Numerical Gradients and Extremum Estimation with U-statistics

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

First version: June 2009

This version: May 2012

Abstract

U-statistics are often used in semiparametric estimation in a variety of settings. An important obstacle to their widespread use is the computational cost — particularly when the kernel of the U-statistic is non-smooth (and smoothing is not desirable since it further increases computational complexity). In this paper we study the properties of an alternative procedure: finite-difference derivative approximations to evaluate the derivatives of U-statistic based objective functions and the use of these approximations in gradient-based optimization routines. We find that the existing guidelines in the literature for the choice of the step size for numerical differentiation are excessively conservative. While most analyses employing U-statistics suggest a choice of the step size of order $n^{-1/4}$, our findings show that uniform consistency of estimated derivatives (and the consistency of the resulting extremum estimator) is maintained even when the step size is chosen to be of order $\log n^2/n^2$. This implies that the step size can be chosen to approach to zero at a rate that is *faster* than the sample size. We illustrate our results with a Monte-Carlo study of a U-statistic based estimator for a semiparametric model with random censoring.

JEL Classification: C14; C52

Keywords: Numerical derivative, entropy condition, stochastic equicontinuity

[*c] Corresponding Author (Email: nekipelov@econ.berkeley.edu), Department of Economics, University of California at Berkeley, Berkeley, California, 94720; [a,b] Department of Economics, Stanford University, Stanford, CA 94305. 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 participants at numerous seminars and conferences for 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.

1 Introduction

U-statistics are widely used in applications ranging from random censoring problems to the analysis of valuation distributions in auctions. However, U-statistic based methods are often computationally intensive particularly with large datasets which are becoming more prevalent in practical applications. This is particularly problematic when the kernel function of the U-statistic is non-smooth, which is relatively common in practice.[2] One can smooth the objective function and use a standard gradient-based numerical routine but this increases computational complexity.

In this paper we consider evaluating the derivative of the objective function numerically. In particular, we examine the properties of finite-difference methods for evaluating gradients as applied to U-statistic based objective functions with non-smooth kernels. Our analysis is most relevant for large samples where traditional approaches to numerical analysis (such as smoothing or polynomial interpolation) of the objective function are computationally costly.

Finite-difference formulas allow us to work directly with the objective function subject to the choice of the step size parameter used for computing the finite differences. We consider the problems of estimating both the gradient and the extremum estimator that is computed using the gradient (using a numerical gradient-based optimization routine). We derive optimal rates for the step size while guaranteeing that the estimator for the gradient is uniformly consistent. We find that our rates are substantially faster than those in the literature on extremum estimation. In particular, our rate for indicator functions is of order log n^2/n^2 which is a considerable improvement over the rate of $n^{-1/4}$ in Sherman (1993). We note that this result is dramatically different from the result that was obtained in Hong, Mahajan, and Nekipelov (2012) where it was shown that numerical differentiation may affect the properties of the classical M-estimators and the step size requred to deliver consistent estimates for the derivatives should have order log n/n.

[2] For instance, the maximum rank correlation estimator (Han (1987) and Sherman (1993)) has an objective function that is characterized by a second-order U-statistic with an indicator kernel function.

We also characterize the properties of the extremum estimators obtained from gradient search routines that use finite-difference methods for gradient evaluation. We find that the choice of step size can affect the convergence rate and the asymptotic distribution of the maximizer of a U-statistic based objective function. These results depend upon the interaction between the order of the numerical differentiation and the properties of the sample objective function. Specifically, we find that if the kernel of the U-statistic is smooth, then the step size can decline arbitrarily fast (bounded from below by a function of the machine precision). We also find that the lower bound for the convergence rate for the step size is substantially faster than that in the existing literature. This implies that precise and robust estimates can be produced without smoothing and thereby opening the door to the use of U-statistic based methods for large samples. We illustrate our findings by analyzing the semiparametric estimation procedure for the random censoring model in Khan and Tamer (2007) and find that the finite-difference optimization routine is well behaved for moderate and large sample sizes.

We also provide a practical guide for choosing the step size, based on the combination of the bias-variance trade-off (as in classical estimation) as well as the trade-off between the variance and the bias arising from computer machine precision. We emphasize, though, that our analysis is aimed at the specific smoothing implied by commonly used numerical optimization routines standard in empirical work. Therefore solutions based on additional kernel smoothing are not considered since they are often not used by numerical optimization routines and they are often computationally costly or infeasbile.[3]

Various aspects of this problem have received some attention in the previous literature, especially in applications to U-statistics. Pakes and Pollard (1989), Newey and McFadden (1994) and Murphy and Van der Vaart (2000) provided sufficient conditions for using numerical derivatives to consistently estimate the asymptotic variance in a parametric model. The properties of numerical derivatives have, however, predominantly been investigated only for smooth models. For instance, Anderssen and Bloomfield (1974) analyzed derivative computations for functions

[3] For instance, the evaluation of the exponent with single precision takes over 5 times CPU time longer than the addition operation, required for the evaluation of the non-smoothed U-statistic.

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. Andrews (1997) considered the relationship between the numerical tolerance for computing GMM-type objective functions and their sample variance. Finally, while Hong, Mahajan, and Nekipelov (2012) consider the use of finite difference formulas for constructing consistent estimators for the derivatives of the objective functions of classical M-estimators, we focus on U-statistics in this paper. To the best of our knowledge, understanding the impact of numerical approximation on the statistical properties of gradient evaluation and extremum estimation in the context of U-statistics remains largely an open question.

The paper is organized as follows. Section 2 analyzes uniformly consistent estimation of numerical derivatives for parametric U-statistics. Section 3 examines the choice of the step size when a numerical gradient-based procedure is employed to compute the approximation to the maximum of the objective function. We also discuss consistency and the distribution theory for the resulting estimator. Section 4 discusses possible approaches to the adaptive choice of the step size. Section 5 demonstrates an application of our analysis to a random censoring problem and presents the Monte Carlo simulation evidence. Finally, section 6 concludes.

2 Estimation of derivatives from non-smooth sample functions

2.1 Definitions

We focus on second order U-statistics since they are the most commonly used in applications. A U-statistic objective function is defined by a symmetric function (the kernel) $f(Z_i, Z_j, \theta)$ as

$$\hat{g}(\theta) = \frac{1}{n\,(n-1)} S_n\,(f) \quad \text{where} \quad S_n\,(f) = \sum_{i \neq j} f\,(Z_i, Z_j, \theta)\,. \tag{2.1}$$

where $\{Z_i\}_{i=1}^n$ are i.i.d.. We denote the expectation with respect to the independent product measure on $\mathcal{Z} \times \mathcal{Z}$ by E_{zz} and the expectation with respect to a single measure by E_z. The population value can then be written as $g(\theta) = E_{zz} f\,(Z_i, Z_j, \theta)\,.$

The object of interest in this section will be the population objective function

$$g(\theta) = E_{zz}\left[f(Z, Z', \theta)\right]$$

as well as its gradient. We consider $\theta \in \Theta \subset \mathbb{R}^d$. Since we want the differentiation operation to be meaningful, we consider the case where the population objective function can be approximated well by a smooth function. Formally,

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

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

We consider the setting where the derivative of the objective function is an interesting object *per se*. However, given the structure of the problem, the closed-form expression for this derivative may not exist, for instance, because $g(\cdot)$ is defined as an implicit function of θ.

