The Annals of Statistics
1994, Vol. 22, No. 1, 1–20

THE ORDER OF THE REMAINDER IN DERIVATIVES OF COMPOSITION AND INVERSE OPERATORS FOR p-VARIATION NORMS[1]

BY R. M. DUDLEY

Massachusetts Institute of Technology

Many statisticians have adopted compact differentiability since Reeds showed in 1976 that it holds (while Fréchet differentiability fails) in the supremum (sup) norm on the real line for the inverse operator and for the composition operator $(F, G) \mapsto F \circ G$ with respect to F. However, these operators are Fréchet differentiable with respect to p-variation norms, which for $p > 2$ share the good probabilistic properties of the sup norm, uniformly over all distributions on the line.

The remainders in these differentiations are of order $\|\cdot\|^{\gamma}$ for $\gamma > 1$. In a range of cases p-variation norms give the largest possible values of γ on spaces containing empirical distribution functions, for both the inverse and composition operators. Compact differentiability in the sup norm cannot provide such remainder bounds since, over some compact sets, differentiability holds arbitrarily slowly.

1. Introduction. The theory of differentiable statistical functionals began with work of von Mises [e.g., von Mises (1936, 1947) and Filippova (1961)]. A nonlinear functional T is defined, for example, on distribution functions. Von Mises differentiated T at a distribution function F along lines. For T to have a (Gâteaux) derivative at F means that, in the direction of a function h,

$$T(F + th) = T(F) + tT'(F)(h) + o(|t|) \quad \text{as} \quad t \to 0. \tag{1.1}$$

Here $T'(F)(\cdot)$ is a bounded linear operator on functions h, for example, of the form

$$T'(F)(h) = \int g\,dh \quad \text{for some function } g \text{ (depending on } F\text{).} \tag{1.2}$$

It was well known that such differentiability is most useful if the $o(|t|)$ is uniform for h in some sets. The object was to analyze the behavior of $T(F_n)$ for empirical distribution functions F_n, and the F_n do not approach F along lines as $n \to \infty$. If the $o(|t|)$ is uniform for h in bounded sets for some norm $\|\cdot\|$, or equivalently,

$$T(F + h) - T(F) - T'(F)(h) = o(\|h\|) \quad \text{as } \|h\| \to 0, \tag{1.3}$$

then T is *Fréchet* differentiable at F for $\|\cdot\|$. For distribution functions on

Received October 1991; revised April 1993.

[1]Research partially supported by NSF grants and a Guggenheim fellowship.

AMS 1991 *subject classifications.* Primary 62G30, 58C20, 26A45; secondary 60F17.

Key words and phrases. Fréchet derivative, compact derivative, Hadamard derivative, Gâteaux derivative, Bahadur–Kiefer theorems, Orlicz variation.

the line, the first natural choice for the norm was the *supremum* (sup) norm $\|h\|_\infty := \sup_x |h(x)|$.

In robust statistics [Huber (1981), Sections 1.5 and 2.5] the derivative, or specifically the g in (1.2), has been called the influence function [Hampel, Ronchetti, Rousseeuw and Stahel (1986)]. Differentiability is also an issue in semiparametric statistics [e.g., Gill (1989) and Bickel, Klaassen, Ritov and Wellner (1993)].

In his landmark thesis Reeds [(1976), Section 6.4] considered two functionals (or *operators*, since their values are also functions): the *composition operator* $(F, G) \mapsto F \circ G$, where $(F \circ G)(x) \equiv F(G(x))$, and the *inverse operator*, $H \mapsto H^{-1}$ where $H^{-1}(y) := \inf\{x: H(x) \geq y\}$. Here F and H are in a space $D[0,1]$ or $D(\mathbb{R})$ of right-continuous functions with left limits. The inverse operator takes a distribution function into a quantile function. Then, applying the composition operator with H^{-1} as G, we get $F \circ H^{-1}$, the quantile–quantile function.

In this introduction, let us consider differentiation at the uniform $U[0,1]$ distribution function $F(x) := x$, $0 \leq x \leq 1$. For a fixed y, say $y = \frac{1}{2}$, the quantile functional $H \mapsto H^{-1}(y)$ is not differentiable even at $U[0,1]$ along lines in $D[0,1]$ (as reviewed in Proposition 2.7), although it is for $C[0,1]$ (see Section 3). Even on $C[0,1]$, the inverse operator is differentiable with the sup norm on its range for functions in the range restricted to $[a,b]$, $0 < a < b < 1$, but not on all of $[0,1]$. Reeds then naturally considered the L^p norms, $\|h\|_p := (\int_0^1 |h(y)|^p dy)^{1/p}$, $1 \leq p < \infty$, on the range of the inverse operator. Then, to get the quantile–quantile function, L^p norms of G are taken on the domain of the composition operator.

Easy examples [Proposition 2.8(b), (c)] show that under these conditions the inverse operator and the composition operator with respect to F are not Fréchet differentiable for the sup norm on their domains. Among Reeds' major results are that in these cases the $o(|t|)$ does hold uniformly over *compact* sets in the sup norm. Although Reeds' (1976) thesis was unpublished, his work is the main ingredient of the survey by Fernholz (1983).

Following the work of Reeds and Fernholz, while Fréchet differentiability is used when it holds, statisticians have worked with compact differentiability as apparently the best kind that held in enough generality. However, if the sup norm is replaced by p-variation norms $\|\cdot\|_{[p]}$ (defined in Section 2), Fréchet differentiability holds for both the composition and inverse operators [Dudley (1991)]. Moreover, the norms $\|\cdot\|_{[p]}$ share the excellent properties of the sup norm of being invariant under all strictly monotone continuous transformations of the line onto itself, and they satisfy (for some values of $p, p < 2$ or $p > 2$ depending on which side of duality one is on) uniform central limit theorems (Donsker properties) [Dudley (1992a)].

The present paper treats the size of the remainder in the Fréchet differentiability and finds bounds for our two operators in (1.3), where $o(\|h\|_{[p]})$ is replaced by $O(\|h\|_{[p]}^\gamma)$ for some $\gamma > 1$. For the inverse operator we get (for $F = U[0,1]$, so $F^{-1} = F$)

$$\|(F+g)^{-1} - F^{-1} + g\|_p = O\big(\|g\|_{[p]}^\gamma\big), \quad \text{for } \gamma = 1 + 1/p, \tag{1.4}$$

as a special case of Corollary 2.4. Theorem 2.2 gives a corresponding remainder bound for joint Fréchet differentiability of the composition operator. The powers γ are shown (in Proposition 2.5 and Theorem 2.2) to be optimal, not only for p-variation norms but for an arbitrary norm on distribution functions. Prior to Dudley (1991) and the present paper, the composition and inverse operators under the given conditions had not been shown to be Fréchet differentiable for any norm, to my knowledge. Note that for twice-differentiable functionals one expects an exponent $\gamma = 2$, so our operators are not that smooth if $p > 1$.

Actually, the remainder in differentiation of the inverse operator had been studied earlier and in precise detail in the main case of statistical interest, where $g = F_n - F$ (F_n empirical), by Bahadur (1966) and Kiefer (1967, 1970). Reeds (1976) and Fernholz (1983) did not cite this work of Bahadur and Kiefer. Take any $p > 2$ in (1.4), giving any $\gamma < \frac{3}{2}$. From $\|F_n - F\|_{[p]} = O_p(n^{-1/2})$ [Dudley (1992a), Corollary 3.8] we get in (1.4) a remainder

$$\left\|F_n^{-1} - F^{-1} + F_n - F\right\|_p = O_p\left(n^{-\gamma/2}\right), \tag{1.5}$$

where $\gamma = 1 + p^{-1}$. Letting $p \downarrow 2$ gives $-\gamma/2 \downarrow -\frac{3}{4}$; or, as $p \uparrow 2$, results of Dudley (1992b) also give $-\frac{3}{4}$ as limiting exponent. Here $-\frac{3}{4}$ is the best possible, as Kiefer proved. See also Theorem 2.6 and Section 5.

