(\cdot)$ such that $\delta \mapsto \phi_n(\delta)/\delta^\gamma$ is non-increasing for some $\gamma > 0$.

A sufficient lower level condition that implies Assumption 4 with $\gamma = 1$ is when $g(Y_i,\theta)$ is Lipschitz continuous in θ with a stochastically bounded Lipschitz constant.

In the above, $\lesssim$ indicates that the left side is bounded by a constant times the right side. Assumption 4 is more stringent than that in Theorem 3.2.5 in Van der Vaart and Wellner (1996), and may fail in cases where Theorem 3.2.5 holds, for example with indicator functions. Theorem 3.2.5 only requires that $E^* \sup_{d(\theta,\theta_0)<\delta} |\mathbb{G}_n(\theta) - \mathbb{G}_n(\theta_0)| \lesssim \phi_n(\delta)$. For i.i.d data, the tail bounds method used in Van der Vaart and Wellner (1996) can be modified to obtain Assumption 4. In particular, define a class of functions $\mathcal{M}_\delta^\epsilon = \{g(Z_i,\theta_1) - g(Z_i,\theta_2), d(\theta_1,\theta_2) \le \delta, d(\theta_1,\theta_0) < \epsilon, d(\theta_2,\theta_0) < \epsilon\}$. Then Assumption 4, which requires bounding $E_P^*||\mathbb{G}_n||_{\mathcal{M}_\delta^\epsilon}$, can be obtained by invoking the maximum inequalities in Theorems 2.14.1 and 2.14.2 in Van der Vaart and Wellner (1996). These inequalities provide that for M_δ^ϵ an envelope function of the class of functions $\mathcal{M}_\delta^\epsilon$,

$$E_P^*||\mathbb{G}_n||_{\mathcal{M}_\delta^\epsilon} \lesssim J\left(1,\mathcal{M}_\delta^\epsilon\right)\left(P^*\left(M_\delta^\epsilon\right)^2\right)^{1/2},$$

$$E_P^*||\mathbb{G}_n||_{\mathcal{M}_\delta^\epsilon} \lesssim J_{[]}\left(1,\mathcal{M}_\delta^\epsilon,L_2(P)\right)\left(P^*\left(M_\delta^\epsilon\right)^2\right)^{1/2},$$

where $J\left(1,\mathcal{M}_\delta^\epsilon\right)$ and $J_{[]}\left(1,\mathcal{M}_\delta^\epsilon,L_2(P)\right)$ are the uniform and bracketing entropy integrals defined in Section 2.14.1 of Van der Vaart and Wellner (1996), and are finite for most parametric classes of functions used in practice. Therefore $\phi_n(\delta)$ depends mostly on the variance of the envelope functions $\left(P^*\left(M_\delta^\epsilon\right)^2\right)^{1/2}$. For reasonably smooth functions that are Hölder-continuous, M_δ^ϵ depends only on δ as required by Assumption 4.

Theorem 2. *Suppose Assumptions 2 and 4 hold and* $\epsilon_n \to 0$, $n\epsilon_n^{2-2\gamma} \to \infty$ *and* $d\left(\hat{\theta},\theta_0\right) = o_p(1)$. *Then,* $L_{1,p}^{\epsilon_n}\hat{g}\left(\hat{\theta}\right) \xrightarrow{p} G(\theta_0)$.

[1] Here and further we use $f(y) = O(g(y))$ notation to indicate that there exists a constant $M < \infty$ such that $|f(y)| \le M|g(y)|$.

Proof. Recall from (2.1) that $L_{1,p}^{\epsilon_n}\hat{g}(\hat{\theta}) - G(\theta_0) = \hat{G}_1(\hat{\theta}) + G_2(\hat{\theta}) + G_3(\hat{\theta})$. As both $\hat{\theta} - \theta_0 \xrightarrow{p} 0$ and $\epsilon_n \to 0$, $G_3(\hat{\theta}) \xrightarrow{p} 0$ and $G_2(\hat{\theta}) \xrightarrow{p} 0$ by their definitions. It remains to consider $\hat{G}_1\left(\hat{\theta}_1\right)$, which can be written as

$$\hat{G}_1\left(\hat{\theta}_1\right) = L_{1,p}^{\epsilon_n}\left(\hat{g}\left(\hat{\theta}\right) - g\left(\hat{\theta}\right)\right) = \frac{1}{\sqrt{n}\epsilon_n}\sum_{l=1}^p c_l\left(\mathbb{G}\left(\hat{\theta} + l\epsilon_n e_j\right) - \mathbb{G}\left(\hat{\theta} - l\epsilon_n e_j\right)\right).$$

By Assumption 4 and the Markov inequality, this is bounded by $O_p\left(\frac{1}{\sqrt{n}\epsilon^{1-\gamma}}\right)$, which approaches 0 when $n\epsilon_n^{2-2\gamma} \to \infty$. □

This result shows that for continuous functions $g(Z_i,\theta)$ that are Lipschitz in θ (for which $\gamma = 1$), the only condition needed for consistency is $\epsilon_n \to 0$. The result of Theorem 2 demonstrates that when the moment function does not have discontinuities, one can choose the step size to decrease as a polynomial rate in the sample size. If the moment function is discontinuous, Theorem 2 does not apply. In this case one has to rely instead on the more general Theorem 1, which prescribes a slower logarithmic rate of decrease in the step size.

Example 1. Consider the simple quantile case where the moment condition is defined by $g(y_i;\theta) = 1(y_i \le \theta) - \tau$. In this case a numerical derivative at θ_0 is given by

$$L_{1,2}^{\epsilon_n}\hat{g}\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \frac{1\left(y_i \le \hat{\theta} + \epsilon_n\right) - 1\left(y_i \le \hat{\theta} - \epsilon_n\right)}{2\epsilon_n}.$$

This is basically the uniform kernel estimate of the density of y_i at $\hat{\theta}$:

$$\hat{f}_y\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2\epsilon_n} 1\left(\frac{|y_i - \hat{\theta}|}{\epsilon_n} \le 1\right) = \frac{1}{n}\sum_{i=1}^n \frac{1}{2\epsilon_n} 1\left(\frac{|y_i - \theta_0 - \left(\hat{\theta} - \theta_0\right)|}{\epsilon_n} \le 1\right).$$

The consistency conditions given in Powell (1984) and Newey and McFadden (1994), both of which require $\sqrt{n}\epsilon_n \to \infty$, are too strong. Under this condition, the second part of the estimation noise due to $\hat{\theta} - \theta_0$, $\frac{\hat{\theta}-\theta_0}{\epsilon_n}$, will vanish. However, for the purpose of consistency this is not necessary. As long as $\hat{f}_y(\theta)$ is uniformly consistent for $f_y(\theta)$ for θ in a shrinking neighborhood of 0, it will follow that $\hat{f}_y\left(\hat{\theta}\right) \xrightarrow{p} f_y(\theta_0)$. Notably, Kato (2012) derives the asymptotic normality of Powell (1984)'s kernel estimator allowing for dependent data and also obtains the optimal rate of convergence.

As this example illustrates, in one dimension, numerical differentiation resembles smoothing using a uniform kernel function. Higher order numerical derivatives are related to higher order kernel functions created by taking linear combinations of uniform kernels with different supports, as they both reduce the bias to either the order of the kernel function or the order of the numerical differentiation which depends on the number of points for taking linear combinations. The relation in the one dimensional case has analogs in the multivariate parameter setting. There is a large literature on replacing non-smooth components in irregular estimators by smooth approximations. Examples include Brown and Wang (2005), Horowitz (1992), Johnson and Strawderman (2009), Seo and Linton (2007), Wang et al. (2009) and Zinde-Walsh (2002), among others. The bandwidth parameter that controls the bias

of the approximation is closely related to the choice of the step size. A key difference is that researchers replacing irregular components with a smooth approximation typically choose both the kernel function and the resulting bandwidth. In numerical differentiation typically only the step size is the parameter of choice, and there is less freedom for choosing among a variety of kernel functions. In addition, the rate condition on the step size resembles one dimensional nonparametric convergence rates.

3. Numerical derivative based optimization of sample functions

3.1. Definitions

Extremum estimators are defined by maximizers of the sample objective function. While many derivative-free methods for optimization exist such as simulated annealing, it is often the case that, either by explicit researcher choice or by an implicit choice made by the maximization routine in the optimization software, the extremum estimator problem is replaced by the search for the zero of the numerically computed gradient. In this section we study the properties of the specific estimators based on numerically solving the first-order conditions for likelihood-type objective functions.

Consider the problem of estimating the parameter θ_0 in a metric space $(\Theta,\ d)$ with the metric d. The true parameter θ_0 is assumed to uniquely maximize the limiting objective function $Q(\theta) = Eg(Y_i;\theta)$. An M-estimator $\hat{\theta}$ of θ_0 is typically defined as

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \hat{\mathrm{Q}}(\theta), \tag{3.2}$$

where $\hat{\mathrm{Q}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Y_i;\theta)$. However, in practice, most sample objective functions $\hat{\mathrm{Q}}(\theta)$ of interest cannot be optimized analytically and are optimized instead through numerical computation. The optimization routine often uses numerical derivatives either explicitly or implicitly. In this section we show that while numerical differentiation does not affect the asymptotic distribution for smooth models (under suitable conditions on the step size sequence), for nonsmooth models a numerical derivative based estimator can translate a nonstandard parametric model into a nonparametric one.

We focus on the class of optimization procedures based on numerical gradients that are evaluated using the finite-difference formulas described in Section 2.1. We start by presenting a finite-difference numerical derivative version of the M-estimator in (3.2). A numerical gradient-based optimization routine effectively substitutes (3.2) with an approximate solution to the non-linear equation

$$\left|\left|L_{1,p}^{\epsilon_n}\hat{\mathrm{Q}}\left(\hat{\theta}\right)\right|\right| = o_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.3}$$

for some sequence of step sizes $\epsilon_n \to 0$. In some cases discussed below the convergence rate in (3.3) can be slower. We do not require the zeros of the first order condition to be exact in order to accommodate nonsmooth models. Many popular optimization packages use $p = 1$, corresponding to $\hat{D}^{\epsilon_n}\left(\hat{\theta}\right) \equiv L_{1,1}^{\epsilon_n}\hat{\mathrm{Q}}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}}\right)$. The cases with $p \geq 2$ correspond to a more general class of estimators that will have smaller asymptotic bias in nonsmooth models. As will be shown, the estimators (3.2) and (3.3) may have the same properties for models with smooth moment functions, but for non-smooth models both the asymptotic distributions and the convergence rates can be substantially different.