2.2 Numerical differentiation using finite differences

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

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

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 $\frac{\partial^k g(x)}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}$ with $L_{k,p}^{\epsilon} g(x)$

will be small:

$$L_{k,p}^{\epsilon}g\left(x\right)-\frac{\partial^{k}g(x)}{\partial x_{1}^{k_{1}}\ldots\partial x_{d}^{k_{d}}}=O(\epsilon^{2p+1-k}).$$

To illustrate the procedure for evaluating the coefficients $c_{l_1,\ldots,l_d}$ consider the case where $d=1$ and $r=2p$. The Taylor expansion is:

$$L_{k,p}^{\epsilon}g\left(x\right)=\frac{1}{\epsilon^{k}}\sum_{l=-p}^{p}c_{l}\left[\sum_{i=0}^{r}\frac{g^{(i)}(x)}{i!}\left(l\epsilon\right)^{i}+O\left(\epsilon^{r+1}\right)\right]=\sum_{i=0}^{r}g^{(i)}(x)\frac{\epsilon^{i}}{\epsilon^{k}}\sum_{l=-p}^{p}\frac{c_{l}l^{i}}{i!}+O\left(\epsilon^{r+1-k}\right).$$

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

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

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

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 restrict most attention to central derivatives.

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

Machine precision also imposes a lower bound on the step size in conjunction with the statistical lower bound (see, e.g. Press, Teukolsky, Vettering, and Flannery (1992)). While for most of the analysis we assume away machine imprecision, we do discuss it and related implementation issues in section 4.

2.3 Consistency of numerical derivatives

Following Serfling (1980), the following decomposition of the objective function into an empirical process and a degenerate U-process component can be used to establish the statistical properties of approximating $G(\theta_0) = \frac{\partial}{\partial\theta} g(\theta)$ by $L_{1,p}^{\varepsilon_n} \widehat{g}\left(\widehat{\theta}\right)$,

$$\hat{g}(\theta) = g(\theta) + \hat{\mu}_n(\theta) + \frac{1}{n(n-1)} S_n(u), \tag{2.2}$$

where

$$\hat{\mu}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mu(Z_i, \theta), \qquad \mu(z, \theta) = E_z f(Z_i, z, \theta) + E_z f(z, Z_i, \theta) - 2g(\theta),$$

and

$$u(z, z', \theta) = f(z, z', \theta) - E_z f(Z_i, z, \theta) - E_z f(z', Z_i, \theta) + g(\theta).$$

The class of kernel functions that we consider in this paper includes discontinuous and non-smooth functions that are of use to practitioners. Our main requirement is that the set of functions is not "too complex" and that we can bound their moments.

ASSUMPTION 2. *Consider functions $f(z, z', \theta)$ contained in class $\mathcal{F} = \{f(\cdot, \cdot, \theta + e_j\,\epsilon),\ \theta \in \Theta\}$ for $\epsilon > 0$. Assume*

(i) All $f \in \mathcal{F}$ are globally bounded such that $\|F\| = \sup_{\theta\in\Theta} |f(\cdot, \cdot, \theta)| < C_1 \ll \infty$.

(ii) The sample moment function has a uniform bound on the variance of its local deviations in some neighborhood of θ_0. That is, for sufficiently small $\epsilon > 0$ there exists a constant C_2 such that for each $j = 1, \ldots, p$

$$\sup_{\theta \in N(\theta_0)} E_{zz}\left[\left(f(Z, Z', \theta + \epsilon e_j) - f(Z, Z', \theta - \epsilon e_j)\right)^2\right] \le C_2 \epsilon.$$

(iii) The graphs of functions from $\mathcal{F}$ form a polynomial class of sets for any $\epsilon \to 0$ (depending on n).

Many functions used in applications fall in this category. By Lemmas 25 and 36 of Pollard (1984), Assumption 2 implies that there exist universal constants $A > 0$ and $V > 0$ such that for any $\mathcal{F}_n \subset \mathcal{F}$ with envelope function $\|F_n\|$,

$$\sup_{\mathcal{Q}} N_1\left(\varepsilon\, \mathcal{Q}F_n,\, \mathcal{Q}, \mathcal{F}_n\right) \leq A\varepsilon^{-V}, \quad \sup_{\mathcal{Q}} N_2\left(\varepsilon\, \left(\mathcal{Q}F_n^2\right)^{1/2},\, \mathcal{Q}, \mathcal{F}_n\right) \leq A\varepsilon^{-V},$$

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

Our discussion of the asymptotic distribution will rely on U-statistic projections which are expressed by a marginal expectation of the kernel function. This expectation is likely to be continuous in the parameters. Formally,

ASSUMPTION 3. *The projections $\mu(z, \theta)$ are Lipschitz-continuous in θ uniformly over z.*

This assumption depends on the distribution of Z_i. For instance, when the kernel of the U-statistic is defined by indicator functions, the expectation of this kernel will be continuous in the parameter for sufficiently smooth distribution of Z_i. Assumption 3controls the impact of numerical differentiation on the projection term by the maximum inequality for Lipschitz-continuous functions (see for example Theorem 3.2.5 in Van der Vaart and Wellner (1996)):

$$E_z \sup_{d(\theta,\theta_0)=o(1)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mu\left(Z_i, \theta+\epsilon_n\right) - \mu\left(Z_i, \theta-\epsilon_n\right) - g(\theta+\epsilon_n) + g(\theta-\epsilon_n)\right) \right| \leq C\epsilon_n,$$

for some $C > 0$.

We now consider the application of the finite difference formula to the objective function of the U-statistic. We start with the case where our goal is only to estimate the gradient of the population objective function at the point of interest θ_0, $G(\theta_0) = \frac{\partial g(\theta_0)}{\partial \theta}$ 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).

For convenience we decompose the error of approximation of $G(\theta_0)$ via the finite-difference formula applied to the sample analog $L_{1,p}^{\epsilon_n}\hat{g}\left(\hat{\theta}\right)$ 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

$$\hat{G}_1(\hat{\theta}) = L_{1,p}^{\epsilon_n}\hat{g}\left(\hat{\theta}\right) - L_{1,p}^{\epsilon_n} g\left(\hat{\theta}\right), \tag{2.3}$$

and

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

The term $G_3(\hat{\theta})$ is responsible for the error of approximating θ_0 with the estimate $\hat{\theta}$, and thus it does not depend on the step size in the finite difference formula ϵ_n. The term $G_2\left(\hat{\theta}\right)$ represents the bias due to approximating the derivative with a finite difference formula and can be controlled if the bias reduction is uniformly small in a neighborhood of θ_0. Finally, the term $\hat{G}_1(\hat{\theta})$ is responsible for replacement of the population objective function with its sample analog.

The following lemma states a set of weak sufficient conditions to ensure consistency of the estimated derivative.

LEMMA 1. *Suppose* $\|F\| = \sup_{\theta \in N(\theta_0)} |f\left(Z_i, Z_j, \theta\right)| \ll C < \infty$. *Under Assumption 2, if* $n^2 \varepsilon_n / \log^2 n \longrightarrow \infty$,

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

Consequently, $\hat{G}_1\left(\hat{\theta}\right) = o_p\left(1\right)$ *if* $d\left(\hat{\theta}, \theta_0\right) = o_p\left(1\right)$, *as defined in* (2.3).