The usual space containing distribution functions on [0, 1]—and the one Reeds and Fernholz treat—is the space $D[0, 1]$ of right-continuous functions with left limits. However, $D[0, 1]$ with supremum norm is nonseparable (has no countable dense subset). In such a space compact differentiability has some substantial drawbacks:

1. *Compact sets—in a nonseparable normed space—are extremely small.* The notion that compact differentiability is good when Fréchet differentiability fails may have come from separable spaces, where indeed the compact sets are about as large as one can expect, short of bounded sets. However, for any sup-norm-compact set K in $D[0,1]$ and for any nonatomic distribution F, the probability that the empirical distribution function F_n is in K is 0. How inconvenient this is may be indicated by the variety of ingenious devices statisticians have applied to getting around it, including smoothing, "tangential" differentiability and use of almost surely convergent realizations.
2. Unlike the Fréchet case, compact differentiability does not give helpful uniform error bounds like (1.4). To the contrary, the rate of $o(|t|)$ in (1.1) depends on the compact set and can be arbitrarily slow, as Proposition 2.1 shows. In other words, in contrast to drawback 1, *some compact sets are inconveniently large.* One can think of them as only extending out in a separable and hence very small set of directions in the nonseparable space, but, in those directions, sticking out far enough to slow down the $o(|t|)$ just short of making it fail. When applied to empirical distribution functions (as it can be but, by drawback 1, only via some modification), compact differentiability would only give a remainder $o_p(n^{-1/2})$ even when a better rate holds.

On the other hand, in separable spaces compact differentiability may be more useful although drawback 2 will still hold there.

There are many statistical functionals or operators for which compact differentiability was the first to be proved. For some of these, Fréchet differentiability (even allowing different norms) has not yet been proved. An important example of such an operator is the product integral [Gill and Johansen (1990)].

One of Reeds' main points was to separate analytical from probabilistic parts of the work, although the asymptotic results obtained may not be the best possible. The p-variation norms evidently come closer to giving optimal results, although there are limitations; (1.4) is sharp as it stands, but when applied to empirical distribution functions in (1.5) it is not: (1.5) holds (as Kiefer proved) with $p = \infty$ on the left and $O_p(n^{-3/4}(\log n)^{1/2})$ on the right.

The composition and inverse operators would seem to be of interest not only in statistics but in analysis more generally. However, analysts seem not to have developed the theory needed. They have treated the composition operator almost exclusively for fixed F, as a special case of the so-called Nemitsky operator, surveyed by Appell and Zabrejko (1990). Also, analysts seem to have given little if any attention to differentiability of the inverse operator in cases where H is not 1–1, but empirical distribution functions are never 1–1.

The main results will be stated in Section 2 and proved in Section 4. Section 3 compares different kinds of differentiability as applied to empirical distribution functions, and Section 5 treats Orlicz variation and martingales.

2. Statements of results. There are a number of possible definitions of differentiability. Let X be a vector space with a norm $\|\cdot\|$ and let T be a functional defined on an open set U in X. (Here we have in mind for X a function space containing distribution functions on $\mathbb{R}$.) Let T take values in another vector space Y with a norm $|\cdot|$. Note that values of the composition and inverse operators are again functions.

T is said to be Gâteaux differentiable at a point $x_0 \in U$ if there is a bounded linear operator $T'(x_0)(\cdot)$ from X into Y such that, for each $v \in X$,

$$|T(x_0 + tv) - T(x_0) - tT'(x_0)(v)| = o(|t|) \quad \text{as } t \to 0. \tag{2.1}$$

The chain rule for composition fails for Gâteaux differentiability. Filippova and von Mises considered Gâteaux differentiability, with further probabilistic hypotheses.

Let $\mathcal{C}$ be a class of bounded subsets of X containing all the finite sets. Then T will be called *$\mathcal{C}$-differentiable* at x_0 if (2.1) holds uniformly for $v \in A$ for all $A \in \mathcal{C}$ [Sebastião e Silva (1956)]. If $\mathcal{C}$ is the family of all bounded sets in X, or equivalently if $\mathcal{C}$ consists of the one set $B_1 := \{x \in X\colon \|x\| \le 1\}$, then T is said to be *Fréchet* differentiable at x_0. Then we have more simply

$$|T(x_0 + u) - T(x_0) - T'(x_0)(u)| = o(\|u\|) \quad \text{as } \|u\| \to 0. \tag{2.2}$$

This is the main kind of differentiability in use among analysts. It is the strongest possible kind of $\mathcal{C}$-differentiability, since sets in $\mathcal{C}$ are bounded. (If $\mathcal{C}$

contains an unbounded set, e.g., in a one-dimensional space X, $\mathcal{C}$-differentiability would imply that T is linear at least in a one-sided neighborhood of x_0.) There seems to be little doubt that Fréchet's is the most useful form of differentiability, when it holds.

Alternatively, if $\mathcal{C}$ is the class of all compact sets in X, we get *compact* or *Hadamard* differentiability, emphasized by statisticians since the work of Reeds (1976) and Fernholz (1983). In the Introduction and in detailed results below are shown the advantages of Fréchet over compact differentiability. However, the Reeds–Fernholz theory can also be extended in a very different direction. One can enlarge the class of compact sets for the sup norm to much larger classes over which $\mathcal{C}$-differentiability still holds: the class of compact sets for the Skorohod topology or the still much larger class of "uniformly Riemann" sets [Dudley (1991), Theorems 4.5 and 5.1]. In other words, referring to drawbacks 1 and 2 of (sup-)norm-compact differentiability in the Introduction, if one is willing to accept the arbitrarily slow remainder rates (drawback 2), it is far from necessary to accept also drawback 1. The Skorohod-compact sets do carry the normalized empirical distribution functions $n^{1/2}(F_n - F)$ with high probability, uniformly in n. The slow rates for compact (and not Fréchet) differentiability are given as follows:

PROPOSITION 2.1. *Let $(X, \|\cdot\|)$ be a normed linear space and $x \in X$. Let T be a function from a neighborhood of x into a normed space $(Y, |\cdot|)$, Gâteaux but not Fréchet differentiable at x. Then, for any sequence $\varepsilon_n \downarrow 0$ (however slowly), there exist a sequence $v_n \to 0$ in X and numbers t_n such that $0 < t_n \le 1/n$ and*

$$\left|T(x + t_n v_n) - T(x) - t_n T'(x)(v_n)\right| \ge \varepsilon_n t_n.$$

(Proofs are given in Section 4.) Compact differentiability for a given norm is presumably of interest mainly when Fréchet differentiability fails. Then, however, since the set $\{0\} \cup \{v_n\}_{n\ge 1}$ from Proposition 2.1 is compact, the remainder in equation (2.1) is $o(|t|)$ only very slowly on some compact sets.