3.2. Consistency

Our first step is to provide a consistency analysis for $\hat{\theta}$. Many commonly used models have multiple local extrema, leading to multiple roots of the first-order condition. To facilitate our analysis we assume that the researcher is able to isolate a subset of the parameter space that uniquely contains the global maximum. For simplicity we will associate this subset with the entire parameter space Θ. The above discussion is formalized in the following identification assumption.

Assumption 5. The map defined by $D(\theta) = \frac{\partial}{\partial\theta}E[g(Y_i,\theta)]$ identifies $\theta_0 \in \Theta$, in the sense that, from $\lim_{n\to\infty}\|D(\theta_n)\| = 0$ it follows that $\lim_{n\to\infty}\|\theta_n - \theta_0\| = 0$ for any sequence $\theta_n \in \Theta$. Moreover, $g(\theta) = E[g(Y_i,\theta)]$ is locally quadratic at θ_0 with $g(\theta) - g(\theta_0) \leq -\bar{H}d(\theta,\theta_0)^2$, for some $0 < \bar{H} < \infty$ and all $\|\theta - \theta_0\| < \delta$ for some $\delta > 0$.

For global consistency we require the population objective function to be sufficiently smooth not only at the true parameter, but also uniformly in the entire parameter space Θ, so that we can rely on a version of Assumption 2 that is uniform over Θ to establish uniform consistency for the estimate of the derivative of the sample moment function. We consider objective functions generated by classes of functions that have polynomial bounds on finite differences. These classes include Lipschitz and Hölder-continuous functions.

The conditions in this section also remain valid for cases where the objective function exhibits substantially irregular behavior; e.g. when its first derivative approaches infinity in the vicinity of the maximum or minimum, for example when $g(Y_i,\theta) = \sqrt{|Y_i - \theta|}$.

We note that on the one hand, a version of Assumption 4 that is uniform over Θ restricts our analysis to functions that have a polynomial envelope on their finite differences. On the other hand, provided that γ can be very close to zero, it allows the finite increment in the parameter to lead to a change in the objective function exceeding the change in the argument. Finite differences of the objective function $g(Y_i,\theta) = \sqrt{|Y_i - \theta|}$ around the origin will be proportional to $1/\sqrt{\epsilon}$ and will not shrink too quickly with the decrease in ϵ. It turns out that this still allows us to provide consistency for the numerical gradient-based estimators.

The following result establishes consistency under a condition on the step size sequence that is a function of the sample size and the modulus of continuity of the empirical process. It is essentially a replica of Theorem 2 and hence stated without proof.

Corollary 1. *Under Assumption 5, as long as $\epsilon_n \to 0$ and $n\epsilon_n^{2-2\gamma} \to \infty$, if Assumptions 2 and 4 hold uniformly over Θ instead of only over $\mathcal{N}(\theta_0)$, then*

$$\sup_{\theta\in\Theta} ||L_{1,p}^{\epsilon_n}\hat{\mathrm{Q}}(\theta) - G(\theta)|| = o_p(1).$$

Consequently, if $||L_{1,p}^{\epsilon_n}\hat{\mathrm{Q}}\left(\hat{\theta}\right)|| = o_p(1)$, then $\hat{\theta} \xrightarrow{p} \theta_0$.

For models that have Lipschitz-continuous sample objective functions (which include models with smooth sample objective functions) where $\gamma = 1$, the restriction $n\epsilon_n^{2-2\gamma} \to \infty$ holds trivially. This implies that for smooth models the sequence of step sizes can approach zero arbitrarily fast.[2]

[2] There are, however, additional problems that are associated with "too fast" convergence of the step size sequence to zero. These problems, however, are not statistical and are connected with the machine computing precision.

3.3. Rate of convergence and asymptotic distribution

In the following theorem we establish the rate of convergence for the extremum estimator with the objective function from the considered class.

Theorem 3. *Suppose* $\hat{\theta} \xrightarrow{p} \theta_0$ *such that* $L_{1,p}^{\epsilon_n}\hat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}\epsilon_n^{1-\gamma}}\right)$. *Under Assumptions 1 and 4, if* $n\epsilon_n^{2-2\gamma} \to \infty$ *and* $\sqrt{n}\epsilon_n^{1-\gamma+2p} = O(1)$, *and suppose that the Hessian matrix* $H(\theta)$ *of* $g(\theta)$ *is continuous, nonsingular and finite at* θ_0, *then* $\sqrt{n}\epsilon_n^{1-\gamma} d\left(\hat{\theta}, \theta_0\right) = O_P(1)$.

For a regular parametric model, $\gamma = 1$, and there is no change in the convergence rate. This result is a Z-estimator version of Theorem 3.2.5 in Van der Vaart and Wellner (1996). Note that given consistency, the conditions required for obtaining the rate of convergence are weaker.

The following theorem, which combines a stochastic equicontinuity condition with verification of the Lindeberg condition, establishes the asymptotic normality of the numerical-derivative based estimator with an additional assumption of the convergence of the variance. Define $H(\theta_0) = \frac{\partial}{\partial\theta} Eg(Y, \theta_0)$ and $\Omega = \lim_{\epsilon\to 0} \epsilon^{2-2\gamma} Var\left(L_{1,p}^{\epsilon} g(Y_i, \theta_0)\right)$.

Theorem 4. *Assume that the conditions of Theorem 3 hold but with* $\sqrt{n}\epsilon_n^{1+2p-\gamma} = o(1)$. *If* $\sqrt{n}\epsilon_n^{2-\gamma} \to \infty$. *Then*

$$\sqrt{n}\epsilon_n^{1-\gamma}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, H(\theta_0)^{-1}\Omega H(\theta_0)^{-1}\right).$$

The additional assumption $\sqrt{n}\epsilon_n^{2-\gamma} \to \infty$, which essentially requires ϵ_n to be larger than the convergence rate of $\frac{1}{\sqrt{n}\epsilon_n^{1-\gamma}}$ established in Theorem 3 in order to show stochastic equicontinuity, turns out to be stronger for smooth models than for nonsmooth models. This is because we are relying on Assumption 4 and the convergence rate result in Theorem 3 to obtain stochastic equicontinuity. When $\gamma = 1$, the conditions are consistent as long as $p \geq 1$, or as long as a two sided central derivative is used. However, for smooth models when $\gamma = 1$, one may impose stronger assumptions on the sample objective function (e.g. Lemma 3.2.19 in Van der Vaart and Wellner (1996)) to weaken this requirement. We demonstrate this in Proposition 1.

Proposition 1. *Suppose the conditions of Theorem 4 hold except* $\sqrt{n}\epsilon_n^{2-\gamma} \to \infty$. *Suppose further that* $g(y_i, \theta)$ *is mean square differentiable in a neighborhood of* θ_0: *for measurable functions* $D(\cdot,\cdot) : \mathcal{Y} \times \Theta \to \mathbb{R}^p$ *such that*

$$E\left[g(Y,\theta) - g(Y,\theta_2) - (\theta_2 - \theta_1)' D(Y,\theta_1)\right]^2 = o\left(\|\theta_1 - \theta_2\|^2\right),$$

$E\|D(Y,\theta_1)\|^2 < \infty$ *for all* θ_1, *and* $\theta_2 \in \mathcal{N}_{\theta_0}$. *Define* $q_\epsilon(z_i,\theta) = L_{1,p}^{\epsilon} g(y_i;\theta) - D(y,\theta)$, *assume that*

$$\sup_{d(\theta_1,\theta_0)=o(1),\epsilon=o(1)} \left[\mathbb{G}_n q_\epsilon(y_i,\theta_1) - \mathbb{G}_n q_\epsilon(y_i,\theta_0)\right] = o_p(1),$$

and $D(y_i,\theta)$ *is Donsker in* $d(\theta_1,\theta_0) \leq \delta$ *(in the sense of* (2.1.1.), *pp* 81, *Van der Vaart and Wellner (1996)), then the conclusion of Theorem 4 holds.*

Note that we still require $\sqrt{n}\epsilon_n^2 \to 0$ to remove the asymptotic bias, and $n\epsilon_n \to \infty$, but we no longer require $\sqrt{n}\epsilon_n \to \infty$. The conditions of this theorem are perhaps best understood in the context of a quantile regression estimator. Consider $g(y,\theta) = |y-\theta|$, and $p = 2$, so that $D(y,\theta) = \text{sgn}(y-\theta)$ and

$$q_\epsilon(y,\theta) = \frac{(y-\theta)}{\epsilon} 1(|y-\theta| \leq \epsilon).$$

Then we can bound $q_\epsilon(y,\theta_1) - q_\epsilon(y,\theta_0)$ by, depending on which of $d(\theta,\theta_0)$ and ϵ is larger, the product between $\frac{1}{\epsilon}\max\left(|y-\theta|, |y-\theta_0|\right)$ and the maximum of $1(|y-\theta|) \leq \epsilon + 1(|y-\theta_0| \leq \epsilon)$, and $[1(\theta-\epsilon \leq y \leq \theta+\epsilon_0) + 1(\theta_0-\epsilon \leq y \leq \theta_0+\epsilon_0)]$. Since $\max(|y-\theta|, |y-\theta_0|) \leq \epsilon$ when $q_\epsilon(y,\theta) - q_\epsilon(y,\theta_0)$ is nonzero, the last condition in Theorem 1 is clearly satisfied by the Euclidean property of the indicator functions. Alternatively, the $q_\epsilon(y,\theta)$ function in the last condition can also be replaced directly by $L_{1,p}^{\epsilon} g(y,\theta)$.

