Extremum Estimation and Numerical Derivatives

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

[a,b] Department of Economics, Stanford University, Stanford, CA 94305.

[c] Department of Economics, University of California at Berkeley, Berkeley, California, 94720

This Version: September 2010

Abstract

Many empirical researchers rely on the use of finite-difference approximation to evaluate derivatives of estimated functions. For instance many optimization routines implicitly use finite-difference formulas for the gradients. Such routines frequently require the choice of step size parameters for finite-difference numerical gradients or similar parameters such as the computing tolerance. This paper investigates the statistical properties numerically evaluated gradients and the properties of extremum estimators computed using numerical gradients. We find that first, for unbiased inference one needs to adjust the step size or the tolerance as a function of the sample size. Second, higher-order finite difference formulas reduce the asymptotic bias analogous to higher order kernels in kernel smoothing. Third, we provide weak sufficient conditions for uniform consistency of the finite-difference approximations for gradients and directional derivatives. Fourth, we analyze the numerical gradient-based extremum estimators and find that the asymptotic distribution of the resulting estimators can be a hybrid between the asymptotics of the original extremum estimators and the asymptotics of the kernel smoothers. Fifth, we state conditions under which the numerical derivative estimator is consistent and asymptotically normal. Sixth, we generalize our results to semiparametric estimation problems. Finally, we show that the theory is also useful in a range of nonstandard estimation procedures.

JEL Classification: C14; C52

Keywords: Numerical derivative, entropy condition, stochastic equicontinuity

---

1 Introduction

The implementation of extremum estimators often involves the use of computational maximization routines. When the anaytical gradient of the objective function is not available these routines use finite-difference approximations to the gradient which involve the use of step size parameters. The statistical noise in this approximation algorithm of the optimization routine is typically ignored in empirical work. In this paper we 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

*Corresponding Author. Email: nekipelov@econ.berkeley.edu.

important implications for the practical use of numerical optimization routines. First, the choice of numerical tolerance and the step size be dependent on the sample size. Second, the asymptotic distribution and, consequently, the shape of the confidence region depends on the particular sequence of step sizes. We focus on numerical gradient-based optimization routines that use finite-difference formulas to approximate the gradient of the objective function. We consider a general framework where the objective function is computed from an i.i.d. sample and can depend on finite or infinite-dimensional unknown parameters.

The importance of numerical derivatives has received some attention in the literature. Newey and McFadden (1994) provided sufficient conditions for using numerical derivatives to consistently estimate the asymptotic variance in a parametric model. The properties of numerical derivatives have predominantly been investigated for very smooth models. For instance, Anderssen and Bloomfield (1974) analyzed derivative computations for functions that are approximated using polynomial interpolation. L'Ecuyer and Perron (1994) considered asymptotic properties of numerical derivatives for the class of general smooth regression models. In a related work Andrews (1997) the author has considered the relationship of numerical tolerance for the computation of GMM-type objective functions and their sample variance. Murphy and Der Vaart (2000) derived some properties of numerical derivatives of Hadamard-differentiable functionals computed in an i.i.d. context. However, to the best of our knowledge there have been no studies of the impact of the numerical optimization routines on the statistical properties of extremum estimators.

Our results include fairly weak sufficient conditions for consistency, rates of convergence and the asymptotic distribution for several classes of numerically computed extremum estimators. We separately study M-estimators, generalized method of moment (GMM) estimators, and estimators that maximize a function involving second-order U-statistics. Numerical M-estimation is considered both for finite-dimensional and infinite-dimensional unknown parameters. We find that the choice of the step size for consistency and convergence to the asymptotic distribution depends on the interplay between the smoothness of the population objective function, the order of chosen approximation, and on the properties of the sample objective function. Specifically, we find that if the sample objective function is at least Lipschitz-continuous, then the step size for numerical differentiation can be chosen to approach zero at an arbitrarily fast rate to provide a consistent estimator that will converge to a normal random variable at a parametric rate. In cases where the objective function is only Hölder-continuous, or where it is discontinuous, the step size has to adapt to the sample size resulting in a slower,nonparametric, rate of convergence for the estimator. For a discontinuous objective function, the asymptotic distribution will be close to normal if the sequence of step sizes converges to zero sufficiently slowly. If the step size converges to zero at an appropriately faster rate, then the asymptotic distribution becomes non-standard. We characterize this distribution and find that it is related to the boundary-crossing time distribution for a non-standard Brownian motion.

We illustrate our findings with several empirical examples. First, we study the system of functional moments that arises models of dynamic discrete games of incomplete information. We find that this model is characterized by a relatively smooth system of moment vectors and numerical differentiation should not affect its asymptotic properties. Second, we analyze extremum estimators where the (non-convex) objective functions behaves like $\sqrt[\alpha]{|Z_i - \theta|}$. We find that numerical gradient-based optimization effectively regularizes the problem and convexifies the objective functions. Finally, we apply numerical gradient-based optimization to the maximum score (Manski (1975)), and find that for an appropriate step size sequence, the behavior of the resulting estimator is similar to the smoothed maximum score estimator of Horowitz (1992).

2 Estimation of derivatives from from non-smooth sample functions

2.1 Derivatives of semiparametric moment functions

We start with presenting a framework for using numerical derivative in obtaining consistent estimators of derivatives of functions whose values are estimated from the data. The derivatives may be used to compute the asymmptotic variance of estimators or to provide the GMM estimators whose minimum is computed via a numerical garident-based routine. While the conditions given in this subsection can be appealing because they accord with the intuition provided in the existing literature, in the next subsection we will show that much weaker conditions are often sufficient for achieving consistency.

Consider a general conditional moment model of the form of $m\left(z,\theta,\eta\left(\cdot\right)\right) = E\left[\rho\left(y,\theta,\eta\left(\cdot\right)\right)|z\right] = 0$, which includes finite dimensional parameters $\theta \in \Theta \subset \mathbb{R}^d$ and infinite dimensional parameters $\eta\left(\cdot\right) \in \mathcal{H}$. This setup includes the unconditional moment as a special case when z is a constant. Semiparametric estimators for this general model and their asymptotic distributions are studied extensively in the literature. In some models, the moment conditions $\rho\left(y,\theta,\eta\left(\cdot\right)\right)$ depends only on the value of the function $\eta\left(\cdot\right)$ at the sample realization of y. In some other models, such as in dynamic discrete choice models and dynamic games, $\rho\left(y,\theta,\eta\left(\cdot\right)\right)$ may depend on the entire function of $\eta\left(\cdot\right)$ in complex ways.

The sieve approach, studied in a sequence of papers by Newey and Powell (2003), Chen and Shen (1998), Ai and Chen (2003) and Chen and Pouzo (2009), approximates class of infinite dimensional functions $\mathcal{H}$ using a parametric family of function $\mathcal{H}_n$ whose dimension increases to infinity as the sample size n increases. If $\eta\left(\cdot\right)$ admits an infinite series expansion $\eta(y) = \sum\limits_{k=1}^{\infty} \gamma_k q^{(k)}(y)$ using basis functions $q^k\left(\cdot\right), k = 0,\ldots,\infty$, a class of approximating functions H_n can take the form of $\eta_n\left(y\right) = \sum\limits_{k=0}^{\kappa_n} \gamma_k q^{(k)}(y)$, where κ_n is the number of sieve terms. Ai and Chen (2003) proposes a semiparametric GMM estimator that essentially takes the form of

$$\left(\hat{\theta},\hat{\eta}\right) = \arg\min_{\theta\in\Theta,\eta\in H_n} \sum_{i=1}^{n} \hat{m}\left(z,\theta,\eta\right)\hat{\Sigma}\left(z\right)^{-1}\hat{m}\left(z,\theta,\eta\right),$$

where $\hat{m}\left(z,\theta,\eta\right)$ is a consistent nonparametric estimator of the conditional expection function $m\left(z,\theta,\eta\right)$ that is usually also based on a series projection of $\rho\left(y,\theta,\eta\left(\cdot\right)\right)$ on an increasing sequence of functions of z. $\hat{\Sigma}(z)$ is a consistent estimate of the conditional covariance matrix $\Sigma(z) = E\left[\rho(y;\,\theta_0,\,\eta_0(\cdot))\,\rho(y;\,\theta_0,\,\eta_0(\cdot))' \mid z\right]$. They demonstrate conditions under which $\left(\hat{\theta},\hat{\eta}\right) \xrightarrow{p} (\theta_0,\eta_0)$, and show that the asymptotic variance for $\hat{\theta}$ depends on the functional derivative of the moment conditions $m(z,\theta,\eta)$ with respect to the finite and infinite dimensional parameters $\alpha = (\theta,\eta)$.

For any $w \in \mathcal{H}$, denote by $\frac{dm(Z,\alpha)}{d\eta}[w]$ the functional derivative of $m(Z,\alpha)$ with respect to the η component in the w direction. Let $w^* \equiv (w_1^*,\ldots,w_d^*)$, where w_j^* is the solution to

$$\min_{w_j \in \mathcal{H}-h_0} E\left\{\left(\frac{dm(Z,\alpha)}{d\theta_j} - \frac{dm(Z,\alpha)}{d\eta}[w_j]\right)' \Sigma(Z)^{-1}\left(\frac{dm(Z,\alpha)}{d\theta_j} - \frac{dm(Z,\alpha)}{d\eta}[w_j]\right)'\right\}.$$

Furthermore define

$$D_{w^*}(Z) \equiv \frac{dm(Z,\alpha_0)}{d\theta'} - \frac{dm(Z,\alpha_0)}{d\eta}[w^*], \text{ where } \frac{dm(Z,\alpha_0)}{d\eta}(w^*) = \left[\frac{dm(Z,\alpha_0)}{d\eta}(w_j^*), j = 1,\ldots,d\right].$$

Then the asymptotic variance for $\sqrt{n}\left(\hat{\theta}-\theta\right)$ is given by V^{-1}, for $V = E\{D_{w^*}(Z)'\,\Sigma(Z)^{-1}\,D_{w^*}(Z)\}$.

A consistent estimate of the asymptotic variance can be formed by replacing each w_j^* with $\hat{w}_j^*$, which minimizes the following sample analog with respect to $w_j \in \mathcal{H}_n - \hat{\eta}$:

$$\sum_{i=1}^{n}\left\{\left(\frac{d\hat{m}(Z,\hat{\alpha})}{d\theta_j} - \frac{d\hat{m}(Z,\hat{\alpha})}{d\eta}[w_j]\right)' \hat{\Sigma}(Z)^{-1}\left(\frac{d\hat{m}(Z,\hat{\alpha})}{d\theta_j} - \frac{d\hat{m}(Z,\hat{\alpha})}{d\eta}[w_j]\right)'\right\}.$$

Then $\hat{V} \xrightarrow{p} V$, where $\hat{V} = \frac{1}{n}\sum_{i=1}^{n} \hat{D}_{\hat{w}^*}(Z_i)'\,\hat{\Sigma}(Z_i)^{-1}\,\hat{D}_{\hat{w}^*}(Z_i)$.

Ackerberg, Chen, and Hahn (2009) further shows that this method of estimating the asymptotic variance for $\hat{\theta}$ essentially amounts to 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 $\hat{\alpha} = \left(\hat{\theta},\hat{\eta}\right)$. However, in many practical estimation problems, the derivatives of $\frac{d\hat{m}(Z,\hat{\alpha})}{d\theta_j}$ and $\frac{d\hat{m}(Z,\hat{\alpha})}{d\eta}[w_j]$ do not have analytic solutions and have to be evaluated numerically. The goal of this paper is to analyze the impact of numerical approximation on statistical properties of the estimator for $D_w(z)$ and the parameter of interest.

2.2 Numerical differentiation using finite differences

Finite difference methods (e.g. Judd (1998)) are often used for 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 jth derivative of $g(x)$ can be estimated by a linear operator, denoted by $L_{k,p}^{\epsilon}g(\theta)$, that makes use of a pth order two-sided formula:

$$L_{k,p}^{\epsilon}g(x) = \frac{1}{\epsilon^j}\sum_{l=-p}^{p} c_l\, g(x+l\epsilon).$$

The usual two sided derivative refers to the case when $p = 1$. When $p \geq 1$, these are called higher order differentiation. For a given p, when the weights $c_l, l = 1, \ldots, p$ are chosen appropriately, the error in approximating $g^{(k)}(x)$ with $L^{\epsilon}_{j,p} g(x)$ will be small:

$$L^{\epsilon}_{k,p} g(x) - g^{(k)}(x) = O(\epsilon^{2p+1-k}).$$

For $r = 2p + 1$, consider the following Taylor expansion:

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

The cofficients c_l are therefore determined by a system of equations where $\delta_{i,k}$ 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 $k = 1$. 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 $j = 1$, $c_1 = 1/12$, $c_{-1} = -1/12$, $c_2 = -2/3$, $c_{-2} = +2/3, c_0 = 0$. In addition to 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 multivariate functions, the notation of pth order central derivatives can also be extended straightforwardly to partial derivatives. Since we are only concerned with $k = 1$, we only need to use $L^{\epsilon, x_j}_{1,p}$ to highlight the element of x for which the linear operator applies to. Also in multivariate functions, $g(x + \epsilon)$ means the vector of $[g(x + \epsilon e_k)], k = 1, \ldots, d$, where e_k is the vector with 1 in the kth position and 0 elsewhere, and d is the dimension of x.

Each jth component of $D_w(z)'$ in the asymptotic variance formula can then be estimated by

$$\tilde{D}^j_{\widehat{w}}(z) = L^{\varepsilon_n, \theta_j}_{1,p} \hat{m}\left(z; \widehat{\theta}, \widehat{\eta}(\cdot)\right) - L^{\tau_n, w_j}_{1,p} \hat{m}\left(z; \widehat{\theta}, \widehat{\eta}(\cdot)\right),$$

where ε_n and τ_n are the relevant step sizes for the numerical derivatives with respect to the finite and infinite-dimensional parameters. In general the step sizes ϵ_n and τ_n can be chosen differently for different elements of the parametric and nonparametric components. It might also be possible to adapt the equal distance grid to a variable distance grid of the form $L^{\epsilon}_{k,p} g(x) = \frac{1}{\epsilon^j} \sum_{l=-p}^{p} c_l g(x + t_l \epsilon)$, where t_l can be different from 1. In addition both the step size and the grid distance can also be made to be dependent on the observations and data driven. These possibilities are left for future research.

2.3 Sufficient conditions for consistency of finite-difference derivatives

Before we strive to obtain the weakest possible sufficient condition for consistency in next section, we first show that the existing sufficient conditions in the literature (e.g. Newey and McFadden (1994) and Powell (1984)) for parametric models can be straighforwardly generalized to semiparametric models.

ASSUMPTION 1. *For a linear operator $\Delta_{p,\theta^{p_1},\delta^{p_2}}[\delta]^{p_1}$ that is p^1th linear in θ, that has a finite second moment and that is linear in each argument, e.g., $\Delta_{2,\theta,\eta}[t\,\delta]\,(\theta-\theta_0) = t\Delta_{2,\theta,\eta}[\delta]\,(\theta-\theta_0)$, the following approximation holds at (θ_0, η_0):*

$$E\left[\left\|m\,(z;\,\theta,\,\eta(\cdot)) - \Delta_{1\theta}\,(\theta-\theta_0) - \Delta_{1\eta}[\delta] - \ldots - \sum_{p_1+p_2=p} \Delta_{p,\theta^{p_1},h^{p_2}}[\delta]^{p_1}\,(\theta-\theta_0)^{p_2}\right\|^2\right]$$
$$= o\left(\|\delta\|^{2p}_{\mathbf{L}^2} + \|\theta-\theta_0\|^{2p}\right).$$

Assumption 1 requires that the conditional moment $m\,(z;\,\theta,\,\eta(\cdot))$ is mean square differentiable in L^2 norm with respect to the distribution of z. The next assumption relates to the rate of convergence of the nonparametric conditional moment estimate. Define U_γ as a neighborhood of θ_0, η_0 with radius γ: $U_\gamma = \{\theta, \eta\,(\cdot) : ||\theta-\theta_0|| < \gamma, |\eta\,(\cdot) - \eta_0\,(\cdot)\,| < \gamma\}$.

ASSUMPTION 2. *For some $k \in \mathbb{N}$, $k \leq 1/2$, uniformly in $z \in \mathcal{Z}$, as $\gamma \to 0$,*

$$\sup_{(\theta,\,\eta(\cdot))\in U_\gamma} \frac{n^{1/k}\,\|\widehat{m}\,(z;\,\theta,\,\eta(\cdot)) - m\,(z;\,\theta,\,\eta(\cdot)) - \widehat{m}\,(z;\,\theta_0,\,\eta_0(\cdot))\|}{1 + n^{1/k}\|\widehat{m}\,(z;\,\theta,\,\eta(\cdot))\,\| + n^{1/k}\|m\,(z;\,\theta,\,\eta(\cdot))\,\|} = o_p(1).$$

For unconditional moment models, typically $k = 1/2$. For conditional moment models, $k < 1/2$. The particular rate will depend on the method and the choice of the tuning parameters used in the estimation procedure.

In addition, the parametric component of the model is assumed to converge at the usual $\sqrt{n}$ rate, while the functional component is assumed to converge at a slower nonparametric rate.

ASSUMPTION 3. *For $k_1 < 1/2$, $n^{1/k_1}\,\|\hat{\eta}(\cdot) - \eta_0(\cdot)\| = O_p(1)$, and $n^{1/2}\left\|\hat{\theta}-\theta_0\right\| = O_p(1)$.*

THEOREM 1. *Under assumptions 1, 2 and 3, if $\varepsilon_n\, n^{1/\max\{k,\,k_1\}} \to \infty$, $\varepsilon_n \to 0$, $\tau_n\, n^{1/\max\{k,\,k_1\}} \to \infty$, $\tau_n \to 0$, then $\sup_{z\in\mathcal{Z}} |\tilde{D}_w\,(z) - D_w\,(z)\,| \xrightarrow{p} 0$.*

The proof of the consistency theorem follows closely the arguments in the literature. The basic idea in the consistency argument is that while the step size should converge to zero to eliminate the bias, it should converge slowly so that the noise in the parameter estimation and in estimating the moment condition should not dominate the step size. In a parametric model, both the noise in the parameter estimation and in estimating the moment condition is of the order of $1/\sqrt{n}$. Therefore as shown in Newey and McFadden (1994) and Powell (1984), sufficiency will hold if $1/\sqrt{n} << \epsilon_n$. The extension of this argument to the semiparametric case is straightforward. The difference is that now the converge rates for both the (infinite-dimensional) parameters and the conditional moment equation are slower than $1/\sqrt{n}$ and therefore imposes a more stringent requirement on rate at which ϵ_n is allowed to converge to zero. However, as we will see in the next section, these sufficient conditions can be substantially weakened because of a local uniformity feature of the variance of the numerical derivatives.

2.4 Weak sufficient conditions for consistency for parametric models

In this section we show that the conditions on the step size for consistent derivative estimation is much weaker than previously known 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 consider the unconditional parametric and conditional semiparametric cases separately to best convey intuitions.

Consider a parametric unconditional moment model defined by the sample and population moment conditions: $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i, \theta)$ and $g(\theta) = Eg(Z_i, \theta)$ where $g(\theta) = 0$ if and only if $\theta = \theta_0$, where 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 $L_{1,p}^{\epsilon_n}\hat{g}(\hat{\theta}) = \left(L_{1,p}^{\epsilon_n,\hat{\theta}_j}\hat{g}(\hat{\theta}), j = 1, \ldots, d\right)$.

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

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

We discuss how to control each of these three terms in turn. Notice first that the 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.

The following assumption is a parametric version of Assumption 1.

ASSUMPTION 4. *A $2p+1$th order mean value expansion applies to the limiting function $g(\theta)$ uniformly in a neighborhood of θ_0. For all ϵ sufficiently small and $r = 2p+1$,*

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

An immediate consequence of this assumption is that $\hat{G}_2\left(\hat{\theta}\right) = O\left(\epsilon^{2p}\right)$. 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 aware of seems to come from a convergence rate result in Pollard (1984).

ASSUMPTION 5. *Consider functions $g(z, \theta)$ contained in class $\mathcal{F} = \{g(\cdot, \theta), \theta \in \Theta\}$. Then*

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

(ii) The sample moment function is Lipschitz-continuous in **mean square** *in some neighborhood of θ_0. That is for sufficiently small $\epsilon > 0$*

$$\sup_{\theta \in N(\theta_0)} E\left[\left(g(Z_i, \theta + \epsilon) - g(Z_i, \theta - \epsilon)\right)^2\right] = O(\epsilon).$$

(iii) The class of graph of functions from $\mathcal{F}$ admits a polynomial degree of discrimination

Most of the functions in econometric applications fall in this category. By Lemmas 25 and 36 of Pollard (1984), assumption 5 implies that there exist universal constants $A > 0$ and $V > 0$ such that for any $\mathcal{F}_n \subset \mathcal{F}$ with envelope function $\|F_n\|$,

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

LEMMA 1. *Under assumption 5, if $n\varepsilon_n/\log n \to \infty$*

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

Consequently, assumption 5 implies that $\hat{G}_1\left(\hat{\theta}\right) = o_p\left(1\right)$ if $d\left(\hat{\theta}, \theta_0\right) = o_p\left(1\right)$.