In this paper, Fréchet differentiability will be considered with respect to ψ-variation norms $\|f\|_{[\psi]} := \|f\|_\infty + \|f\|_{(\psi)}$, where $\|f\|_\infty := \sup_x |f(x)|$, $\|f\|_{(\psi)} := \inf\{K > 0\colon\ v_\psi(f/K) \le 1\}$ and

$$v_\psi(f) := \sup\left\{\sum_i \psi\big(|f(x_i) - f(x_{i-1})|\big)\colon x_0 < x_1 < \cdots\right\},$$

and where ψ is assumed to be a convex, strictly increasing function on $[0, \infty[$, with $\psi(0) = 0$, $\lim_{x\downarrow 0} \psi(x)/x = 0$ and $\lim_{x\uparrow\infty} \psi(x)/x = +\infty$. Such a ψ will be called an *Orlicz function*. Here $v_\psi(f)$ is called the *ψ-variation* of f. $\|\cdot\|_{[\psi]}$ is indeed a norm: subadditivity is proved, for example, in Musielak and Orlicz [(1959), Section 3.03], and the other properties of a norm are clear.

The best known cases are where $\psi(x) = x^p, 0 \le x < \infty, 1 \le p < \infty$. Here x^p is an Orlicz function for $1 < p < \infty$, while $\|f\|_{(1)}$ is the usual total variation of f. Then ψ-variation is called p-variation and written $v_p(f)$, and we set $\|f\|_{(p)} := \|f\|_{(\psi)} = v_p(f)^{1/p}$, $\|f\|_{[p]} := \|f\|_{[\psi]}$.

The notions of Orlicz-variation and p-variation are not very familiar. I did not know of them in advance, but found that they seemed to arise naturally in the research being reported here and in Dudley (1991), (1992a) and (1992b). One can arrive at them perhaps as follows. For the composition operator, one has a remainder term $f(G+g) - f(G)$ which one would like to be $o(|f| + |g|)$ in some sense. In showing differentiability one might think f itself should be differentiable, but a moment's thought suggests that one only needs some continuity property of f so that $f(G+g) - f(G)$ is smaller than f itself. Having a bounded derivative is equivalent to satisfying a Lipschitz condition $|f(x) - f(y)| \leq C|x-y|$. For present purposes it will be enough if f satisfies a weaker *Hölder condition* of order α, $|f(x) - f(y)| \leq C|x-y|^{\alpha}$ for some $\alpha, 0 < \alpha < 1, C < \infty$. (This is one way of defining differentiability of fractional order α.)

Next, if we want a function space suitable for nonparametric statistics, we would like a space, with a norm, invariant under all increasing homeomorphisms—in other words strictly increasing, continuous functions from the line onto itself—just as the space $D(\mathbb{R})$ of bounded, right-continuous functions with left limits and supremum norm is invariant. Also, to deal with empirical distribution functions and the like, we would like functions that are not necessarily continuous. We can satisfy both these desires by taking the set of all functions $f \circ g$ where f is α-Hölder and g is nondecreasing and bounded (not necessarily continuous). The set of such $f \circ g$ is exactly the set of functions of bounded p-variation for $p = 1/\alpha$ [e.g., Dudley (1992a), proof of Theorem 2.1].

The test of a function space and norm, like any other mathematical objects, is how well they serve their purposes, and the results given here and in Dudley (1991) and (1992a) seem to show that p-variation spaces and norms serve well for differentiability of some interesting functionals of empirical distribution functions.

One may then ask why one needs the more complicated-looking Orlicz variation beyond just p-variation. There are two reasons. One is that Orlicz functions allow one to refine and interpolate between the p-variation norms; for example, x^p can be multiplied by logarithmic factors. The best possible Orlicz function may give somewhat better results than any particular x^p, as will be seen in Section 5. Second, one can take Orlicz functions ψ decreasing to 0 as $x \downarrow 0$ faster than any power. Classes of bounded ψ-variation can then be larger than classes of bounded p-variation. It turns out that each set in a very large class of sets called *uniformly regulated* sets, among which are all compact sets for the Skorohod topology (hence also for the supremum norm) in $D[0,1]$, is of bounded ψ-variation for some ψ [Dudley (1991), Theorems 2.2 and 2.6], which is not the case for p-variation.

Returning to the present development, an Orlicz function ψ is said to satisfy the Δ_2 condition if, for some constant $K < \infty, \psi(2x) \leq K\psi(x)$ for all $x > 0$. For the space of measurable functions f on a finite measure space such that $\int \psi(|f|)\, d\mu < \infty$ (Orlicz class), the Δ_2 condition is only important for large x, say $x \geq x_0 > 0$ [see, e.g., Krasnosel'skii and Rutickii (1961), page 23]. For ψ-variation spaces, the condition is only needed for small x, say $0 < x \leq x_0$ [Musielak and Orlicz (1959)], but here, for simplicity, it is assumed for all

$x > 0$. Clearly, the functions x^p satisfy $\Delta_2, 1 < p < \infty$.

If an Orlicz function ψ satisfies Δ_2, then by convexity we have a "quasi-subadditivity" property: for any $x, y > 0$,

$$\psi(x+y) = \psi((2x+2y)/2) \leq (\psi(2x)+\psi(2y))/2 \leq K(\psi(x)+\psi(y))/2.$$

Fréchet differentiability holds for the operators $(f,g) \mapsto (F+f)\circ(G+g)$ and $f \mapsto (F+f)^{-1}$ with respect to the $\|\cdot\|_{[\psi]}$-norm of f under suitable conditions on F, G and g [Dudley (1991), Theorems 4.5 and 5.1]. Here, in some cases, the $o(\cdot)$ condition will be improved to $O(\|f\|_{[p]}^{\eta})$ for some $\eta > 1$ depending on p and other hypotheses. For the inverse operator there are bounds for the remainder in terms of p- and ψ-variation (Theorems 2.3 and 2.6 and Section 5).

Let $V_p(\mathbb{R})$ be the set of all real-valued functions f on $\mathbb{R}$ for which $\|f\|_{(p)} < \infty$. In the differentiation of the composition operator, one of the two remainder terms [Dudley (1991), Section 5] is $f(G+g) - f(G)$. If f is Lipschitz, with $\|f\|_L = K$, for example, if f has a bounded derivative with $\|f'\|_\infty = K$, and if g is bounded, then we have

$$\|f(G+g) - f(G)\|_\infty \leq K\|g\|_\infty \leq (\|f\|_L^2 + \|g\|_\infty^2)/2.$$

I have found in the literature outside of statistics two papers on joint differentiability of composition of nonlinear functions, with respect to norms: Gray (1975) and Brokate and Colonius (1990). Both papers took g bounded and f with bounded derivative, at least locally, with f and g both possibly Banach-valued. Reeds (1976) and Fernholz (1983) treat the much deeper case where f need not be Lipschitz and g varies in an L^p space and so may be unbounded. As far as I know, only statisticians have worked on these cases.

Let μ be a finite measure. In a composition $f\circ g$ where f is bounded, the size of large values of g is not important. It will turn out that for g the pseudometric

$$\rho_p(g_1, g_2) := \left(\int (\min(1, |g_1 - g_2|))^p \, d\mu\right)^{1/p}$$

works the same as the usual $L^p(\mu)$ distance for $1 \leq p < \infty$. The function $m(y) := \min(y, 1)$ for $y > 0$ is nondecreasing, concave and so subadditive. By the Minkowski inequality [e.g., Dudley (1989), Theorem 5.1.5], ρ_p is indeed a pseudometric. It is easily seen to metrize convergence in measure.