4. Monte Carlo evidence

To illustrate the theoretical relationship between the smoothness of the moment conditions and the convergence rate of the step size, we consider the problem of estimating the standard errors for M-estimators using the standard "sandwich" formula where the Hessian of the objective function is computed numerically. Consider the objective function that is defined by a continuously distributed scalar covariate Z

$$Q(\theta) = E\left[|Z-\theta|^{1+\gamma}\right],$$

where $\gamma \in (0,1)$ is a known constant and θ is the parameter of interest. We note that if Z has unbounded support that does not depend on θ and the density of Z is continuously differentiable, then the objective function $Q(\cdot)$ has a continuous second derivative.

The corresponding sample analog is constructed from the sample $\{z_i\}_{i=1}^n$ such that

$$\widehat{Q}(\theta) = \frac{1}{n}\sum_{i=1}^{n} |z_i - \theta|^{1+\gamma}.$$

The sample objective function is continuously differentiable when $\gamma > 0$, which leads to the equation that determines the estimator

$$\frac{\partial\widehat{Q}(\widehat{\theta})}{\partial\theta} = \widehat{g}(\widehat{\theta}) = -(1+\gamma)\frac{1}{n}\sum_{i=1}^{n} \text{sign}\left(z_i - \widehat{\theta}\right)|z_i - \widehat{\theta}|^{\gamma} = 0. \tag{4.4}$$

Assuming the population objective function is differentiable, the resulting estimator is consistent and converges to the maximizer of $Q(\theta)$ at the parametric rate. If the Hessian of the population objective function $H(\theta_0)$ were known, then the variance of the estimator could be obtained from

$$V_\theta = H(\theta_0)^{-1} V_g H(\theta_0)^{-1},$$

where $V_g = (1+\gamma)^2 E\left[|Z-\theta|^{2\gamma}\right]$. However, the first derivative of the sample objective function is not differentiable. Instead, it is Hölder-continuous with the degree of Hölder-continuity γ. Thus, the Hessian cannot be obtained by a straightforward twice differentiation of the sample objective function. To obtain the Hessian we use the finite difference formulas:

$$\widehat{H}_1^{+}\left(\widehat{\theta}\right) = L_{1,1}^{\epsilon_n}\widehat{g}\left(\widehat{\theta}\right) = \frac{\widehat{g}\left(\widehat{\theta}+\epsilon_n\right) - \widehat{g}\left(\widehat{\theta}\right)}{\epsilon_n},$$

for the right derivative, and for the left derivative formula.

$$\widehat{H}_1^{-}\left(\widehat{\theta}\right) = L_{1,1}^{\epsilon_n}\widehat{g}\left(\widehat{\theta}\right) = \frac{\widehat{g}\left(\widehat{\theta}\right) - \widehat{g}\left(\widehat{\theta}-\epsilon_n\right)}{\epsilon_n}.$$

The second-order formula is

$$\widehat{H}_2\left(\widehat{\theta}\right) = L_{1,2}^{\epsilon_n}\widehat{g}\left(\widehat{\theta}\right) = \frac{\widehat{g}\left(\widehat{\theta}+\epsilon_n\right) - \widehat{g}\left(\widehat{\theta}-\epsilon_n\right)}{2\epsilon_n},$$

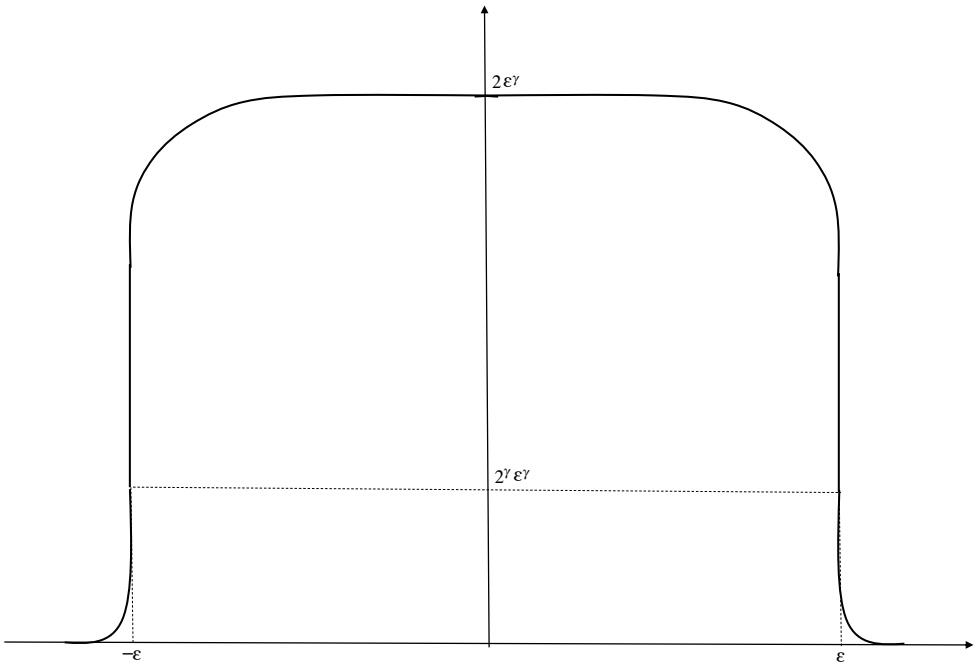

Fig. 1. The shape of the envelope function for class $\mathcal{F}_{\delta,\gamma,\epsilon}$ centered at a given y.

and the third-order formula is

$$\widehat{H}_3\left(\widehat{\theta}\right) = L_{1,3}^{\epsilon_n}\widehat{g}\left(\widehat{\theta}\right) = \frac{-\widehat{g}\left(\widehat{\theta}-2\epsilon_n\right)+8\widehat{g}\left(\widehat{\theta}-\epsilon_n\right)-8\widehat{g}\left(\widehat{\theta}+\epsilon_n\right)+\widehat{g}\left(\widehat{\theta}+2\epsilon_n\right)}{12\epsilon_n}.$$

The idea of the Monte Carlo will be the following. We generate the data for Z from the standard normal distribution, meaning that the true minimizer of the objective function $Q(\cdot)$ is $\theta_0 = 0$. Using (4.4) we obtain the parameter estimate $\widehat{\theta}$. Then using the sample analog of the formula for V_g as well as each of the formulas for the numerical derivatives, we construct the estimator for V_θ. Then we form a t-statistic to test the null hypothesis $H_0 : \theta_0 = 0$. We compare the empirical test size with the nominal size by computing the probability $\Pr\left(|\widehat{t}| > q_{1-\alpha/2}\right)$, where $q_{1-\alpha/2}$ is $1-\alpha/2$ quantile of the standard normal distribution. We consider the designs with $\gamma = 0.25, 0.5, 0.75$. For the step size we use the sequence $\epsilon_n = C_1 \log n/n^r$. with $r = 0.5, 1, 1.5, 2$. The constant C_1 is calibrated such that the step size coincides for all specifications for the minimal chosen sample size.

First, we provide the expression for the Hessian of the population objective function. To express the Hessian we notice that

$$\begin{aligned} H(\theta) &= \frac{\partial}{\partial\theta}\left\{(1+\gamma)E\left[\text{sign}(y-\theta)\,|y-\theta|^\gamma\right]\right\} \\ &= \frac{\partial}{\partial\theta}\left\{(1+\gamma)\int_{-\infty}^{+\infty}\phi(\theta+t)\text{sign}(t)\,|t|^\gamma\,dt\right\} \\ &= -(1+\gamma)\int_{-\infty}^{+\infty}\phi(\theta+t)\text{sign}(t)\,(t+\theta)|t|^\gamma\,dt, \end{aligned}$$

where $\phi(\cdot)$ is the density of the standard normal distribution. As a result, $H(\theta_0) = -(1+\gamma)E\left[|Y|^{1+\gamma}\right]$. To determine the theoretical properties of the estimator for the Hessian, consider the class of functions

$$\begin{aligned} \mathcal{F}_{\delta,\gamma,\epsilon} = \{&\text{sign}(\cdot-\theta-\epsilon)|\cdot-\theta-\epsilon|^\gamma \\ &-\text{sign}(\cdot-\theta+\epsilon)|\cdot-\theta+\epsilon|^\gamma \;:\; \|\theta-\theta_0\|\le\delta\}. \end{aligned}$$

Provided that this class is indexed by a compact set, its entropy is a power function of δ, meaning that the corresponding covering integral is finite. This class also has a finite envelope $F_{\delta,\gamma,\epsilon}(y)$. This envelope can be evaluated by noticing that for each θ, functions in $\mathcal{F}_{\delta,\gamma,\epsilon}$ will take the form of the "smoothed step function". In fact, note that for each y we can express each function $f \in \mathcal{F}_{\delta,\gamma,\epsilon}$ as

$$-f(y;\theta,\epsilon) = \begin{cases} |y-\theta+\epsilon|^\gamma - |y-\theta-\epsilon|^\gamma, & \text{if } y-\theta>\epsilon, \\ |\theta+\epsilon-y|^\gamma + |y-\theta+\epsilon|^\gamma, & \text{if } y-\theta\in[-\epsilon,\epsilon] \\ |\theta+\epsilon-y|^\gamma - |\theta-\epsilon-y|^\gamma, & \text{if } y-\theta<-\epsilon. \end{cases}$$

Note that $|f|$ is decreasing when $\theta > y$ and increasing when $\theta < y$. It is strictly convex when $|y-\theta| > \epsilon$ and strictly concave when $|y-\theta| < \epsilon$. The Hessian of $|f|$ when $|y-\theta| < \epsilon$ is equal to

$$-\gamma(1-\gamma)(|\theta+\epsilon-y|^{\gamma-2} + |y-\theta+\epsilon|^{\gamma-2})$$

which approaches zero as $\gamma \to 0$ meaning that $|f|$ approaches to a step function on $|y-\theta| < \epsilon$ for small γ. $|f|$ attains its maximum at $\theta = y$ with value $2\,\epsilon^\gamma$. When $\epsilon \ll 1$, then the maximum of $|f|$ increases as $\gamma \to 0$. This means that $|f|$ approaches to an indicator function on $[-\epsilon,\epsilon]$ with the height of the step size $2\,\epsilon^\gamma$ for small γ. In Fig. 1 we demonstrate the general shape of the envelope.