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 $\mathcal{F}_n = \{\epsilon_n L_{1,p}^{\epsilon_n} g\left(\cdot, \theta\right), \theta \in N\left(\theta_0\right)$, with envelope function F, such that $PF \leq C$. Then we can write

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

For each $f \in \mathcal{F}_n$, note that $Ef^2 = E\left(\epsilon_n L_{1,p}^{\epsilon_n} g\left(\cdot,\theta\right)\right)^2 = O\left(\epsilon_n\right)$ because of assumption 5.(ii). The lemma then follows immediately by taking $\alpha_n = 1$ and $\delta_n^2 = \epsilon_n$ in Theorem 2.37 in Pollard (1984). $\square$

THEOREM 2. *Under Assumptions 4 and 5, $L_{1,p}^{\epsilon_n}\hat{g}\left(\hat{\theta}\right) \xrightarrow{p} G\left(\theta_0\right)$ if $\epsilon_n \to 0$ and $n\epsilon_n/\log n \to \infty$, and if $d\left(\hat{\theta}, \theta_0\right) = o_p\left(1\right)$.*

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

The result of Theorem 2 improved if we are willing to impose the following stronger assumption, which holds for smoother functions such as those that are Hölder-continuous.

ASSUMPTION 6. *Consider functions $g(z, \theta)$ contained in class $\mathcal{F} = \{g(\cdot, \theta),\, \theta \in \Theta\}$. Then*

(i) For some $\gamma \in (0, 1]$ and sufficiently small $\epsilon > 0$ functions in $\mathcal{F}$ admit polynomial envelopes in ϵ on their finite differences in the small neighborhood of $\theta_0 \in \Theta$:

$$\sup_{\theta\in N(\theta_0)} \|g\left(Z_i, \theta + \epsilon\right) - g\left(Z_i, \theta - \epsilon\right)\| \leq C\, \epsilon^{2\gamma-2}.$$

(ii) For all sufficiently small ϵ and all $\theta \in \Theta$, if we define $\mathbb{G}_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(g(Z_i, \theta) - g(\theta)\right)$, then

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

for functions $\phi_n(\cdot)$ such that $\delta \mapsto \phi_n(\delta)/\delta^\gamma$ is non-increasing for γ defined in part (i).

Assumption 6 is more stringent than 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

$$E^* \sup_{d(\theta, \theta_0) < \delta} |\mathbb{G}(\theta_1) - \mathbb{G}(\theta_2)| \lesssim \phi_n(\delta). \tag{2.2}$$

For i.i.d data, the tail bounds method used in Van der Vaart and Wellner (1996) to compute (2.2) can be modified to obtain assumption 6. In particular, define a class of functions $\mathcal{M}_\delta^\epsilon = \{g(Z_i, \theta_1) - g(Z_i, \theta_2), d(\theta_1, \theta_2) \leq \delta, d(\theta_1, \theta_0) < \epsilon, d(\theta_2, \theta_0) < \epsilon\}$. Then assumption 6, which requires bounding $E_P^* ||G_n||_{\mathcal{M}_\delta^\epsilon}$, 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 $\mathcal{M}_\delta^\epsilon$,

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

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

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

THEOREM 3. *Under assumptions 4 and 6, $L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right) \xrightarrow{p} G(\theta_0)$ if $\epsilon_n \to 0$ and $n\epsilon^{2-2\gamma} \to \infty$, and if $d\left(\hat{\theta}, \theta_0\right) = o_p(1)$.*

This result shows that for continuous functions $g(Z_i, \theta)$ that are Lipschitz in θ, the only condition needed for consistency is $\epsilon_n \to 0$. The result of Theorem 3 demonstrates that as long as the sample moment function does not have discontinuities, one can pick the step size to decrease at the polynomial rate with the sample size. If the moment function is descontinuous, Theorem 2 needs to be applied instead of Theorem 3 prescribing a slower logarithmic rate of decrease in the step size.

Example Consider the simple quantile case where the moment condition is defined by $g(z_i; \theta) = 1(z_i \leq \theta) - \tau$. In this case the numerical derivative estimate of the density of z_i at θ is given by

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

This is basically the uniform kernel estimate of the density of z_i at θ:

$$\hat{f}\left(\hat{\theta}-\theta_0\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\epsilon}1\left(\frac{|z_i-\hat{\theta}|}{\epsilon}\leq 1\right)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2\epsilon}1\left(\frac{|z_i-\theta_0-\left(\hat{\theta}-\theta_0\right)|}{\epsilon}\leq 1\right).$$

The consistency conditions given in Powell (1984) and Newey and McFadden (1994), both of which require $\sqrt{n}\epsilon\to\infty$, are too strong. The intuition reason for this is because under this condition, the second part of the estimation noise due to $\hat{\theta}-\theta_0$, $\frac{\hat{\theta}-\theta_0}{\epsilon}$, will vanish. However, for the purpose of consistency this is not necessary. As long as $\hat{f}(x)$ is uniformly consistent for $f(x)$ for x in a shrinking neighborhood of 0 of size $n^{-\eta}$, it will follow that

$$\hat{f}\left(\hat{\theta}-\theta_0\right)\xrightarrow{p}f(0)=f_z(\theta_0).$$

2.5 Optimal rates for derivative estimation

The optimal choice of the step size depends typically on the smoothness of the empirical process indexed by γ and the smoothness of the population moment function indicated by the magnitude of the "Taylor residual" p. For the choice of the optimal rate for the step size of numerical differentiation, one can consider decomposition of the numerical derivatives into components $\hat{G}_1$, G_2 and G_3 corresponding to the variance, deterministic and stochastic bias components. The optimal choice of the step size will provide the minimum mean-squared error for the estimated derivative by balancing the bias and the variance. When the sample moment function is discontinuous, conditions of Theorem 2 apply, delivering the logarithmic rate of decay for the variance component $\hat{G}_1\left(\hat{\theta}\right)=O_p\left(\sqrt{\frac{\log n}{n\epsilon}}\right)$. On the other hand, application of Assumption 4 to the population moment leads to $\hat{G}_2\left(\hat{\theta}\right)=\epsilon^{2p}$. Under conditions of Theorem 3, the variance term has a polynomial depenence from the step size with $\hat{G}_1\left(\hat{\theta}\right)=O_p\left(\frac{1}{\sqrt{n\epsilon^{1-\gamma}}}\right)$, while the bias term is still determined by Assumption 4. We note that for Lipschitz-continuous or differentiable models, in which generally $\gamma=1$, there is no trade off between the variance and the bias, in which case the smaller the step size ϵ, the smaller the bias term. However, in this case the order of the root mean square is bounded from below by the variance term of $O\left(1/\sqrt{n}\right)$ for sufficiently smaller ϵ_n. The next theorem formalizes.

THEOREM 4. *Under the conditions of Theorem 2, if $\hat{\theta}-\theta_0=O_p\left(1/\sqrt{n}\right)$, the optimal rate of ϵ satisfies $\epsilon=O\left(\left(\log n/n\right)^{4p+1}\right)$, in which case the mean-squared error is $O_p\left(\left(\log n/n\right)^{\frac{4p}{4p+1}}\right)$. When the conditions of theorem 3 hold instead, the optimal rate of $\epsilon=O\left(n^{-\frac{1}{2(1-\gamma+2p)}}\right)$, in which case the error is $O_p\left(n^{-\frac{2p}{2(1-\gamma+2p)}}\right)$.*

2.6 Uniform consistency of directional derivatives for semiparametric models

This subection extends the weak consistency condition to directional derivatives of semiparametric conditional moment models. As in section 2.1, semiparametric conditional moment models are usually defined by the conditional moment function $m(\theta,\eta;\,z)=E\left[\rho\left(Y_i,\theta,\eta\right)\mid Z_i=z\right]$. In this section we focus on two special cases where the conditional moment function is estimated nonparametrically using orthogonal series and when using kernel smoothing. The infinite-dimensional parameter η is assumed to be estimated using series or sieves. The series estimator used to recover the conditional moment function is based on the vector of basis functions $p^N(z) = (p_{1N}(z), \dots, p_{NN}(z))'$,

$$\widehat{m}\,(\theta, \eta, z) = p^{N\prime}(z) \left(\frac{1}{n} \sum_{i=1}^{n} p^N(z_i) p^{N\prime}(z_i) \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} p^N(z_i) \rho\,(\theta, \eta; y_i)\,. \tag{2.3}$$

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

$$\widehat{m}\,(\theta, \eta, z) = \left(\frac{1}{n b_n^{d_z}} \sum_{i=1}^{n} K\left(\frac{z_i - z}{b_n^{d_z}} \right) \right)^{-1} \frac{1}{n b_n^{d_z}} \sum_{i=1}^{n} K\left(\frac{z_i - z}{b_n} \right) \rho\,(\theta, \eta; y_i)\,. \tag{2.4}$$

In either case, we will denote the resulting estimate by $\widehat{m}\,(\theta, \hat{\eta};\, x)$. It turns out that the numerical derivative consistency results for η apply without any modification to the parametric component θ. Therefore with no loss of generality below we will focus on differentiating with respect to η.

The directional derivative of m in the direction $w \in \mathcal{H} - \eta_0$ with respect to η, $G_w = \left.\frac{d\, m(\theta_0, \eta_0 + \tau w, x)}{d\,\tau}\right|_{\tau=0}$, 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,

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

Given that the direction w itself has to be estimated from the data as in section 2.1, 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).

ASSUMPTION 7. *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

(ii) For some $C > 0$, $\sup_{z \in \mathcal{Z}} \|p^N(z)\| \leq C < \infty$.

(iii) The population conditional moment belongs to the completion of the sieve space and

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

Assumption 7[ii] is convenient because $\rho\,(\cdot)$ is uniformly bounded. It can be relaxed to allow for a sequence of constants $\zeta_0(N)$ with $\sup_{z \in \mathcal{Z}} \|p^N(z)\| \leq \zeta_0\,(N)$, where $\zeta_0\,(N)$ grows at appropriate rates such as $\zeta_0(N)^2 N/n \to 0$ as $n \to \infty$. We will comment on how this can be taken into account.

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

ASSUMPTION 8. *(i) Uniformly bounded moment functions:* $\sup_{\theta,\eta} \|\rho(\theta,\eta,\cdot)\| \leq C$.

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

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

(iii) For each n*, the class of functions* $\mathcal{F}_n = \{\rho\left(\theta,\eta+\epsilon_n w;\cdot\right) - \rho\left(\theta,\eta-\epsilon_n w;\cdot\right), \theta\in\Theta, \eta, w \in \mathcal{H}_n\}$ *is Euclidean whose coefficients depend on the number of sieve terms. In other words, there exist constants* A, V*, and* $0 \leq r_0 < \frac{1}{2}$ *such that the covering entropy satisfies*

$$\log N\left(\delta, \mathcal{F}_n, \mathbf{L}_\infty\right) \leq A\, n^{2r_0}\, \delta^{-V}.$$

The hardest condition to verify is 8 (iii). This assumption imposes a joint restriction both on the class of functions $\mathcal{F}_n$ containing sieve estimators for η and the class of conditional moment functions parametrized both by θ and η. An example where this assumption holds is when $\rho(\cdot)$ is (weakly) monotone in η for each θ and $\mathcal{H}_n$ is a orthogonal basis of dimensionality $K(n)$. For example, $\rho(\cdot)$ can be an indicator in nonparametric quantile regression. Lemma 5 in Shen and Wong (1994) suggests that the uniform metric entropy of the class of sieve $\mathcal{F}_n$ has order $K(n)\log\frac{1}{\varepsilon} \leq K(n)\varepsilon^{-1}$ for sufficiently small $\varepsilon > 0$ and $\|\eta_n - \eta_0\|_\infty < \varepsilon$. Then by Lemma 2.6.18 in Van der Vaart and Wellner (1996), if 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, Linton, and Van Keilegom (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 8 (iii).

The following the result is formulated in the spirit of Theorem 37 of Pollard (1984). A related idea for unconditional sieve estimation has been used in Zhang and Gijbels (2003).

LEMMA 2. *Under assumptions 7 and 8*

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

uniformly in z *and* w*, provided that* $\epsilon_n \to 0$ *and* $\frac{n\epsilon_n}{N^2\,\log\,n} \to \infty$.

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

ASSUMPTION 9. *Kernel function* $K(\cdot)$ *integrates to 1, it is bounded and its square has a finite integral.*

Then we can formulate the following lemma that replicates the result of Lemma 2 for the case of the kernel estimator.

LEMMA 3. *Under assumptions 8 and 9*

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

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

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

THEOREM 5. *Under assumptions 4, 8, and either 7 or 9, $L_{1,p}^{\epsilon_n,w} \widehat{m}\left(\hat{\theta},\hat{\eta},z\right) \xrightarrow{p} \frac{\partial m(\theta,\eta,z)}{\partial \eta}[w]$, uniformly in z and w, if $\frac{n\epsilon_n}{N^2 \log n} \to \infty$ for series estimator, and $b_n \to 0$, $\frac{n\epsilon_n b_n}{\log n} \to \infty$ for kernel-based estimator, provided that $d\left(\hat{\theta},\theta_0\right) = o_p(1)$ and $d(\hat{\eta},\eta_0) = o_p(1)$.*

This theorem allows us to use finite-difference formulas for evaluation of directional derivatives. An interesting feature of this result is that it only indirectly depend on the rate of convergence of the infinite-dimensional parameter through our assumption 8[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.

2.7 Optimal rates for semiparametric estimators

In this section we consider a special case where finite differences of the moment function have a non-trivial envelope. Examples of such functions include Lipshitz and Hölder continuous functions. We introduce the following modification to Assumption 8(i):

ASSUMPTION 8.

(i') For any sufficiently small $\epsilon > 0$

$$E\left[\sup_{(\theta,\eta)\in\Theta\times\mathcal{H},w\in\mathcal{H},|w|<C} \left\|\rho\left(\theta,\eta+w\epsilon,Z\right)-\rho\left(\theta,\eta+w\epsilon,Z\right)\right\| \,\middle|\, X=x\right] \lesssim \phi_n(\epsilon),$$

where $\phi_n(\epsilon)$ is a function, that is non-increasing in n and $\phi_n(\epsilon)/\epsilon^\gamma$ is decreasing in ϵ for some $0 < \gamma \leq 1$.

Note that if $\rho(\cdot)$ is γ-Hölder continuous in η then the inequality in Assumption 8 (i') applies with power envelope of order ϵ^γ. This modification allows us to improve the rate results for the considered classes of functions. This result of particular interest because it also implies appropriate properties of the numerical directional derivative in cases where the moment function is at least Lipschitz continuous.

LEMMA 4. *Under either pair of assumptions 7 and 8(i'),(ii)-(iv), or 9 and 8 (i'), (ii)-(iv)*

$$\sup_{d(\hat{\theta},\theta_0)=o_p(1),d(\hat{\eta},\eta_0)=o_p(1)} \left| L_{1,p}^{\epsilon_n,w}\widehat{m}\left(\hat{\theta},\hat{\eta},x\right) - L_{1,p}^{\epsilon_n,w} m\left(\theta_0,\eta_0,x\right)\right| = o_p(1)$$

pointwise in x, provided that $\epsilon_n \to 0$ *and* $\frac{n}{N^2\,\epsilon_n^{2(1-\gamma)}\,\log\,n} \to \infty$ *for series estimator, and* $b_n \to 0$, $\frac{nb_n}{\epsilon_n^{2(1-\gamma)}\,\log\,n} \to \infty$ *for kernel estimator.*

The consistency of the numerical derivative is a direct consequence of this lemma. We note that conditions of Lemma 4 are weaker than the conditions for the functions with trivial (constant) envelopes. We can note that for functions that are Lipschitz-continuous the power $\gamma = 1$. This means that Lemma 4 will only set the requirement for the estimator of the conditional moment, but not the step size. This means that the use of the finite-difference formula will not affect the properties of the estimated conditional moment. On the other hand, we can note that consistency of the estimator for the directional derivative will be assured when $b_n = O\left(\epsilon_n^{2(1-\gamma)}\right)$ (among other cases). One interesting observation from this case is when the moment function becomes substantially non-smooth (i.e. when γ is close to zero), the step size for numerical differentiation should be much larger than the bandwidth.

Example: nonparametric LAD regression Consider the problem which is defined by the LAD objective and the regression function $\eta\left(X\right) \in C^{\alpha}\left(\mathcal{X}\right)$ ($\mathcal{X} \subset \mathbb{R}^d$). In this case the objective function takes the form $Q\left(\eta,x\right) = E\left[\left|Z_i - \eta\left(X_i\right)\right| \,\middle|\, X_i = x\right]$. Minimization of this objective with respect to functions $\eta(\cdot)$ leads to the moment equation

$$m\left(\eta,x\right) = E\left[\mathbf{1}\left\{Z_i \leq \eta\left(X_i\right)\right\} - \frac{1}{2} \,\middle|\, X_i = x\right] = 0$$

The object of interest will be the directional derivative of this moment equation at the true estimator in some direction $h(\cdot)$. Noting that

$$m\left(\eta + \tau\,h,x\right) = \int\limits_{\mathcal{Z}} \left[\mathbf{1}\left\{z \leq \eta(x) + \tau\,h(x)\right\} - \frac{1}{2}\right] f_{z|x}\left(z\right)\,dz,$$

we find the directional derivative of interest as $\frac{\partial m}{\partial \eta}[h] = \left.\frac{\partial m(\eta+\tau\,h,x)}{\partial \tau}\right|_{\tau=0} = f_{z|x}\left(\eta(x)\right)h(x)$. To solve this problem for a particular sample, we construct an empirical moment equation

$$\widehat{m}\left(\eta,x\right) = \frac{1}{n\,b_n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{b_n}\right)\left[\mathbf{1}\left\{z_i \leq \hat{\eta}\left(x_i\right)\right\} - \frac{1}{2}\right] = 0.$$

Then we can set a uniform grid over $\mathcal{X}$ and represent $\eta(x)$ by its values on the grid. If such grid contains G_n points, then $\eta_n(x) = \sum\limits_{k=1}^{G_n} \gamma_k \mathbf{1}\{x \in \Delta_k\}$ where Δ_k is the k-th segment of the grid. Then in each segment Δ_k the model becomes parametric. Conditional on $x \in \Delta_k$

$$\widehat{m}\left(\eta, x\right) = \frac{1}{n\, b_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b_n}\right) \left[\mathbf{1}\left\{z_i \leq \gamma_k^0\right\} - \frac{1}{2}\right] + \hat{\gamma}_k \frac{1}{n\, b_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{b_n}\right) f_{z|x}\left(\gamma_k^0\right) + O_p\left(\frac{1}{G_n^{\alpha} b_n\, n}\right)$$

Therefore, the estimator can be expressed as

$$\hat{\gamma}_k = -\frac{\frac{1}{n^2\, b_n} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \mathbf{1}\{x_j \in \Delta_k\} K\left(\frac{x_j - x_i}{b_n}\right) \left[\mathbf{1}\left\{z_i \leq \gamma_k^0\right\} - \frac{1}{2}\right]}{\frac{1}{n^2\, b_n} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \mathbf{1}\{x_j \in \Delta_k\} K\left(\frac{x_j - x_i}{b_n}\right) f_{z|x}\left(\gamma_k^0\right)} + O_p\left(G_n^{-\alpha}\right).$$

Then the estimator takes the form

$$\hat{\eta}\left(x\right) = \sum_{k=1}^{G_n} \hat{\gamma}_k \mathbf{1}\{x \in \Delta_k\}.$$

Then the estimator will converge at rate $\sqrt{n/G_n}$ and will not interfere with the kernel estimator as long as $G_n\, b_n \to \infty$ and $G_n^{1+2\alpha}/n \to \infty$.

Next we can approximate the directional derivative by a finite-difference formula. Consider $\widehat{\frac{\partial m}{\partial \eta}}[h] = \frac{\widehat{m}(\hat{\eta} + \varepsilon_n\, h, x) - \widehat{m}(\hat{\eta} - \varepsilon_n\, h, x)}{2\varepsilon_n}$. Using the estimator for the regression function approximated on the grid as above, we can transform the expression for the directional derivative to

$$\widehat{\frac{\partial m}{\partial \eta}}[h] = \frac{1}{n} \sum_{i=1}^{n} h(x_i) \frac{1}{b_n} K\left(\frac{x - x_i}{b_n}\right) \frac{1}{\varepsilon_n\, h(x_i)} U\left(\frac{z_i - \hat{\eta}\left(x_i\right)}{h(x_i) \varepsilon_n}\right),$$

where $U(\cdot)$ is a uniform kernel. For $h(x) > 0$ on $x \in \mathcal{X}$ and $\|h(x)\| = 1$ we can consider a stochastic bandwidth $q_n = \varepsilon_n\, h(x_i) = O_p\left(\varepsilon_n\right)$. As a result, the expression above can be re-written as

$$\widehat{\frac{\partial m}{\partial \eta}}[h] = \frac{1}{n} \sum_{i=1}^{n} h(x_i) \frac{1}{b_n} K\left(\frac{x - x_i}{b_n}\right) \frac{1}{q_n} U\left(\frac{z_i - \hat{\eta}\left(x_i\right)}{q_n}\right).$$

We note that as long as $\frac{\eta(x_i) - \hat{\eta}(x_i)}{q_n} \to 0$, the variance of the combined kernel is determined by $\|b_n q_n\| \leq b_n \varepsilon_n$. Then the summand is dominated by a regular multiplicative kernel. Uniform convergence will be provided by the restriction on the rate of decrease of the product bandwidth $\frac{n b_n \varepsilon_n}{\log\, n} \to \infty$. This coinsides with the condition that we provided in the previous section that assures the uniform convergence $\widehat{\frac{\partial m}{\partial \eta}}[h] \xrightarrow{p} \frac{\partial m}{\partial \eta}[h]$.