The other remainder term, $F(G+g) - F(G) - F'(G)g$, does not depend on f, so the following will give an explicit bound for the dependence of the remainder on f, and some optimality properties of the bound:

THEOREM 2.2. *Let $\beta > 0$. Let G be an increasing function on $[0,1]$ with $G(y) - G(x) \geq \beta(y-x)$, for $0 \leq x \leq y \leq 1$. Let $1 \leq p < \infty$ and $1 \leq s < \infty$. Then there is a constant $C < \infty$ such that, for $f \in V_p(\mathbb{R})$ and $g \in L^s[0,1]$,*

$$\|f(G+g) - f(G)\|_p \leq C\|g\|_s^{s/(p(1+s))}\|f\|_{[p]} \leq C(\|g\|_s^{\eta} + \|f\|_{[p]}^{\eta}),$$

where $\eta := 1 + s/(p(1+s)) > 1$. *Conversely, if* $\|\cdot\|$ *is any norm on a space* V *of functions containing the function* $h(x) = 1_{\{x>0\}}, U(x) = x, 0 \le x \le 1$, *and there are constants* C, α *and* γ *such that*

$$\|f \circ (U+g) - f \circ U\|_p \le C\|g\|_s^\alpha \|f\|^\gamma,$$

for $\|f\| \le 1$, *then* $\gamma \le 1$ *and* $\alpha \le s/(p(1+s))$. *In both parts of this theorem,* $\|\cdot\|_p$ *can be replaced by* $\rho_p(0,\cdot)$ *and/or* $\|g\|_s$ *by* $\rho_s(0,g)$.

Next, consider the inverse operator. Let G be an increasing, continuous function from $[0,1]$ onto itself, having a derivative everywhere on the open interval $]0,1[$ which extends to a continuous, strictly positive function G' on $[0,1]$. Such a G will be called an increasing *diffeomorphism* of $[0,1]$. Let $g \in R[0,1]$, the space of all bounded real functions on $[0,1]$, continuous almost everywhere for Lebesgue measure [Dudley (1991)]. Then it is known [e.g., Dudley (1991), Theorem 4.5] that the derivative of the inverse operator $g \mapsto (G+g)^{-1}$ at $g = 0$ is the linear operator $g \mapsto -(g \circ G^{-1})/(G' \circ G^{-1})$, bounded for the sup norm on g since by assumption $\inf_{[0,1]} G' > 0$. The remainder in the differentiation is

$$R_g := (G+g)^{-1} - G^{-1} + (g \circ G^{-1})/(G' \circ G^{-1}) \quad \text{on } [0,1].$$

Let us recall a definition of modulus of continuity. For a function f and $\delta > 0$, let $\omega(\delta, f) := \sup\{|f(x) - f(y)|: |x-y| \le \delta\}$. We then have the following.

THEOREM 2.3. *For any increasing diffeomorphism* G *of* $[0,1]$, *so that* $0 < \beta := \inf G' \le \eta := \sup G' < \infty$, *and for any Orlicz function* ψ *satisfying* Δ_2, *there are constants* $C_1, C_2 < \infty$ *such that, for* $\gamma := \|g\|_\infty$,

$$\int_0^1 \psi(|R_g(y)|)\,dy \le C_1 \left\{ \gamma \left[v_\psi\left(\frac{g}{\beta}\right) + \psi(C_2\gamma) \right] + \psi\left(2\gamma\beta^{-2}\omega\left(\frac{10\gamma}{\beta}, G'\right)\right)\right\}.$$

For $\psi(u) \equiv u^p, 1 \le p < \infty$, *we have*

$$\|R_g\|_p \le C\left\{ \gamma^{1/p}\|g\|_{[p]} + \frac{\gamma}{\beta^2}\omega\left(\frac{10\gamma}{\beta}, G'\right)\right\}.$$

If G satisfies a stronger smoothness condition, specifically a Hölder condition on G', then, taking $y(x) = x^p$, Theorem 2.3 yields the following.

COROLLARY 2.4. *If in addition to the hypotheses of Theorem 2.3 we have* $|G'(x) - G'(y)| \le D(y-x)^{1/p}$ *for* $0 \le x \le y \le 1$, *where* D *is a constant, then there is a constant* $K < \infty$ *such that*

$$\|R_g\|_p \le K\|g\|_{[p]}^{1+1/p}, \quad \text{for } g \in V_p[0,1], \|g\|_{[p]} \le 1.$$

Now, an example will show that for a given value of p and any norm $\|\cdot\|$, $\|R_g\|_p \le K_p\|g\|^\alpha$ implies $\alpha \le 1 + 1/p$, since a remainder in Gâteaux differentiability along one line is of order $t^{1+1/p}$. Thus, by Corollary 2.4, $\|\cdot\| = \|\cdot\|_{[p]}$ achieves the largest possible power of the norm for the remainder in this case.

PROPOSITION 2.5. *Let $f = 1_{[a,b]}, 0 < a < b < 1$. Then for $0 < t < \min(b-a, 1-b)$ and $U(x) \equiv x, 0 \le x \le 1$,*

$$\left\|(U+tf)^{-1} - U + tf\right\|_p = C_p t^{1+1/p},$$

for some constant C_p. Thus, for any normed space $(y, |\cdot|)$ of functions on $[0,1]$ containing U and $1_{[a,b]}$, if the inverse operator is Fréchet differentiable at U from $(y,|\cdot|)$ to $L^p[0,1]$ with $\|(U+g)^{-1} - U + g\|_p = O(|g|^\alpha)$ as $|g| \to 0$ in Y, then $\alpha \le 1 + 1/p$.

For a probability distribution function F and its empirical distribution functions F_n, $n^{1/2}\|F_n - F\|_{[p]}$ is bounded in probability uniformly in n and F for $p > 2$ [Dudley (1992a), Corollary 3.8]; another proof is given in Section 5. "Bahadur–Kiefer theorems" treat the size of the remainder R_g in Theorem 2.3 when $G \equiv F$ and $g \equiv g_n \equiv F_n - F$ [Bahadur (1966) and Kiefer (1967, 1970)]. Most of the results treat the case where F is the uniform distribution function U on $[0,1]$. Here, $G' \equiv 1$ on $[0,1]$, so $\omega(\cdot, G') \equiv 0$. So Corollary 2.4 gives $\|R_{g_n}\|_2 \le \|R_{g_n}\|_p = O_p(n^{(-1/2)(1+1/p)})$. Letting $p \downarrow 2$, we get $\|R_{g_n}\|_2 = O_p(n^{-t})$ for any $t < \frac{3}{4}$. A more precise result, not claimed as new, gives the size of the remainder for the L^2 norm in probability:

THEOREM 2.6. *For the L^2 norm of the Bahadur–Kiefer remainder we have $\|R_{g_n}\|_2 = O_p(n^{-3/4})$.*

Kiefer [1967, (1.6)] showed that, for each y, $n^{3/4}R_{g_n}(y)$ has a specific nontrivial limiting distribution (depending on y). It follows that $O_p(n^{-3/4})$ cannot be improved. For the supremum norm, Kiefer (1970) found a limiting distribution for $n^{3/4}\|R_{g_n}\|_\infty/(\log n)^{1/2}$ and proved for $G = U$ that

$$\limsup_{n\to\infty} \|R_{g_n}\|_\infty n^{3/4}(\log n)^{-1/2}(\log\log n)^{-1/4} = 2^{-1/4}$$

almost surely. Shorack (1982) gave another proof. Let $\alpha_n := n^{1/2}(G_n - G)$. Kiefer (1970) proved that $n^{3/4}\|R_{g_n}\|_\infty/(\|\alpha_n\|_\infty \log n)^{1/2} \to 1$ in probability. Shorack (1982) in one direction, and Deheuvels and Mason (1990) in the other, showed that this convergence is almost sure. Thus the asymptotic magnitude of $\|R_{g_n}\|_\infty$ follows from that of $\|\alpha_n\|_\infty$. Kiefer (1970) also proved his results for G with second derivative G'' bounded above as well as G' bounded below by $\beta > 0$, replacing R_{g_n} by $(G' \circ G^{-1})R_{g_n}$. See also Csörgő and Révész [(1978), Theorem 4, page 891].