We now evaluate the variance of the envelope for $\mathcal{F}_{\delta,\gamma,\epsilon}$. We do so by considering separately the regions $|y-\theta| > \epsilon$ and $|y-\theta| < \epsilon$. We start with the region $|y-\theta| > \epsilon$. Consider some fixed y. We note that as $|\theta| \to \infty$, we can make the following representation

$$\begin{aligned} &|\theta+\epsilon-y|^\gamma - |\theta-\epsilon-y|^\gamma \\ &= |\theta|^\gamma\left|1-\frac{y-\epsilon}{\theta}\right|^\gamma - |\theta|^\gamma\left|1-\frac{y+\epsilon}{\theta}\right|^\gamma \\ &= |\theta|^{\gamma-1}2\gamma\epsilon + o(|\theta|^{\gamma-1}), \end{aligned}$$

where we used the Taylor expansion $|1+y|^\gamma = 1+\gamma\,y+o(x)$ for $y \to 0$. As a result $|f|/\epsilon$ approaches zero in the limit at the rate $2\gamma|\theta|^{1-\gamma}$. As a result, the variance of $|f|$ (and, therefore, its envelope) for $|x-\theta|^\gamma$ can be majorized by a constant multiple of $\gamma^2|\theta_0|^{2-2\gamma}\epsilon^2$.

Now we consider the variance of the envelope for $|y-\theta| < \epsilon$. Provided that for small γ $|f|$ approaches the step function on $[-\epsilon,\epsilon]$ of height ϵ^γ, the variance of the envelope can be majorized

Table 1
Rejection rates for H_0 with $\gamma = 0.25$, $\alpha = 0.05$.

n	ϵ_n			
	$\log n/n^2$	$\log n/n^{1.5}$	$\log n/n$	$\log n/n^{0.5}$
$H_1^+(\theta_0)$				
10	0.2690	0.2690	0.2690	0.2690
20	0.2370	0.2365	0.2355	0.2350
50	0.1795	0.1790	0.1780	0.1740
100	0.1530	0.1525	0.1515	0.1485
200	0.1360	0.1355	0.1345	0.1280
500	0.1115	0.1120	0.1120	0.1060
1 000	0.0945	0.0940	0.0930	0.0820
2 000	0.0910	0.0895	0.0875	0.0780
5 000	0.0865	0.0865	0.0865	0.0750
10 000	0.0805	0.0810	0.0790	0.0695
20 000	0.0715	0.0712	0.0685	0.0720
50 000	0.0605	0.0614	0.0602	0.0601
100 000	0.0510	0.0505	0.0500	0.0502
$H_1^-(\theta_0)$				
10	0.2705	0.2705	0.2705	0.2705
20	0.2345	0.2360	0.2350	0.2365
50	0.1800	0.1790	0.1785	0.1765
100	0.1525	0.1525	0.1525	0.1535
200	0.1360	0.1350	0.1335	0.1300
500	0.1115	0.1095	0.1065	0.1015
1 000	0.0945	0.0945	0.0945	0.0865
2 000	0.0910	0.0900	0.0885	0.0815
5 000	0.0865	0.0865	0.0865	0.0785
10 000	0.0805	0.0805	0.0800	0.0634
20 000	0.0715	0.0712	0.0691	0.0712
50 000	0.0603	0.0614	0.0602	0.0601
100 000	0.0510	0.0505	0.0500	0.0510

by the area under the step function supported between $-\epsilon$ and ϵ with the step height $\propto \epsilon^{2\gamma}$. When $\gamma \geq 1/2$ such an evaluation will not be accurate because the envelope will move away from the step function. In particular, we note that on $|y-\theta| < \epsilon$, any envelope for $\gamma > 1/2$ will lie below the envelope for $\gamma = 1/2$. At the same time for $|y-\theta| < \epsilon$

$$\begin{aligned}\left(|\theta+\epsilon-y|^{1/2}+|y-\theta+\epsilon|^{1/2}\right)^2 &\leq \theta+\epsilon-y+y-\theta+\epsilon \\ &= 2\epsilon.\end{aligned}$$

This gives a more accurate characterization of the square of the envelope for $\gamma \geq 1/2$.

Now we combine our variance evaluations for $|y-\theta| > \epsilon$ and $|y-\theta| \leq \epsilon$: $E\left[F^2_{\delta,\gamma,\epsilon}(Y)\right]^{1/2} \leq C_1\epsilon + C_2\epsilon^{1/2+\gamma} \leq (C_1+C_2)\epsilon^{1/2+\gamma}$ whenever $\gamma < 1/2$ and $E\left[F^2_{\delta,\gamma,\epsilon}(Y)\right]^{1/2} \leq C_1\epsilon + C_3\epsilon$ whenever $\gamma \geq 1/2$. The component $C_1\epsilon$ comes from the majorant on the variance for $|y-\theta| > \epsilon$ while the second component comes from the evaluation on $|y-\theta| < \epsilon$.

Provided that the covering integral is finite, the class of functions with such an envelope admits the evaluation in Assumption 4 which means that the empirical process associated with the objective function can be bounded in outer expectation by a constant multiple of $E\left[F^2_{\delta,\gamma,\epsilon}(Y)\right]^{1/2}$, which allows us to apply Theorem 2. As a result, the sufficient condition for consistency will require the step size sequence ϵ_n to be chosen such that $n\,\epsilon_n^{1-2\gamma} \to \infty$ whenever $\gamma < 1/2$. For $\gamma \geq 1/2$ there is no restriction on the choice of the step size sequence.

We note that, given that the estimator for θ_0 obtained from solving the first-order condition is consistent, the estimator for the Hessian, and thus, the standard errors will be consistent, as long as $\epsilon_n^{1-2\gamma} n \to \infty$. In Tables 1–3 we show the results of the Monte Carlo simulations for quantile $\alpha = 0.05$. Smaller values of γ correspond to less smooth objective functions. Table 1 shows that for the least smooth case, the chosen step size sequences lead to a decline in the rejection probability towards the nominal rejection rate. However, the best results are observed for the slowest sequences of step sizes. Even for the slowest chosen sequences the nominal rate is not achieved even in samples of size 10,000. As the objective function becomes smoother, the test rejection rate becomes closer to the nominal one as seen in Tables 2 and 3. Moreover, Table 3 shows that for smoother objective function the choice of a particular step size sequence becomes less relevant leading to the rejection

Table 2
Rejection rates for H_0 with $\gamma = 0.25$, $\alpha = 0.05$.

n	ϵ_n			
	$\log n/n^2$	$\log n/n^{1.5}$	$\log n/n$	$\log n/n^{0.5}$
$H_2(\theta_0)$				
10	0.2710	0.2710	0.2710	0.2710
20	0.2365	0.2355	0.2355	0.2355
50	0.1805	0.1805	0.1785	0.1780
100	0.1530	0.1530	0.1525	0.1540
200	0.1370	0.1360	0.1360	0.1330
500	0.1115	0.1120	0.1115	0.1130
1 000	0.0945	0.0945	0.0950	0.0900
2 000	0.0910	0.0900	0.0910	0.0840
5 000	0.0865	0.0865	0.0875	0.0775
10 000	0.0805	0.0805	0.0810	0.0696
20 000	0.0715	0.0710	0.0702	0.0720
50 000	0.0611	0.0610	0.0605	0.0601
100 000	0.0508	0.0502	0.0502	0.0502
$H_3(\theta_0)$				
10	0.2700	0.2700	0.2700	0.2700
20	0.2370	0.2350	0.2330	0.2350
50	0.1805	0.1800	0.1775	0.1760
100	0.1525	0.1515	0.1520	0.1495
200	0.1370	0.1360	0.1335	0.1310
500	0.1115	0.1115	0.1105	0.1110
1 000	0.0945	0.0940	0.0920	0.0905
2 000	0.0910	0.0895	0.0900	0.0835
5 000	0.0865	0.0860	0.0865	0.0821
10 000	0.0805	0.0805	0.0795	0.0731
20 000	0.0703	0.0685	0.0685	0.0680
50 000	0.0593	0.0590	0.0592	0.0590
100 000	0.0500	0.0502	0.0502	0.0501

Table 3
Rejection rates for H_0 with $\gamma = 0.5$, $\alpha = 0.05$.

n	ϵ_n			
	$\log n/n^2$	$\log n/n^{1.5}$	$\log n/n$	$\log n/n^{0.5}$
$H_1^+(\theta_0)$				
10	0.1326	0.1326	0.1326	0.1326
20	0.1006	0.0996	0.0996	0.0996
50	0.0771	0.0771	0.0781	0.0781
100	0.0626	0.0626	0.0621	0.0631
200	0.0581	0.0566	0.0566	0.0566
500	0.0506	0.0501	0.0501	0.0486
1 000	0.0470	0.0470	0.0465	0.0460
2 000	0.0575	0.0575	0.0575	0.0580
5 000	0.0515	0.0515	0.0510	0.0515
10 000	0.0540	0.0540	0.0540	0.0555
20 000	0.0515	0.0512	0.0515	0.0515
50 000	0.0505	0.0505	0.0505	0.0501
100 000	0.0500	0.0500	0.0500	0.0500
$H_1^-(\theta_0)$				
10	0.1356	0.1356	0.1356	0.1356
20	0.1011	0.1011	0.1006	0.1001
50	0.0771	0.0771	0.0781	0.0786
100	0.0626	0.0626	0.0631	0.0621
200	0.0576	0.0576	0.0576	0.0586
500	0.0511	0.0511	0.0506	0.0506
1 000	0.0470	0.0470	0.0470	0.0460
2 000	0.0575	0.0575	0.0575	0.0570
5 000	0.0515	0.0515	0.0515	0.0510
10 000	0.0540	0.0540	0.0540	0.0540
20 000	0.0515	0.0512	0.0515	0.0525
50 000	0.0505	0.0505	0.0505	0.0499
100 000	0.0500	0.0500	0.0500	0.0500