3 Numerical optimization of non-smooth sample functions

3.1 Parametric extremum estimation: definitions

In this section we study the properties of estimators based on numerically solving the first-order conditions for likelihood-type objective functions. The original estimator of interest maximizes the sample objective function. However, either by explicit researcher choice or by the implicit choice of the maximization routine in the optimization software, the original problem is replaced by the search for the zero of the numerically computed gradient. Consider the problem of estimating parameter θ_0 in a metric space (Θ, d) with the metric d. The true parameter θ_0 is assumed to uniquely maximize the limiting objective function $Q(\theta) = Eg(Z_i; \theta)$. An M-estimator $\hat{\theta}$ of θ_0 is typically defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \hat{Q}(\theta), \tag{3.5}$$

where $\hat{Q}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i; \theta)$. However, in practice, most sample objective functions $\hat{Q}(\theta)$ 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 numerical differentiation can sometimes lead to a model that is different from the one usually studied under M-estimation . 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 translate a nonstandard parametric model into a nonparametric one.

We focus on the class of optimization procedures that are based on numerical gradients, that are evaluated using the finite-difference formulas which we described in Section 2.2. We start by presenting a finite-difference numerical derivative version of the, M-estimator in (3.5). A numerical gradient-based optimization routine effectively substitutes (3.5) by a solution to the non-linear equation

$$||L_{1,p}^{\varepsilon_n} \hat{Q}_n\left(\hat{\theta}\right)|| = o_p\left(\frac{1}{\sqrt{n}}\right), \tag{3.6}$$

for some sequence of step sizes $\varepsilon_n \to 0$ and $\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i, \theta)$. 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 $\hat{D}_n^{\varepsilon}\left(\hat{\theta}\right) \equiv L_{1,1}^{\varepsilon} Q_n\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}}\right)$. 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 will argue, the estimators (3.5) and (3.6) 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 of extremum estimators for parametric models

Our first step is to provide consistency of $\hat{\theta}$. The consistency analysis is based on the premise that the population problem has a unique maximum and the first-order condition has an isolated well-defined root corresponding to the global maximum. 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 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 10. *The map* $\Theta \mapsto \mathbb{R}^k$ *defined by* $D(\theta) = \frac{\partial}{\partial\theta} E[g(Z_i, \theta)]$ *is identified at* $\theta_0 \in \Theta$. *In other words 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,* $g(\theta) = E[g(Z_i, \theta)]$ *is locally quadratic at* θ_0 *with* $g(\theta) - g(\theta_0) \lesssim -d(\theta, \theta_0)^2$.

The next assumption maintains suitable measurability requirements.

ASSUMPTION 11. *The parameter space* Θ *has a compact cover. For each* n, *there exists a countable subset* $T_n \subset \Theta$ *such that*

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

where P^* stands for the outer measure. In general, 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. If the moment function is continuous, it trivially satisfies this condition, but it also allows us to consider the moments defined by discontinuous functions. More precisely, Assumption 11 is a sufficient condition for the moment function to be *image admissible Suslin*. As it is discussed in Dudley (1999) and Kosorok (2008) this property will be required to establish the functional uniform law of large numbers needed for consistency.

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 for which we can rely on Assumption 4 that we previously used to establish uniform consistency for the estimate of the derivative of the sample moment function.

We will organize the discussion below by different classes of functions that can be used in practice. We start with functions that have absolutely locally bounded finite differences. Indicator functions and other functions with finite jumps fall into this category. Then we consider a class of functions that have polynomial bounds on finite differences. This class includes Lipschitz and Hölder-continuous functions which can have “mildly explosive” finite differences (those which have infinite jumps but approach infinity slower than some power of the distance to the point of discontinuity).

3.2.1 Functions with absolutely bounded finite differences

For the proof of consistency of the exremum estimator we need to provide primitive conditions for the uniform convergence in probability of the numerical derivative of the sample objective function to the derivative of the population objective function. This proof usually invokes the use of maximum inequalities that can bound the expectation of extreme deviations of the sample objective function in small neighborhoods of the parameter space. It is hard to work with the maximum inequality directly when the sample objective function experiences finite jumps: in this case small deviations of the parameter may lead to finitely large changes in the objective function. However, establishing the uniform convergence in probability still remains possible if we are willing to analyze more delicate properties of the function class under consideration. In our analysis we focus on the class of functions outlined by Assumption 5.

The following theorem establishes the consistency of numerical gradient-based extremum estimators for the class of possibly discontinuous functions described by Assumption 5. It is a corollary of Theorem 2 following directly from the uniform convergence in probability as in Amemiya (1985) and, therefore, we omit the proof.

THEOREM 6. *Under assumptions 10, 11, 4, and 5, as long as* $\varepsilon_n \to 0$ *and* $\frac{n\varepsilon_n}{\log n} \to \infty$*,*

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

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

This Theorem suggests that even though the sample objective function can be discontinuous with finite jumps, as long as it is Lipschitz-continuous in the mean square, one can use the numerical gradient-based routine for its optimization as long as the step size decays logarithmically with the sample size.

3.2.2 Functions with polynomial envelopes for finite differences

Our results in the previous subsection refer to the classes of objective functions for which the changes in the values of the objective function may not be directly related to the magnitude of the parameter changes. In this subsection we consider the case where such connection can be established. Surprisingly, our results are also valid for the cases of substantially irregular behavior of the objective function when its first derivative approach infinity in the vicinity of the maximum or minimum. A case in point is the objective function defined by $g(Z_i, \theta) = \sqrt{|Z_i - \theta|}$ for which local changes in θ around the origin lead to inversely proportional changes in the moment function. It turns out that this explosive behavior can be compensated by an appropriate choice of the sequence of step sizes for the numerical derivative. It turns out that the objective functions of this type belong to a class of functions outlined in Assumption 6 which includes the Hölder-continuous functions.

We note that Assumption 6 (i) restricts our analysis to the 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 differences of functions to be locally "explosive". For instance, if we consider finite differences of the objective function $g\left(Z_i, \theta\right) = \sqrt{|Z_i - \theta|}$ around the origin, we note that they will be proportional to $1/\sqrt{\epsilon}$ and will not shrink with the decrease in ϵ. It turns out that this still allows us to provide consistency for the numerical gradient-based estimators.

The following theorem 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.

THEOREM 7. *Under Assumptions 4, 6, 10, and 11, as long as* $\varepsilon_n \to 0$ *and* $\frac{n\varepsilon_n^{2-2\gamma}}{\log n} \to \infty$,

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

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

Theorem 7 establishes the lower bound on the rate of approach of the sequence of step sizes ε_n to zero. For models that have Lipschitz-continuous sample objective functions (which include models with smooth sample objective functions) $\gamma = 1$. In this case the restriction $\frac{n\varepsilon_n^{2-2\gamma}}{\log n} \to \infty$ holds trivially. This implies that for smooth models the sequence of step sizes can approach zero arbitrarily fast.[2] In cases of Hölder-continuous functions, the Theorem prescribes slower rate of approach of the step size to zero with the sample size for the objective functions with smaller Hölder index. The slowest rate of approach of the step size to zero is required when the finite differences of the objective function diverge to infinity as the step size shrinks (implying that the first derivatives can approach to infinity at the point of interest). The intuition for this result is that for non-smooth functions, the numerical derivative operator works as a smoothing device and uniform consistency requires balancing the variance and the bias due to smoothing.

3.3 Rate of convergence and asymptotic distribution in parametric case

3.3.1 Functions with absolutely bounded finite differences

In the previous section we provided sufficient conditions that determine consistency of the estimator that equates the finite-difference approximation of the of gradient of objective function to zero. For the classes where local parameter changes do not lead to proportional changes in the sample objective function we restricted our attention to the functions with absolutely bounded finite differences forming Euclidean classes. Our condition provided the result that the numerical derivative of the sample objective function converges to the derivative of the population objective function uniformly in probability. In addition, making use of the result of Lemma 1 we can establish the precise rate of convergence for the objective. For $\hat{\theta} \xrightarrow{p} \theta_0$

$$\sup_{\theta \in N(\theta_0)} \sqrt{\frac{n\varepsilon_n}{\log\, n}} ||L_{1,p}^{\epsilon_n} \hat{Q}(\theta) - L_{1,p}^{\epsilon_n} \hat{Q}(\theta)|| = O_p(1),$$

under Assumption 5 and $n\varepsilon_n / \log\, n \to \infty$. This result will be useful for localization of the estimated parameter.

LEMMA 5. *Suppose* $\hat{\theta} \xrightarrow{p} \theta_0$ *and* $L_{1,p}^{\varepsilon} \hat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)$. *Under Assumptions of Theorem 6, if* $n\varepsilon_n / \log n \to \infty$, *and* $\sqrt{n\varepsilon^{1+p}} = O(1)$, *then* $\sqrt{\frac{n\varepsilon_n}{\log\, n}} d\left(\hat{\theta}, \theta_0\right) = o_{P^*}(1)$.

[2] In Section 7 we point at some 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.

Once we have "located" the parameter, we can investigate the behavior of the sample objective function in shrinking neighborhoods of size $\left(\frac{\log\, n}{n\varepsilon_n}\right)$. It turns out, that in such neighborhoods we can improve the rate result by choosing the step size in accordance with the radius of the neighborhoods containing the true parameter.

LEMMA 6. *Under conditions of theorem 6 with* $\sqrt{n\varepsilon_n^{1+p}} = o\,(1)$*, and* $\frac{n\varepsilon_n^3}{\log\, n} \to \infty$ *we have*

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

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

THEOREM 8. *Suppose* $\hat{\theta} \xrightarrow{p} \theta_0$ *and* $L_{1,p}^{\varepsilon}\hat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)$*. Under Assumptions of Theorem 6, if* $n\varepsilon_n/\log n \to \infty$*,* $\varepsilon_n = o\left(\sqrt{\frac{\log\, n}{n\varepsilon_n}}\right)$*, and* $\sqrt{n\varepsilon^{1+p}} = O\,(1)$*, then* $\sqrt{n\varepsilon_n} d\left(\hat{\theta},\theta_0\right) = O_P^*\,(1)$*.*

We have established the convergence rate for the finite-difference based extremum estimators with bounded finite differences. Under additional assumptions we can also establish the normality of the asymptotic distribution,which requires showing stochastic equicontinuity as a corollary of the rate of convergence.

COROLLARY 1. *Under conditions of theorem 6 with* $\sqrt{n\varepsilon_n^{1+p}} = o\,(1)$*, and* $n\varepsilon_n^3 \to \infty$ *we have*

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

Proof. The result in the corollary follows directly from the result in Lemma 6 if one notices that $n\varepsilon_n^3 \to \infty$ guarantees that $\varepsilon_n = o\left(\frac{\log\, n}{n\varepsilon_n}\right)$. Then by using parametrization $\theta_n = \theta_0 + \frac{t_n}{\sqrt{n\varepsilon_n}}$ we replicate the argument of Lemma 6. □

In the proof of Lemma 6 we found that a convenient normalization for finite differences of the sample objective function is $(g(\cdot,\theta+\varepsilon_n) - g(\cdot,\theta+\varepsilon_n))/\sqrt{\varepsilon_n}$. We can impose an assumption on this object that assures its normality at θ_0 such as the standard Lindeberg condition which is often used to show pointwise asymptotic normality of kernel smoothers.

ASSUMPTION 12. *A CLT holds: As* $n \to \infty$ *and* $\epsilon_n \to 0$*,*

$$\frac{\mathbb{G}_n\left(\theta_0+\varepsilon_n\right) - \mathbb{G}_n\left(\theta_0-\varepsilon_n\right)}{\sqrt{\varepsilon_n}} \xrightarrow{d} N\left(0,\Omega\right).$$

Based on this intuition we can provide the following Theorem.

THEOREM 9. *Assume that the conditions of theorem 6 hold but with* $\sqrt{n\varepsilon_n^{1+p}} = o\,(1)$*. In addition, suppose that the Hessian matrix* $H\left(\theta\right)$ *of* $g\left(\theta\right)$ *is continuous, nonsingular and finite at* θ_0*. Then if Assumption 12 holds with* $n\varepsilon_n^3 \to \infty$

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

3.3.2 Functions with polynomial envelopes for finite differences

In case where functions of interest permit power envelopes on the finite differences, we can establish the rate of convergence and describe the asymptotic distribution of the resulting estimator. Next, we establish the rate of convergence and the asymptotic distribution of the numerical derivative based M-estimator for the functions that admits polynomial envelopes on finite differences. We provide the following general result.

THEOREM 10. *Suppose* $\hat{\theta} \xrightarrow{p} \theta_0$ *and* $L_{1,p}^{\varepsilon} \hat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}\varepsilon_n^{1-\gamma}}\right)$. *Under Assumptions 11 and 6, if*

$$n\varepsilon_n^{2-2\gamma} / \log n \to \infty$$

and $\sqrt{n}\varepsilon_n^{1-\gamma+2p} = O\left(1\right)$, *then* $\sqrt{n}\varepsilon_n^{1-\gamma} d\left(\hat{\theta}, \theta_0\right) = O_P^*\left(1\right)$.

This result is a Z-estimator version of Theorem 3.2.5 in Van der Vaart and Wellner (1996). Note that given the consistency assumption, the conditions required for obtaining the rate of convergence are weaker. For a typical two sided derivative $\nu = 2$. In this case, for a regular parametric model where $\gamma = 1$, the condition $\sqrt{n}\varepsilon^2 \to 0$ is needed to obtain the usual asymptotic distribution without bias. This rate is compatible with $\sqrt{n}\epsilon \to \infty$ and also allows for an ε sequence that delivers a consistent, albeit non-optimal, estimator of the asymptotic variance.

We now proceed with the analysis of the asymptotic distribution of the estimator that we obtain by equating the numerical derivative to zero. The following simplifying assumption suggests the asymptotic normality of the numerical derivative of the sample objective function at the true parameter value. It can be established for a sufficiently slow step size convergence rate by verifying the Lindeberg conditions and the finiteness of the covariance function. Section 3.4 establishes a more general form of the distribution of the estimator for the cases where the step sizes approaches zero at different rates. Asymptotic normality turns out to be a special case for the more general distribution.

ASSUMPTION 13. *A CLT holds: As* $n \to \infty$ *and* $\epsilon_n \to 0$,

$$\frac{\mathbb{G}_n\left(\theta_0 + \varepsilon_n\right) - \mathbb{G}_n\left(\theta_0 - \varepsilon_n\right)}{\varepsilon_n^{\gamma}} \xrightarrow{d} N\left(0, \Omega\right).$$

The following theorem establishes the asymptotic normality of the numerical derivative-based estimator with an additional assumption regarding the derivatives of the population objective function.

THEOREM 11. *Assume that the conditions of theorem 10 hold but with* $\sqrt{n}\varepsilon_n^{1+2p-\gamma} = o\left(1\right)$. *In addition, suppose that the Hessian matrix* $H\left(\theta\right)$ *of* $g\left(\theta\right)$ *is continuous, nonsingular and finite at* θ_0 *with Assumption 13 valid. Furthermore,* $\sqrt{n}\varepsilon_n^{2-\gamma} \to \infty$. *Then*

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

The additional assumption $\sqrt{n}\varepsilon_n^{2-\gamma} \to \infty$ turns out to be stronger for smooth models than for nonsmooth models. This is an artifact that we are relying on Assumption 6 and the convergence rate result in theorem 10 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 $\gamma = 1$, we might be willing to 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. The next theorem states such an alternative result.

THEOREM 12. *Suppose the conditions of theorem 11 hold except $\sqrt{n}\varepsilon_n^{2-\gamma} \to \infty$. Suppose $n\varepsilon_n \to \infty$. Suppose further that $g(z_i, \theta)$ is mean square differentiable in a neighborhood of θ_0: for measurable functions $D(\cdot,\cdot) : \mathcal{Z} \times \Theta \to \mathbb{R}^p$ such that*

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

$E\|D(Z,\theta_1)\|^2 < \infty$ for all θ_1, and $\theta_2 \in \mathcal{N}_{\theta_0}$. Define $q_\varepsilon(z_i,\theta) = L_{1,p}^{\varepsilon} g(z_i;\theta) - D(z,\theta)$, Assume that

$$\sup_{d(\theta,\theta_0)=o(1),\varepsilon=o(1)} \left[\mathbb{G}q_\varepsilon(z_i,\theta_1) - \mathbb{G}q_\varepsilon(z_i,\theta_0)\right] = o_p(1),$$

and $D(z_i,\theta)$ is Donsker in $d(\theta,\theta_0) \leq \delta$, then the conclusion of theorem 11 holds.

Note that we still require $\sqrt{n}\varepsilon_n^2 \to 0$ to remove the asymptotic bias, and $n\varepsilon_n \to \infty$ to eliminate the variance in assumption 13, but we no longer require $\sqrt{n}\varepsilon \to \infty$. The conditions of this theorem are best understood in the context of a quantile regression estimator. Consider $g(z_i,\theta) = |z_i - \theta|$, and $p = 2$, so that $D(z,\theta) = \operatorname{sgn}(z_i - \theta)$ and

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

Then we can bound $q_\varepsilon(z_i,\theta_1) - q_\varepsilon(z_i,\theta_0)$ by, depending on which of $d(\theta,\theta_0)$ and ε is larger, the product between $\frac{1}{\varepsilon}\max(|z_i - \theta|, |z_i - \theta_0|)$ and the maximum of $1(|z_i - \theta|) \leq \varepsilon + 1(|z_i - \theta_0| \leq \varepsilon)$, and $[1(\theta - \varepsilon \leq z_i \leq \theta + \varepsilon_0) + 1(\theta_0 - \varepsilon \leq z_i \leq \theta_0 + \varepsilon_0)]$. Since $\max(|z_i - \theta|, |z_i - \theta_0|) \leq \varepsilon$ when $q_\varepsilon(z_i,\theta) - q_\varepsilon(z_i,\theta_0)$ is nonzero, the last condition in theorem 12 is clearly satisfied by the euclidean property of the indicator functions. Alternatively, the $q_\varepsilon(z_i,\theta)$ function in the last condition can also be replaced directly by $L_{1,p}^{\varepsilon} g(z_i,\theta)$.

3.4 General distribution results

The previous asymptotic normality result turns out to be an artifact of an excessively slow rate of approach of the sequence of step sizes ε_n to zero. Our previous assumption required that we choose $n\varepsilon_n^3 \to \infty$ for the discontinuous case and $\frac{\varepsilon_n^{\gamma-2}}{\sqrt{n}} = o(1)$ for continuous case. This assumption can be relaxed, at a cost of making the asymptotic distribution non-standard. However, this weakening also demonstrates that the numerical derivative-based estimators for non-smooth sample objective functions have interesting parallels with the cube-root asymptotics of Kim and Pollard (1990).

The following assumption has a simple implication for the covariance function of the sample objective function. It requires that the pairwise products of the sample objective function computed at different points in a vanishing neighborhood of the true parameter value have continuous expectations. Moreover, the variance of the numerical derivative of the sample objective function is infinitesimal at the points where pointwise derivative of the sample objective function may not exist. We first present the assumption for the case of discontinuous moment function.

ASSUMPTION 14. *Suppose that for each sequence $\varepsilon_n \to 0$ with $\sqrt{n\varepsilon_n^{1+p}} = o(1)$*

(i) The covariance function

$$H_{n,\varepsilon_n}(s,t) = \lim_{\alpha\to\infty} E\left[\frac{\alpha}{\varepsilon_n}\left(g\left(Z_i,\theta_0+\varepsilon_n+\frac{s}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{s}{\alpha}\right)\right)\right.$$
$$\left.\times\left(g\left(Z_i,\theta_0+\varepsilon_n+\frac{t}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{t}{\alpha}\right)\right)\right]$$

is finite and has a finite limit as $n\to\infty$ for $s,t\in\mathbb{R}$ and $\varepsilon_n = O\left(\frac{1}{\sqrt[3]{n}}\right)$.

(ii) For each t and each $\delta>0$

$$\lim_{\alpha\to\infty} E\left[\frac{\alpha}{\varepsilon_n}\left(g\left(Z_i,\theta_0+\varepsilon_n+\frac{t}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{t}{\alpha}\right)\right)^2\right.$$
$$\left.\times\mathbf{1}\left\{\|g\left(Z_i,\theta_0+\varepsilon_n+\frac{t}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{t}{\alpha}\right)\|>\alpha\,\delta\right\}\right]=0.$$

Similarly, we can impose an assumption for the case of Hölder-continuous objective functions.

ASSUMPTION 15. *Suppose that for each sequence $\varepsilon_n \to 0$ with $\sqrt{n}\varepsilon_n^{1+2p-\gamma} = o(1)$*

(i) The covariance function

$$H_{n,\varepsilon_n}(s,t) = \lim_{\alpha\to\infty} E\left[\frac{\alpha}{\varepsilon_n^{2\gamma}}\left(g\left(Z_i,\theta_0+\varepsilon_n+\frac{s}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{s}{\alpha}\right)\right)\right.$$
$$\left.\times\left(g\left(Z_i,\theta_0+\varepsilon_n+\frac{t}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{t}{\alpha}\right)\right)\right]<\infty$$

is finite and has a finite limit as $n\to\infty$ for $s,t\in\mathbb{R}$ and $\varepsilon_n = O\left(n^{-1/2(2-\gamma)}\right)$.

(ii) For each t and each $\delta>0$

$$\lim_{\alpha\to\infty} E\left[\frac{\alpha}{\varepsilon_n^{2\gamma}}\left(g\left(Z_i,\theta_0+\varepsilon_n+\frac{t}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{t}{\alpha}\right)\right)^2\right.$$
$$\left.\times\mathbf{1}\left\{\|g\left(Z_i,\theta_0+\varepsilon_n+\frac{t}{\alpha}\right)-g\left(Z_i,\theta_0-\varepsilon_n+\frac{t}{\alpha}\right)\|>\alpha\,\delta\right\}\right]=0.$$

We combine Assumptions 14 and 15 with Assumptions 5 and 6 that restrict the attention to particular (large) parametric classes of functions.