Now let $\|\cdot\|$ be any norm on a function space containing the centered empirical distribution functions $F_n - F$ and such that $n^{1/2}\|F_n - F\|$ is measurable and does not converge to 0 in probability. (This seems a rather mild assumption on $\|\cdot\|$.) Then if $\|F_n - F\| = O_p(n^{-\beta})$, we have $\beta \le \frac{1}{2}$. This and the second half of Proposition 2.5 show that for $p > 2$ (unlike $p < 2$) a separation of analytic and probabilistic methods does not given optimal results: For $2 < p < \infty$, Kiefer's correct order of magnitude (in probability),

$$\|U_n^{-1} - U + (U_n - U)\|_p = O_p(n^{-3/4}),$$

does not follow by combining Fréchet differentiability of the inverse operator in the norm $\|\cdot\|$ (even with the best possible power bound for the remainder) with an estimate $\|U_n - U\| = O_p(n^{-\beta})$.

Next, it will be seen that the inverse operator, even at such a smooth distribution F as the uniform, is not differentiable in a neighborhood of F (even in the Gâteaux sense). So the inverse operator is not C^1 and has no second derivative. The last statement (at least) in the following proposition was essentially known [Fernholz (1983), page 66, gives a related example], but it recalls a notable fact.

PROPOSITION 2.7. *Let* $U(x) = x$, $0 \le x \le 1$, *and* $f(x) := 1_{[0,1/2]}(x)$. *Then, for any fixed* $t > 0$, $g \mapsto (U + tf + g)^{-1}$ *is not Lipschitz for the* L^p *norm on the range,* $1 < p < \infty$, *even along the same direction* $g = uf$ *as* $u \downarrow 0$, *and so is not Gâteaux differentiable in that direction. For* $p = \infty$ *the operator is not continuous. For* $p = 1$ *it is Lipschitz along the given line but still not Gâteaux differentiable. Allowing also* $t < 0$ *and for fixed* $y = \frac{1}{2}$, $g \mapsto (U + g)^{-1}(y)$ *is not Gâteaux differentiable along the line* $g = tf$, $t \to 0$.

So, by the way, despite the remainder bound, the inverse operator on $D[0,1]$ at U does not satisfy (in any norm) the definition of "smooth" given in Wong and Severini [(1991), page 610].

Next are easy examples showing non-Fréchet differentiability of the composition and inverse operators with respect to the sup norm, variously on domains and/or ranges. Such examples must have been known although I do not have references for them.

PROPOSITION 2.8. *At* $F = G = U[0,1]$, *the following hold.*

(a) *The composition operator is not jointly Gâteaux differentiable on* $D[0,1]$ *for the* sup *norm on the range: for some* $f, g \in D[0,1]$,

$$\|tf \circ (G + sg) - tf \circ G\|_\infty \ne o(|t| + |s|) \quad \text{as } s, t \to 0.$$

(b) *The composition operator is not jointly differentiable into* L^2 *for the* sup *norm on the domain*

$$\|f \circ (G + g) - f \circ G\|_2 \ne o(\|f\|_\infty + \|g\|_\infty) \quad \text{as } \|f\|_\infty + \|g\|_\infty \to 0.$$

(c) *The inverse operator is not Fréchet differentiable at* G *for the* sup *norm on the domain and the* L^2 *(or* sup*) norm on the range*

$$\|(G + g)^{-1} - G^{-1} + g\|_2 \ne o(\|g\|_\infty) \quad \text{as } \|g\|_\infty \to 0.$$

3. Kinds of differentiability. This section will list several forms of differentiability. The question will be which forms are better adapted to proving asymptotic normality of suitable functionals of the empirical d.f. F_n, say on $[0, 1]$. Let us consider differentiation at a sufficiently regular distribution such as the uniform distribution $U[0, 1]$. Three kinds of Fréchet differentiability are as follows:

(FF) for functionals with finite-dimensional values, Fréchet differentiability for functions between finite-dimensional spaces;
(FP) Fréchet differentiability for norms $\|\cdot\|$ stronger than the sup norm, such as p-variation norms, on subspaces of $D[0, 1]$ such that $\|F_n - F\|$ is finite, or better still $O_p\left(n^{-1/2}\right)$;
(FS) Fréchet differentiability for the sup norm on $D[0, 1]$.

Here (FS), when it holds, always implies (FP), but neither necessarily implies nor is implied by (FF), since no relation is assumed between the norm on the finite-dimensional space and any norm on distribution functions.

(FF) has the advantage, when it holds, that in finite dimensions we can often hope for more than one derivative, giving multiterm Taylor expansions, while the nonlinear functionals of interest on infinite-dimensional spaces tend to have no more than one derivative, as seen in Section 2.

Three kinds of differentiability of compact type are as follows:

(CC) compact differentiability in $C[0, 1]$ for the sup norm, applied to empirical distribution functions by replacing them, for example, by piecewise-linear rather than piecewise-constant functions;
(CD) compact differentiability in $D[0, 1]$ for the sup norm, applied by way of almost surely convergent realizations;
(CR) $\mathcal{C}$-differentiability, where $\mathcal{C}$ may be the class of sets compact for the Skorohod topology, or the still larger uniformly regulated or uniformly Riemann sets as defined in Dudley (1991).

Here (CR) implies (CD) implies (CC) and neither converse holds. All the compact forms share the disadvantage of slow convergence of remainders to 0, as shown in Proposition 2.1, when Fréchet differentiability fails. When all three hold, (CR) has the advantage that normalized empirical distribution functions $n^{1/2}(F_n - F)$ belong, with high probability uniformly in n, to sets of the classes mentioned in (CR), without the need for a.s. convergent realizations as in (CD) or smoothing as in (CC).

The forms of differentiability will be compared for some functionals of interest.

3.1. *Specific quantiles and the sup norm on the range.* The functional $F \mapsto F^{-1}(y)$ for a fixed y, as seen in the last statement of Proposition 2.7, is not even Gâteaux differentiable on $D[0, 1]$, so (CD) and (FP) and the stronger forms (FS) and (CR) must all fail. (FF) applies as follows: If F and G are any two distribution functions with $F(x_\alpha) = G(u_\alpha) = \alpha$, $F'(x_\alpha) > 0$ and $G'(u_\alpha) > 0$, then asymptotic normality of the sample α quantile of F follows from that for

G by the one-dimensional delta-method. One interesting choice of G is standard exponential, when the order statistics are partial sums of independent exponential variables with different scale parameters, as in a classic paper of Rényi (1953). Rényi's representation helps to treat the joint distribution of several order statistics. The more standard choice is to take G as uniform $U[0,1]$. Then the distribution function at x of the rth order statistic for sample size n and a general distribution function F is the beta$(r, n-r+1)$ distribution function at $F(x)$ [e.g., Kendall and Stuart (1977), Section 14.2, page 347]. The simple normal approximation to the beta can be much improved by a Cornish–Fisher expansion [e.g., Pratt (1968), page 1467, Molenaar (1970), page 72, or Holt (1986)]. The expansion can then be combined with a Taylor expansion of F around x_α.