The consistency of the numerical derivatives of U-statistics then follows directly from Lemma 1.

THEOREM 1. *Under assumptions 1 and 2 and the conditions of lemma 1,* $L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right) \xrightarrow{p} G\left(\theta_0\right)$ *if* $\epsilon_n \rightarrow 0$ *and* $n\epsilon_n^2 / \log^2 n \rightarrow \infty$, *and if* $d\left(\hat{\theta}, \theta_0\right) = o_p\left(1\right)$.

As in the case of the empirical process, this theorem establishes a very weak condition on the step size for numerical differentiation when the envelope function of the differenced moment function does not necessarily decrease with the shrinking step size. We note that the resulting condition for the step size is weaker in the case of the U-statistics relative to the case of the empirical sums. This is an artifact of the property that the projection of U-statistics tends to be smoother than the kernel function itself leading to a smaller modulus of continuity of the U-process.

We further note that these rate conditions for derivative estimation are substantial improvements over those previously found in the literature. For instance, Sherman (1993) uses the rate of $n^{-1/4}$.

Our rate of log n^2/n^2 is, for instance, compatible with the rate $n^{-3/2}$ thus allowing the step size to approach to zero *faster* than the sample size.

3 Numerical gradient-based estimation with U-statistics

3.1 Numerical optimization and numerical gradients

Consider extremum estimation where a population objective function is replaced with its sample analog. In cases where the sample analog represents a one-fold summation of the kernel function, the corresponding estimator is called an M-estimator. The maximizers of U-statistics represent a different class of extremum estimators where the sample analog represents a two-fold summation.

A common method for approachin the maximization problem is by solving a system of nonlinear equations represented by the first-order condition. In this section we analyze the problem of finding the solution of the first-order condition when the gradient of the objective function is replaced by its finite difference approximation and the objective function itself is replaced with its sample analog.

Consider the problem of estimating the parameter θ_0 in a metric space $(\Theta,\, d)$ with metric d. The true parameter θ_0 is assumed to uniquely maximize the limiting objective function $Q\left(\theta\right) = E_{zz} f\left(Z, Z'; \theta\right)$. We define an extremum estimator $\hat{\theta}$ of θ_0 as the maximizer of the U-statistic corresponding to the above expectation

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{g}\left(\theta\right), \tag{3.4}$$

where $\hat{g}\left(\theta\right) = \frac{1}{n(n-1)} S_n\left(f\right)$. Most sample objective functions $\hat{g}\left(\theta\right)$ used in practice do not have a closed form solution and maximization has to be performed numerically. Computerized optimization routines often use numerical derivatives either explicitly or implicitly. In this section we show that numerical differentiation leads to an implicit smoothing of the objective function and thus numerical derivatives can be used in place of the actual derivatives of the smoothed objective function, thus, facilitating the computations of the estimators. In particular, 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 possess different statistical properties than the maximizer of the objective function.

We consider the solution of the numerical first-order condition where the gradient is replaced with its finite-difference approximation described in Section 2.2. We explicitly represent the new estimator as a finite-difference approximation of the first order condition for problem (3.4). A numerical gradient-based optimization routine effectively substitutes (3.4) with the approximate solution to the non-linear equation

$$||L_{1,p}^{\varepsilon_n}\hat{g}\left(\hat{\theta}\right)|| = o_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.5}$$

for some sequence of step sizes $\varepsilon_n \to 0$. 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 $\hat{D}^{\varepsilon_n}\left(\hat{\theta}\right) \equiv L_{1,1}^{\varepsilon_n}\hat{g}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}}\right)$. The cases with $p \geq 2$ correspond to a more general class of estimators that will have smaller asymptotic bias in nonsmooth models. As we show, estimators (3.4) and (3.5) can have the same properties for models with continuous moment functions, but for non-smooth models both their asymptotic distributions and the convergence rates can be substantially different.

3.2 Consistency

Our first step is to provide a consistency result 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. This discussion is formalized in the following identification assumption.

ASSUMPTION 4. *The map defined by $D(\theta) = \frac{\partial}{\partial\theta}E_{zz}\left[f\left(Z, Z', \theta\right)\right]$ identifies $\theta_0 \in \Theta$. In other words from $\lim\limits_{n\to\infty} \|D(\theta_n)\| = 0$ it follows that $\lim\limits_{n\to\infty} \|\theta_n - \theta_0\| = 0$ for any sequence $\theta_n \in \Theta$. Moreover, $g(\theta) = E_{zz}\left[f\left(Z, Z', \theta\right)\right]$ is locally quadratic at θ_0 with $g(\theta) - g(\theta_0) \leq -\bar{H}d\left(\theta, \theta_0\right)^2$, for some $0 < \bar{H} < \infty$ amd 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 the previous Assumption 1 to establish uniform consistency for the estimate of the derivative of the sample moment function. In some cases when the objective function is not continuous, the value that sets the first-order condition to zero might not exist, in which case we choose the point that will set the first-order condition very close to zero. Note that in this section we will only consider the distribution results regarding the first numerical derivative.

First, we establish the basic result for consistency of the maximizer of the U-statistic which is based on the result of Lemma 1.

THEOREM 2. *Under Assumptions 2 and 3, and 4, provided that* $\sqrt{\varepsilon_n}n/\log n \to \infty$ *and* $\sqrt{n\varepsilon_n^{1+p}} = O(1)$, *for any sequence* $\hat{\theta}$ *such that* $||L_{1,p}^{\varepsilon_n}\hat{g}\left(\hat{\theta}\right)|| = o_p\left(\frac{1}{\sqrt{n}}\right)$*:*

$$d(\hat{\theta}, \theta_0) = o_p(1).$$

Having established consistency, we next investigate the convergence rate of the resulting estimator.

3.3 Asymptotic distribution

To estiablish the convergence rate for the maximizer of the U-statistic, we maintain Assumption 2 and 3 for the class of kernels of the U-statistic and the identification assumption 4. We note that we can improve upon the result of Lemma 1, which establishes that as long as $n^2\varepsilon_n/\log^2 n \to \infty$,

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

Moreover, we can apply Lemma 10 in Nolan and Pollard (1987) which states that for $t_n \geq \max\{\epsilon_n^{1/2}, \frac{\log n}{n}\}$ we have (for some constant $\beta > 0$)

$$P\left(\sup_{\mathcal{F}_n} |S_n(f)| > \beta^2 n^2 t_n^2\right) \leq 2A\,\exp\left(-nt_n\right)$$

However, we note that provided that log $n\sqrt{\varepsilon_n}/n \to \infty$, we can strengthen this result. In fact, provided that for sufficiently large n $t_n = \sqrt{\varepsilon_n}$, we can show that

$$\sup_{d(\theta,\theta_0)=o(1)} \frac{n^2\varepsilon_n}{\log^2 n} \| L^{\epsilon_n}_{1,p}\hat{g}\left(\theta\right) - L^{\epsilon_n}_{1,p}g\left(\theta\right)\| = O_p(1).$$

We next repeat the steps that we construct a "nesting" argument to find the rate of convergence for the estimator of interest. In this argument, in the first step we find the largest size of the shrinking neighborhood of the true parameter that contains the estimator. Then we establish the local stochastic equicontinuity result adapted to that shrinking neighborhood. The latter result will further be used to find the convergence rate for the maximizer of the U-statistic.