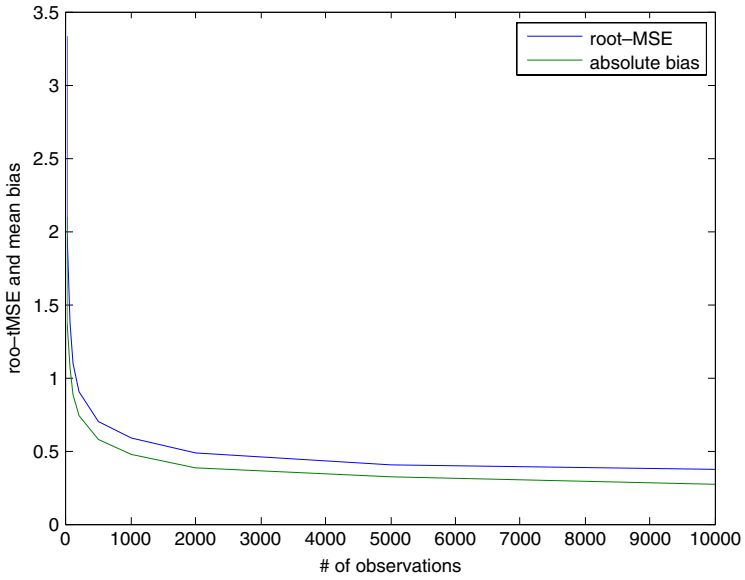

Fig. 2. Root-mean squared error and bias for formula H_3 with sequence log n/n^2 and $\gamma = 0.25$.

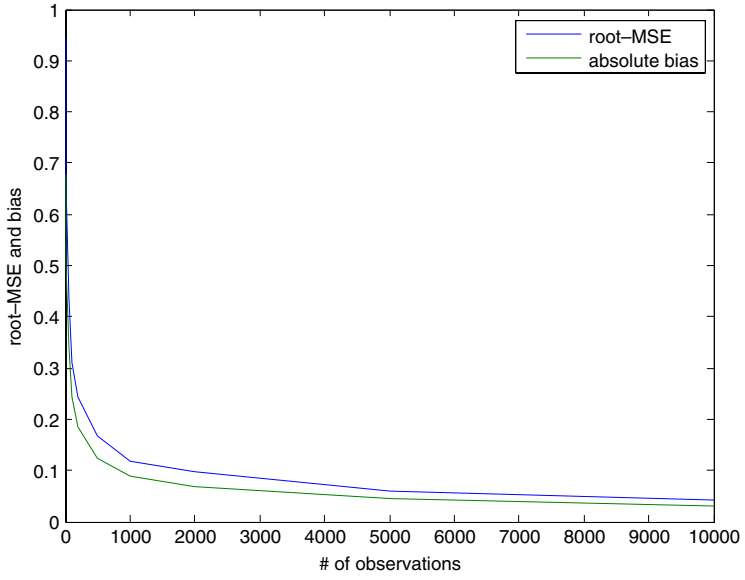

Fig. 3. Root-mean squared error and bias for formula H_3 with sequence log n/n^2 and $\gamma = 0.5$.

probabilities that are the same up to the fourth significant digit. This is compatible with our theoretical results that connect the smoothness of the objective function and the importance of the choice of the step size sequence.

We also analyze the performance of the root-mean square error, standard deviation and bias of the obtained estimator for the variance. To do so we notice that if Z is a standard normal random variable, then the maximizer of the population objective function is $\theta_0 = 0$. As a result, we can express

$$V_g = (1+\gamma)^2 E\left[|Y|^{2\gamma}\right].$$

Thus, the formula for the variance of the estimator can be written as

$$V_\theta = E\left[|Y|^{1+\gamma}\right]^{-2} E\left[|Y|^{2\gamma}\right]. \tag{4.5}$$

We computed the variance using formula (4.5) via Monte Carlo integration with 500,000 random draws. Then we compared the result with the empirical estimates. Figs. 2–4 demonstrate the decline of the root-mean squared error and the absolute bias with the sample size for the third-order finite difference formula used to compute the Hessian. It is clear that the decline occurs more rapidly for the smoother objective function (compare the rate of decline for $\gamma = 0.25$ versus $\gamma = 0.75$). The absolute bias declines at the same rate as the root-mean squared error. The precision of the estimates substantially improves towards the large samples with the rate of improvement close to geometric, (see Tables 4–6).

5. Conclusion

In this paper we study the impact of using numerical finite-difference approximations on the properties of estimates of derivatives of estimated functions, focusing on the case where the function is computed from a cross-sectional data sample. We provide weak sufficient conditions for uniformly consistent estimation of the gradients and the directional derivatives of semiparametric moments. Such results can be used to examine numerically evaluated Hessians (in estimating asymptotic variances), efficient weighting matrices and optimal instruments used

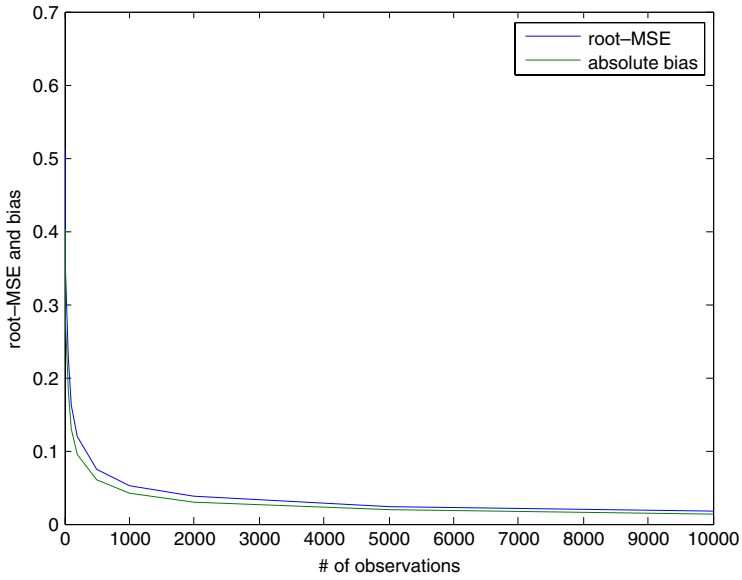

Fig. 4. Root-mean squared error and bias for formula H_3 with sequence $\log n/n^2$ and $\gamma = 0.75$.

Table 4
Rejection rates for H_0 with $\gamma = 0.5$, $\alpha = 0.05$.

n	ϵ_n			
	$\log n/n^2$	$\log n/n^{1.5}$	$\log n/n$	$\log n/n^{0.5}$
$H_2(\theta_0)$				
10	0.1341	0.1341	0.1341	0.1341
20	0.1011	0.1011	0.1006	0.1006
50	0.0771	0.0771	0.0776	0.0776
100	0.0626	0.0631	0.0631	0.0631
200	0.0576	0.0571	0.0571	0.0576
500	0.0506	0.0506	0.0501	0.0501
1 000	0.0470	0.0470	0.0465	0.0465
2 000	0.0575	0.0575	0.0575	0.0580
5 000	0.0515	0.0515	0.0515	0.0510
10 000	0.0540	0.0540	0.0540	0.0540
20 000	0.0515	0.0512	0.0515	0.0525
50 000	0.0505	0.0505	0.0505	0.0499
100 000	0.0500	0.0500	0.0500	0.0500
$H_3(\theta_0)$				
10	0.1341	0.1341	0.1341	0.1341
20	0.1011	0.1011	0.1006	0.1006
50	0.0771	0.0771	0.0776	0.0776
100	0.0626	0.0631	0.0631	0.0621
200	0.0576	0.0576	0.0571	0.0576
500	0.0506	0.0506	0.0506	0.0496
1 000	0.0470	0.0470	0.0465	0.0470
2 000	0.0575	0.0575	0.0575	0.0575
5 000	0.0515	0.0515	0.0515	0.0510
10 000	0.0540	0.0540	0.0540	0.0550
20 000	0.0515	0.0512	0.0515	0.0505
50 000	0.0505	0.0505	0.0505	0.0500
100 000	0.0500	0.0500	0.0500	0.0500

Table 5
Rejection rates for H_0 with $\gamma = 0.75$, $\alpha = 0.05$.

n	ϵ_n			
	$\log n/n^2$	$\log n/n^{1.5}$	$\log n/n$	$\log n/n^{0.5}$
$H_1^+(\theta_0)$				
10	0.0861	0.0861	0.0861	0.0861
20	0.0726	0.0725	0.0725	0.0725
50	0.0570	0.0570	0.0570	0.0570
100	0.0565	0.0565	0.0565	0.0565
200	0.0560	0.0560	0.0560	0.0560
500	0.0495	0.0495	0.0495	0.0495
1 000	0.0535	0.0535	0.0535	0.0535
2 000	0.0535	0.0535	0.0535	0.0535
5 000	0.0520	0.0520	0.0520	0.0520
10 000	0.0510	0.0510	0.0510	0.0510
10 000	0.0510	0.0510	0.0510	0.0510
20 000	0.0500	0.0500	0.0500	0.0500
50 000	0.0500	0.0500	0.0505	0.0500
100 000	0.0500	0.0500	0.0500	0.0500
$H_1^-(\theta_0)$				
10	0.0856	0.0856	0.0856	0.0856
20	0.0726	0.0726	0.0726	0.0726
50	0.0570	0.0570	0.0570	0.0570
100	0.0565	0.0565	0.0565	0.0565
200	0.0560	0.0560	0.0560	0.0560
500	0.0495	0.0495	0.0495	0.0495
1 000	0.0535	0.0535	0.0535	0.0535
2 000	0.0535	0.0535	0.0535	0.0535
5 000	0.0520	0.0520	0.0520	0.0520
10 000	0.0510	0.0510	0.0510	0.0510
10 000	0.0510	0.0510	0.0510	0.0510
20 000	0.0500	0.0500	0.0500	0.0500
50 000	0.0500	0.0500	0.0505	0.0500
100 000	0.0500	0.0500	0.0500	0.0500

in applied research. We study M-estimators where the optimization routine uses a finite-difference approximation to the gradient. Finite-point approximation formulas use tuning parameters such as the step size. We find that the presence of such parameters may affect the statistical properties of the resultant extremum estimator, and that the properties of the estimator obtained from the numerical optimization routine depend on the interaction between the precision of approximation, and the smoothness of the population and sample objective functions. Furthermore, in ongoing research we also extend the results in this paper to rank-type criterion functions that involve U-statistics with multiple sample summations (Hong et al. (2012)).