The compact differentiability (CC) also applies to individual quantiles as shown by Reeds [(1976), page 127] and Esty, Gillette, Hamilton and Taylor (1985). Gill [(1989), page 107] proves compact differentiability in D, tangentially to C, a form intermediate between (CC) and (CD) which applies to empirical distributions. The compact forms seem to give less precise information than (FF).

The supremum norm on the range of the operator $F \mapsto F^{-1}$ can be treated in the same way if one restricts to an interval $[a, b]$, $0 < a < b < 1$ [e.g., Esty, Gillette, Hamilton and Taylor (1985) and Gill (1989)]. It has also been treated on $(0, 1)$ via classical (stochastic) methods [e.g., Csörgő and Révész (1978)].

3.2. *The inverse operator, and composition operator with respect to f, for L^p norms on the range.* Reeds showed that here (CD) holds while (FS) fails. We now have that (FP) holds with the error bounds given in Section 2. Also, all the given forms of (CR) hold for these operators [Dudley (1991)]. The main theme of the present paper is that (FP) is preferable because of the remainder bounds.

3.3. *The operator $(F, G) \mapsto \int F\,dG$.* Gill [(1989), pages 110–111] showed that (CD) applies in this case, while (FS) fails. Now, (FP) also holds, much as in 3.2, and gives error bounds stronger than those from (CD) and nearly but not quite the strongest possible bounds [Dudley (1992b), Corollary 3.5]. The conclusion only follows for some (not all) p-variation norms on F and G. It would seem then that the (CR) forms could only apply to one of F, G, while stronger conditions are put on the other; I have at this point no precise result to report for (CR) in this case.

3.4. *The product integral.* This integral had been applied in other fields, especially differential equations [Dollard and Friedman (1979)], but seems to have been almost unknown to statisticians until recent years, when Gill and Johansen (1990) proved that (CD) applies under suitable conditions. As with Reeds' work on the composition and inverse operators, it seems that again statisticians were the first to prove a functional differentiability fact which should be of interest to nonlinear analysts. Gill and Johansen's work was an outstanding success for compact differentiability.

The product integral is related to $\int F\,dG$, where F and G are operator-valued functions, on which see Krabbe (1961). Freedman [(1983), Theorem 5.1] shows that a product integral exists for operator-valued functions of bounded p-variation for $p < 2$. I do not know at this writing whether the (FP) differentiability holds for some p.

In the (many) cases where Fréchet differentiability has not (yet) been proved for suitable norms, there are a number of useful positive results for compact differentiability, such as validity of the bootstrap [van der Vaart and Wellner (1994)] and preservation of asymptotic efficiency of estimators of (infinite-dimensional) parameters [van der Vaart (1991)]. About efficiency, slow remainders (Proposition 2.1) may be a concern.

One useful property of a norm is the central limit theorem (Donsker property) which has been proved for the ordinary empirical process in p-variation norms for some values of p in Dudley (1992a). Vervaat (1972) showed that, for any sequence $\{G_n\}_{n\geq 1}$ of nondecreasing stochastic processes on $[0,\infty)$, $I(t) \equiv t$, and any sequence $[\delta_n]$ of positive random variables converging to 0 in probability, $(G_n - I)/\delta_n$ converges weakly in $D[0,\infty)$ to a continuous stochastic process ξ with $\xi(0) = 0$ if and only if $(G_n^{-1} - I)/\delta_n$ converges weakly in the same space to $-\xi$. Whitt (1980) noted that $\xi(0) = 0$ was needed. Vervaat and later others applied his and further results to processes including partial sum and counting processes. I do not know at this writing what can be done with p-variation in such situations.

4. Proofs. We first have the following.

PROOF OF PROPOSITION 2.1. Since T is not Fréchet differentiable, there are $\delta > 0$ and $u_k \in X$ with $\|u_k\| \leq 1$ such that $u_k \to 0$ as $k \to \infty$ and

$$|T(x+u_k) - T(x) - T'(x)(u_k)| \geq \delta\|u_k\|.$$

We can assume that $\delta < 1$ and $\varepsilon_n \geq 1/n$ for all n. Take a subsequence u_{k_n} such that $\|u_{k_n}\| \leq 1/n^2$ for all n. Let $t_n := \|u_{k_n}\|\delta/\varepsilon_n$. Then $0 < t_n \leq \delta/n < 1/n$. Let $v_n := u_{k_n}/t_n$. Then $\|v_n\| = \varepsilon_n/\delta \to 0$, and

$$\|T(x+t_n v_n) - T(x) - t_n T'(x)(v_n)\| \geq \varepsilon_n t_n. \qquad \square$$

Next, to help in the proof of Theorem 2.2, we have the following.

LEMMA 4.1. *Let f be a real-valued function on $\mathbb{R}$ and let G be a real-valued function on $[0,1]$. Suppose that, for some $\beta > 0$, $G(y) - G(x) \geq \beta(y-x)$ for $0 \leq x \leq y \leq 1$. Let J be the smallest integer greater than or equal to $1/\beta$. Let m be a positive integer, $k = 1,\dots,m$, and $I_{m,k} := [(k-1)/m, k/m[$. Then, for any Orlicz function ψ,*

$$\sum_{k=J+1}^{m-J} \sup\left\{\psi\left(f(G(x)) - f(G(x)+g_k)\right): g_k \leq \frac{1}{m}, x \in I_{m,k}\right\} \leq (2J+1)v_\psi(f).$$

PROOF. If $x_i \in I_{m,i}$, $i = k,j$, and $j - k \geq 2J + 1$, then $x_j - x_k \geq 2J/m$ and $G(x_j) - G(x_k) \geq 2/m$, so the intervals with endpoints $G(x_i)$ and $G(x_i) + g_i$ are nonoverlapping for $i = j,k$. It follows that the sum in the statement restricted to indices $J+1, 3J+2, 5J+3, \ldots$ is at most $v_\psi(f)$. Likewise for the restrictions to indices $J+i, J+i+2J+1, J+i+4J+2, \ldots$, for $i = 1,2,\ldots,2J+1$, and the conclusion follows. □

Recall that $v_\psi(f) = v_p(f) = \|f\|_{(p)}^p$ when $\psi(u) \equiv u^p$.

PROOF OF THEOREM 2.2. For $f \in V_p(\mathbb{R})$, f is continuous except for at most countably many jumps, so f is Borel measurable, and functions $f(G)$ and $f(G+g)$ are measurable. Let $\delta := \|g\|_s$ and $\alpha := s/(1+s)$. We can assume $\delta \leq 1$, or take $C = 2$. Then $|g| \leq \delta^\alpha$ on a set $A \subset [0,1]$ whose complement has Lebesgue measure at most $\delta^{s(1-\alpha)} = \delta^\alpha$. The same holds if $\delta = \rho_s(0,g)$. Take a positive integer m such that $1/(2m) \leq \delta^\alpha \leq 1/m$. We can assume that $\beta \leq 1$. Let J be the smallest integer greater than or equal to $1/\beta$ and $B := A \cap [J/m, 1-J/m]$. Now

$$\int_{[0,1]\setminus B} |f(G+g) - f(G)|^p(x)\,dx \leq 2^p\|f\|_\infty^p\left(\delta^\alpha + \frac{2J}{m}\right). \tag{4.1}$$