LEMMA 2. *Suppose* $\hat{\theta} \xrightarrow{p} \theta_0$ *and* $L^{\varepsilon}_{1,p}\hat{g}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)$. *Suppose that Assumptions 2 and 3 hold*

(i) If $n\sqrt{\varepsilon_n}/\log n \to \infty$, *and* $\sqrt{n\varepsilon^{1+p}} = O\left(1\right)$, *then* $\frac{n^2\varepsilon_n}{\log^2 n} d\left(\hat{\theta},\theta_0\right) = o_P\left(1\right)$.

(ii) If $\sqrt{n\varepsilon_n^{1+p}} = o\left(1\right)$, *and* $\frac{n\varepsilon_n}{\log n} \to \infty$ *we have*

$$\sup_{d\left(\hat{\theta},\theta_0\right)=O\left(\frac{\log^2 n}{n^2\varepsilon_n}\right)} \left(L^{\epsilon_n}_{1,p}\hat{g}\left(\hat{\theta}\right) - L^{\epsilon_n}_{1,p}\hat{g}\left(\theta_0\right) - L^{\epsilon_n}_{1,p}g\left(\hat{\theta}\right) + L^{\epsilon_n}_{1,p}g\left(\theta_0\right)\right) = o_p\left(\frac{1}{n}\right).$$

In this lemma we, first establish the "maximum" radius of the shrinking neighborhood containing the parameter. In the next step we considered the behavior of the objective function in the small neighborhood of order $O\left(\frac{\log^2 n}{n^2\varepsilon_n}\right)$ of the true parameter. As we show below, we can improve upon the rate of the objective function using the envelope property.

We can use this result to establish the rate of convergence of the resulting estimator.

THEOREM 3. *Suppose* $\hat{\theta} \xrightarrow{p} \theta_0$ *and* $L^{\varepsilon}_{1,p}\hat{g}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}}\right)$. *Under Assumptions 1, 2 and 3, if* $n\varepsilon_n/\log n \to \infty$, *and* $\sqrt{n\varepsilon^{1+p}} = O\left(1\right)$, *then* $\sqrt{n}d\left(\hat{\theta},\theta_0\right) = O_P\left(1\right)$.

Proof. We note that by Lemma 2 in small neighborhoods of the true parameter the U-statistic part is of stochastic order $o_p\left(\frac{1}{n}\right)$. As a result, the sum will be dominated by the projection term. Provided that the projection is Lipschitz-continuous, we can apply the standard rate result in

Newey and McFadden (1994) which gives the stochastic order for the first term $O_p\left(\frac{1}{\sqrt{n}}\right)$ and thereby ensures parametric convergence. $\square$

The last relevant result concerns the asymptotic distribution of the estimator.

ASSUMPTION 5. *Suppose that for any θ_1 and θ_2 in the neighborhood of θ_0 there exists a function $\dot{\mu}(\cdot)$ such that*

$$|\mu(z,\theta_1)-\mu(z,\theta_2)| \leq \dot{\mu}(z)\|\theta_1-\theta_2\|,$$

with $E\left[\dot{\mu}(Z)\dot{\mu}(Z)'\right]=\Omega<\infty$. Moreover, for this function

$$E\left[\left(\mu(\theta,Z)-\mu(\theta_0,Z)-(\theta-\theta_0)'\,\dot{\mu}(Z)\right)^2\right]=o\left(\|\theta-\theta_0\|^2\right).$$

This assumption allows us to obtain the following characterization of the asymptotic distribution of the estimator corresponding to the zero of the numerical gradient of the U-statistic.

THEOREM 4. *Suppose Assumptions 1, 2, 3 and 5 hold. In addition, assume that the Hessian matrix $H(\theta)$ of $g(\theta)$ is nonsingular. Then*

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

Proof. Consider the problem $L_{1,p}^{\varepsilon}\hat{g}\left(\hat{\theta}\right)=o_p\left(\frac{1}{\sqrt{n}}\right)$. Using the result in Lemma 2 we can replace the numerical derivative of the sample objective function with

$$L_{1,p}^{\epsilon_n}\hat{g}\left(\theta_0\right)+L_{1,p}^{\epsilon_n}g\left(\hat{\theta}\right)-L_{1,p}^{\epsilon_n}g\left(\theta_0\right)=o_p\left(\frac{1}{\sqrt{n}}\right).$$

Then by the property of the finite-difference formula $L_{1,p}^{\epsilon_n}g\left(\theta_0\right)=O(\varepsilon_n^{2p})$. It follows that

$$L_{1,p}^{\epsilon_n}g\left(\hat{\theta}\right)=D(\hat{\theta})+O(\varepsilon_n^{2p})=H(\theta_0)(\hat{\theta}-\theta_0)+O(\frac{1}{n}+\varepsilon_n^{2p})$$

by Theorem 3. Finally, by the U-statistic projection result,

$$\hat{g}(\theta_0)=\frac{1}{n}\sum_{i=1}^{n}\mu(z_i,\theta_0)+o_p\left(\frac{1}{\sqrt{n}}\right),$$

which we combine with Assumptions 5 and 3 to conclude that

$$L^{\epsilon_n}_{1,p}\frac{1}{n}\sum_{i=1}^{n}\mu(z_i,\theta_0)=\frac{1}{n}\sum_{i=1}^{n}\dot{\mu}(z_i)+o_p\left(\frac{1}{\sqrt{n}}+\varepsilon_n^{2p}\right).$$

Assembling terms, we obtain that

$$H(\theta_0)(\hat{\theta}-\theta_0)+\frac{1}{n}\sum_{i=1}^{n}\dot{\mu}(z_i)=o_p\left(\frac{1}{\sqrt{n}}\right),$$

provided that $\sqrt{n\varepsilon_n^{p+1}}=o(1)$. Then given that the data are i.i.d. and the function $\dot{\mu}(\cdot)$ has a finite second moment, we can apply the Lindeberg-Levy CLT and the result follows. □

4 Adaptive Choice of the Step Size

The choice of step size for computing numerical derivatives is an important practical question. The choice of the step size is akin to the choice of the smoothing parameter in nonparametric analysis. A survey of works on the choice of bandwidth for density estimation can be found in Jones, Marron, and Sheather (1996) with related results for non-parametric regression estimation and estimation of average derivatives in Hardle and Marron (1985) and Hart, Marron, and Tsybakov (1992) among others. The rest of this section focuses on consistent estimation of the derivatives.