THEOREM 13. *1. Suppose that assumptions 4, 5 and 15 hold. The population objective has a finite Hessian* $H(\theta_0)$ *at* θ_0 *and* $\frac{n^3}{\varepsilon_n} = O(1)$. *Then*

$$\sqrt{\frac{\varepsilon_n}{n}} \sum_{i=1}^{n} L_{1,p}^{\epsilon_n} g\left(Z_i, \theta_0 + \frac{t}{\sqrt{n\varepsilon_n}}\right) \rightsquigarrow Z(t).$$

2. Suppose that assumptions 4, 6 and 15 hold. The population objective has a finite Hessian $H(\theta_0)$ *at* θ_0 *and* $\frac{\varepsilon_n^{\gamma-2}}{\sqrt{n}} = O(1)$. *Then*

$$\frac{\varepsilon_n^{1-\gamma}}{\sqrt{n}} \sum_{i=1}^{n} L_{1,p}^{\epsilon_n} g\left(Z_i, \theta_0 + \frac{t}{\varepsilon_n^{1-\gamma}\sqrt{n}}\right) \rightsquigarrow Z(t).$$

In these expressions $Z(t)$ *is a mean-zero Gaussian process with covariance function* $H(s,t) = \sum_l = -p^p l^2 c_l^2 \lim_{n\to\infty} H_{n,l\varepsilon_n}(s,t)$. *Then* $\sqrt{n}\varepsilon_n^{1-\gamma}\left(\widehat{\theta} - \theta_0\right) \rightsquigarrow \hat{t}$, *where* $\hat{t}$ *is defined by* $Z(\hat{t}) = H(\theta_0)\,\hat{t}$. *In one dimension,* $\hat{t}$ *can be interpreted as a boundary-crossing distribution.*

In the special case where $\sqrt{n}\varepsilon_n^{2-\gamma} \to \infty$, $Z(t)$ *is normal and does not depend on* t.

Remark: Theorem 13 establishes that the lowest rate at which ε_n approaches zero is $n^{-1/2(2-\gamma)}$ for the case of Hölder-continuous moment functions and $n^{-1/3}$ for the case of discontinuous objective functions. A faster approach of the step size to zero leads to a loss of stochastic equicontinuity. This condition also provides the slowest convergence rate for the estimator of $n^{1/(2(2-\gamma))}$ for the Hölder case and $n^{1/3}$ for the discontinuous case.

Theorem 13 shows that depending on the step size sequence chosen, the asymptotic distribution is a "hybrid" between the distribution of the original extremum estimators and the distribution of smoothed estimators. The asymptotics in case of the "under-smoothed" numerical derivative is non standard and is driven by the boundary-crossing distribution of the limiting process (a discussion of this distribution can be found, for instance in Durbin (1971, 1985)).

4 Semiparametric extremum estimators

4.1 Definitions

Our approach of transforming the problem of extremum estimation to the problem of solving a numerical first-order condition can be extended to the case where the parameter space is either entirely infinite-dimensional or contains an infinite-dimensional component. We consider a metric product space $\Theta \times \mathcal{H}$ where Θ is a compact subset of a Euclidean space $\mathbb{R}^p$ and $\mathcal{H}$ is a functional Banach space. Semiparametric extremum estimators typically arize from conditional moment equations. We consider a population moment equation

$$m(\eta, \theta, z) = E\left[\rho(\theta, \eta, Y_i)\,|Z_i = z\right] = 0$$

with $\rho : \Theta \times \mathcal{H} \times \mathcal{Y} \mapsto \mathcal{M} \subset \mathbb{R}^k$. The estimation problem is re-casted into an optimization problem by defining the objective $Q(\theta, \eta) = E\left[m(\theta, \eta, Z_i)' \, W(Z_i) \, m(\theta, \eta, Z_i)\right]$ using a $k \times k$ positive semi-definite (almost everywhere in $\mathcal{Z}$) weighting matrix $W(\cdot)$. The estimator minimizes the sample objective function with the infinite-dimensional component η over sieve space $\mathcal{H}_n$

$$\left(\hat{\theta}, \hat{\eta}\right) = \arg \min_{(\theta,\eta) \in \Theta \times \mathcal{H}_n} \widehat{Q}_n(\theta, \eta) = \frac{1}{n} \sum_{i=1}^{n} \hat{m}(\theta, \eta, z_i)' \, \widehat{W}(z_i) \, \hat{m}(\theta, \eta, z_i) \, .$$

A typical set of the necessary conditions for the optimum of $Q(\theta, \eta)$ can be found, for example, in the general class of mathematical programming problems in Pshenichnyi (1971). Consider a cone K in $\Theta \times \mathcal{H}$. If $(\theta_0, \, \eta_0) \in \Theta \times \mathcal{H}$ optimizes $Q(\theta, \eta)$, then there exists a number $\lambda_0 \geq 0$ such that $\theta(\lambda) = \theta_0 + \lambda \, \delta \in \Theta$ and $\eta(\lambda) = \eta_0 + \lambda \, w \in \mathcal{H}$ for all $(\delta, w) \in K$ and $\lambda \in \mathbb{R}$. Moreover, $\lambda_0 \, \zeta(\delta, w) = 0$, and

$$\lim_{\lambda \to +0} \frac{Q(\theta(\lambda), \eta(\lambda)) - Q(\theta_0, \eta_0)}{\lambda} \leq \zeta(\delta, w) \, ,$$

where $\zeta(\delta, w)$ is a functional which is convex with respect to (δ, w). If we assume that the objective functional is strictly concave at (θ_0, η_0) then $\lambda_0 > 0$. This transforms the necessary conditions to

$$\lim_{\lambda \to +0} \frac{Q(\theta_0 + \lambda \, \delta, \eta_0 + \lambda \, w) - Q(\theta_0, \eta_0, x)}{\lambda} = 0.$$

In particular, this directional derivative should be equal to zero for all directions within the cone K. Specifically, if $\Theta \times \mathcal{H}$ is a linear space, then this should be valid for all directions in $(\Theta - \theta_0) \times (\mathcal{H} - \eta_0)$. If the functional is Frechèt differentiable in (θ, η), then the directional derivative exists in all directions in K and we can write the necessary condition for the extremum in the simple form:

$$\frac{d}{d\tau} Q(\theta_0 + \tau \, \delta, \eta_0 + \tau \, w) \mid_{\tau=0} = 0,$$

in all directions $w \in \mathcal{H} - \eta_0$. In particular if Θ is a finite-dimensional vector space and $\mathcal{H}$ is a finite-dimensional functional space, then we can construct a system of first-order condition that exactly identifies a parameter pair (θ, η) as

$$\begin{aligned} &\tfrac{\partial Q(\theta,\eta)}{\partial \theta_k} = 0, \quad \text{for} \;\; k = 1, \ldots, p, \\ &\tfrac{\partial Q(\theta,\eta)}{\partial \eta} [\psi_j] = 0, \quad \text{for} \;\; j = 1, \ldots, G, \end{aligned} \tag{4.7}$$

where $\psi_j(\cdot)$ is a system of distinct elements of $\mathcal{H}$. In cases where the functional space $\mathcal{H}$ is infinite-dimensional, then we define the population solution to the system of first-order condition as a limit of sequence of solutions in the finite-dimensional sive spaces $\mathcal{H}_n$ such that $\mathcal{H}_n \subseteq \mathcal{H}_{n+1} \subseteq \mathcal{H}$ for all n.

4.2 Consistency

We consider the class of models where such substitution of maximization of the functional to finding a solution to the system of functional equations is possible. We formalize this concept by the following assumption.

ASSUMPTION 16. *Suppose that (θ_0, η_0) is the maximizer of the functional $Q(\theta, \eta)$ and $\mathcal{H}_n$ is the sieve space such that $\mathcal{H}_n \subset \mathcal{H}_{n+1} \subset \mathcal{H}$. The set $\mathcal{H}_\infty$ is complete in $\mathcal{H}$. The sets $\mathcal{H}_n$ share the same basis $\{\psi_j\}_{j=0}^{\infty}$ and $\langle \eta_0, \psi_j \rangle \to 0$ as $j \to \infty$. We assume that the left-hand side of (4.7) is continuous with respect to the strong product norm in $\Theta \times \mathcal{H}$. Suppose that (θ_n, η_n) solves (4.7). Then for any sequence of the sieve spaces $\mathcal{H}_n$ satisfying the above conditions the corresponding system of solutions (θ_n, η_n) converges to (θ_0, η_0) in the strong norm. Moreover, $Q(\cdot)$ is locally quadratic at (θ_0, η_0), that is for some neighborhod of (θ_0, η_0) there exist positive constants C_1 and C_2 such that*

$$Q\left(\theta_0, \eta_0\right) - Q\left(\theta, \eta\right) \leq C_1 \|\theta - \theta_0\|^2 + C_2 \|\eta - \eta_0\|^2.$$

This identification condition establishes the properties of the population objective function. We will now consider the properties of the sample objective function which can be expressed as

$$\widehat{Q}\left(\theta, \eta\right) = \frac{1}{n} \sum_{i=1}^{n} \hat{m}\left(\theta, \eta, z_i\right)' \, \widehat{W}\left(z_i\right) \, \hat{m}\left(\theta, \eta, z_i\right).$$

Following our analysis in Section 2.6 where we focused on computing directional derivatives of semiparametric models via finite-difference methods, we consider two cases where the estimator for the conditional moment function is obtained via series approximation (2.3) or kernel smoothing (2.4). The transformed system of equations that needs to be solved to obtain the estimator $\hat{\eta}$ and $\hat{\theta}$ can be obtained by directly applying the directional derivative to the objective function. Without loss of generality to simplify the algebra we focus on the case where the finite-dimensional parameter θ is scalar, the conditional moment function is one-dimensional, and the weighting matrix is a non-negative weighting function. We will use the notation for the step size ϵ_n for differentiation with respect to the finite-dimensional parameter and the notation τ_n for the step size of differentiation with respect to the infinite-dimensional parameter. This leads us to the expression for the numerical first-order condition in the form

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} L_{1,p}^{\varepsilon_n} \hat{m}\left(\theta, \eta, z_i\right)' \, \widehat{W}\left(z_i\right) \, \hat{m}\left(\theta, \eta, z_i\right) = o_p(1), \\ &\frac{1}{n} \sum_{i=1}^{n} L_{1,p}^{\tau_n, \psi_j} \hat{m}\left(\theta, \eta, z_i\right)' \, \widehat{W}\left(z_i\right) \, \hat{m}\left(\theta, \eta, z_i\right) = o_p(1), \end{aligned} \tag{4.8}$$

for $j = 1, \ldots, G_n$.

In the following Theorem we provide the uniform convergence result. As we notice, under standard assumptions regarding the series expansion and the kernel estimator, there is an interference between the step size for numerical differentiation and the choice of the tuning parameter (number of terms in the expansion or the bandwidth).

THEOREM 14. *For some neighborhood U around (θ_0, η_0)*

$$\sup_{(\theta,\eta)\in U} \left| L_{1,p}^{\epsilon_n} \widehat{Q}\left(\theta, \eta\right) - \frac{\partial Q\left(\theta, \eta\right)}{\partial \theta} \right| = o_p(1), \qquad \sup_{(\theta,\eta)\in U} \left| L_{1,p}^{\tau_n, \psi_j} \widehat{Q}\left(\theta, \eta\right) - \frac{\partial Q\left(\theta, \eta\right)}{\partial \eta}[\psi_j] \right| = o_p(1),$$

for each ψ_j *for* $j = 1, \ldots, G_n$*, if* $\hat{m}(\cdot)$ *is estimated via a series estimator under Assumption 7 and 8, provided that* $\epsilon_n \to 0$ *and* $\frac{n^2 \epsilon_n}{N \log^2 n} \to \infty$*, and* $\frac{n^2 \tau_n}{N \log^2 n} \to \infty$*. Equivalently we obtain the same result for the kernel estimator for the moment function under assumptions 8 and 9, provided that* $\epsilon_n \to 0$*,* $b_n \to 0$*,* $\frac{n^2 \tau_n b_n}{\log^2 n} \to \infty$*, and* $\frac{n^2 \epsilon_n b_n}{\log^2 n} \to \infty$*.*

Consequently, for both the kernel-based and sieve-based estimators we can conclude that $\left(\hat{\theta}, \hat{\eta}\right) \xrightarrow{p} (\theta_0, \eta_0)$ *provided that system (4.8) is satisfied.*

We note that the result in Theorem 14 regarding the required rate of approach of the step size to zero is similar to that in the case of the extremum estimators based on U-statistic. The intuition behind this is that the numerical first-order condition for the generalized minimum distance estimator can be re-casted as a function of a second-order U-statitic. Hence, theorem 14 is analogous to the results for uniform consistency of derivatives of U-statistics and the extremum estimators in section 5.

4.3 Convergence rate

We will proceed with establishing the rate of convergence for the estimator of interest. For continuity of the moment function itself we rely on Assumption 1 which guarantees that the standard U-statistics projection results hold. We begin with the analysis of the rate of convergence for the finite-dimensional parameter. The next result will be a direct application of our results regarding the U-statistics.

LEMMA 7. *Suppose* $\left(\hat{\theta},\, \hat{\eta}\right) \xrightarrow{p} (\theta_0,\, \eta_0)$ *and system of equations (4.8) is satisfied with the right-hand side being* $o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)$ *for the derivative with respect to* θ *and* $o_p\left(\frac{1}{\sqrt{n\tau_n}}\right)$ *for each of the directional derivatives with respect to* η *. Under Assumptions 1, 5, and 8, for series estimator under Assumption 7 and for kernel estimator under Assumption 9 with* $\sqrt{n \min\left[\varepsilon_n,\, \tau_n\right]^{1+p}} = o\left(1\right)$*, and* $\frac{n\varepsilon_n}{N \log n} \to \infty$*,* $\frac{n\tau_n}{N \log n} \to \infty$ *for sieve case and* $\frac{nb_n\varepsilon_n}{\log n} \to \infty$*,* $\frac{nb_n\tau_n}{\log n} \to \infty$ *for the kernel case we have*

$$\sup_{d\left(\hat{\theta},\theta_0\right)=O\left(\frac{\log^2 n}{n^2 b_n \varepsilon_n}\right),\, d(\hat{\eta},\eta_0)=O\left(\frac{\log^2 n}{n^2 b_n \tau_n}\right)} \Bigg\| L_{1,p}^{\epsilon_n} \hat{Q}\left(\hat{\theta}, \hat{\eta}\right) - L_{1,p}^{\epsilon_n} \hat{Q}\left(\theta_0, \hat{\eta}\right) - L_{1,p}^{\epsilon_n} Q\left(\hat{\theta}, \hat{\eta}\right) + L_{1,p}^{\epsilon_n} Q\left(\theta_0, \hat{\eta}\right) \Bigg\| = o_p\left(\frac{1}{n}\right).$$

This Theorem can be interpreted as a corollary of Lemma 7 for U-statistics. The result regarding the rate of convergence for the finite dimensional parameter follows directly from Lemma 7 and it is completely analogous to Theorem 18 for U-statistics.

THEOREM 15. *Suppose* $\left(\hat{\theta},\, \hat{\eta}\right) \xrightarrow{p} (\theta_0,\, \eta_0)$ *and system of equations (4.8) is satisfied with the right-hand side being* $o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)$ *for the derivative with respect to* θ *and* $o_p\left(\frac{1}{\sqrt{n\tau_n}}\right)$ *for each of the directional derivatives with respect to* η *. Under Assumptions 1, 5, and 8, for series estimator under Assumption 7 and for kernel estimator under Assumption 9 with* $\sqrt{n \min\left[\varepsilon_n,\, \tau_n\right]^{1+p}} = o\left(1\right)$*, and*

$\frac{n\varepsilon_n}{N \log n} \to \infty$, $\frac{n\tau_n}{N \log n} \to \infty$ *for sieve case and* $\frac{nb_n\varepsilon_n}{\log n} \to \infty$, $\frac{nb_n\tau_n}{\log n} \to \infty$ *for the kernel case, then* $\sqrt{n}d\left(\hat{\theta}, \theta_0\right) = O_P^*(1)$.

Establishing the rate for the convergence of the infinite-dimensional parameter can be done through the use of Assumption 16 that states the completeness of the limiting space $\mathcal{H}_\infty$. We note that if the sieve space $\mathcal{H}_n$ is fixed as the sample size increases, then the system of equations (4.8) gives a consistent estimator of the projection of the infinite-dimensional parameter on G_n first component of the basis in $\mathcal{H}_n$. Also, given that if G_n is fixed, the projection coefficients are essentially finite-dimensional parameters. As a result, provided that

$$\hat{\eta}(y) = \sum_{j=1}^{G_n} \hat{a}_j \psi_j(y),$$

we can establish that $d\left(\hat{a}_j,\, \langle \eta_0, \psi_j \rangle\right) = O_p\left(\frac{1}{\sqrt{n}}\right)$. As a result, we can evaluate the $\mathbf{L}^2$ distance between $\hat{\eta}$ and η_0 as

$$\|\hat{\eta} - \eta_0\|_{\mathbf{L}^2} = \sqrt{\sum_{j=1}^{G_n} d\left(\hat{a}_j,\, \langle \eta_0, \psi_j \rangle\right) \|\psi_j\|} = O_p\left(\sqrt{\frac{G_n}{n}}\right).$$

We note that there will be no interference between the convergence of the numerical derivative and the convergence of the infinite-dimensional component if $n\tau_n/G_n \to \infty$.

5 Numerical derivatives of U-statistics

5.1 Consistency of numerical derivatives

Numerical derivatives can also be used for objective functions that are based on U-statistics, an example of which is the maximum rank correlation estimator of Sherman (1993). The model considered in this section is parametric. Second order U-statistics, which we focus on, are the most commonly used in applications. A U-statistic objective function is defined from an i.i.d. sample $\{Z_i\}_{i=1}^n$ by a symmetric function $g(Z_i, Z_j, \theta)$ as as

$$\hat{g}(\theta) = \frac{1}{n\,(n-1)} S_n\,(f) \quad \text{where} \quad S_n\,(f) = \sum_{i \neq j} f\left(Z_i, Z_j, \theta\right). \tag{5.9}$$

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 $g(\theta) = E_{zz} f\left(Z_i, Z_j, \theta\right)$. This population objective function satisfies Assumption 4.

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 $G\left(\theta_0\right) = \frac{\partial}{\partial \theta} g\left(\theta\right)$ by $L_{1,p}^{\varepsilon_n} \widehat{g}\left(\widehat{\theta}\right)$,

$$\hat{g}\left(\theta\right) = g\left(\theta\right) + \hat{\mu}_n(\theta) + \frac{1}{n\,(n-1)} S_n\left(u\right), \tag{5.10}$$

where

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

and

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

The condition for controlling the degenerate U-process will be weaker than that needed for the empirical process term because of its faster convergence rate. We maintain Assumption 5, with the new interpretation of functions $g\left(\cdot, \cdot, \theta\right)$. We also make the following additional assumptions.

ASSUMPTION 17. *The projections $\mu\left(z, \theta\right)$ are Lipschitz-continuous in θ uniformly over z.*

This assumption depends on the distribution of Z_i. For instance, when the kernel is defined by indicator functions, the expectation will be continuous in parameter for sufficiently smooth distribution of Z_i. It controls the impact of numerical differentiation on the projection term by the maximum inequality for Lipschitz-continuous functions:

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

for some $C > 0$.

ASSUMPTION 18. *For a neighborhood $N\left(\theta_0\right)$ around θ_0,*

$$\sup_{\theta \in N(\theta_0), z \in \mathcal{Z}} E\left|g(Z_i, z, \theta + \epsilon) - g(Z_i, z, \theta - \epsilon)\right|^2 = O(\epsilon).$$

This assumption allows us to establish an analog of Lemma 1 for the case of the U-processes, which is presented below.

LEMMA 8. *Suppose $\|F\| = \sup_{\theta \in N(\theta_0)} \left|g\left(Z_i, Z_j, \theta\right)\right| \ll C < \infty$. Under Assumptions 5 and 18, if $n^2 \varepsilon_n / \log^2 n \to \infty$,*

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

Consequently, assumption 4 implies that $\hat{G}_1\left(\hat{\theta}\right) = o_p\left(1\right)$ if $d\left(\hat{\theta}, \theta_0\right) = o_p\left(1\right)$, as defined in (2.1).

The consistency of the numerical derivatives of U-statistics follows directly from Lemma 8.

THEOREM 16. *Under assumptions, 4, 5 and the conditions of lemma 8, $L_{1,p}^{\epsilon_n} \hat{g}\left(\hat{\theta}\right) \xrightarrow{p} G\left(\theta_0\right)$ if $\epsilon_n \to 0$ and $n\epsilon_n^2 / \log^2 n \to \infty$, 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 the weakest possible condition on the step size of numerical differentiation when the envelope function of the differenced moment function does not necessarilydecrease with the shrinking step size. We note that the resulting condition for the step size is weaker in the case of the U-statistics versus 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 scale of the U-statistic.

5.2 Optimal rate for derivative estimation for U-statistics

We can provide tighter conditions on the step size of the numerical differentiation if the function of interest permits non-trivial envelopes on its finite differences. Hölder-continuous functions, for instance, fall in this category.

Consider the following assumption

ASSUMPTION 19. *Define* $\mathbb{U}(\theta) = \frac{n}{n(n-1)} \sum\limits_{1 \leq i \neq j \leq n} u(Z_i, Z_j, \theta)$. *There exists some* $\epsilon > 0$*, such that for all* δ *sufficiently small,*

$$E^* \sup_{d(\theta_1,\theta_2)<\delta, d(\theta_1,\theta_0)<\epsilon, d(\theta_2,\theta_0)<\epsilon} |\mathbb{U}(\theta_1) - \mathbb{U}(\theta_2)| \lesssim \phi_n(\delta),$$

for functions $\phi_n(\cdot)$ *such that* $\delta \mapsto \phi_n(\delta)/\delta^{\psi}$ *is decreasing for some* $\psi > 0$ *and* $\psi \leq 1$*. Furthermore, assumptions 6 and 13 hold for the projection function* $g(Z_i, \theta) = E_z f(z, Z_i, \theta)$.