Also, letting $B_{m,k} := B \cap I_{m,k}$, with $I_{m,k} := [(k-1)/m, k/m[$ as before, we have

$$\begin{aligned}\int_B |f(G+g) - f(G)|^p\,dx &= \sum_{k=J+1}^{m-J}\int_{B_{m,k}} |f(G+g) - f(G)|^p\,dx\\ &\leq \sum_{k=J+1}^{m-J}\frac{1}{m}\sup\left\{|f(G+g) - f(G)|^p(x) : x \in B_{m,k}\right\}\\ &\leq \left(\frac{2J+1}{m}\right)\|f\|_{(p)}^p \quad \text{(by Lemma 4.1)}\\ &\leq (2J+1)\,2\delta^\alpha\|f\|_{(p)}^p.\end{aligned}$$

It follows that $\int_0^1 |f(G+g) - f(G)|^p\,dx \leq C_1\|g\|_s^\alpha\|f\|_{[p]}^p$, where

$$C_1 := 4J + 2 + 2^p(1+4J) \leq \frac{4}{\beta} + 6 + 2^p\left(5 + \frac{4}{\beta}\right) \leq 2^p\left(11 + \frac{8}{\beta}\right),$$

so the first inequality in the statement holds, with $C := 22 + 16/\beta$.

Next, for any $A, B, \rho > 0$, we have $AB^\rho \leq A^{1+\rho} + B^{1+\rho}$ since $1 \leq (A/B)^\rho + B/A$. The second inequality in the statement then follows. Also, we have $\rho_p(0,\eta) \leq \|\eta\|_p$ for any L^p function η.

Now in the converse direction, take $f = th$ as $t \to 0$. For fixed g, we get $\gamma \leq 1$. Then let $g = -\delta 1_{[0,\delta]}$ as $\delta \downarrow 0$, so $\|g\|_s = \rho_s(0,g) = \delta^{1+1/s}$ and

$$\rho_p(f \circ (U+g), f \circ U) = \|f \circ (U+g) - f \circ U\|_p = \delta^{1/p}.$$

The conclusions follow. □

PROOF OF THEOREM 2.3. For any real function f on an interval $[a,b]$ let $\mathrm{osc}_{[a,b]}f := (\sup - \inf)_{[a,b]}(f)$. For a given y let $\xi := G^{-1}(y)$. Take a positive integer m such that $1/(2m) \leq \gamma := \|g\|_\infty \leq 1/m$. The assumptions imply that $\beta \leq 1$. Let $a := \xi - 1/(m\beta)$ and $b := \xi + 1/(m\beta)$. If $a \geq 0$ and $b \leq 1$, then by Dudley [(1991), Lemma 4.2],

$$|R_g(y)| \leq \mathrm{osc}_{[a,b]}(g)/\beta + \mathrm{osc}_{[a,b]}G'/(m\beta^2). \tag{4.2}$$

Again, consider the intervals $I_{m,k} := [(k-1)/m, k/m[$. Let J be the least integer greater than or equal to $1/\beta$. Suppose $J+1 \leq k \leq m-J$. Then by (4.2), for $y \in I_{m,k}$ and $I(m,k,J) := [(k-1-J)/m, (k+J)/m[$, we have $a \geq 0$, $b \leq 1$ and

$$\begin{aligned} |R_g(y)| &\leq \beta^{-1}\mathrm{osc}_{I(m,k,J)}\, g + \mathrm{osc}_{I(m,k,J)}G'/(m\beta^2) \\ &\leq \beta^{-1}\mathrm{osc}_{I(m,k,J)}\, g + \omega((2J+1)/m, G')/(m\beta^2), \end{aligned} \tag{4.3}$$

and we note that $(2J+1)/m \leq (3+2/\beta)/m$. Recall that ψ satisfies Δ_2 and is therefore quasi-subadditive. Then as in the proof of Lemma 4.1, there is a $K < \infty$ such that

$$\sum_{k=J+1}^{m-J} \sup\{\psi(|R_g(y)|): y \in I_{m,k}\} \leq K\left\{(2J+1)v_\psi\left(\frac{g}{\beta}\right) + m\psi(\Delta)\right\},$$

where $\Delta := \omega((3+2/\beta)/m, G')/(m\beta^2)$. Thus

$$\int_{J/m}^{1-J/m} \psi(|R_g(y)|)\, dy \leq Km^{-1}(2J+1)v_\psi(g/\beta) + K\psi(\Delta).$$

Noting that $m^{-1} \leq 2\gamma$, the latter integral is bounded as in the statement of the theorem. Next, for $0 \leq y \leq J/m$, we have $0 \leq (G+g)^{-1}(y) \leq (J+1)/(\beta m)$ since $G(x) \geq \beta x$, so $(G+g)(x) \geq \beta x - 1/m \geq J/m$ for $x \geq (J+1)/(\beta m)$.

Also, $0 \leq G^{-1}(y) \leq J/(\beta m)$, so $|(G+g)^{-1}(y) - G^{-1}(y)| \leq (J+1)/(\beta m)$, while $\sup|(g \circ G^{-1})/(G' \circ G^{-1})| \leq 1/(m\beta)$, so $|R_g(y)| \leq (J+2)/(m\beta) \leq (\beta^{-1}+3)/(m\beta)$. Thus

$$\int_0^{J/m} \psi(|R_g(y)|)\, dy \leq \left(\frac{J}{m}\right)\psi\left(\frac{\beta^{-1}+3}{m\beta}\right) \leq \frac{4\|g\|_\infty}{\beta}\psi\left(\left(\frac{6}{\beta} + \frac{2}{\beta^2}\right)\|g\|_\infty\right).$$

Since $1/m \leq 2\gamma$ and a similar bound holds for the interval $[1-J/m, 1]$, the first statement in the theorem follows. Then for $\psi(u) \equiv u^p$, taking pth roots and since $(A+B)^{1/p} \leq A^{1/p} + B^{1/p}$ for $A, B \geq 0, 1 \leq p < \infty$, Theorem 2.3 is proved. □

PROOF OF PROPOSITION 2.5. It is easy to check that, for $0 < y < 1$ and $0 < t < \min(1-b, b-a)$,

$$\text{Rem} := (U+tf)^{-1}(y) - y + tf(y) = \begin{cases} -t, & b < y < b+t, \\ a-y+t, & a < y < a+t, \\ 0, & \text{otherwise} \end{cases}$$

(except at finitely many endpoints). Thus

$$\|\text{Rem}\|_p = \left(t^{p+1} + \int_0^t u^p \, du\right)^{1/p} = \left(1 + \frac{1}{p+1}\right)^{1/p} t^{1+1/p}.$$

The second part of the proposition then follows directly. □

PROOF OF THEOREM 2.6. Apply (4.3), where now $\beta = J = 1$ and G' is constant. So we have, for $2 \le k \le m-1$ and $y \in [(k-1)/m, k/m[=: I(m,k)$, that $|R_g(y)| \le \text{osc}_{I(m,k,1)} g$. Let $g = F_n - F$ and square both sides. I claim that for the uniform distribution as here, $F(x) \equiv x$ for $0 \le x \le 1$, the distribution of $\text{osc}_{[a,b]} g$ for $0 \le a < b \le 1$ depends only on $b - a$. The number s of observations X_i in $[a,b]$ is binomial$(n, b-a)$. Given s, the observations are distributed uniformly and independently in $[a,b]$. F_n has a jump of height $1/n$ at each X_i, and $-F$ is decreasing linearly with slope -1 on the interval. The claim follows.