Following the nonparametric estimation literature, we use the integrated mean-squared error as the criterion for the choice of the step size (although we can also consider other loss functions). Previously we considered the decomposition: $L^{\varepsilon_n}_{1,p}\hat{g}\left(\hat{\theta}\right)-G(\theta_0)=\hat{G}_1\left(\hat{\theta}\right)+G_2\left(\hat{\theta}\right)+G_3\left(\hat{\theta}\right)$. We now consider the problem of the optimal constant choice using the mean-squared error as the criterion for the choice of the step size

$$\mathrm{MSE}\left(\varepsilon\right)=E\|L^{\varepsilon_n}_{1,p}\hat{g}\left(\hat{\theta}\right)-G(\theta_0)\|^2,$$

which we approximate by the leading terms G_1 and G_2 since G_3 does not depend on the step size. Assuming that the function $g(\cdot)$ has at least $p+1$ derivatives,

$$L^{\epsilon_n}_{1,p}g(\theta)=\frac{1}{\varepsilon}\sum_{l=-p}^{p}c_l g\left(\theta+l\varepsilon\right)=g'(\theta)+\varepsilon_n^p g^{(p+1)}\left(\theta\right)\sum_{l=-p}^{p}\frac{c_l\, l^p}{(p+1)!}+o\left(\varepsilon_n^p\right).$$

$\hat{G}_1(\hat{\theta})$ can be approximated upto first order by $\hat{G}_1(\theta_0)$ and $G_2(\hat{\theta})$ can be approximated by $G_2(\theta_0)$. Thus $G_2(\theta) = \varepsilon_n^p g^{(p+1)}(\theta) \sum\limits_{l=-p}^{p} \frac{c_l\, l^p}{(p+1)!} + o(\varepsilon_n^p)$. We can evaluate the variance of $\hat{G}_1(\theta)$ as

$$\mathrm{Var}\,(G_1) = E_z \left[\frac{1}{\varepsilon_n} \sum_{l=-p}^{p} c_l \left(\mu\left(Z, \theta + l\varepsilon_n\right) - g\left(\theta + l\varepsilon_n\right)\right)\right]^2 = \frac{V_{\varepsilon_n}}{n^2 \varepsilon_n} = O(\varepsilon_n^{-1} n^{-2}),$$

where

$$V_{\varepsilon_n} = E_z \left[\sum_{l=-p}^{p} c_l \left(\mu\left(Z, \theta + l\varepsilon_n\right) - g\left(\theta + l\varepsilon_n\right)\right)\right]^2 \geq \min_{\theta} E_z \left[\left(\mu\left(Z, \theta\right) - g\left(\theta\right)\right)\right]^2 \sum_{l=-p}^{p} c_l^2 > 0.$$

Another component that needs to be considered is the operating precision of the computer. This error is known and fixed and we denote it $[\delta\, g]$. The overall bias will contain two components: the bias due to the approximation of the derivative by the finite-difference formula and the bias due to machine rounding:

$$\mathrm{bias}\,(\varepsilon) \approx \varepsilon_n^{2p} \left(g^{(p+1)}\left(\theta\right) \sum_{l=-p}^{p} \frac{c_l\, l^p}{(p+1)!} \right)^2 + O\left(\frac{[\delta\, g]}{\varepsilon_n} \sum_{l=-p}^{p} c_l |l|^p \right).$$

The square-root of mean-squared error is given by

$$\mathrm{MSE}^{1/2}\,(\varepsilon) \approx \varepsilon_n^{2p} \left(g^{(p+1)}\left(\theta\right) \sum_{l=-p}^{p} \frac{c_l\, l^p}{(p+1)!} \right)^2 + \varepsilon_n^{-1/2} n^{-1} V_{\varepsilon_n}^{1/2} + O\left(\frac{[\delta\, g]}{\varepsilon_n} \sum_{l=-p}^{p} c_l |l|^p \right).$$

Therefore, we can see that the mean-squared error will start increasing whenever $\varepsilon_n = O([\delta g]^{1/(2p+1)})$. Thus we can choose $\varepsilon_n = \max\{\frac{C}{n^r},\, [\delta g]^{1/(2p+1)}\}$, where r is the selected rate for ε_n. We next consider the choice of C. In most applications, however, the derivative $g^{(p+1)}$ is unknown. One simple way of choosing C is an analog of biased cross-validation. We choose a simple first-order formula for $g^{(p+1)}$ and pick a preliminary (over-smoothed) step size ε_n^{**} then evaluate

$$\widehat{g^{(p+1)}}\left(\theta\right) = \frac{1}{\varepsilon_n^{**}} \sum_{k=0}^{[p/2]} g\left(\theta + (-1)^k \varepsilon_n^{**} \right).$$

Plugging this expression into the expression for the mean-squared error, we can obtain the optimal step sizes. Provided that the order of the (mean-square) variance term is $n^{-1}\varepsilon_n^{-1/2}$, this term will always be dominated by the machine precision term. As a result, the choice of the step size is necessarily determined by the computer approximation error and the constant needs to be selected such that

$$C^{**} = \left(\frac{p!\;(p+1)!\;[\delta\, g] \sum\limits_{l=-p}^{p} c_l}{\left(\widehat{g^{(p+1)}}\,(\theta) \sum\limits_{l=-p}^{p} c_l\, l^p \right)^2} \right)^{1/(2p+1)}$$

which minimizes the mean-squared error. We note that this finding is in a striking contrast with that in Hong, Mahajan, and Nekipelov (2012) where it was found that the optimal choice of the step size of numerical differentiation for the objective functions of M-estimators may be determined by the statistical properties of those objective functions rather than the machine precision.

In case where one can compute the function in a relatively straightforward way, the calibration of the constants of interest can be performed by minimizing the approximate expression for the mean-squared error with respect to C.This approach is equivalent to the plug-in approach in the bandwidth selection literature.

5 Application and Monte-Carlo Evidence

We illustrate our results with a semiparametric panel data model with random censoring considered in Khan and Tamer (2007) and consider a simplified version of that setup applied to cross-sectional settings. Khan and Tamer (2007) consider estimation of the linear index coefficients in a fixed effects model with (potentially endogenous) random censoring. They provide restrictions on the support of the censoring variables that allow identification of the parameters of interest in case of arbitrary correlation between the censoring points and the regressors. Censoring with this structure is commonplace in the proportional hazard models as well as in the competing risks and survival analysis literatures.

Khan and Tamer (2007) introduce a distribution-free estimator and specify an objective function involving a second order U-statistic. Since economic theory rarely provides specific functional form relationships, such distribution free estimators are attractive from a purely economic perspective as well and are also tractable since they do not suffer from the curse of dimensionality that is common to fully non-parametric estimation procedures.

We consider a simplified version of the estimator used in Khan and Tamer (2007). In our setup the latent variable Y^* is generated by the index equation

$$Y^* = X_1 + \theta\, X_2 + \varepsilon,$$

where X_1, X_2 and ε follow a standard normal distribution. We consider the case of independent censoring where the censoring point C has a standard normal distribution. The observed variables can be characterized by the pair

$$\begin{aligned} Y &= Y^* \,(1 - D) + C\, D, \\ D &= \mathbf{1}\{Y^* > C\}. \end{aligned}$$

We then consider the problem of estimating a single parameter θ. We denote $z = (y, x_1, x_2)'$ and define the objective function with the maximum rank correlation structure with kernel

$$f(z_i, z_j, \theta) = \mathbf{1}\{y_i > y_j\}\mathbf{1}\{x_{1i} + \theta\, x_{2i} > x_{1j} + \theta\, x_{2j}\}.$$

The objective function can be then represented by $\hat{g}(\theta) = \frac{1}{n(n-1)} S_n(f)$. Note that the projection of the U-statistic kernel is smooth due to the smoothness of the joint distribution of Z.