We make this assumption by noting a similarity in the asymptotic results for U-processes to the results for empirical processes noted in Nolan and Pollard (1987), Nolan and Pollard (1988) and Arcones and Gine (1993). Assumption 19 maintains the spirit of assumption underlying Theorem 6 of Nolan and Pollard (1987). Nolan and Pollard (1987) and Nolan and Pollard (1988) consider the concept of random covering numbers. In fact, we can define the class of functions

$$\mathcal{F}_\delta^\epsilon = \{f(Z_i, Z_j, \theta_1) - f(Z_i, Z_j, \theta_2) \, : \, d(\theta_1, \theta_2) < \epsilon, \, d(\theta_1, \theta_0) < \delta, \, d(\theta_2, \theta_0) < \delta\}$$

with the envelope F_δ^ϵ. Consider the empirical measure

$$T_n f = \sum_{i \neq j, \, i \neq j - [n/2], \, i \neq j + [n/2]} f(\xi_i, \xi_j).$$

Theorem 6 of Nolan and Pollard (1987) states that

$$E \sup_{\mathcal{F}} \|S_n(f)\| \leq C \, E\left[\theta_n + \tau_n \, J(\theta_n/\tau_n, T_n, \mathcal{F}, F)\right],$$

where $\tau_n = \left(T_n F^2\right)^{1/2}$ and $\theta_n = \frac{1}{4} \sup\limits_{\mathcal{F}} \left(T_n f^2\right)^{1/2}$, and $J(\cdot)$ is a covering integral defined by the random measure T_n. Application of the Jensen's inequality and the Cauchy-Schwartz inequality allows one to evaluate the supremum of the empirical process as

$$E \sup_{\mathcal{F}} \sqrt{n} \|S_n(f)\| \leq C \left(E_{zz} F^2\right)^{1/2} \left[1 + \left(E \, J(1, T_n, \mathcal{F}, F)^2\right)^{1/2}\right].$$

According to Nolan and Pollard (1987) the covering integral $J(1, T_n, \mathcal{F}, F)$ can be bounded by a sum covering integrals of classes of one-dimensional functions $E_z\, f(\cdot, Z_i, \theta)$ which are finite for most used classes of functions. As a result, the behavior of the supremum will be mainly associated with the envelope function F.

We extend the assumption 13 to the case of the U-statistic part of the objective function. Note that the corresponding U-statistic is degenerate. Then we can apply the analog of the Central Limiting Theorem for U-processes in Arcones and Gine (1993). This allows us to formulate the following assumption

ASSUMPTION 20. *We can confine the asymptotic behavior of $\mathbb{U}$ for $n \to \infty$ and $\epsilon_n \to 0$, for each collection of $(t_1, \ldots, t_k)$ as*

$$\left[\frac{\mathbb{U}(\theta_0 + t_j\epsilon_n) - \mathbb{U}(\theta_0 - t_j\epsilon_n)}{\epsilon_n^{\psi}}, j = 1, \ldots, k\right] \xrightarrow{d} U_f,$$

where distribution U_f is a mixture of chi-squared random variables as in Serfling (1980) and indexed by the kernel function f.

This assumption implies that the standard deviation of the symmetrized and de-meaned U-statistic decreases at the rate n, as it should for a degenerate U-statistic by Serfling (1980).

The original objective function admits a decomposition

$$\hat{g}(\theta) - g(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mu(Z_i, \theta) + S_n(u).$$

Then noting that $E_z\, \mu(Z_i, \theta) = 0$ by construction, we can consider the empirical process $\mathbb{G}'(\theta) = \frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} \mu(Z_i, \theta)$. The following theorem establishes the general consistency result for the numerical derivatives of problems defined by U-processes.

THEOREM 17. *Under assumptions 4 and 19 as well as 6 and 13 satisfied for the function $(E_z\, f(z, Z_i, \theta) - E_{zz}\, f(Z_j, Z_i, \theta))$ with $\gamma \geq \psi$, if $\hat{\theta} - \theta_0 = O_p(n^{-\eta})$ for $\eta > 0$, then $L^{\epsilon}_{m,p}\hat{g}\left(\hat{\theta}\right) \xrightarrow{p} \frac{\partial^m g(\theta_0)}{\partial\theta_1^{m_1}\ldots\partial\theta_k^{m_k}}$, if $n\,\epsilon_n^{2-2\gamma} \to \infty$ with $\psi \geq 2\gamma - 1$ or $n\,\epsilon_n^{1-\psi} \to \infty$ with $\psi \leq 2\gamma - 1$.*

if γ or $\psi < 1$, provided that $\eta \geq \frac{1}{1-\gamma+2p}$ and $\sqrt{n}\epsilon^{\gamma-\psi} \to \infty$, the best rate of convergence between $L^{\epsilon}_{1,p}\hat{g}\left(\hat{\theta}\right)$ and G is achieved when $\epsilon_n = O\left(n^{-1/(1-\gamma+2p)}\right)$ in which case

$$||L^{\epsilon}_{1,p}\hat{g}\left(\hat{\theta}\right) - G||^2 = O\left(n^{-\frac{2p}{1-\gamma+2p}}\right).$$

In case where $\sqrt{n}\epsilon^{\gamma-\psi} \to 0$ provided that $\eta \geq \frac{1}{1-\psi+p}$ then the optimal rate for ϵ is $\epsilon_n = O\left(n^{-1/(1-\psi+p)}\right)$ with the mean-squared error

$$||L^{\epsilon}_{1,p}\hat{g}\left(\hat{\theta}\right) - G||^2 = O\left(n^{-\frac{2p}{1-\psi+p}}\right).$$

This theorem allows us to construct consistent estimates for the Jacobi matrix and the Hessian of in the problem defined by a U-statistic. Our conditions are more general than those provided in Sherman (1993) and extend to much richer classes of functions. However, for instance, for the case of the U-statistic defined by an indicator function in the maximum rank correlation estimator one needs to use a more general Theorem 16 which needs a srtonger condition for the step size. The reason for this is that the finite differences between the indicator functions have trivial (constant) envelopes. As opposed to Sherman (1993), it still provides a weaker restriction on the step size with $n^2\epsilon_n/\log^2 n \to 0$ as opposed to $n\sqrt{\epsilon_n} \to 0$.

5.3 Numerical gradient-based estimation with U-statistics

We consider the solution of the empirical first-order condition $\hat{\theta}$ defined by $||L_{1,p}^{\varepsilon_n}\hat{Q}_n\left(\hat{\theta}\right)|| = o_p\left(\frac{1}{\sqrt{n}}\right)$. In some cases when the objective function is not continuous, the value that sets the first-order condition to zero might not exist, so we propose to choose the point that will set the first-order condition very close to zero. In this section we will only consider the distribution results regarding the first numerical derivative.

The structure of the consistency argument for the U-statistic defined estimation problem is similar to that for the standard sample means. In particular, when the kernel function is "sufficiently" smooth, the behavior of the objective function will be dominated by the empirical process component. In that case the analysis from our previous discussion will be valid for the objective function defined by the sample mean of $E_z\, f\left(z, Z_i, \theta\right)$. We maintain Assumption 10 applied to the map $D\left(\theta\right) = \frac{\partial}{\partial\theta}E_{zz}\left[f\left(Z_i, Z_j, \theta\right)\right]$. We also keep Assumption 11 following Arcones and Gine (1993) where the authors state that this assumption, along with the finiteness of the absolute moment of the U-statistic, constitute a sufficient measurability requirement. A deeper discussion of applicability of these conditions can be found in Section 10 of Dudley (1999).

5.4 U-statistics with kernels with absolutely bounded finite differences

We maintain Assumption 5, 17 and 18 for the class of kernels of the U-statistic.

Lemma 8 establishes the consistency result for this objective function which implies that under $n^2\varepsilon_n/\log^2 n \to \infty$,

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

Moreover, we can apply Lemma 10 in Nolan and Pollard (1987). This lemma 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 note that we can provide condition

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

We next repeat the steps that we followed to determine the rate of convergence of the estimators given by sample means.

LEMMA 9. *Suppose* $\hat{\theta} \stackrel{p}{\longrightarrow} \theta_0$ *and* $L_{1,p}^{\varepsilon} \hat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)$*. Under Assumptions 5 (i) and (ii), 17 and 18, if* $n\sqrt{\varepsilon_n}/\log n \to \infty$*, and* $\sqrt{n\varepsilon^{1+p}} = O(1)$*, then* $\frac{n^2\varepsilon_n}{\log^2 n} d\left(\hat{\theta}, \theta_0\right) = o_{P^*}(1)$*.*

In the next step we consider 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, we can improve upon the rate of the objective function using the envelope property.

LEMMA 10. *Under conditions of Lemma 9 with* $\sqrt{n\varepsilon_n^{1+p}} = o(1)$*, and* $\frac{n\varepsilon_n}{\log n} \to \infty$ *we have*

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

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

THEOREM 18. *Suppose* $\hat{\theta} \stackrel{p}{\longrightarrow} \theta_0$ *and* $L_{1,p}^{\varepsilon} \hat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}}\right)$*. Under Assumptions of Lemma 9, if* $n\varepsilon_n/\log n \to \infty$*, and* $\sqrt{n\varepsilon^{1+p}} = O(1)$*, then* $\sqrt{n} d\left(\hat{\theta}, \theta_0\right) = O_P^*(1)$*.*

Proof. We note that by Lemma 10 in the small neighborhoods of the true parameter the U-statistic part has a 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 gives the corresponding parametric convergence. □

5.5 Functions with polynomial envelopes on the finite differences

We keep Assumption 4 that applies to the expectation of the objective function and implies a sufficient degree of smoothness. We also keep Assumption 6 applied to the one-dimensional projection of the kernel of the U-statistic establishing the polynomial bound on its modulus of continuity. For the U-process part we need to make a uniform assumption regarding its modulus of continuity as well. Define $\mathbb{U}(\theta) = \frac{n}{n(n-1)} \sum_{1 \le i \ne j \le n} u(Z_i, Z_j, \theta)$.

ASSUMPTION 21. *The function* $E_z f(z, Z_i, \theta)$ *satisfies Assumption 6. Moreover, for the symmetrized kernel* $u(\cdot,\cdot,\theta)$ *and all sufficiently small* δ*,*

$$E^* \sup_{d(\theta_1,\theta_2)<\delta, \theta_1\in\Theta, \theta_2\in\Theta} |\mathbb{U}(\theta_1) - \mathbb{U}(\theta_2)| \lesssim \phi_n(\delta),$$

for functions $\phi_n(\cdot)$ *such that* $\delta \mapsto \phi_n(\delta)/\delta^{\psi}$ *is non-increasing for some* $\psi \le 1$*.*

We make this assumption by noting a similarity in the asymptotic results for U-processes to the results for empirical processes noted in Nolan and Pollard(1987,1988) and Arcones and Gine (1993). We can consider the class of functions

$$\mathcal{F}_{\delta}^{\varepsilon}=\left\{f\left(Z_i,Z_j,\theta_1\right)-f\left(Z_i,Z_j,\theta_2\right)\;:\;d\left(\theta_1,\theta_2\right)<\varepsilon,\,d\left(\theta_1,\theta_0\right)<\delta,\,d\left(\theta_2,\theta_0\right)<\delta\right\}.$$

For some moment functions $f(\cdot)$ the minimal envelope function for this class is a constant. In this case we can apply the result of Lemma 10 in Nolan and Pollard (1987) to establish the bound in Assumption 21. We take $\varepsilon_n \to 0$. Also take $f(\cdot)$ be one-dimensional indicator functions. From the previous observations we establish that

$$\sup_{\mathcal{F}_{\delta}^{\varepsilon}}\sup_{z} E\left(f\left(z,Z_j,\theta_1\right)-f\left(z,Z_j,\theta_2\right)\right)^{1/2}=O\left(\min\{\delta,\,\varepsilon\}^{1/2}\right).$$

For $\varepsilon_n \to 0$ the bound will be determined by ε_n. Then we find $t_n = O\left(\max\{\varepsilon_n,\,\frac{\log\,n}{n}\}\right)$. From Lemma 10 in Nolan and Pollard (1987) we find that almost surely $\sup\limits_{\mathcal{F}} S_n(f) < \beta^2 n^2 t_n^2$. We consider a numerical derivative of the normalized objective function. This allows us to re-write the almost sure inequality

$$\sup_{\mathcal{F}}\|L_{1,p}^{\epsilon_n}\hat{g}(\theta)\| < C\,\varepsilon_n \max\left\{1,\,\frac{\log^2\,n}{n^2\varepsilon_n^2}\right\}.$$

As a result, provided that $\frac{n\varepsilon_n}{\log\,n}\to\infty$, the bound on the U-process is defined by ε_n when the kernel of the U-process is a one-dimensional indicator.

Next we impose an assumption on the kernel function (that allows us to construct uniform bounds on the kernel function and its marginal expectations) and state the finite variance condition. We need this assumption to validate the use of exponential inequalities that will play the central role in the consistency argument.

ASSUMPTION 22. *Assume that $E_z\, f\left(z,Z_i,\theta\right)$ satisfies 6 with parameter ψ. Moreover, the kernel function itself admits a power envelope such that*

$$\sup_{\theta\in\Theta,,\,(z,z')\in\mathcal{Z}\times\mathcal{Z}}||f\left(z,z',\theta+\varepsilon\right)-f\left(z,z',\theta-\varepsilon\right)||<C\varepsilon^{2\psi-1},\quad \textit{for } \varepsilon \textit{ small enough.}$$

We assume that the population objective function $E_{zz}\, f\left(Z_i,Z_j,\theta\right)$ satisfies 4 uniformly in Θ. Finally for some sufficiently small $\varepsilon>0$

$$\sup_{\theta\in\Theta} E\left(\frac{f\left(Z_i,Z_j,\theta+\varepsilon\right)-f\left(Z_i,Z_j,\theta-\varepsilon\right)}{\varepsilon^{\psi}}\right)^2<\infty.$$

This assumption will only be relevant for substantially non-smooth kernels. If the kernel is smooth, then we can use a weaker assumption which would compare the modulus of continuity of the degenerate U-process defined by the symmetrized and de-meaned kernel and the modulus of continuity of the empirical process defined by the marginal expectation of the kernel with respect to the argument.

Assumptions 22 remain relatively weak and is satisfied for most "conventional" classes of functions such as Hölder-continuous functions and indicator functions. Moreover, for models with uniformly finite moments of the objective function, the empirical process corresponding to the projection of the U-statistic dominates the U-process which leads to a high degree of smoothness in the behavior of the original U-statistic.

These assumptions allow us to characterize the local behavior of the empirical process of the conditional mean separately from the U-process corresponding to the residual U-statistic. For a fixed δ, Assumptions 21 and 6 together imply that $E^* \sup_{d(\theta_1,\theta_2)<\delta,\theta_1\in\Theta,\theta_2\in\Theta} \sqrt{n}\|\widehat{Q}(\theta_1) - \widehat{Q}(\theta_1)\| \sim \delta^\gamma + \frac{\delta^\psi}{\sqrt{n}}$.

THEOREM 19. *Under assumptions 10, 11, 22 and 21, as long as* $\varepsilon \to 0$, $\frac{n\varepsilon^{2-2\gamma}}{\log n} \to \infty$ *and* $n\varepsilon^{2\gamma-2\psi}\log n \to \infty$, *and* $\sqrt{n}\varepsilon^{\gamma-2\psi} \to \infty$, *and* $n^{1-6\lambda}\varepsilon^{10\gamma-8\psi-2} \to 0$,

$$\sup_{\theta\in\Theta} ||L^{\varepsilon}_{1,p}\hat{Q}\left(\theta\right) - G\left(\theta\right)|| = o_p\left(1\right).$$

Consequently, $\hat{\theta} \xrightarrow{p} \theta_0$ *if* $||L^{\varepsilon}_{1,p}\hat{Q}\left(\hat{\theta}\right)|| = o_p\left(1\right)$ *and if* $G\left(\theta\right) = 0$ *uniquely at* $\theta = \theta_0$.

5.6 Rate of Convergence and Asymptotic Distribution

Next, we establish the rate of convergence and the asymptotic distribution of the numerical derivative based M-estimator. We provide the following general result.

THEOREM 20. *Suppose* $\hat{\theta} \xrightarrow{p} \theta_0$ *and* $L^{\varepsilon}_{1,p}\hat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n}\varepsilon^{1-\gamma}}\right)$. *Under assumption 6, if*

$$n\varepsilon^{2-2\gamma}/\log n \to \infty$$

and $\sqrt{n}\varepsilon^{-\gamma+2p} = O\left(1\right)$ *and* $\frac{1}{\sqrt{n}}\varepsilon_n^{\psi-\gamma} = O\left(1\right)$, *then* $\sqrt{n}\varepsilon^{1-\gamma}d\left(\hat{\theta},\theta_0\right) = O_P^*\left(1\right)$.

If $\sqrt{n}\varepsilon_n^{\gamma-\psi} \to 0$, *then* $n\varepsilon^{1-\psi}d\left(\hat{\theta},\theta_0\right) = O_P^*\left(1\right)$.

Now we can proceed with the analysis of the asymptotic distribution of the estimator that we obtain by equating the numerical derivative to zero. This is ensured by imposing an additional assumption requiring the existence of two continuous derivatives of the population objective function.

THEOREM 21. *Assume that the conditions of theorem 10 hold but with* $\sqrt{n}\varepsilon^{p-\gamma} = o\left(1\right)$, *in addition, suppose assumption 13 holds, and the continuous Hessian matrix of* $Q\left(\theta\right)$, *denoted* $H\left(\theta\right)$, *is nonsingular and finite at* θ_0. *If* $\sqrt{n}\varepsilon_n^{\gamma-\psi} \to \infty$ *and* $\sqrt{n}\varepsilon^{2-\gamma} \to \infty$, *then*

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

If $\sqrt{n}\varepsilon_n^{\gamma-\psi} \to 0$ *with* $\psi < 1$, *and* $n\varepsilon^{2-\psi} \to \infty$, *then*

$$n\varepsilon^{1-\psi}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} H\left(\theta_0\right)^{-1}U_f,$$

where U_f *is a mixture of chi-squared random variables.*

6 Applications

6.1 Irregular models with continuous moment functions

In the following example the objective function has a well-defined minimum, but the population objective does not have a finite Hessian. Consider an extremum estimation problem where the objective function in the population is given by

$$Q(\theta) = E\left[|x_i - \theta|^{\alpha}\right],$$

where $\alpha < 1/2$. The irregularity of this problem is easy to see if one considers x_i distributed normally. Then, as random variable $\alpha|x_i - \theta|^{\alpha-1}$ does not have a finite variance, estimation of θ with a parametric convergence rate will not be possible. Consider a sample problem of minimizing

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \theta|^{\alpha}.$$

Assume that the distribution of x_i is symmetric and it has an absolutely continuous density function. We use a numerical derivative to form a sample moment condition. Then, by linearity of the numerical differentiation operator

$$L_{1,p}^{\varepsilon_n} Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} L_{1,p}^{\varepsilon_n} |x_i - \theta|^{\alpha}.$$

Consider for simplicity the numerical differentiation operator of order 2. Then, if $\|x_i - \theta\| > \varepsilon_n$, then using the Taylor expansion for small θ and ε_n:

$$L_{1,p}^{\varepsilon_n} |x_i - \theta|^{\alpha} = \frac{\operatorname{sign}(x_i - \theta)}{2\varepsilon_n}\left[|x_i|^{\alpha} + \alpha|x_i|^{\alpha-1}(\theta + \varepsilon_n) - |x_i|^{\alpha} - \alpha|x_i|^{\alpha-1}(\theta - \varepsilon_n) + o(\theta + \varepsilon_n)\right].$$

For $|x_i - \theta| < \varepsilon_n$

$$L_{1,p}^{\varepsilon_n} |x_i - \theta|^{\alpha} = \frac{\operatorname{sign}(x_i - \theta)}{2\varepsilon_n}\left[|x_i|^{\alpha} + \alpha|x_i|^{\alpha-1}(\theta + \varepsilon_n) - |x_i|^{\alpha} - \alpha|x_i|^{\alpha-1}(-\theta + \varepsilon_n) + o(\theta + \varepsilon_n)\right].$$

Combining the two cases we can write:

$$L_{1,p}^{\varepsilon_n} |x_i - \theta|^{\alpha} = \operatorname{sign}(x_i)\,\alpha|x_i|^{\alpha-1}\left[\mathbf{1}\{|x_i| > \varepsilon_n\} + 2\theta\, U\left(\frac{x_i - \theta}{\varepsilon_n}\right)\right] + o(1),$$

where $U(\cdot)$ is a uniform kernel. Given that by assumption the distribution of the covariate is symmetric, the solution for the population objective $\theta_0 = 0$. Expansion at θ_0 for the sample objective gives

$$L_{1,p}^{\varepsilon_n} Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \alpha|x_i|^{\alpha-1}\operatorname{sign}(x_i)\,\mathbf{1}\{|x_i| > \varepsilon_n\} + 2\widehat{\theta}\frac{1}{\varepsilon_n n}\sum_{i=1}^{n} \alpha|x_i|^{\alpha-1}\operatorname{sign}(x_i)\, U\left(\frac{x_i - \theta_0}{\varepsilon_n}\right) + o_p(r_n) = 0,$$

with $r_n = \max\{\theta, \varepsilon_n\}/n$. Then we can express the parameter of interest

$$\widehat{\theta} = \frac{\frac{1}{n}\sum_{i=1}^{n} |x_i|^{\alpha-1}\text{sign}\,(x_i)\,\mathbf{1}\,\{|x_i| > \varepsilon_n\}}{\frac{1}{\varepsilon_n n}\sum_{i=1}^{n} |x_i|^{\alpha-1}\text{sign}\,(x_i)\,U\left(\frac{x_i}{\varepsilon_n}\right)} + o_p\left(r_n\right).$$