Acknowledgments

We would like to thank Tim Armstrong for insightful comments and crucial inputs in obtaining the most general version of the main theorem of the paper. We also thank the coeditor, the associate editor, two anonymous referees, and participants at numerous seminars and conferences for helpful comments. Generous supports from the National Science Foundation (SES 1024504), the University of California at Berkeley and Stanford University are acknowledged. The usual disclaimer applies.

Table 6
Rejection rates for H_0 with $\gamma = 0.75$, $\alpha = 0.05$.

n	ϵ_n			
	$\log n/n^2$	$\log n/n^{1.5}$	$\log n/n$	$\log n/n^{0.5}$
$H_2(\theta_0)$				
10	0.0861	0.0861	0.0861	0.0861
20	0.0726	0.0726	0.0726	0.0726
50	0.0570	0.0570	0.0570	0.0570
100	0.0565	0.0565	0.0565	0.0565
200	0.0560	0.0560	0.0560	0.0560
500	0.0495	0.0495	0.0495	0.0495
1 000	0.0535	0.0535	0.0535	0.0535
2 000	0.0535	0.0535	0.0535	0.0535
5 000	0.0520	0.0520	0.0520	0.0520
10 000	0.0510	0.0510	0.0510	0.0510
10 000	0.0510	0.0510	0.0510	0.0510
20 000	0.0500	0.0500	0.0500	0.0500
50 000	0.0500	0.0500	0.0505	0.0500
100 000	0.0500	0.0500	0.0500	0.0500
$H_3(\theta_0)$				
10	0.0861	0.0861	0.0861	0.0861
20	0.0726	0.0726	0.0726	0.0726
50	0.0570	0.0570	0.0570	0.0570
100	0.0565	0.0565	0.0565	0.0565
200	0.0560	0.0560	0.0560	0.0560
500	0.0495	0.0495	0.0495	0.0495
1 000	0.0535	0.0535	0.0535	0.0535
2 000	0.0535	0.0535	0.0535	0.0535
5 000	0.0520	0.0520	0.0520	0.0520
10 000	0.0510	0.0510	0.0510	0.0510
10 000	0.0510	0.0510	0.0510	0.0510
20 000	0.0500	0.0500	0.0500	0.0500
50 000	0.0500	0.0500	0.0505	0.0500
100 000	0.0500	0.0500	0.0500	0.0500

Appendix

A.1. Uniform consistency of directional derivatives for semiparametric models

This subsection extends the weak consistency condition to directional derivatives of semiparametric conditional moment models.

Consider a general conditional moment model defined by, for known $m(\cdot)$ and unknown $\rho(\cdot)$,

$$m(z, \theta, \eta(\cdot)) = E[\rho(Y, \theta, \eta(\cdot)) | Z = z] = 0,$$
$$\text{if and only if} \quad (\theta, \eta(\cdot)) = (\theta_0, \eta_0(\cdot)).$$

The parameters above comprise a finite-dimensional component, $\theta \in \Theta \subset \mathbb{R}^d$, and an infinite-dimensional component $\eta(\cdot) \in \mathcal{H}$ that is contained in the Banach space $\mathcal{H}$. This setup includes the unconditional parametric moment as a special case where the "instrument" z is a constant and $\eta(\cdot)$ is not present. Because the moment condition $m(\cdot)$ can be multi-dimensional, this setup also includes two step and multi-step step estimators, when some of the moment conditions corresponding to initial stage estimators only depend on the infinite-dimensional functions $\eta(\cdot)$. Semiparametric estimators for this general model and their asymptotic distributions are studied extensively in the literature. In some models, the moment conditions $\rho(y, \theta, \eta(\cdot))$ depend only on the value of the function $\eta(\cdot)$ evaluated at the argument y. In some other models, such as in dynamic discrete choice models and dynamic games, $\rho(y, \theta, \eta(\cdot))$ may depend on the entire function of $\eta(\cdot)$ in complex ways, e.g. through value function iterations.

The sieve approach, studied in a sequence of papers by Chen and Shen (1998), Ai and Chen (2003) and Chen and Pouzo (2009), approximates the class of infinite dimensional functions $\mathcal{H}$ using a parametric family of functions $\mathcal{H}_n$ whose dimension increases to infinity with the sample size n. Given the complexity of conditional moment functions in many relevant applications, finding a closed form expression of the variance of the finite-dimensional parameters in semiparametric problems can be challenging or infeasible.

A variance formula for the semiparametric minimum distance estimators in Ai and Chen (2003) [hereafter AC] can illustrate the use of numerical derivative approximation. For any $w \in \mathcal{H}$ and $\alpha = (\theta, \eta)$, denote by

$$\frac{\partial m(Z, \alpha)}{\partial \eta}[w] = \left.\frac{dm(Z, \theta, \eta + \tau w)}{d\tau}\right|_{\tau=0}$$

the directional derivative of $m(Z, \alpha)$ with respect to the η component in the w direction. Under conditions given in AC, $\hat{\theta}$ (Equation (4), pp 1798 in AC, where $\hat{\eta}$ is also defined) is $\sqrt{n}$ consistent and asymptotically normal while $\hat{\eta}$ obtains the optimal nonparametric convergence rate for η. Consistent inference for θ requires estimation of the ordinary derivative of the objective function with respect to the finite-dimensional parameter and the directional derivative with respect to the infinite-dimensional parameter

$$D_{w_j}(z) \equiv \frac{\partial m(Z, \alpha)}{\partial \theta_j} - \frac{\partial m(Z, \alpha)}{\partial \eta}[w_j] \tag{A.6}$$

uniformly consistently in various directions w_j. Ackerberg et al. (2012) further show that treating the entire estimation procedure for α as parametric and reading off the variance of $\hat{\theta}$ from the upper-left block of an estimate of the asymptotic variance–covariance matrix of $\hat{\alpha} = \left(\hat{\theta}, \hat{\eta}\right)$ will give consistent estimates of the asymptotic variance of the parametric component. A related method for consistent asymptotic variance estimation when kernel methods are used to estimate the nonparametric component was developed in Newey (1994). However, in many practical estimation problems, the derivatives of $\frac{\partial \hat{m}(Z, \hat{\alpha})}{\partial \theta_j}$ and $\frac{\partial \hat{m}(Z, \hat{\alpha})}{\partial \eta}[w_j]$ do not have analytic solutions and have to be evaluated numerically. See, e.g., Aradillas-Lopez (2010) and Hong and Shum (2010). This might be the case even if $\rho(\cdot)$ is linear in the infinite-dimensional function $\eta(\cdot)$. For example in dynamic models typically $\rho(y; \theta_0, \eta_0(\cdot)) = \eta_0(z) - \Gamma(\eta_0(x)) - f(x, z; \theta)$ for a known parametric function $f(x, z; \theta)$ and $y = (x, z)$. Typically the unknown conditional expectation function $m(z, \theta, \eta(\cdot))$ needs to be estimated from the data nonparametrically.

In this section we focus on two special cases where the conditional moment function is estimated nonparametrically using orthogonal series and when it is estimated using kernel smoothing. The infinite-dimensional parameter η is assumed to be estimated using sieves. The series estimator used to recover the conditional moment function is based on the vector of basis functions $p^N(z) = (p_{1N}(z), \ldots, p_{NN}(z))'$,

$$\widehat{m}(\theta, \eta, z) = p^{N\prime}(z)\left(\frac{1}{n}\sum_{i=1}^{n} p^N(z_i)p^{N\prime}(z_i)\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} p^N(z_i)\rho(y_i, \theta, \eta). \tag{A.7}$$

The kernel estimator is defined using a multi-dimensional kernel function $K(\cdot)$ and a bandwidth sequence b_n as

$$\widehat{m}(\theta, \eta, z) = \left(\frac{1}{nb_n^{d_z}}\sum_{i=1}^{n} K\left(\frac{z_i - z}{b_n^{d_z}}\right)\right)^{-1}\frac{1}{nb_n^{d_z}} \times \sum_{i=1}^{n} K\left(\frac{z_i - z}{b_n}\right)\rho(y_i, \theta, \eta). \tag{A.8}$$

In either case, we will denote the resulting estimate by $\widehat{m}(\theta, \eta, z)$. It turns out that the numerical derivative consistency results for η apply without any modification to the parametric

component θ. Therefore without loss of generality below we will focus on differentiating with respect to η.

The directional derivative of m with respect to η in the direction $w \in \mathcal{H} - \eta_0$, $G_w = \frac{d\, m(\theta_0, \eta_0 + \tau w, z)}{d\tau}\Big|_{\tau=0}$, is estimated using $L_{1,p}^{\epsilon_n, w}\widehat{m}\left(\hat{\theta}, \hat{\eta}, z\right)$, where an additional index is used to emphasize the direction for which the derivative is taken,

$$L_{1,p}^{\epsilon_n, w}\widehat{m}\left(\hat{\theta}, \hat{\eta}, z\right) = \frac{1}{\epsilon_n}\sum_{l=-p}^{p} c_l \widehat{m}\left(\hat{\theta}, \hat{\eta} + lw\,\epsilon_n, z\right).$$

Given that the direction w itself has to be estimated from the data, we desire consistency results that hold uniformly both around the true parameter value and the directions of numerical differentiation. As in our analysis of parametric models, we focus on i.i.d data samples. We also impose standard assumptions on the basis functions as in Newey (1997). Well known conditions that satisfy Assumption 6 are available in, for example, the handbook chapter by Chen (2007).