We use a numerical gradient approach to approximate the extremum for this objective function. To construct the numerical gradient, we use finite difference formulas of different orders. The step size ε_n depends on the sample size. In particular, the first-order right derivative formula is

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

and analogously for the left derivative formula. The second-order formula is

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

and the third-order formula is

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

The estimator is then re-defined as a solution to the numerical first-order condition

$$\widehat{D}_k\left(\widehat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}}\right), \tag{5.6}$$

which mimics the solution obtaind using a numerical gradient-based maximization routine. We can anticipate the properties of the analyzed estimator by analyzing its behavior analytically. For illustration we can use the numerical derivative formula $\widehat{D}_2\left(\widehat{\theta}\right)$. Application of this formula to the sample objective function leads to the expression

$$\widehat{D}_2\left(\widehat{\theta}\right) = \frac{1}{n\varepsilon_n}\sum_{i=1}^{n}\mathbf{1}\{y_i \geq y_j\}U\left(\frac{1}{\varepsilon_n}\left(\theta+\frac{x_{1i}-x_{1j}}{x_{2i}-x_{2j}}\right)\right).$$

It is clear that in small samples where the step size of numerical differentiation is "small" the sample first-order condition can have multiple roots. Given the structure of the objective functions the roots will either be contained in disjoint convex compact sets or will be singletons. To facilitate root finding, we use a dense grid over the state space of the model. For the step size ε_n we choose the size of the grid cell to be $O\left(\varepsilon_n/\log n\right)$. This will guarantee that the error (measured as the Hausdorff distance between the true set of roots and the set of roots on the grid) will vanish at a faster rate than the numerical error from approximating the gradient using a finite-difference formula. For simplicity we use a uniform grid on $[-2, 2]$ such that the cell size is $\Delta_n = C\,\frac{\varepsilon_n}{\log n}$, the number of grid points is $N_{\Delta_n} = \left[\frac{2\log n}{C\,\varepsilon_n}\right]+1$ and the grid points can be obtained as $\theta_g = -1+\Delta\left(g-1\right)$ and so forming the set $G_{\Delta_n} = \{\theta_g\}_{g=1}^{N_{\Delta_n}}$. The grid search algorithm will identify the set of points

$$Z_n = \left\{\theta \in G_{\Delta_n} \,:\, \left|\widehat{D}_k\left(\theta\right)\right| \leq \sqrt{\frac{\log n}{\varepsilon_n n}}\right\}.$$

We call this set the set of roots of the numerical first-order condition on a selected grid. Our Monte-Carlo study will analyze the structure of the set of roots on the grid to evaluate the performance of the numerical gradient-based estimator. The Monte-Carlo study proceeds in the following steps.

1. We generate 1000 Monte-Carlo samples with the number of observations from 10 to 1000. Each simulation sample is indexed by s and the sample size is denoted n_s.

2. We choose sample-adaptive step of numerical differentiation as $\varepsilon = C\,(n_s)^q$. We choose $C = 2$ and q from 0.2 to 2.

3. Using this step size, we set up the function that we associate with the empirical first-order condition with $\widehat{D}_k\left(\widehat{\theta}^s\right)$ for different orders of numerical derivatives.

4. Using the grid over the support $[-1,\,1]$ (which we described above) we find all solutions satisfying (5.6). This will form the set of roots on the grid Z_{n_s}.

5. We store all roots on the grid and report the statistics averaged across the roots.

6. If $\#Z_{n_s}$ is the number of roots found in simulation s, we evaluate the mean-squared errors of estimation as:

$$\text{MSE}\left(\widehat{\theta}\right) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\frac{1}{\#Z_{n_s}}\sum_{r=1}^{\#Z_{n_s}}\left(\widehat{\theta}_{rs}-\theta_0\right)^2}$$

Our simulation results are represented in Tables 1 and 2. The tables show the trade-offs between bias and variance for different choices of the rates at which the step size approaches zero. The constants were chosen on using cross-validation. The tables demonstrate that the bias of the estimates is lower if one uses a higher-order formula for the numerical derivative. The variance, on the other hand, remains stable across different formulas. We can see a slight increase in the variance towards the higher-order derivative formulas if the step size sequence approaches zero slowly and the sample is small.

This is an indication that in the cases of objective functions defined by the U-statistics, there is no dramatic manifestation of the bias-variance tradeoff in cases of "standard" choices of the step sizes. Moreover, our analysis demonstrates that the distribution of estimates does not change even when one chooses the step size to approach to zero at the same rate as the sample size. This is evidence that the previously existing guidelines for choosing the step size for numerical differentiation to be of order $n^{-1/4}$ are excessive. One can choose the step size to be much smaller at little to no cost in terms of the impact on the resulting asymptotic distribution.

[Table 1 about here.]

[Table 2 about here.]

6 Conclusion

In this paper we analyze the use of numerical finite-difference approximations for computing derivatives and solving the first-order conditions corresponding to objective functions defined by second-order U-statistics. Using finite-difference approximations is computationally attractive because both smoothing the U-statistic kernel and evaluating the objective function directly can be computationally infeasible particularly in large data settings.

We establish sufficient conditions on the step size of the finite difference formulas that guarantee uniform consistency of the resulting estimators of the derivatives of the objective function. We find that the lower bound on the rate at which the step size sequence converges to zero is of order $\log n^2/n^2$ which substantially improves a widely used practical guideline of order $n^{-1/4}$. From this it follows that numerical derivatives can be precisely evaluated even if the step size approaches zero at the same rate or faster than (the inverse of) the sample size.

We also find that such a step size sequence yields consistent estimators of the extremum estimators defined by minimizing U-statistic based objective functions. We consider an estimation procedure that replaces maximum search with the solution to a finite-difference approximation to the first order conditions. This opens the door to a much simpler way of computing extremum estimators defined by U-statistics: instead of smoothing the objective function and employing a gradient-based procedure with a very small step size, one can use a numerical gradient-based procedure with a data-dependent step size. As a result, U-statistics based estimators can be applied to large samples, something that was impractical using previous approaches.

References

Anderssen, R., and P. Bloomfield (1974): "Numerical differentiation procedures for non-exact data," *Numerische Mathematik*, 22, 157–182.

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

Han, A. (1987): "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics*, 35(2), 303–316.

Hardle, W., and J. Marron (1985): "Optimal bandwidth selection in nonparametric regression function estimation," *The Annals of Statistics*, pp. 1465–1481.

Hart, J., J. Marron, and A. Tsybakov (1992): "Bandwidth choice for average derivative estimation," *Journal of the American Statistical Association*, 87, 218–226.

Hong, H., A. Mahajan, and D. Nekipelov (2012): "Extremum Estimation and Numerical Derivatives," *Stanford and UC Berkeley Working Paper*.

Jones, M., J. Marron, and S. Sheather (1996): "A brief survey of bandwidth selection for density estimation.," *Journal of the American Statistical Association*, 91.

Judd, K. (1998): *Numerical Methods in Economics*. MIT Press.

Khan, S., and E. Tamer (2007): "Partial rank estimation of duration models with general forms of censoring," *Journal of Econometrics*, 136(1), 251–280.

L'Ecuyer, P., and G. Perron (1994): "On the Convergence Rates of IPA and FDC Derivative Estimators," *Operations Research*, 42, 643–656.

Murphy, S., and A. Van der Vaart (2000): "On Profile Likelihood.," *Journal of the American Statistical Association*, 95.

NEWEY, W., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2113–2241. North Holland.

NOLAN, D., AND D. POLLARD (1987): "U-processes:rates of convergence," *The Annals of Statistics*, pp. 780–799.

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027–1057.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer Verlag.