Note that for each ε_n the variable $|x_i|^{\alpha-1}\mathbf{1}\,\{|x_i| > \varepsilon_n\}$ has a finite second moment. Therefore, the step size for numerical derivative plays the role of the trimming sequence. This estimator will have a slow convergence rate given by the convergence rate for the numerator for a non-zero denominator. To choose an appropriate step for numerical differentiation, note that we need to ensure that the denominator does not vanish. Then

$$E\left[\left\|\frac{1}{\varepsilon_n}|x_i|^{\alpha-1}\text{sign}\,(x_i)\,U\left(\frac{x_i}{\varepsilon_n}\right)\right\|\right] = \int \frac{1}{\varepsilon_n}\varepsilon_n^{\alpha-1}|u|^{\alpha-1}U\left(u\right)f_X\left(u\varepsilon_n\right)\,du = \frac{2\varepsilon_n^{\alpha-2}}{\alpha}f_X\left(0\right) + o\left(\varepsilon_n^{\alpha-2}\right).$$

Next, we need to balance the variance of the constructed estimator. Suppose that δ_n is its stabilizing factor. Then the order of the variance of $\delta_n\widehat{\theta}$ can be evaluated as

$$nE\left|\delta_n\widehat{\theta}\right|^2 \approx \frac{\alpha\delta_n^2}{2\varepsilon_n^{\alpha-2}f_X(0)}E\left[|x_i|^{2\alpha-2}\mathbf{1}\,\{|x_i| > \varepsilon_n\}\right] \approx \frac{\alpha\delta_n^2\varepsilon_n^{\alpha}}{2f_X(0)} = O(1).$$

Therefore, it will suffice to choose $\delta_n \sim \varepsilon_n^{-\alpha/2}$ for the finiteness of the variance. In the latter expression we used that $|x|^{\alpha-1}$ increases towards infinity at $x = 0$, therefore

$$E\left[|x_i|^{2\alpha-2}\mathbf{1}\,\{|x_i| > \varepsilon_n\}\right] \leq \varepsilon_n^{2\alpha-2}.$$

Lastly, we need to verify the Lindeberg condition. We can provide the order of the envelope for $\widehat{\theta}$ which is

$$\delta_n\frac{\alpha\varepsilon_n^{\alpha-1}}{2\varepsilon_n^{\alpha-2}f_X(0)} = O\left(\delta_n\varepsilon_n\right).$$

Then the stabilizing sequence has to be chosen such that

$$\delta_n\varepsilon_n\mathbf{1}\left\{\delta_n\varepsilon_n > \tau\sqrt{n}\right\} \to 0,$$

for all $\tau > 0$. As a result, we have two conditions for the step size of numerical differentiation. First, from the Lindeberg condition and the finiteness of the variance: $\varepsilon_n^{1-\alpha/2} \to 0$. The other condition comes from the need to balance bias. Given that the higher-order term in the Taylor expansion is $O\left(\varepsilon_n^2\right)$, then we need to ensure that $\sqrt{n}\varepsilon_n^{\alpha/2+2} \to 0$. Therefore, we can choose $\varepsilon_n = o\left(n^{-\frac{1}{4+\alpha}}\right)$. The convergence rate for $\widehat{\theta}$ is $\varepsilon_n^{\alpha/2}\sqrt{n}$.

6.2 Models with discontinuous moment functions

Our Theorem 7 is valid in the settings where the sample moment function is discontinuous. One vivid example of consistency of the estimator defined by the numerical derivative is its application to the maximum score estimator of Manski (1975) and Manski (1985). This estimator was re-considered in Horowitz (1992), where it was shown that substitution of the indicator function by kernel smoothers leads to a faster converging estimator once the rate at which the bandwidth parameter is approaching zero is chosen correctly. Here we illustrate that one can use the numerical first-order condition for the original maximum score objective of Manski and for an appropriately chosen step size, such an estimator will have properties similar to those in Horowitz (1992).

Consider the population objective function of Manski's maximum score estimator

$$Q(\theta) = E\left[\left(y - \frac{1}{2}\right)\mathbf{1}\left\{x'\theta > 0\right\}\right]$$

with the sample analog

$$\widehat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\mathbf{1}\left\{x_i'\theta > 0\right\}.$$

The sample objective is non-smooth in θ which may complicate the search for the maximum of the objective with respect to θ. We can apply the numerical gradient-based approach to construct a more manageable estimation technique. We consider the situation where the moment function

$$g(y, x, \theta) = \left(y - \frac{1}{2}\right)\mathbf{1}\left\{x'\theta > 0\right\}$$

has p continuous mean-square derivatives.

Horowitz (1992) imposes the normalization where the coefficient of the regressor that has a continuous density is normalized to 1. In our case if there is only one regressor this implies $x'\theta = x + \theta$. We consider the use of the first numerical derivative operator of order 2. Then from its linearity, it follows that application of this operator leads to the system of the first-order conditions

$$L_{1,2}^{\varepsilon_n}\widehat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\frac{1}{\varepsilon_n}U\left(\frac{x_i + \theta}{\varepsilon_n}\right) = 0,$$

where $U(\cdot)$ is a uniform kernel. Consider the use of the smoothed maximum score procedure applied in the same case. Then the indicator is substituted by the cumulative kernel. For instance, one can use cumulative uniform kernel:

$$K(z) = \mathbf{1}\{z \in [-1, 1]\}z + \mathbf{1}\{z > 1\}.$$

For the bandwidth parameter h_n we can write the objective function as

$$\widehat{Q}_n^s(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)K\left(\frac{x_i + \theta}{h_n}\right).$$

The corresponding first-order condition for the uniform kernel is

$$\frac{\partial}{\partial\theta}\widehat{Q}_n^s(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\frac{1}{h_n}U\left(\frac{x_i + \theta}{h_n}\right) = 0.$$

One can see that the equations corresponding to the numerical gradient and the smoothed maximum score are identical if the step size for the numerical differentiation $\varepsilon_n = h_n$. As a result, we can apply the result of Horowitz (1992) which implies that the estimator $\widehat{\theta}$ solving the first-order condition with the numerical gradient converges at the rate $\sqrt{\varepsilon_n\, n}$ to a normal distribution.

In a more general case, there is a vector of non-constant regressors $\left(x^1, x^2\right)$, where x^1 is a scalar regressor with continuous density, and the single index has the form $x^1 + \theta_1 + x^{2\prime}\theta_2$. Then, in addition to the numerical derivative with respect to θ_1 which will have the same form as before. The derivative with respect to components of θ_2 will lead to

$$L_{1,2}^{\varepsilon_n,k}\widehat{Q}_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\frac{x_{ik}^2}{x_{ik}^2\varepsilon_n}U\left(\frac{x^1 + \theta_1 + x^{2\prime}\theta_2}{x_{ik}^2\varepsilon_n}\right) = 0,$$

where x_k^2 is the k-th component of x^2. Denote $h_{ik}^2 = x_{ik}^2\varepsilon_n$. We can treat h_k^2 as stochastic bandwidth sequence. Then

$$L_{1,2}^{\varepsilon_n,k}\widehat{Q}_n\left(\theta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\frac{x_{ik}^2}{h_{ik}^2}U\left(\frac{x^1 + \theta_1 + x^{2\prime}\theta_2}{h_{ik}^2}\right) = 0.$$

Given that $\varepsilon_n \to 0$ at an appropriate rate, under regularity conditions in Horowitz (1992) we can guarantee that $h_{ik}^2 \to 0$ a.s. The condition above will be numerically equivalent to the condition in Horowitz (1992) if one substitutes the fixed bandwidth sequence by a stochastic sequence in our case. Let $h_k^2 = E\left[h_{ik}^2\right]$. Noting that

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\frac{x_{ik}^2}{h_{ik}^2}U\left(\frac{x^1 + \theta_1 + x^{2\prime}\theta_2}{h_{ik}^2}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{2}\right)\frac{x_{ik}^2}{h_k^2}U\left(\frac{x^1 + \theta_1 + x^{2\prime}\theta_2}{h_k^2}\right) + \Delta_n,$$

where $\Delta_n \to 0$ a.s., we obtain equivalence of the smoothed maximum score objective and the numerical first-order condition.

7 The choice of the magnitude of the step size

Our asymptotic results are concerned with the optimal choice of the rate of the step size for numerical differentiation. An important practical question will be the choice of the magnitude of the step size for a particular data sample. In the non-parametric estimation literature there are approaches to the choice of the bandwidth for kernel smoothing. Survey of work 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.

To a large extent, we can obtain the results for the optimal choice of constants simpler than in the case of non-parametric estimation because we will not be interested in the "uniform" step size. Previously we considered the decomposition: $L_{1,p}^{\varepsilon_n}\hat{g}\left(\hat{\theta}\right) - G = \hat{G}_1 + \hat{G}_2 + G_3 + G_4$, where

$$G_1 = \left[L_{1,p}^{\varepsilon_n}\hat{g}\left(\hat{\theta}\right) - L_{1,p}^{\varepsilon_n}g\left(\hat{\theta}\right)\right] - \left[L_{1,p}^{\varepsilon_n}\hat{g}\left(\theta_0\right) - L_{1,p}^{\varepsilon_n}g\left(\theta_0\right)\right]$$

and

$$G_2 = L_{1,p}^{\varepsilon_n} \hat{g}(\theta_0) - L_{1,p}^{\varepsilon_n} g(\theta_0)$$

and

$$G_3 = L_{1,p}^{\varepsilon_n} g\left(\hat{\theta}\right) - L_{1,p}^{\varepsilon_n} g(\theta_0), \qquad G_4 = L_{1,p}^{\epsilon_n} g(\theta_0) - G.$$

We proved that $L_{1,p}^{\varepsilon_n} \hat{g}\left(\hat{\theta}\right) - G = O_p\left(\hat{G}_2 + G_4\right)$. We can now consider the problem of the optimal constant choice. We consider the mean-squared error as the criterion for the choice of the step size, i.e. the function of interest is

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

which we approximate by the leading terms G_2 and G_4. We note that

$$L_{1,p}^{\epsilon_n} g(\theta) = \frac{1}{\varepsilon} \sum_{k=1}^{p} a_k g(\theta + t_k \varepsilon).$$

Assuming that function $g(\cdot)$ has at least $p+1$ derivatives, we can evaluate the result of application of the numerical derivative as

$$L_{1,p}^{\epsilon_n} g(\theta) = g'(\theta) + \varepsilon_n^p g^{(p+1)}(\theta) \sum_{k=1}^{p} \frac{a_k\, t_k^p}{(p+1)!} + o(\varepsilon_n^p).$$

Thus $G_4 = \varepsilon_n^p g^{(p+1)}(\theta) \sum\limits_{k=1}^{p} \frac{a_k\, t_k^p}{(p+1)!} + o(\varepsilon_n^p)$. We can evaluate the variance of G_2 as

$$\text{Var}(G_2) = \frac{1}{n} E\left[\frac{1}{\varepsilon_n} \sum_{k=1}^{p} a_k \left(g(\theta + t_k \varepsilon_n,\, Z_i) - g(\theta + t_k \varepsilon_n)\right)\right]^2$$

$$= \varepsilon_n^{-(2-2\gamma)} n^{-1} \left[\sum_{k=1}^{p} a_k^2 \text{Var}\left(\varepsilon_n^{-\gamma} g(\theta + t_k \varepsilon_n,\, Z_i)\right) + \sum_{k,m=1}^{p} a_k a_m \text{Cov}\left(\varepsilon_n^{-\gamma} g(\theta + t_k \varepsilon_n,\, Z_i), \varepsilon_n^{-\gamma} g(\theta + t_m \varepsilon_n,\, Z_i)\right)\right]$$

$$= \varepsilon_n^{-(2-2\gamma)} n^{-1} V_g(\varepsilon_n).$$

In case where $\gamma = 1$, the variance of G_2 will not affect the estimation. However, there will still be the numerical error corresponding to the operating precision of computer operations. This error is known and fixed. We denote it $[\delta\, g]$. Then the total error can be evaluate as

$$\text{MSE}_1(\varepsilon) \approx \varepsilon_n^{2p} \left(g^{(p+1)}(\theta) \sum_{k=1}^{p} \frac{a_k\, t_k^p}{(p+1)!}\right)^2 + \frac{[\delta\, g]}{\varepsilon_n} \sum_{k=1}^{p} a_k.$$

In case where $\gamma < 1$ the numerical error will be exceeded by the sampling error. As a result, we can compute

$$\text{MSE}_{<1}(\varepsilon) \approx \varepsilon_n^{2p} \left(g^{(p+1)}(\theta) \sum_{k=1}^{p} \frac{a_k\, t_k^p}{(p+1)!}\right)^2 + \varepsilon_n^{-(2-2\gamma)} n^{-1} V_g(\varepsilon_n).$$

Then we can choose $\varepsilon_n = \frac{C}{n^r}$, where r is the optimal rate for ε_n if $\gamma < 1$ and $r = 0$ otherwise. The problem is to choose C. In most applications, however, the derivative $g^{(p+1)}$ is unknonw. One simple way of choosing C is the analog of biased cross-validation. We can choose a simple first-order formula for $g^{(p+1)}$ and pick a preliminary (over-smootheed) step size $\varepsilon_n^{**} = \frac{(p+1)\mathrm{Var}(\theta, Z_i)}{n^{1/2(p+1)}}$ 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. Then for $\gamma = 1$ we find that

$$C^{**} = \left(\frac{p!\ (p+1)!\ [\delta\, g] \sum\limits_{k=1}^{p} a_k}{\left(\widehat{g^{(p+1)}}\left(\theta\right) \sum\limits_{k=1}^{p} a_k\, t_k^p\right)^2} \right)^{1/(2p+1)}.$$

For $\gamma < 1$ we find

$$C^{**} = \left(\frac{(2-2\gamma)p!\,(p+1)!\,V_g\left(\varepsilon_n^{**}\right)}{\left(\widehat{g^{(p+1)}}\left(\theta\right) \sum\limits_{k=1}^{p} a_k\, t_k^p\right)^2} \right)^{1/(2p+2-2\gamma)}.$$

Note that if the function $g(\cdot)$ is intensive to compute, the choice of these constants allows one to use a relatively small subsample to calibrate the step sizes. Then one can use these constants to initialize the step sizes on a large scale using the entire sample.

In case where one can compute the function in a relatively straigntforward way, calibration of the constants of interest can be performed by minimizing the approximate expression for the mean-squared error with respect to C, taking into account that the step size will enter both in the expression for the derivative $g^{(p+1)}$ and $V_g\left(\varepsilon_n\right)$. This approach is equivalent to the solve-the-equation plug-in approach in the bandwidth selection literature.

8 Monte-carlo evidence

9 Conclusion

In this paper we study the impact that the use of numerical finite-point approximation can make on the estimate of the gradient of a function. We focus on the case where the function is computed from a cross-sectional data sample. We find weak sufficient conditions that allow us to provide uniformly consistent estimates for the gradients of functions and the directional derivatives of semiparametric moments. Such results may be used to compute the Hessians of sample function to estimate the asymptotic variance or they can be used as inputs in efficient two-step estimation procedures. We further investigate the role of finite-point approximation in computation of extremum estimators. Finite-point approximation formulas use tuning parameters such as step size and in this paper we find that the presence of such parameters may affect statistical properties of the original extremum estimation. We study extremum estimation for classical M-estimators, U-statistics, and semiparametric generalized minimum distance estimators where the optimization routine uses a finite-difference approximation to the numerical gradient. We find that the properties of the estimator obtained from the numerical optimization routine depends on the interference between the smoothness of the population objective function, the precision of approximation, and the smoothness of the sample objective function. While in smooth models the choice of the sequence of step sizes may not affect the properties of the estimator, in non-smooth models the presence of numerical optimization routine can alter both the rate of convergence of the estimator and its asymptotic distribution.

A Proofs

A.1 Proof of Theorem 1

Proof. We need to verify that uniformly in $z \in \mathcal{Z}$, $(\theta, \eta(\cdot)) \in \Theta \times \mathcal{H}$ and $w_j \in \mathcal{H}_n$ the numerical derivative will be converging to the population derivative at $(\theta_0, \eta_0(\cdot))$. We begin with noting that

$$\begin{aligned}\hat{m}\left(z;\, \hat{\theta}+e_j\varepsilon_n, \hat{\eta}(\cdot)\right) &= \hat{m}\left(z;\, \hat{\theta}+e_j\varepsilon_n, \hat{\eta}(\cdot)\right) - m\left(z;\, \hat{\theta}+e_j\varepsilon_n, \hat{\eta}(\cdot)\right)\\ &-\hat{m}\left(z;\, \theta_0, \eta_0(\cdot)\right) + \hat{m}\left(z;\, \theta_0, \eta_0(\cdot)\right) - m\left(z;\, \theta_0, \eta_0(\cdot)\right)\\ &+m\left(z;\, \hat{\theta}+e_j\varepsilon_n, \hat{\eta}(\cdot)\right) - m\left(z;\, \theta_0+e_j\varepsilon_n, \hat{\eta}(\cdot)\right)\\ &+m\left(z;\, \theta_0+e_j\varepsilon_n, \hat{\eta}(\cdot)\right) - m\left(z;\, \theta_0+e_j\varepsilon_n, \eta_0(\cdot)\right)\\ &+m\left(z;\, \theta_0+e_j\varepsilon_n, \eta_0(\cdot)\right) - m\left(z;\, \theta_0, \eta_0(\cdot)\right)\end{aligned}$$

Using the expansion representation above, we conclude that

$$\begin{aligned}&\hat{m}\left(z;\, \hat{\theta}+e_j\varepsilon_n, \hat{\eta}(\cdot)\right) \stackrel{\mathbf{L}^2}{=} O_p\left(n^{-1/k}\right) + \Delta_{1\theta}\left(\hat{\theta}-\theta_0\right) + \Delta_{1\eta}[\hat{\eta}-\eta_0]\\ &+\left(\hat{\theta}-\theta_0\right)'\Delta_{2\theta^2}\left(\hat{\theta}-\theta_0\right) + \Delta_{2\eta^2}[\hat{\eta}-\eta_0]^2 + \Delta_{2\theta\eta}[\hat{\eta}-\eta_0]\left(\hat{\theta}-\theta_0\right)\\ &+\Delta^j_{1\theta}\varepsilon_n + \Delta^{jj}_{2\theta^2}\varepsilon_n^2 + o_p\left(\|\hat{\eta}-\eta_0\|^2_{\mathbf{L}^2}\right) + o_p\left(\|\hat{\theta}-\theta_0\|^2\right).\end{aligned}$$

Using a similar technique we can represent

$$\begin{aligned}&\hat{m}\left(z;\, \hat{\theta}-e_j\varepsilon_n, \hat{\eta}(\cdot)\right) \stackrel{\mathbf{L}^2}{=} O_p\left(n^{-1/k}\right) + \Delta_{1\theta}\left(\hat{\theta}-\theta_0\right) + \Delta_{1\eta}[\hat{\eta}-\eta_0]\\ &+\left(\hat{\theta}-\theta_0\right)'\Delta_{2\theta^2}\left(\hat{\theta}-\theta_0\right) + \Delta_{2\eta^2}[\hat{\eta}-\eta_0]^2 + \Delta_{2\theta\eta}[\hat{\eta}-\eta_0]\left(\hat{\theta}-\theta_0\right)\\ &-\Delta^j_{1\theta}\varepsilon_n + \Delta^{jj}_{2\theta^2}\varepsilon_n^2 + o_p\left(\varepsilon_n^2\right).\end{aligned}$$

As a result, we can evaluate

$$\frac{\hat{m}\left(z;\,\hat{\theta}+e_j\varepsilon_n,\hat{\eta}(\cdot)\right)-\hat{m}\left(z;\,\hat{\theta}-e_j\varepsilon_n,\hat{\eta}(\cdot)\right)}{2\varepsilon_n}\overset{\mathbf{L}^2}{=}\Delta^j_{1\theta}+O_p\left(\varepsilon_n^{-1}\left(n^{-1/k}+n^{-1/2}+n^{-1/k_1}\right)\right)$$

$$+o_p\left(\varepsilon_n\right).$$

Note that $\varepsilon_n\, n^{1/\max\{k,\,k_1\}}\to\infty$ and $\varepsilon_n\to 0$ will assure uniform convergence of the moment function to its derivative. Next, we provide the result for the uniform convergence of the directional derivative with respect to the infinite-dimensional parameter. Consider a particular direction w_j (which in practice will be an element of the sieve space containing the approximation $\hat{\eta}(\cdot)$), then:

$$\hat{m}\left(z;\,\hat{\theta},\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)=\hat{m}\left(z;\,\hat{\theta},\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)-m\left(z;\,\hat{\theta},\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)$$

$$-\hat{m}\left(z;\,\theta_0,\eta_0(\cdot)\right)+\hat{m}\left(z;\,\theta_0,h_0(\cdot)\right)-m\left(z;\,\theta_0,\eta_0(\cdot)\right)$$

$$+m\left(z;\,\hat{\theta},\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)-m\left(z;\,\theta_0,\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)$$

$$+m\left(z;\,\theta_0,\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)-m\left(z;\,\theta_0,\eta_0(\cdot)+\tau_n w_j(\cdot)\right)$$

$$+m\left(z;\,\theta_0,\eta_0(\cdot)+\tau_n w_j(\cdot)\right)-m\left(z;\,\theta_0,\eta_0(\cdot)\right).$$