Assumption 6. For the basis functions $p^N(z)$ the following holds:

(i) The smallest eigenvalue of $E\left[p^N(Z_i)\, p^{N\prime}(Z_i)\right]$ is bounded away from zero uniformly in N.[3]

(ii) For some $\zeta_0(N)$ such that $\zeta_0(N)^2 N/n \to 0$, $\sup_{z\in\mathcal{Z}} \|p^N(z)\| \leq \zeta_0(N)$.

(iii) The population conditional moment belongs to the completion of the sieve space and for some $\alpha > 0$,

$$\sup_{(\theta,\eta)\in\Theta\times\mathcal{H}} \sup_{z\in\mathcal{Z}} \left\| m(\theta, \eta, z) - proj\left(m(\theta, \eta, z) \mid p^N(z)\right)\right\| = O\left(N^{-\alpha}\right).$$

When all the basis functions are uniformly bounded, typically $\zeta_0(N) = \sqrt{N}$. In the above

$$proj\left(m(\theta, \eta, z) \mid p^N(z)\right) = p^N(z)'\left(Ep^N(z)p^N(z)'\right)^{-1} Ep^N(z)m(\theta, \eta, z).$$

The following assumption on the moment function $\rho(\cdot)$ does not require smoothness or continuity (Shen and Wong, 1994; Zhang and Gijbels, 2003).

Assumption 7. (i) The moment functions are uniformly bounded: $\sup_{\theta,\eta,y} \|\rho(y, \theta, \eta)\| \leq C$. The density of covariates Z is uniformly bounded away from zero on its support.

(ii) Suppose that $0 \in \mathcal{H}_n$ and for $\epsilon_n \to 0$ and some $C > 0$,

$$\sup_{\substack{z\in\mathcal{Z},\, \eta,\, w\in\mathcal{H}_n,\, |\eta|,\, |w| < C,\\ \theta\in\mathcal{N}(\theta_0)}} \mathrm{Var}\left(\rho(Y_i, \theta, \eta + \epsilon_n w) - \rho(Y_i, \theta, \eta - \epsilon_n w) \mid Z_i = z\right) = O(\epsilon_n),$$

(iii) For each n, the class of functions $\mathcal{F}_n = \{\rho(\cdot, \theta, \eta + \epsilon_n w) - \rho(\cdot, \theta, \eta - \epsilon_n w),\ \theta \in \Theta,\ \eta, w \in \mathcal{H}_n\}$ is Euclidean whose graphs form a polynomial class of sets and whose coefficients depend on the number of sieve terms. There exist constants A, and $0^+ \leq r_0 < \frac{1}{2}$ such that the covering number satisfies

$$\log N(\delta, \mathcal{F}_n, \mathbf{L}_1) \leq A n^{2r_0} \log\left(\frac{1}{\delta}\right),$$

and for $r_0 = 0^+$, n^{0+} is defined as $\log n$.

[3] We note that the considered series basis may not be orthogonal with respect to the semi-metric defined by the distribution of Z_i.

Sufficient primitive conditions. Assumption 7(iii) is a high level condition and its verification will depend upon the application at hand. In general, this assumption imposes a joint restriction both on the class of functions $\mathcal{H}_n$ containing sieve estimators for η and the class of conditional moment functions parameterized both by θ and η. It is possible to provide lower level sufficient conditions. In some cases the entropy bounds required in Assumption 7(iii) can be provided in terms of the entropy of the class $\mathcal{H}_n$. This includes the case when $\rho(\cdot)$ is (weakly) monotone in η for each θ and $\mathcal{H}_n$ is an orthogonal basis of dimensionality $K(n)$. For example, $\rho(\cdot)$ can be an indicator function in a nonparametric quantile regression. Lemma 5 in Shen and Wong (1994) suggests that the L_1-metric entropy of the class of sieve $\mathcal{F}_n$ has order $K(n)\log\frac{1}{\epsilon}$ for sufficiently small $\epsilon > 0$ and $\|\eta_n - \eta_0\|_{\mathbf{L}_1} < \epsilon$. Then by Lemma 2.6.18 in Van der Vaart and Wellner (1996), if the function $\rho(\cdot)$ is monotone, its application to η (for fixed θ) does not increase the metric entropy. In addition, the proof of Theorem 3 in Chen et al. (2003) shows that the metric entropy for the entire class $\mathcal{F}_n$ is a sum of metric entropies that are obtained by fixing η and θ. The choice $K(n) \sim n^{2r_0}$ delivers condition 7(iii).

Denote $\pi_n\eta = \arg\inf_{\eta'\in\mathcal{H}_n} \|\eta' - \eta\|$. Let $d(\cdot)$ be the metric generated by the $\mathbf{L}^1$ norm. The following result extends Theorem 37 of Pollard (1984) to the case of sieve estimators. A related idea for unconditional sieve estimation has been used in Zhang and Gijbels (2003).

Lemma 2. *Suppose that* $d(\pi_n\eta, \eta) = O\left(n^{-\phi}\right)$. *Under Assumptions 6 and 7*

$$\sup_{d(\theta,\theta_0)=o(1), d(\eta,\eta_0)=o(1), \eta\in\mathcal{H}_n} \left|L_{1,p}^{\epsilon_n, w}\widehat{m}(\theta, \eta, z) - L_{1,p}^{\epsilon_n, w}m(\theta, \eta, z)\right| = o_p(1)$$

uniformly in z and w, provided that $\epsilon_n \to 0$ and $\min\{N^\alpha, n^\phi\}\epsilon_n \to \infty$, and $\frac{n\epsilon_n}{\zeta_0(N)^2 N n^{2r_0}\log n} \to \infty$.

Next we provide a similar result for the case where the conditional moment function is estimated via a kernel estimator. We begin with formulating the requirement on the kernel.

Assumption 8. The kernel function $K(\cdot)$ satisfies condition (iii) in Assumption 3. Moreover, it integrates to 1, is bounded and of qth order, and is square-integrable.

We formulate the following lemma replicating the result of Lemma 2 for the case of the kernel estimator. For uniformity we rely on Assumption 7(i) that requires the density of covariates to be uniformly bounded away from zero.

Lemma 3. *Under Assumptions 7 and 8*

$$\sup_{d(\theta,\theta_0)=o(1), d(\eta,\eta_0)=o(1), \eta\in\mathcal{H}_n} \left|L_{1,p}^{\epsilon_n, w}\widehat{m}(\theta, \eta, z) - L_{1,p}^{\epsilon_n, w}m(\theta, \eta, z)\right| = o_p(1)$$

uniformly in w and z where $f(z)$ is strictly positive for the kernel estimator provided that $\epsilon_n \to 0$, $b_n \to 0$, $\epsilon_n \min\{b_n^{-q}, n^\phi\} \to \infty$ and $\frac{n\epsilon_n b_n^{d_z}}{n^{2r_0}\log n} \to \infty$.

Using Lemmas 2 and 3 we can formulate the consistency result for the directional derivative.

Theorem 5. *Under Assumptions 2 and 7, and either 6 or 8,* $L_{1,p}^{\epsilon_n, w}\widehat{m}\left(\hat{\theta}, \hat{\eta}, z\right) \xrightarrow{p} \frac{\partial m(\theta,\eta,z)}{\partial\eta}[w]$, *uniformly in z and w, if $\epsilon_n \to 0$, $N \to \infty$, $\epsilon_n \min\{N^\alpha, n^\phi\} \to \infty$, and $\frac{n\epsilon_n}{\zeta_0(N)^2 N n^{2r_0}\log n} \to \infty$ for series estimator, and $\epsilon_n, b_n \to 0$, $\epsilon_n \min\{b_n^{-q}, n^\phi\} \to \infty$, and $\frac{n\epsilon_n b_n^{d_z}}{n^{2r_0}\log n} \to \infty$ for kernel-based estimator, provided that $d\left(\hat{\theta}, \theta_0\right) = o_p(1)$ and $d\left(\hat{\eta}, \eta_0\right) = o_p(1)$.*

This theorem allows us to use finite-difference formulas to evaluate directional derivatives. An interesting feature of this result is that it only depends on the rate of convergence of the infinite-dimensional parameter indirectly through Assumption 7(iii) which implicitly bounds the number of sieve terms that one can use by n^{2r_0} with $r_0 < \frac{1}{2}$, i.e. it has to increase slower than the sample size.

Remark. Our results in this section apply to the case where one is interested in obtaining a finite-difference based estimator for the directional derivative that is uniformly consistent over z. Such a need may arise where the direction of differentiation is also estimated, an example of which is the efficient sieve minimum distance estimator in Ai and Chen (2003). If one only needs to estimate the numerical derivative pointwise the conditions on the choice of the step size can be weakened. Such results may be relevant when one is interested in estimating the directional derivative at a point and a given direction.