PRESS, W., S. A. TEUKOLSKY, W. VETTERING, AND B. FLANNERY (1992): *Numerical Recipes in C, The Art of Scientific Computing*. Cambridge.

SERFLING, R. (1980): *Approximation Theorems in Mathematical Statistics*. John Wiley and Sons.

SHERMAN, R. P. (1993): "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, 61, 123–137.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak convergence and empirical processes*. Springer-Verlag, New York.

A Appendix

A.1 Proof of Lemma 1

The result of the theorem can be obtained by using the argument in the proof of Theorem 9 of Nolan and Pollard (1987). We define the class of functions $\mathcal{F}_n = \{\epsilon_n L_{1,p}^{\epsilon_n} g\left(\cdot,\cdot,\theta\right), \theta \in N\left(\theta_0\right)$, with envelope function F, such that $PF \leq C$. Then we can write

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

Noting (2.2), lemma 1 can be shown separately for the $\hat{\mu}_n(\theta)$ and $S_n(u)/n(n-1)$ components of the decomposition. Therefore the result of lemma 1 holds for the $\hat{\mu}_n(\cdot)$ component as long as ϵ_n satisfies the rate assumptions $n\sqrt{\varepsilon_n}/\log\, n \to \infty$. Given this, without loss of generality we focus on $S_n(u)$ component and assume that the U-statistic's kernel $f(\cdot,\cdot,\theta)$ is degenerate.

Due to Assumption 2, for each $f \in \mathcal{F}_n$, $E|f|^2 = E\left|\epsilon_n L_{1,p}^{\epsilon_n} g(\cdot,\theta)\right|^2 = O(\epsilon_n)$. Define $t_n \geq \max\{\epsilon_n^{1/2}, \frac{\log\, n}{n}\}$ as in Lemma 10 of Nolan and Pollard (1987). Under the condition $n\sqrt{\epsilon_n}/\log n \to \infty$ in lemma 1, for large enough n, $t_n = \epsilon_n$. Denote $\delta_n = \mu\, t_n^2\, n^2$. By the Markov inequality,

$$P\left(\sup_{f\in\mathcal{F}_n} |S_n(f)| > \delta_n\right) \leq \delta_n^{-1} P \sup_{f\in\mathcal{F}_n} |S_n(f)|.$$

By assumption 2, the covering integral of $\mathcal{F}_n$ is bounded by a constant multiple of $H(s) = s\,[1 + \log(1/s)]$. The maximum inequality in Theorem 6 of Nolan and Pollard (1987) implies that

$$P \sup_{f\in\mathcal{F}_n} |S_n(f)|/n \leq C\, P\, H\left[\sup_{f\in\mathcal{F}_n} |T_n\, f^2|^{1/2}/n\right].$$

where T_n is the symmetrized U-statistic constructed analogously to the symmetrized empirical process from the Radamacher sequences (see Nolan and Pollard (1987)). The right-hand side can be further bounded by Lemma 10 in Nolan and Pollard (1987). This lemma states that there exists a constant β such that

$$P\left(\sup_{f\in\mathcal{F}_n} |S_{2n}(f)| > \beta^2\, 4n^2\, t_n^2\right) \leq 2\, A\, \exp(-2n\, t_n),$$

where A is the Euclidean constant in assumption 2. Since $f(\cdot)$ is globally bounded, $|f(\cdot)|^2 \leq B|f(\cdot)|$ for a constant B. In addition, note that $|S_{2n}(f)| \geq |T_n f|$. Therefore, we find that $|T_n f^2| \leq B|S_{2n}(f)|$, which implies

$$P\left(\sup_{f\in\mathcal{F}_n} |T_n\, f^2| > \frac{4\beta^2}{B}\, n^2\, t_n^2\right) \leq 2\, A\, \exp(-2n\, t_n).$$

Also note that $H[\cdot]$ achieves its maximum at 1 and is increasing for its argument less than 1. For sufficiently large n the term $\frac{4\beta^2}{B}\, t_n^2 \ll 1$. Then

$$\begin{aligned}
P\, H\left[\sup_{f\in\mathcal{F}_n} |T_n\, f^2|^{1/2}/n\right] &= P\Bigg(H\left[\frac{1}{n}\sup_{f\in\mathcal{F}_n} |T_n\, f^2|^{1/2}\right]\mathbf{1}\left\{\sup_{f\in\mathcal{F}_n}|T_n\, f^2| > \frac{4\beta^2}{B}\, n^2\, t_n^2\right\} \\
&+ H\left[\frac{1}{n}\sup_{f\in\mathcal{F}_n} |T_n\, f^2|^{1/2}\right]\mathbf{1}\left\{\sup_{f\in\mathcal{F}_n}|T_n\, f^2| < \frac{4\beta^2}{B}\, n^2\, t_n^2\right\}\Bigg) \\
&\leq 1\cdot P\left(\sup_{f\in\mathcal{F}_n}|T_n\, f^2| > \frac{4\beta^2}{B}\, n^2\, t_n^2\right) + H\left[\frac{2\beta}{\sqrt{B}}\, t_n\right]\cdot 1 \\
&\leq 2\, A\, \exp\left(-2n\, t_n\right) + H\left(\frac{2\beta}{\sqrt{B}}\, t_n\right).
\end{aligned}$$

Substituting this result into the maximum inequality one can obtain

$$\begin{aligned}
P\left(\sup_{f\in\mathcal{F}_n}|S_n(f)| > \delta_n\right) \leq & n\delta_n^{-1}\left(H\left(\frac{2\beta}{\sqrt{B}}\, t_n\right) + 2\, A\, \exp\left(-2n\, t_n\right)\right) \\
= & (t_n\, n)^{-1} + (nt_n)^{-2}\exp\left(-2n\, t_n\right) - (t_n\, n)^{-1}\log\, t_n.
\end{aligned}$$

By assumption $t_n n >> \log n \to \infty$, the first term vanishes. The second term also vanishes by showing that $n^{-1}t_n^{-2}\exp\left(-2n\, t_n\right) \to 0$, because it is bounded by, for some $C_n \to \infty$, $1/\left(\log n n^{C_n} t_n\right)$. Finally, considering the term $t_n^{-1}n^{-1}\log\, t_n$, we note that it can be decomposed into $t_n^{-1}n^{-1}\log\left(nt_n\right) - t_n^{-1}n^{-1}\log n$. Both terms converge to zero because both $t_n n \to \infty$ and $\frac{t_n n}{\log\, n} \to \infty$. We have thus shown that for any $\mu > 0$

$$P\left(\sup_{f\in\mathcal{F}_n}|\frac{1}{n\,(n-1)}S_n(f)| > \mu\, \varepsilon_n\right) = o(1).$$

This proves the statement of the theorem. □

A.2 Proof of Lemma 2

Proof. **(i)**

We note that for the projection part

$$\sup_{d(\theta,\theta_0)=o(1)} \frac{1}{\sqrt{n}}\|L_{1,p}^{\epsilon_n}\hat{\mu}\left(\theta\right) - L_{1,p}^{\epsilon_n}\mu\left(\theta\right)\| = o_p(1).$$

As a result the U-process part will dominate and the convergence rate will be determined by its order $\frac{\log^2\, n}{n^2\varepsilon_n}$.