Using the local $\mathbf{L}^2$-representation, we can approximate the expansion above as

$$\hat{m}\left(z;\,\hat{\theta},\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)\overset{\mathbf{L}^2}{=}O_p\left(n^{-1/k}\right)+\Delta_{1\theta}\left(\hat{\theta}-\theta_0\right)+\Delta_{1\eta}[\hat{\eta}-\eta_0]$$

$$+\left(\hat{\theta}-\theta_0\right)'\Delta_{2\theta^2}\left(\hat{\theta}-\theta_0\right)+\Delta_{2\eta^2}[\hat{\eta}-\eta_0]^2+\tau_n\Delta_{2h^2}[\hat{\eta}-\eta_0,\,w_j]$$

$$+\Delta_{2\theta\eta}[\hat{\eta}-\eta_0]\left(\hat{\theta}-\theta_0\right)+\tau_n\Delta_{2\theta\eta}[w_j]\left(\hat{\theta}-\theta_0\right)$$

$$+\tau_n\Delta_{1\eta}[w_j]+\tau_n^2\Delta_{2\eta^2}[w_j]^2+o_p\left(\|\hat{\eta}-\eta_0\|^2_{\mathbf{L}^2}\right)+o_p\left(\|\hat{\theta}-\theta_0\|^2\right).$$

We can write a similar expression for $\hat{m}\left(z;\,\hat{\theta},\hat{\eta}(\cdot)-\tau_n w_j(\cdot)\right)$. As a result, the symmetrized numerical directional derivative will be approximated locally by

$$\frac{\hat{m}\left(z;\,\hat{\theta},\hat{\eta}(\cdot)+\tau_n w_j(\cdot)\right)-\hat{m}\left(z;\,\hat{\theta},\hat{\eta}(\cdot)-\tau_n w_j(\cdot)\right)}{2\tau_n}\overset{\mathbf{L}^2}{=}\Delta_{1\eta}[w_j]+\Delta_{2\eta^2}[\hat{\eta}-\eta_0,\,w_j]$$

$$+\Delta_{2\theta\eta}[w_j]\left(\hat{\theta}-\theta_0\right)+O_p\left(\tau_n^{-1}\left(n^{-1/k}+n^{-1/2}+n^{-1/k_1}\right)\right)+o_p\left(\tau_n\right).$$

Note that $\left\|\Delta_{2\eta^2}[\hat{\eta}-\eta_0,\,w_j]\right\|_{\mathbf{L}^2}=O_p\left(n^{-1/k_1}\right)$. For $k_1>2$ this term will dominate and determine the lower bound on the sub-parametric convergence rate for the numerical derivative. The conditions for τ_n will be similar to the conditions for ε_n, that is $\tau_n\, n^{1/\max\{k,\,k_1\}}\to\infty$ and $\tau_n\to 0$. Moreover, it is clear that the convergence rate for τ_n is slower than n^{-1/k_1}. This result assures that for $z\in\mathcal{Z}$, $(\theta,\,\eta(\cdot))\in\Theta\times\mathcal{H}$, and $w_j\in\mathcal{W}$, $\widetilde{D}_{w_j}(z)\xrightarrow{p}D_{w_j}(z)$. □

A.2 Proof of Lemma 2

It follows directly from assumption 7.[iii] that

$$\left|L^{\epsilon_n,w}_{1,p}m\left(\theta,\eta,z\right)-proj\left(L^{\epsilon_n,w}_{1,p}m\left(\theta,\eta,z\right)|p^N\left(z\right)\right)\right|=o\left(1\right)$$

uniformly in θ, η, z. Therefore it suffices to prove Lemma 2 for

$$(*) = \left| L_{1,p}^{\epsilon_n, w} \widehat{m}\left(\theta, \eta, z\right) - proj\left(L_{1,p}^{\epsilon_n, w} m\left(\theta, \eta, z\right) | p^N\left(z\right) \right) \right|.$$

As demonstrated in Newey (1997), for $P = \left(p^N(z_1), \ldots, p^N(z_n)\right)'$ and $\hat{Q} = P'P/n$

$$\|\hat{Q} - I\| = O_p\left(\frac{N}{n}\right), \quad \text{where his } \zeta_0\left(N\right) = C.$$

Hence the smallest eigenvalue of $\hat{Q}$ will converge to 1. Following Newey (1997) we use the indicator 1_n to indicate the cases where the smallest eigenvalue of $\hat{Q}$ is above $\frac{1}{2}$ to avoid singularities. Introduce the vector $\Delta = \left(\rho\left(\theta, \eta + \epsilon_n w; Y_i\right) - \rho\left(\theta, \eta - \epsilon_n w; Y_i\right)\right)_{i=1}^n$. Let β be the vector of coefficients in the linear projection of $E\left[\Delta \,|\, z\right]$ on the sieve space. Also define $G = \left(E\left[\Delta \,|\, Z_i\right]\right)_{i=1}^n$. Then $(*)$ equals to a linear combination of $1_n |p^{N\prime}(z)\left(\hat{\beta} - \beta\right)|/\epsilon_n$. Note that

$$p^{N\prime}(z)\left(\hat{\beta} - \beta\right) = p^{N\prime}(z)\left(\hat{Q}^{-1}\, P'\left(\Delta - G\right)/n + \hat{Q}^{-1}\, P'\left(G - P\beta\right)/n\right). \tag{A.11}$$

For the first term in (A.11), since the eigenvalues of $\hat{Q}$ are converging to one, $\|\hat{Q}^{-1}\| \leq C - o_p(1)$,

$$\left| p^{N\prime}(z)\hat{Q}^{-1}P'\left(\Delta - G\right) \right| \leq C \sum_{k=1}^N \left\| \sum_{i=1}^n p_{Nk}(z_i)\left(\Delta_i - G_i\right) \right\|.$$

Denote $\mu_n = \mu N^{-1}\epsilon_n$. Next we adapts the arguments for proving Theorem 37 in Pollard (1984) to provide the bound for $P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\|p^{N\prime}(z)\hat{Q}^{-1}P'\left(\Delta - G\right)\| > N\mu_n\right)$. For N non-negative random variables Y_i we note that

$$P\left(\sum_{i=1}^N Y_i > c\right) \leq \sum_{i=1}^N P\left(Y_i > c/N\right).$$

Using this observation, we can find that

$$P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\|p^{N\prime}(z)\hat{Q}^{-1}P'\left(\Delta - G\right)\| > N\mu_n\right) \leq \sum_{k=1}^N P\left(\sup_{\mathcal{F}_n}\left\| \frac{1}{n}\sum_{i=1}^n p_{Nk}(z_i)\left(\Delta_i - G_i\right)\right\| > \mu_n\right)$$

The variance of each $p_{Nk}(z_i)\left(\Delta_i - G_i\right)$ is bounded by ϵ_n. Then application of Theorem 37 in Pollard (1984) leads to

$$\begin{aligned} &P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\left\|\sum_{i=1}^n p_{Nk}(z_i)\left(\Delta_i - G_i\right)\right\| > 8\mu_n\right) \\ &\leq 2A\mu_n^{-V} n^{2r_0} \exp\left(-\frac{1}{128}\frac{n}{N^2}\mu^2\epsilon_n\right) + R(n), \end{aligned}$$

where $R(n)$ is of smaller order of magnitude than the first term. As a result, we find that

$$P\left(\sup_{\mathcal{F}_n} \frac{1}{n}\|p^{N\prime}(z)\hat{Q}^{-1}P'\left(\Delta - G\right)\| > N\mu_n\right) \leq 2AN\mu_n^{-V} n^{2r_0} \exp\left(-\frac{1}{128}\frac{n}{N^2}\mu^2\epsilon_n\right) + NR(n)$$

Consider the log of the first term,

$$V\log\left(N/\left(\mu\epsilon_n\right)\right) - \frac{1}{128}\frac{n}{N^2}\mu^2\epsilon_n + 2r_0\log n + \log N.$$

This expression converges to $-\infty$ if $\frac{n\zeta_0(N)^2\epsilon_n}{N\ \log\ n} \to \infty$, because one can always find an $\alpha > 0$ large enough so that both $-\frac{1}{128}\frac{n}{N^2}\mu^2\epsilon_n + (2r_0 + \alpha)\log n$ and $V\log(N/(\mu\epsilon_n)) + \log N - \alpha\log n$ converge to $-\infty$. Hence the first term is of $o(1)$. So is the second term above.

Finally, the second term in (A.11) can be shown to converge to 0 if $n\epsilon^2/N \to \infty$. □

A.3 Proof of Lemma 3

Recall the definition of the kernel estimator

$$\widehat{m}(\theta, \eta, z) = \left(\frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{z - z_i}{b_n}\right)\right)^{-1} \frac{1}{nb_n}\sum_{i=1}^{n} \rho(\theta, \eta, z_i) K\left(\frac{z - z_i}{h_n}\right)$$

For the expression of interest, we can consider

$$\frac{\widehat{m}(\theta, \eta + \epsilon_n w, z) - \widehat{m}(\theta, \eta - \epsilon_n w, z)}{\epsilon_n} = \left(\frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{z - z_i}{b_n}\right)\right)^{-1}$$
$$\times \frac{1}{nb_n\epsilon_n}\sum_{i=1}^{n} \left[\rho(\theta, \eta + \epsilon_n w, y_i) - \rho(\theta, \eta - \epsilon_n h, y_i)\right] K\left(\frac{z - z_i}{b_n}\right).$$

Then we can consider a class of functions

$$\mathcal{G}_n = \{\left[\rho(\theta, \eta + \epsilon_n h, y_i) - \rho(\theta, \eta - \epsilon_n h, y_i)\right] K\left(\frac{z - z_i}{b_n}\right)\}$$

and directly apply Theorem 37 in Pollard (1984). This leads to condition $\frac{n\epsilon_n b_n}{\log\ n} \to \infty$. □

A.4 Proof of Lemma 4

We utilize our proof of Lemma 2 and start with the analysis of the series estimator for the conditional moment function. The idea behind the proof for the series estimator was to apply the exponential inequality to each projection of the moment function on the sieve terms used to estimate the conditional moment function. We use $\Delta = (\rho(\theta, \eta + \epsilon_n h; Z_i) - \rho(\theta, \eta - \epsilon_n h; Z_i))_{i=1}^n$ and $G = (E[\Delta \mid X_i])_{i=1}^n$. Then we note that

$$E\,\|\frac{1}{\epsilon_n}p_{Nk}(x_i)(\Delta_i - G_i)\| \leq C\zeta_0(N)\,\epsilon_n^{\gamma-1}$$

due to Assumptions 7 and 8(i'). This implies that the upper bound on the variance is proportional to $\zeta_0^2(N)\,\epsilon_n^{2(\gamma-1)}$. Then denoting $\gamma_n = \mu N^{-1}\epsilon_n^{2(\gamma-1)}\zeta_0^2(N)$ we use our previous derivations to obtain

$$P\left(\frac{1}{n\epsilon_n}\sum_{i=1}^{n} \|p_{Nk}(x_i)(\Delta_i - G_i)\| > 8\mu_n\right)$$
$$\leq 2A\mu_n^{-V} n^{2r_0} \exp\left(-\frac{1}{128}\frac{n}{N}\zeta_0(N)^2\epsilon_n^{2(\gamma-1)}\right) + R(n),$$

where $R(n)$ is of smaller order of magnitude than the first term. We note that the argument in the exponent converges to $-\infty$ if $\frac{n\zeta_0(N)^2}{N\ \epsilon_n^{2(1-\gamma)}\ \log\ n} \to \infty$.

Consider now similar conditions for kernel-based estimators. As before, we can represent the finite-dioefference formula for the directional derivative of interest as

$$\widehat{m}\left(\theta,\eta+\epsilon_n h,x\right)-\widehat{m}\left(\theta,\eta-\epsilon_n h,x\right)=\left(\frac{1}{nb_n}\sum_{i=1}^{n}K\left(\frac{x-x_i}{b_n}\right)\right)^{-1}$$
$$\times\frac{1}{nb_n}\sum_{i=1}^{n}\left[\rho\left(\theta,\eta+\epsilon_n h,z_i\right)-\rho\left(\theta,\eta-\epsilon_n h,z_i\right)\right]K\left(\frac{x-x_i}{b_n}\right)$$

We note that the variance of the summand is bounded from above by $C\,b_n\,\epsilon_n^{2(\gamma-1)}$ due to the bound on the conditional expectation in Assumption 8 (i') and the Cauchy-Schwartz inequality. We can then apply Theorem 37 in Pollard (1984) and pick $\gamma_n=\mu\epsilon_n^{2(\gamma-1)}b_n$ for the class of functions

$$\mathcal{G}_n=\{\frac{1}{\epsilon_n}\left[\rho\left(\theta,\eta+\epsilon_n h,z_i\right)-\rho\left(\theta,\eta-\epsilon_n h,z_i\right)\right]\frac{1}{b_n}K\left(\frac{x-x_i}{b_n}\right)\}.$$

As a result we formulate the combined condition for the bandwidth and the step size for of the numerical differentiation as $\frac{nb_n}{\epsilon_n^{2(1-\gamma)}\log\,n}\to\infty$.

□

A.5 Proof of theorem 7

Proof. We note that according to assumption 4, $\sup_{\theta\in\Theta}||EL_{1,p}^{\varepsilon}\hat{Q}\left(\theta\right)-G\left(\theta\right)||=O\left(\varepsilon^{\nu}\right)$.

We can partition the parameter space into $j=1,\ldots,J$ grids of interval lengths η such that $\eta=o\left(\varepsilon\right)$ but that it still satisfies $n\eta/\log n\to\infty$. Denote the center of the jth grid by θ_j. Then we can write, for any δ and $b_n=\sqrt{\frac{\log n}{n\varepsilon^{2-2\gamma}}}\to 0$,

$$P\left(\sup_{\theta\in\Theta}|L_{1,p}^{\varepsilon}\hat{Q}\left(\theta\right)-EL_{1,p}^{\varepsilon}\hat{Q}\left(\theta\right)|>\delta b_n\right)<\sum_{j=1}^{J}P\left(|L_{1,p}^{\varepsilon}\hat{Q}\left(\theta_j\right)-EL_{1,p}^{\varepsilon}\hat{Q}\left(\theta_j\right)|>\frac{\delta b_n}{2}\right)$$
$$+P\left(\sup_{j=1,\ldots,J}\sup_{|\theta-\theta_j|\leq\eta}|L_{1,p}^{\varepsilon}\hat{Q}\left(\theta\right)-EL_{1,p}^{\varepsilon}\hat{Q}\left(\theta\right)-\left(L_{1,p}^{\varepsilon}\hat{Q}\left(\theta_j\right)-EL_{1,p}^{\varepsilon}\hat{Q}\left(\theta_j\right)\right)|>\frac{\delta b_n}{2}\right)$$

The second term on the right hand size can be written as a linear combination of

$$G_n^t\equiv\sup_{j=1,\ldots,J}\sup_{|\theta-\theta_j|\leq\eta}\frac{1}{\sqrt{n\varepsilon}}\left(\mathbb{G}\left(\theta+t\varepsilon\right)-\mathbb{G}\left(\theta-t\varepsilon\right)-\mathbb{G}\left(\theta_j+t\varepsilon\right)+\mathbb{G}\left(\theta_j-t\varepsilon\right)\right)|$$

By Assumption 6 (ii) through the Markov inequality,

$$P\left(G_n^t\geq\frac{\delta b_n}{C}\right)\leq\frac{O\left(\min\left(\eta,\varepsilon\right)^{\gamma}\right)}{\delta b_n\sqrt{n\varepsilon}}\lesssim O\left(\frac{\eta^{\gamma}}{\delta b_n\sqrt{n}\varepsilon}\right)=o\left(1\right).$$

Next the first term is bounded by an (Hoeffding) exponential inequality using the upper bound from Assumption 6 (ii):

$$\sum_{j=1}^{J}P\left(|L_{1,p}^{\varepsilon}\hat{Q}\left(\theta_j\right)-EL_{1,p}^{\varepsilon}\hat{Q}\left(\theta_j\right)|>\frac{\delta b_n}{2}\right)\leq CJ\exp\left(-\frac{n\left(\delta b_n/4\right)^2}{2C\varepsilon^{2\gamma-2}}\right)$$

We can further bound the above probability by, noting that $J \leq n^{\alpha}$ for some constant α,

$$J \exp\left(-\frac{n\left(\delta b_n/4\right)^2}{2C\varepsilon^{2\gamma-2}}\right) = \exp\left(-\left(\delta C-\alpha\right)\log n\right) \overset{\delta\to\infty}{\longrightarrow} 0.$$

The remaining statements of the theorem are standard arguments. □

A.6 Proof of Lemma 5

Proof. To find the convergence rate for the estimator $\hat{\theta}$ we need to find the "balancing" sequence ρ_n that assures that $\rho_n d\left(\hat{\theta},\theta_0\right) = O_P^*(1)$. Using the assumption of the compactness of the parameter space, we cover it using a grid with cells $S_{j,n} = \left\{\theta : 2^{j-1} < \rho_n\, d\left(\theta,\theta_0\right) < 2^j\right\}$. The idea of the proof is to show that for any given $\kappa > 0$ we can find finite β such that the probability of event $\rho_n d\left(\hat{\theta},\theta_0\right) > \beta$ is below κ. Using a partition of the parameter space, we can pick a finite integer M such that $2^M < \beta$. Then we evaluate the probability of a large deviation $\rho_n d\left(\widehat{\theta},\theta_0\right) > 2^M$. We know that the estimator solves

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

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 cells $S_{j,n}$, $\sqrt{n\varepsilon_n} L_{1,p}^{\varepsilon_n} Q_n\left(\theta\right)$ achieves a distance as close as desired to zero. Hence, for every $\delta > 0$,

$$P\left(\rho_n\, d\left(\widehat{\theta},\theta_0\right) > 2^M\right) \leq \sum_{\substack{j \geq M \\ 2^j < \delta\rho_n}} P\left(\sup_{\theta\in S_{j,n}}\left(-\left\|L_{1,p}^{\varepsilon_n}\widehat{Q}\left(\theta\right)\right\|\right) \geq -o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)\right) + P\left(2\, d\left(\widehat{\theta},\theta_0\right) \geq \delta\right)$$

Then we evaluate the population objective, using the fact that it has p mean-square derivatives with Taylor residual of order ν:

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

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

$$\left\|L_{1,p}^{\varepsilon_n} Q\left(\theta\right) - L_{1,p}^{\varepsilon_n}\widehat{Q}\left(\theta\right)\right\| \geq \left\|L_{1,p}^{\varepsilon_n} Q\left(\theta\right)\right\| - \left\|L_{1,p}^{\varepsilon_n}\widehat{Q}\left(\theta\right)\right\|.$$

Therefore

$$\sup_{\theta\in S_{j,n}}\left(-\left\|L_{1,p}^{\varepsilon_n}\widehat{Q}\left(\theta\right)\right\|\right) \geq \sup_{\theta\in S_{j,n}}\left\|L_{1,p}^{\varepsilon_n} Q\left(\theta\right) - L_{1,p}^{\varepsilon_n}\widehat{Q}\left(\theta\right)\right\| - \sup_{\theta\in S_{j,n}}\left\|L_{1,p}^{\varepsilon_n} Q\left(\theta\right)\right\|$$

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

$$P\left(\sup_{\theta\in S_{j,n}}\left\|L_{1,p}^{\varepsilon_n} Q\left(\theta\right) - L_{1,p}^{\varepsilon_n}\widehat{Q}\left(\theta\right)\right\| \geq C\, d\left(\theta,\theta_0\right) + C'\varepsilon_n^{\nu-1} + o\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)\right)$$

$$\leq \frac{E^*\left[\sup_{\theta\in S_{j,n}}\left\|L_{1,p}^{\varepsilon_n} Q\left(\theta\right) - L_{1,p}^{\varepsilon_n}\widehat{Q}\left(\theta\right)\right\|\right]}{C\, d\left(\theta,\theta_0\right) + C'\varepsilon_n^{\nu-1} + o\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)} = O\left(\rho_n\sqrt{\frac{\log\, n}{n\varepsilon_n}}\right)2^{-(j-1)}.$$

Thus if $\rho_n = o\left(\sqrt{\frac{n\varepsilon_n}{\log\, n}}\right)$ then the probability of interest is $o(1)$. □

A.7 Proof of Lemma 6

Proof. Consider a class of functions

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

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 5 (ii). Note that

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

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

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

and for the second group

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

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

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

Next we denote $\delta_n = \min\left\{\varepsilon_n,\, \sqrt{\frac{\log\, n}{n\varepsilon_n}}\right\}$. Invoking Lemma 33 in Pollard (1984) and using Assumption 5 (iii) we obtain that

$$P\left(\sup_{\mathcal{G}_n} P_n f^2 > 64\delta_n\right) \leq 4A\left(\delta_n\right)^{-V} \exp\left(-n\delta_n\right) = 4A \exp\left(-n\delta_n + V\log\left(\frac{1}{\delta_n}\right)\right).$$

If $n\varepsilon_n^3/\log\, n \to \infty$, then this allows us to conclude that $\sup\limits_{\mathcal{G}_n} P_n f^2 = o_p\left(\delta_n\right)$. Next we can apply the maximum inequality from Theorem 2.14.1 in Van der Vaart and Wellner (1996) which implies that for the functions with constant envelopes

$$\sqrt{\frac{n}{\varepsilon_n}} \sup_{\mathcal{G}_n} \left|P_n\, f - P\, f\right| \lesssim \frac{1}{\sqrt{\varepsilon_n}} J\left(\sup_{\mathcal{G}_n} P_n f^2,\, \mathcal{G}_n\right),$$

where $J(\cdot)$ is a covering integral:

$$J\left(\delta, \mathcal{F}\right) = \sup_{Q} \int_0^{\delta} \sqrt{1 + N\left(\epsilon, \mathcal{F}, \mathbf{L}^2(Q)\right)}\, d\,\epsilon.$$