A.2. Proof of Theorem 3

Proof. The rate of convergence adapts the proof of Theorem 3.2.5 of Van der Vaart and Wellner (1996) to our case. Denote the rate of convergence for the estimator $\widehat{\theta}$ by ρ_n. Then we can partition the parameters space into sets $S_{j,n} = \left\{\theta : 2^{j-1} < \rho_n d(\theta, \theta_0) < 2^j\right\}$. Then we evaluate the probability of a large deviation $\rho_n d\left(\widehat{\theta}, \theta_0\right) > 2^M$ for some integer M, where $\rho_n = \sqrt{n}{\epsilon_n}^{1-\gamma}$. We know that the estimator solves, for $r_n = {\epsilon_n}^{1-\gamma}$

$$\sqrt{n} r_n L_{1,p}^{\epsilon_n} Q_n\left(\widehat{\theta}\right) = o_p(1).$$

If $\rho_n d\left(\hat{\theta}, \theta_0\right)$ is larger than 2^M for a given M, then over the θ in one of the shells $S_{j,n}$, $\sqrt{n} r_n L_{1,p}^{\epsilon_n} Q_n(\theta)$ achieves a distance as close as desired to zero. Hence, for every $\delta > 0$,

$$\begin{aligned} &P\left(\rho_n d\left(\widehat{\theta}, \theta_0\right) > 2^M\right) \\ &\leq \sum_{\substack{j \geq M \\ 2^j < \delta \rho_n}} P\left(\sup_{\theta \in S_{j,n}} \left(-\left\|L_{1,p}^{\epsilon_n} Q_n(\theta)\right\|\right) \geq -o_p\left(\frac{1}{\sqrt{n} r_n}\right)\right) \\ &\quad + P\left(2 d\left(\widehat{\theta}, \theta_0\right) \geq \delta\right). \end{aligned}$$

Note that mean square differentiability implies that for every θ in a neighborhood of θ_0, $g(\theta) - g(\theta_0) \lesssim -d^2(\theta, \theta_0)$. Then we evaluate the population objective, using the fact that it has p mean-square derivatives:

$$\left\|L_{1,p}^{\epsilon_n} Q(\theta)\right\| \geq C d(\theta, \theta_0) + C' \epsilon_n^{\nu-1},$$

where θ_0 is the zero of the population first-order condition and the approximated derivative has a known order of approximation $\|L_{1,p}^{\epsilon_n} Q(\theta_0)\| = C' \epsilon_n^{\nu-1}$ for some constant C'. Substitution of this expression into the argument of interest leads to

$$\begin{aligned} \left\|L_{1,p}^{\epsilon_n} Q(\theta) - L_{1,p}^{\epsilon_n} Q_n(\theta)\right\| &\geq \left\|L_{1,p}^{\epsilon_n} Q(\theta)\right\| - \left\|L_{1,p}^{\epsilon_n} Q_n(\theta)\right\| \\ &\geq C d(\theta, \theta_0) + C' \epsilon_n^{\nu-1} + o_p\left(\frac{1}{\sqrt{n} r_n}\right). \end{aligned}$$

Then applying the Markov inequality to the re-centered process for $\theta \in S_{j,n}$

$$\begin{aligned} &P\left(r_n \sqrt{n}\left\|L_{1,p}^{\epsilon_n} Q(\theta) - L_{1,p}^{\epsilon_n} Q_n(\theta)\right\| \geq C r_n \sqrt{n} d(\theta, \theta_0)\right. \\ &\quad \left. + C' r_n \sqrt{n} \epsilon_n^{\nu-1} + o(1)\right) \leq C' r_n^{-1} n^{-1/2} \left(\frac{2^j}{\rho_n}\right)^{-1}. \end{aligned}$$

Then $\rho_n = \sqrt{n}$ in the regular case and $\rho_n = r_n \sqrt{n}$ in cases where $\gamma \neq 1$.

Finally also note that the evaluation for the expectation holds for $\theta = \theta_0 \pm t_k \epsilon_n$, as shown above. By Markov's inequality according to Theorem 2.5.2 from van der Vaart and Wellner (1998) it follows that the process $r_n \sqrt{n} L_{1,p}^{\epsilon_n} Q_n(\theta_0)$ indexed by ϵ_n is P-Donsker. □

A.3. Proof of Theorem 4

Proof. The result will follow if we can demonstrate that

$$\sqrt{n} r_n\left(L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right) - L_{1,p}^{\epsilon_n} \hat{g}(\theta_0) - G\left(\hat{\theta}\right) + G(\theta_0)\right) = o_p(1). \quad \text{(A.9)}$$

Because of the assumption that $\sqrt{n}{\epsilon_n}^{\nu-\gamma} \to \infty$, the bias is sufficiently small. Therefore this is equivalent to showing that

$$\begin{aligned} &\sqrt{n} r_n\left(L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right) - L_{1,p}^{\epsilon_n} \hat{g}(\theta_0) - E L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right) + E L_{1,p}^{\epsilon_n} \hat{g}(\theta_0)\right) \\ &= o_p(1). \end{aligned}$$

Because of the convergence rate established in Theorem 3, this will be implied by, with $r_n = {\epsilon_n}^{1-\gamma}$:

$$\begin{aligned} \sup_{d(\theta,\theta_0) \lesssim O\left(\frac{1}{\sqrt{n}{\epsilon_n}^{1-\gamma}}\right)} &\sqrt{n} r_n\left(L_{1,p}^{\epsilon_n} \hat{g}(\theta) - L_{1,p}^{\epsilon_n} \hat{g}(\theta_0) - E L_{1,p}^{\epsilon_n} \hat{g}(\theta)\right. \\ &\left. + E L_{1,p}^{\epsilon_n} \hat{g}(\theta_0)\right) = o_p(1). \end{aligned}$$

The left hand side can be written as a linear combination of the empirical processes:

$$\begin{aligned} \sup_{d(\theta,\theta_0) \lesssim O\left(\frac{1}{\sqrt{n}{\epsilon_n}^{1-\gamma}}\right)} &\sqrt{n} \frac{r_n}{\epsilon_n}\left[\mathbb{G}_n(\theta + t\epsilon_n) - \mathbb{G}_n(\theta - t\epsilon_n)\right. \\ &\left. - \mathbb{G}_n(\theta_0 + t\epsilon_n) - \mathbb{G}_n(\theta_0 - t\epsilon_n)\right]. \end{aligned}$$

Because of Assumption 4, it is bounded stochastically by

$$O_p\left(\frac{r_n}{\epsilon_n} \min\left(d(\theta, \theta_0), \epsilon_n\right)^{\gamma}\right).$$

When $\sqrt{n}{\epsilon_n}^{2-\gamma} \to \infty$, $d(\theta, \theta_0) \lesssim O\left(\frac{1}{\sqrt{n}{\epsilon_n}^{1-\gamma}}\right) = o(\epsilon_n)$. Hence the above display is $o_p(1)$. Therefore (A.9) holds.

Recall that $\hat{\theta}$ is defined by $\sqrt{n} r_n(L_{1,p}^{\epsilon_n} \hat{g}(\hat{\theta})) = o_p(1)$. Then (A.9) implies that, using a first order Taylor expansion of $G(\theta)$:

$$\begin{aligned} &\sqrt{n} r_n\left(L_{1,p}^{\epsilon_n} \hat{g}(\theta_0) - E L_{1,p}^{\epsilon_n} \hat{g}(\theta_0)\right) + H(\theta_0) \sqrt{n} r_n\left(\hat{\theta} - \theta_0\right) \\ &= o_p(1). \quad \square \end{aligned}$$

References

Ackerberg, D., Chen, X., Hahn, J., 2012. A practical asymptotic variance estimator for two-step semiparametric estimators. Rev. Econom. Statist. 94 (2), 481–498.

Ai, C., Chen, X., 2003. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica 71, 1795–1843.

Anderssen, R., Bloomfield, P., 1974. Numerical differentiation procedures for non-exact data. Numer. Math. 22, 157–182.

Andrews, D., 1997. A stopping rule for the computation of generalized method of moments estimators. Econometrica 65 (4), 913–931.

Aradillas-Lopez, A., 2010. Semiparametric estimation of a simultaneous game with incomplete information. J. Econometrics 157 (2), 409–431.

Brown, B., Wang, Y., 2005. Standard errors and covariance matrices for smoothed rank estimators. Biometrika 92 (1), 149–158.

Chen, X., 2007. Large sample sieve estimation of semi-nonparametric models. Handbook Econom. 6, 5549–5632.

Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. Econometrica 71, 1591–1608.

Chen, X., Pouzo, D., 2009. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. J. Econometrics 152 (1), 46–60.

Chen, X., Shen, X., 1998. Sieve extremum estimates for weakly dependent data. Econometrica 66, 289–314.

Dudley, R., 1999. Uniform Central Limit Theorems. university press, Cambridge.

Hong, H., Mahajan, A., Nekipelov, D., 2012, Numerical Gradients and Extremum Estimation with U-statistics. Working Paper, UC Berkeley, UCLA and Stanford University.
Hong, H., Shum, M., 2010. Pairwise-difference estimation of a dynamic optimization model. Rev. Econ. Stud. 77 (1), 273–304.
Horowitz, J., 1992. A smoothed maximum score estimator for the binary response models. Econometrica 60.
Johnson, L., Strawderman, R., 2009. Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. Biometrika 96 (3), 577–590.
Judd, K., 1998. Numerical Methods in Economics. MIT Press.
Kato, K., 2012. Asymptotic normality of Powell's kernel estimator. Ann. Inst. Statist. Math. 64 (2), 255–273.
Kosorok, M., 2008. Introduction to Empirical Processes and Semiparametric Inference. Springer Verlag.
Kristensen, D., Salanié, B., 2010, 2013, Higher-order properties of approximate estimators, CAM Working Papers, University of Copenhagen, Columbia University, University College London.
L'Ecuyer, P., Perron, G., 1994. On the convergence rates of IPA and FDC derivative estimators. Oper. Res. 42, 643–656.
Murphy, S., Van der Vaart, A., 2000. On profile likelihood. J. Amer. Statist. Assoc. 95.
Newey, W., 1997. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79, 147–168.
Newey, W., McFadden, D., 1994. Large Sample Estimation and Hypothesis Testing. In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. North Holland, pp. 2113–2241.
Newey, W.K., 1994. Kernel estimation of partial means and a general variance estimator. Econometric Theory 10 (02), 1–21.
Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57 (5), 1027–1057.
Pollard, D., 1984. Convergence of Stochastic Processes. Springer Verlag.
Powell, J.L., 1984. Least absolute deviations estimation for the censored regression model. J. Econometrics 25, 303–325.
Seo, M., Linton, O., 2007. A smoothed least squares estimator for threshold regression models. J. Econometrics 141 (2), 704–735.
Shen, X., Wong, W., 1994. Convergence rate of sieve estimates. Ann. Statist. 22, 580–615.
Van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer-Verlag, New York.
Wang, Y., Shao, Q., Zhu, M., 2009. Quantile regression without the curse of unsmoothness. Comput. Stat. Data Anal. 53 (10), 3696–3705.
Zhang, J., Gijbels, I., 2003. Sieve empirical likelihood and extensions of the generalized least squares. Scand. J. Stat. 1–24.
Zinde-Walsh, V., 2002. Asymptotic theory for some high breakdown point estimators. Econometric Theory 18 (5), 1172–1196.