(ii)

Consider a class of functions

$$\mathcal{G}_n = \left\{ g\left(\cdot, \theta_n + \varepsilon_n\right) - g\left(\cdot, \theta_n - \varepsilon_n\right) - g\left(\cdot, \theta_0 + \varepsilon_n\right) + g\left(\cdot, \theta_0 - \varepsilon_n\right), \, \theta_n = \theta_0 + t_n \frac{\log^2 n}{n^2 \varepsilon_n} \right\},$$

with $\varepsilon_n \to 0$ and $t_n = O(1)$. We can evaluate the L^2 norm of the functions from class $\mathcal{G}_n$ using Assumption 2 (ii). Note that

$$E\left[\left(g\left(Z_i, z, \theta_n + \varepsilon_n\right) - g\left(Z_i, z, \theta_n - \varepsilon_n\right)\right)^2\right] = O\left(\varepsilon_n\right),$$

with the same evaluation for the second term. On the other hand, we can change the notation to $\theta_{1n} = \theta_0 + \varepsilon_n + \frac{t_n}{2} \frac{\log^2 n}{n^2 \varepsilon_n}$ and $\theta_{1n} = \theta_0 + \frac{\varepsilon_n}{2} + t_n \frac{\log^2 n}{n^2 \varepsilon_n}$. The we can group the first term with the third and the second one with the fourth. For the first group this leads to

$$E\left[\left(g\left(Z_i, z, \theta_{1n} + \frac{t_n}{2} \frac{\log^2 n}{n^2 \varepsilon_n}\right) - g\left(Z_i, z, \theta_{1n} - \frac{t_n}{2} \frac{\log^2 n}{n^2 \varepsilon_n}\right)\right)^2\right] = O\left(\frac{\log^2 n}{n^2 \varepsilon_n}\right),$$

and for the second group

$$E\left[\left(g\left(Z_i, z, \theta_{2n} + \frac{\varepsilon_n}{2}\right) - g\left(Z_i, z, \theta_{2n} - \frac{\varepsilon_n}{2}\right)\right)^2\right] = O\left(\varepsilon_n\right).$$

Thus, two different ways of grouping the terms allow us to obtain two possible bounds on the norm of the entire term. As a result, we find that

$$P f^2 = O\left(\min\left\{\varepsilon_n, \, \frac{\log^2 n}{n^2 \varepsilon_n}\right\}\right), \quad f \in \mathcal{G}_n.$$

Next we denote $\delta_n = \min\left\{\varepsilon_n, \, \frac{\log^2 n}{n^2 \varepsilon_n}\right\}$.

Due to assumptions of the theorem, for each $f \in \mathcal{F}_n$, $E|f|^2 = E\left|\epsilon_n L_{1,p}^{\epsilon_n} g\left(\cdot, \theta\right)\right|^2 = O\left(\epsilon_n\right)$. Define $t_n \geq \max\{\delta_n^{1/2}, \, \frac{\log n}{n}\}$ as in Lemma 10 of Nolan and Pollard (1987) then for $n\sqrt{\delta_n}/\log n \to \infty$

$$\sup_{\mathcal{F}_n} \| \frac{1}{n(n-1)} T_n(f^2) \| = o_p\left(\delta_n^2\right),$$

where T_n is the symmetrized measured defined in Nolan and Pollard (1987). By Assumption 2 (iii), the covering integral of $\mathcal{F}_n$ is bounded by a constant multiple of $H(s) = s\left[1 + \log\left(1/s\right)\right]$. The maximum inequality in Theorem 6 of Nolan and Pollard (1987) implies that

$$P \sup_{f \in \mathcal{F}_n} |S_n(f)|/n \leq C \, P \, H\left[\sup_{f \in \mathcal{F}_n} |T_n \, f^2|^{1/2}/n\right].$$

Then the stochastic order of $\frac{1}{n\varepsilon_n} \sup_{f \in \mathcal{F}_n} |S_n(f)|$ can be evaluated as

$$\frac{\sqrt{n}}{\varepsilon_n} \frac{1}{n\varepsilon_n} \sup_{f \in \mathcal{F}_n} |S_n(f)| = O_p\left(\frac{\delta_n}{\varepsilon_n} \log\, \delta_n\right) = O_p\left(\frac{\log\left(\frac{n^2\varepsilon_n}{\log\, n}\right)}{\frac{n^2\varepsilon_n^2}{\log\, n}}\right) = o_p(1).$$

This delivers the result in the Lemma.

□

Table 1: Variance, MSE and Bias of Estimated Parameters

	Sample Size				
	10	100	200	500	1000
$\varepsilon_n \sim n^{-1}$					
Variance					
Derivative (1)	0.317	0.229	0.195	0.160	0.133
Derivative (2)	0.317	0.227	0.192	0.158	0.129
Derivative (3)	0.302	0.202	0.172	0.137	0.092
Derivative (4)	0.288	0.211	0.182	0.145	0.111
MSE					
Derivative (1)	0.327	0.320	0.301	0.260	0.214
Derivative (2)	0.327	0.320	0.299	0.260	0.200
Derivative (3)	0.328	0.303	0.276	0.217	0.128
Derivative (4)	0.329	0.318	0.284	0.243	0.160
Abs(Bias)					
Derivative (1)	0.102	0.302	0.327	0.317	0.285
Derivative (2)	0.096	0.305	0.327	0.320	0.266
Derivative (3)	0.160	0.318	0.323	0.283	0.190
Derivative (4)	0.201	0.326	0.320	0.313	0.220
$\varepsilon_n \sim n^{-1/2}$					
Variance					
Derivative (1)	0.282	0.112	0.062	0.022	0.011
Derivative (2)	0.307	0.152	0.065	0.022	0.010
Derivative (3)	0.289	0.106	0.055	0.021	0.010
Derivative (4)	0.288	0.115	0.057	0.021	0.011
MSE					
Derivative (1)	0.372	0.168	0.106	0.044	0.023
Derivative (2)	0.351	0.154	0.067	0.026	0.013
Derivative (3)	0.369	0.112	0.062	0.024	0.012
Derivative (4)	0.357	0.119	0.060	0.023	0.011
Abs(Bias)					
Derivative (1)	0.301	0.238	0.211	0.147	0.113
Derivative (2)	0.210	0.040	0.052	0.068	0.056
Derivative (3)	0.283	0.080	0.087	0.057	0.048
Derivative (4)	0.264	0.059	0.054	0.035	0.026

Table 2: Variance, MSE and Bias of Estimated Parameters

	Sample Size				
	10	100	200	500	1000
	$\varepsilon_n \sim n^{-1/4}$				
Variance					
Derivative (1)	0.284	0.160	0.105	0.032	0.012
Derivative (2)	0.303	0.181	0.113	0.035	0.012
Derivative (3)	0.301	0.149	0.081	0.026	0.010
Derivative (4)	0.304	0.161	0.094	0.029	0.011
MSE					
Derivative (1)	0.340	0.245	0.139	0.039	0.014
Derivative (2)	0.334	0.232	0.128	0.037	0.012
Derivative (3)	0.362	0.184	0.093	0.029	0.011
Derivative (4)	0.355	0.207	0.111	0.031	0.012
Abs(Bias)					
Derivative (1)	0.236	0.292	0.184	0.084	0.046
Derivative (2)	0.176	0.224	0.124	0.037	0.010
Derivative (3)	0.247	0.188	0.108	0.048	0.022
Derivative (4)	0.225	0.215	0.127	0.047	0.025