For Euclidean class as in Assumption 5 (iii) we can evaluate $J(\delta, \mathcal{G}_n) = O\left(\delta\sqrt{\log\left(\frac{1}{\delta}\right)}\right)$. Using the expression for $\sup\limits_{\mathcal{G}_n} P\, f^2$, we can evaluate

$$\sqrt{\frac{n}{\varepsilon_n}} \sup_{\mathcal{G}_n} \left|P_n\, f - P\, f\right| = o_p\left(\frac{\delta_n}{\sqrt{\varepsilon_n}}\sqrt{\log\left(\frac{1}{\delta_n}\right)}\right).$$

Then, provided that $\frac{n\varepsilon_n}{\log\, n} \to \infty$, we see that $\delta_n = \sqrt{\log\, n/(n\,\varepsilon_n)}$ and

$$\sqrt{\frac{n}{\varepsilon_n}} \sup_{\mathcal{G}_n} |P_n\, f - P\, f| = o_p\left(\sqrt{\frac{\log\left(\frac{n\varepsilon_n}{\log\, n}\right)}{\frac{n}{\log\, n}}}\right) = o_p(1).$$

The statement of the Lemma follows directly from this result. □

A.8 Proof of theorem 8

Proof. This proof will replicate the steps of proof of Lemma 5. We perform the triangulation of the parameter space according to the balancing rate ρ_n into segments $S_{j,n} = \left\{\theta\,:\, 2^{j-1} < \rho_n\, d\,(\theta,\theta_0) < 2^j\right\}$. For every $\delta > 0$,

$$P\left(\rho_n\, d\left(\widehat{\theta},\theta_0\right) > 2^M\right) \leq \sum_{\substack{j \geq M \\ 2^j < \delta\rho_n}} P\left(\sup_{\theta\in S_{j,n}}\left(-\left\|L_{1,p}^{\varepsilon_n}\widehat{Q}\,(\theta)\right\|\right) \geq -o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right)\right) + P\left(2\,d\left(\widehat{\theta},\theta_0\right) \geq \delta\right)$$

We can then follow the steps of Lemma 5 to evaluate the upper bound for the elements in the sum on the right-hand side as

$$\begin{aligned} &P\left(\sup_{\theta\in S_{j,n}}\left\|L_{1,p}^{\varepsilon_n}Q\,(\theta) - L_{1,p}^{\varepsilon_n}\widehat{Q}\,(\theta) - L_{1,p}^{\varepsilon_n}Q\,(\theta_0) + L_{1,p}^{\varepsilon_n}\widehat{Q}\,(\theta_0)\right\| \geq O\left(\frac{2^j}{\rho_n}\right)\right) \\ &\leq \frac{E^*\left[\sup_{\theta\in S_{j,n}}\left\|L_{1,p}^{\varepsilon_n}Q\,(\theta) - L_{1,p}^{\varepsilon_n}\widehat{Q}\,(\theta) - L_{1,p}^{\varepsilon_n}Q\,(\theta_0) + L_{1,p}^{\varepsilon_n}\widehat{Q}\,(\theta_0)\right\|\right]}{\frac{2^j}{\rho_n}} = O\left(\rho_n\frac{1}{\sqrt{n\varepsilon_n}}\right)2^{-(j-1)}. \end{aligned}$$

Thus if $\rho_n = O\left(\sqrt{n\varepsilon_n}\right)$ then the probability of interest is $O(1)$. □

A.9 Proof of theorem 9

Proof. We can note that the scaled bias due to numerical approximation can be evaluated as

$$\sqrt{n\,\varepsilon_n}\left(L_{1,p}^{\epsilon_n}Q\,(\theta) - G\,(\theta)\right) = O_p\left(\sqrt{n\varepsilon_n^{1+\nu/2}}\right) = o_p\,(1)\,.$$

We know that the estimator solves

$$L_{1,p}^{\epsilon_n}\widehat{Q}\left(\hat{\theta}\right) = o_p\left(\frac{1}{\sqrt{n\varepsilon_n}}\right).$$

Then

$$\sqrt{n\varepsilon_n}\left(L_{1,p}^{\epsilon_n}\hat{Q}\left(\hat{\theta}\right) - L_{1,p}^{\epsilon_n}Q(\hat{\theta}) + G\left(\hat{\theta}\right) + L_{1,p}^{\epsilon_n}Q\left(\hat{\theta}\right) - G\left(\hat{\theta}\right)\right) = o_p\,(1)\,.$$

This means that locally

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

From Corollary 1 it follows that

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

Then we can apply the CLT to obtain the desired result. □

A.10 Proof of theorem 10

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 $\rho_n = \sqrt{n}\varepsilon_n^{1-\gamma}$. We know that the estimator solves

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

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 $S_{j,n}$, $\sqrt{n} r_n L_{1,p}^{\varepsilon_n} Q_n(\theta)$ achieves a distance as close as desired to zero. Hence, for every $\delta > 0$,

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

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:

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

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

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

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

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

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

A.11 Proof of theorem 11

Proof. The result will follow if we can demonstrate that

$$\sqrt{n} r_n \left(L_{1,p}^{\varepsilon} \hat{g}\left(\hat{\theta}\right) - L_{1,p}^{\varepsilon} \hat{g}(\theta_0) - G\left(\hat{\theta}\right) + G(\theta_0)\right) = o_p(1). \tag{A.12}$$

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

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

Because of the convergence rate established in theorem 10, this will implied by, with $r_n = \varepsilon^{1-\gamma}$:

$$\sup_{d(\theta,\theta_0)\lesssim O\left(\frac{1}{\sqrt{n}\varepsilon^{1-\gamma}}\right)} \sqrt{n} r_n \left(L^{\varepsilon}_{1,p}\hat{g}\left(\theta\right) - L^{\varepsilon}_{1,p}\hat{g}\left(\theta_0\right) - EL^{\varepsilon}_{1,p}\hat{g}\left(\theta\right) + EL^{\varepsilon}_{1,p}\hat{g}\left(\theta_0\right)\right) = o_p\left(1\right).$$

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

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

Because of assumption 6, it is bounded stochastically by

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

When $\sqrt{n}\varepsilon^{2-\gamma} \to \infty$, $d\left(\theta,\theta_0\right) \lesssim O\left(\frac{1}{\sqrt{n}\varepsilon^{1-\gamma}}\right) = o\left(\varepsilon\right)$. Hence the above display is $o_p\left(1\right)$. Therefore (A.12) holds.

Recall that $\hat{\theta}$ is defined by $\sqrt{n}r_n(L^{\varepsilon}_{1,p}\hat{g}\left(\hat{\theta}\right) = o_p\left(1\right)$. Then (A.12) implies that, using a first order taylor expansion of $G\left(\theta\right)$:

$$\sqrt{n}r_n\left(L^{\varepsilon}_{1,p}\hat{g}\left(\theta_0\right) - EL^{\varepsilon}_{1,p}\hat{g}\left(\theta_0\right)\right) + H\left(\theta_0\right)\sqrt{n}r_n\left(\hat{\theta}-\theta_0\right) = o_p\left(1\right).$$

□

A.12 Proof of Theorem 14

Denote $\hat{A}(\theta,\eta,z_i) = L^{\varepsilon_n}_{1,p}\hat{m}\left(\theta,\eta,z_i\right)'\,\widehat{W}\left(z_i\right)$, and $\hat{B}_j(\theta,\eta,z_i) = L^{\tau_n,\psi_j}_{1,p}\hat{m}\left(\theta,\eta,z_i\right)'\,\widehat{W}\left(z_i\right)$ and $A(\theta,\eta,z_i)$ and $B_j(\theta,\eta,z_i)$ their population analogs. Then we can decompose

$$\hat{A}(z_i,\theta,\eta)\hat{m}\left(\theta,\eta,z_i\right) - A(z_i,\theta,\eta)m\left(\theta,\eta,z_i\right) = \left(\hat{A}(z_i,\theta,\eta) - A(z_i,\theta,\eta)\right)m\left(\theta,\eta,z_i\right)$$
$$+ A(z_i,\theta,\eta)\left(\hat{m}\left(\theta,\eta,z_i\right) - m\left(\theta,\eta,z_i\right)\right) + \left(\hat{A}(z_i,\theta,\eta) - A(z_i,\theta,\eta)\right)\left(\hat{m}\left(\theta,\eta,z_i\right) - m\left(\theta,\eta,z_i\right)\right).$$

We can provide the same expansion for $\hat{B}(\cdot)$. We start with an easier case of the kernel-based estimator for the moment function. We note that due to Assumption 8

$$\sup_{(\theta,\eta)\in U}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{B}(z_i,\theta,\eta) - B(z_i,\theta,\eta)\right)m\left(\theta,\eta,z_i\right)\right\| \le C\sup_{(\theta,\eta)\in\Theta\times\mathcal{H}}\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{B}(z_i,\theta,\eta) - B(z_i,\theta,\eta)\right)\right\|.$$

For the kernel-based estimator we can express

$$\frac{1}{n^2 b_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{nb_n}\sum_{j=1}^{n}K\left(\frac{z_i-z_j}{b_n}\right)\right)^{-1}K\left(\frac{z_i-z_j}{b_n}\right)\left(L^{\tau_n,\psi_j}_{1,p}\rho\left(\theta,\eta,y_i\right) - L^{\tau_n,\psi_j}_{1,p}m\left(\theta,\eta,z_i\right)\right).$$

We note that this expression represents a second-order U-statistic with kernel $K\left(\frac{z_i-z_j}{b_n}\right)L^{\tau_n,\psi_j}_{1,p}\left(\rho\left(\theta,\eta,y_i\right) - m\left(\theta,\eta,z_i\right)\right)$. It corresponds to the numerical derivative of the U-statistic with kernel $K\left(\frac{z_i-z_j}{b_n}\right)\left(\rho\left(\theta,\eta,y_i\right) - m\left(\theta,\eta,z_i\right)\right)$. We have investigate the properties of U-statistics of this structure in Section 5. Given that

$$E\Bigg[K\left(\frac{Z_i-Z_j}{b_n}\right)^2\Bigg(\rho\left(\theta,\eta+w\tau_n,Y_i\right) - m\left(\theta,\eta+w\tau_n,Z_i\right)$$
$$-\rho\left(\theta,\eta-w\tau_n,Y_i\right) - m\left(\theta,\eta-w\tau_n,Z_i\right)\Bigg)^2\Bigg|Z_i=z\Bigg] = O\left(b_n\tau_n\right),$$

we can provide the evaluation for the supremum of interest using results of Lemma 9, which is

$$\sup_{(\theta,\eta)\in U}\left\|\frac{1}{n^2b_n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{1}{nb_n}\sum_{j=1}^{n}K\left(\frac{z_i-z_j}{b_n}\right)\right)^{-1}K\left(\frac{z_i-z_j}{b_n}\right)\right.$$
$$\left.\left(L_{1,p}^{\tau_n,\psi_j}\rho\left(\theta,\eta,y_i\right)-L_{1,p}^{\tau_n,\psi_j}m\left(\theta,\eta,z_i\right)\right)\right\|=O_p\left(\frac{\log^2 n}{n^2\tau_n b_n}\right).$$

We can evaluate the supremum of $B(z_i,\theta,\eta)\left(\hat{m}\left(\theta,\eta,z_i\right)-m\left(\theta,\eta,z_i\right)\right)$ from the uniform convergence rate for the kernel estimator with $O_p\left(\sqrt{\frac{\log n}{nb_n}}\right)$. The last term can be evaluated from the uniform convergence rates for the kernel estimator and the numerical derivative. This provides the combined rate $O_p\left(\frac{\log n}{nb_n\sqrt{\tau_n}}\right)$. We note that provided that $nb_n/\log n\to\infty$, the expression is dominated by the first term. As a result, we assure that the supremum of the entire expression is of order $o_p(1)$ if $\frac{\log^2 n}{n^2\tau_n b_n}\to 0$. A similar result will follow for the derivative with respect to the parametric component.

We can perform the same analysis for the series-based estimator. We can express the first component of the decomposition above as

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(p^{N\prime}(z_i)\left(\frac{1}{n}\sum_{i=1}^{n}p^N(z_i)p^{N\prime}(z_i)\right)^{-1}p^N(z_j)L_{1,p}^{\tau_n,\psi_j}\rho\left(\theta,\eta,y_j\right)-L_{1,p}^{\tau_n,\psi_j}E\left[\rho\left(\theta,\eta,Y_j\right)|Z_i=z_i\right]\right).$$

We note that this is also a U-statistic. We find that we can evaluate the mean square by denoting $G=E\left[\rho\left(\theta,\eta+w\tau_n,Y_j\right)-\rho\left(\theta,\eta-w\tau_n,Y_j\right)\middle|Z_i=z_i\right]$

$$E\Bigg[p^{N\prime}(z_i)\left(\frac{1}{n}\sum_{i=1}^{n}p^N(z_i)p^{N\prime}(z_i)\right)^{-1}p^N(z_i)p^{N\prime}(z_i)\left(\frac{1}{n}\sum_{i=1}^{n}p^N(z_i)p^{N\prime}(z_i)\right)^{-1}p^N(z_i)$$
$$\left(\rho\left(\theta,\eta+w\tau_n,Y_j\right)-\rho\left(\theta,\eta-w\tau_n,Y_j\right)-G\right)^2\Bigg|Z_i=z\Bigg]=O\left(N\tau_n\right),$$

which follows from Cauchy-Schwartz inequality and our result in Lemma 2. Having established this result, we notice that the remaining manipulations will be identical to the case where the moment function is estimated using kernels. This gives the sufficient condition for convergence with $\frac{N\log^2 n}{n^2\tau_n}\to 0$. □

A.13 Proof of Lemma 8

The result of the theorem can be obtained by adapting the argument in Lemma 1 to 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 $PF\le C$. Then we can write

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

Noting (5.10), lemma 1 can be shown separately for the $\hat{\mu}_n\left(\theta\right)$ and $S_n\left(u\right)/n(n-1)$ components of the decomposition. Because assumption 17 is a special case of assumption 6, Theorem 3 applies with $\gamma=1$. Therefore the result of lemma 8 holds for the $\hat{\mu}_n\left(\cdot\right)$ component as long as ϵ_n. We will hence with no loss of generality focus on $S_n(u)$ and assume that $g\left(\cdot,\cdot,\theta\right)$ is degenerate.

Due to Assumption 18, for each $f \in \mathcal{F}_n$, $E|f|^2 = E\left|\epsilon_n L^{\epsilon_n}_{1,p} g(\cdot,\theta)\right|^2 = O(\epsilon_n)$. 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 8, 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 5, the covering integral of $\mathcal{F}_n$ is bounded by a constant multiple of $H(s) = s\left[1 + \log(1/s)\right]$. the maximum inequality in Theorem 6 of Nolan and Pollard (1987) implies that

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

where T_n is the symmetrized measured defined in 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 5. 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 $\frac{4\beta^2}{B}\, t_n^2 \lll 1$. Then

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

Substituting this result into the maximum inequality one can obtain

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

By assumption $t_n n >> \log n \to \infty$, the first term vanishes. The second term also vanishes by showing that $n^{-1} t_n^{-2} \exp(-2n\, t_n) \to 0$, 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 $t_n^{-1} n^{-1} \log(n t_n) - t_n^{-1} n^{-1} \log n$. 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.14 Proof of Lemma 9

Proof. We note that for the projection part

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

As a result the U-process part will dominate and the convergence rate will be determined by its order $\frac{\log^2 n}{n^2 \varepsilon_n}$. The rest follows from the proof of Lemma 5. □

A.15 Proof of Lemma 10

Proof. The proof of this lemma largely relies on the proof of Lemma 6. Consider a class of functions

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

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 5 (ii). Note that

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

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

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

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 Assumption 18, for each $f \in \mathcal{F}_n$, $E|f|^2 = E\left| \epsilon_n L^{\epsilon_n}_{1,p} g(\cdot, \theta) \right|^2 = O(\epsilon_n)$. 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 5 (iii), the covering integral of $\mathcal{F}_n$ is bounded by a constant multiple of $H(s) = s\left[1 + \log(1/s)\right]$. 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.

□

A.16 Proof of Theorem 17

Proof. We can analyze the empirical process $\mathbb{G}'(\theta)$ using the same tools that we used for the case of numerical derivatives of the standard empirical processes. As before we can analyze the difference

$$G_m^t \equiv \mathbb{G}'\left(\hat{\theta}+t\epsilon_n\right) - \mathbb{G}'\left(\hat{\theta}-t\epsilon_n\right) - \mathbb{G}'\left(\theta_0+t\epsilon_n\right) + \mathbb{G}'\left(\theta_0-t\epsilon_n\right).$$

As a consequence, assumption 6 applied to $\mu\left(Z_i,\theta\right)$ implies immediately that the "stochastic residual" term corresponding to the empirical process part of the decomposition of statistic $S_N(f)$ is

$$G_1' = O_p\left(\frac{1}{\sqrt{n}\epsilon_n}\min\left(|\hat{\theta}-\theta_0|,\epsilon_n\right)^{\gamma}\right),$$

where m is the order of the numerical derivative taken. We can provide a similar evaluation for the U-process part, which leads to the stochastic order

$$U_1 = O_p\left(\frac{1}{n\epsilon_n}\min\left(|\hat{\theta}-\theta_0|,\epsilon_n\right)^{\psi}\right).$$

The variance components can be evaluated from our analogs of the CLT. In particular, from Assumption 20 it follows that $U_2 = O_p\left(\frac{1}{n\epsilon_n^{1-\psi}}\right)$. Similar to our previous analysis we can note that $G_1' \lesssim O_p\left(G_2'\right)$ and similarly $U_1 \lesssim O_p\left(U_2\right)$. Hence

$$L_{m,p}^{\epsilon}\hat{g}\left(\hat{\theta}\right) - G = O_p\left(G_2'+G_4'+U_2+U_4\right) = O_p\left(\epsilon_n^{\nu} + \frac{1}{\sqrt{n}\epsilon_n^{1-\gamma}} + \frac{1}{n\epsilon_n^{1-\psi}}\right).$$

If the second and the third term approach zero, it is sufficient that $n\,\epsilon_n^{2-2\gamma}\to\infty$ and $n\,\epsilon_n^{1-\psi}\to\infty$. This can be equivalently stated as $n\,\epsilon_n^{2-2\gamma}\to\infty$ with $\psi\geq 2\gamma-1$ or $n\,\epsilon_n^{1-\psi}\to\infty$ with $\psi\leq 2\gamma-1$.

Optimality of the step size choices can be considered analogously to the case of the empirical processes. In particular, in case where the second term dominates the third the calculations will be identical to the case of the empirical process. Domination of the second term will be assured by the condition $\sqrt{n}\varepsilon^{\gamma-\psi}\to\infty$ which leads to $\frac{\gamma-\psi}{1-\gamma+2\nu} > \frac{1}{2}$. The latter is consistent with the condition $2\psi < 2\nu+\gamma-1$. When the third term dominates, calculations will be the same with the substitution of the $\sqrt{n}$ term by n term. □

Acknowledgments: The authors acknowledges generous research supports from the National Science Foundation, the University of California at Berkeley and Stanford University. We thank Tim Armstrong and numerous conference and seminar participants for insightful comments. The usual disclaimer applies.

References

Ackerberg, D., X. Chen, and J. Hahn (2009): "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators," unpublished manuscript, UCLA and Yale University.

Ai, C., and X. Chen (2003): "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

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.

Arcones, M., and E. Gine (1993): "Limit theorems for U-processes," *The Annals of Probability*, 21, 1494–1542.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608.

Chen, X., and D. Pouzo (2009): "Estimation of nonparametric conditional moment models with possibly nonsmooth moments," SSRN working paper.

Chen, X., and X. Shen (1998): "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314.

Dudley, R. (1999): *Uniform central limit theorems*. Cambridge university press.

Durbin, J. (1971): "Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test," *Journal of Applied Probability*, pp. 431–453.

——— (1985): "The first-passage density of a continuous Gaussian process to a general boundary," *Journal of Applied Probability*, pp. 99–122.

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.

Horowitz, J. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Models," *Econometrica*, 60.

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.

Kim, J., and D. Pollard (1990): "Cube root asymptotics," *Ann. Statist.*, 18, 191–219.

Kosorok, M. (2008): *Introduction to empirical processes and semiparametric inference*. Springer Verlag.

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

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

——— (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of Maximum Score Estimator," *Journal of Econometrics*, 27, 205–228.

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

Newey, W. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.

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.

Newey, W., and J. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71, 1565–1578.

Nolan, D., and D. Pollard (1987): "U-processes:rates of convergence," *The Annals of Statistics*, pp. 780–799.

——— (1988): "Functional limit theorems for U-processes," *The Annals of Probability*, pp. 1291–1298.

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

Powell, J. L. (1984): "Least Absolute Deviations Estimation For The Censored Regression Model," *Journal of Econometrics*, 25, 303–325.

Pshenichnyi, B. (1971): *Necessary conditions for an extremum*. CRC.

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

Shen, X., and W. Wong (1994): "Convergence rate of sieve estimates," *The Annals of Statistics*, 22, 580–615.

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.

Zhang, J., and I. Gijbels (2003): "Sieve empirical likelihood and extensions of the generalized least squares," *Scandinavian Journal of Statistics*, pp. 1–24.