So, consider $k = 2$, and set $[a,b] = A := I(m,2,1) = [0, 3/m]$. We have

$$\begin{aligned}(\text{osc}_A(F_n - F))^2 &= \left(\text{osc}_A\left(F_n - \frac{ms}{3n}F + \left(\frac{ms}{3n} - 1\right)F\right)\right)^2 \\ &\le 2\left(\text{osc}_A\left(F_n - \frac{sm}{3n}F\right)\right)^2 + 2\left(\frac{ms}{3n} - 1\right)^2\left(\frac{9}{m^2}\right).\end{aligned}$$

The expectation of the latter term is $(6/mn)(1-3/m)$. Now $F_n - msF/(3n)$ can be written as $(s/n)(G_s - G)$, where G is the uniform distribution on A. Thus, by the Dvoretzky–Kiefer–Wolfowitz inequality [Dvoretzky, Kiefer and Wolfowitz (1956); see also Shorack and Wellner (1986), page 354], there is an absolute constant C such that the conditional expectation of the former term given s is at most Cs/n^2, so its expectation is less than $3C/(mn)$. So for a constant $C', E((\text{osc}_A(F_n - F))^2) \le C'/(mn)$ and the sum over $m-2$ intervals is at most C'/n. We can deal with the intervals $[0, 1/m[$ and $[1-1/m, 1[$ as in the proof of Theorem 2.3, getting an upper bound of $K\|g\|_\infty^3$ for a constant K. So

$$E\left(\int_0^1 R_{g_n}(y)^2 \, dy\right) = O(n^{-3/2}) \quad \text{and} \quad \|R_{g_n}\|_2 = O_p(n^{-3/4}). \qquad \square$$

PROOF OF PROPOSITION 2.7. Let $b := \frac{1}{2}$. We have for $t, u > 0$,

$$(U+tf)^{-1}(y) = \begin{cases} y, & b+t \le y \le 1, \\ y-t, & t \le y < b+t, \\ 0, & 0 \le y < t; \end{cases}$$

$$(U+(t+u)f)^{-1}(y) - (U+tf)^{-1}(y) = \begin{cases} 0, & b+t+u < y \le 1, \\ -t-u, & b+t < y < b+t+u, \\ -u, & t+u < y < b+t, \\ t-y, & t < y < t+u, \\ 0, & 0 \le y < t. \end{cases}$$

As $u \downarrow 0$ for fixed $t > 0$, we see from the range $b+t < y < b+t+u$ that the operator is not continuous in the $\|\cdot\|_\infty$ norm. The difference equals $-u1_{\{t<y<b+t\}}$, which is linear in u, plus a term $\eta(t,u,y) = (-t-u)1_{\{b+t<y<b+t+u\}}$, plus another term on a disjoint interval. For $1 \le p < \infty$ we then have $\|\eta(t,u,\cdot)\|_p \ge tu^{1/p}$ (so the "remainder" is larger than the "derivative"!). Thus, for $p > 1$, the inverse operator is not Lipschitz and is not Gâteaux differentiable at $U+tf$. For $p = 1$, suppose it were Gâteaux differentiable. Then, for some function $\gamma(t,\cdot)$,

$$\|\eta(t,u,\,\cdot\,) - u\gamma(t,\,\cdot\,)\|_1 = o(|u|) \quad \text{as } |u| \to 0.$$

Then $\gamma(t,y) = 0$ (almost everywhere) for $y > b+t$, but this yields a contradiction.

Lastly, $(U+tf)^{-1}\left(\frac{1}{2}\right) \equiv \frac{1}{2}$ for $t < 0$ and $\frac{1}{2} - t$ for $t > 0$, showing non-differentiability at $t = 0$. □

PROOF OF PROPOSITION 2.8. (a) Let $f \equiv 1_{[0,1/2]}$, $g \equiv 1$ and $s \downarrow 0$. Then $|f(G+s)(x) - f(x)| = 1$ for $\frac{1}{2} - s < x < \frac{1}{2}$.

(b) For $n = 1, 2, \ldots$, let $g_n(x) = 1/(2n)$ for $j/n \le x < (2j+1)/(2n)$, $j = 0, \ldots, n-1$, and $g_n(x) = 0$ otherwise. Let $f_n = g_n$. Then $\|f_n\|_\infty = \|g_n\|_\infty = 1/(2n)$, $f_n \circ (G+g_n) \equiv 0$ and $\|f_n \circ G\|_2 = 1/(2^{3/2}n)$.

(c) For the same g_n and for $(2j+1)/(2n) < y < (j+1)/n, j = 0, \ldots, n-1$, $(G+g_n)^{-1}(y) = y - 1/(2n)$ while $g_n(y) = 0$, so

$$\|(G+g_n)^{-1} - G^{-1} + g_n\|_2 \ge 1/(2^{3/2}n). \qquad \square$$

5. Orlicz variation and martingales. This section will give a proof [quite different from the one in Dudley (1992a), Corollary 3.8] that $n^{1/2}\|F_n - F\|_{(p)}$ is bounded in probability uniformly in n for $p > 2$, and we also mention an Orlicz function to come as close as possible to 2-variation. [In the proof of Theorem 2.6, a kind of 2-variation was finite because the lengths of the intervals, $1/m = O(n^{-1/2})$, went to 0 fast enough.]

Let $\psi_1(u) := u^2/\log\log(1/u)$, for $0 < u \le e^{-e}$. It can be checked by derivatives that ψ_1 can be defined for $u > e^{-e}$ to be an Orlicz function satisfying Δ_2. For the Brownian process $X_\bullet\colon t \mapsto X_t$ on a bounded interval $0 \le t \le T < \infty$, Taylor (1972) showed that $v_{\psi_1}(X_\bullet) < \infty$ a.s., while if $\psi_1(u)/\psi(u) \mapsto 0$ as $u \downarrow 0$, for example, $\psi(u) \equiv u^2/(\log\log(1/u))^\alpha$, $\alpha < 1$, then $v_\psi(X_\bullet) = +\infty$ a.s.

Monroe (1972) showed that any right-continuous martingale process M_t having left limits can be written as X_{T_t} for an increasing family of stopping times $T_t < \infty$. Thus any such M_t has bounded ψ_1-variation on bounded intervals $0 \le t \le s$ a.s. [Monroe (1976), page 134]. If, moreover, the martingale M_t has mean 0 and $EM_s^2 < \infty$, then $ET_s = EM_s^2$ [Monroe (1972), Theorems 5 and 11]. Thus for martingales $M^{(n)}$ on $0 \le t \le s$ having $E((M_s^{(n)})^2)$ uniformly bounded, the stopping times T_s will be bounded in probability, uniformly in n, and likewise for $T_t \le T_s, 0 \le t \le s$. So the ψ_1-variations of $M^{(n)}$ will be bounded in probability uniformly in n.

Let F_n be empirical distribution functions of the uniform distribution $F(t) = t, 0 \le t \le 1$. For each $n, M^{(n)}(t) := n^{1/2}(F_n(t)-t)/(1-t), 0 \le t < 1$, is a martingale [e.g., Shorack and Wellner (1986), page 4 and page 133, Proposition 1]. Then $M^{(n)}$ has mean 0 and variances bounded for $0 \le t \le s := \frac{1}{2}$, uniformly in n. Likewise, symmetrically, $n^{1/2}(F_n(t) - t)/t, \frac{1}{2} \le t \le 1$, are reversed martingales with uniformly bounded variances in the given range.

Now the following will be helpful. Krabbe (1961) treats the p-variation case. Lacking a reference for general ψ-variation, I will sketch a proof:

LEMMA 5.1. *Let ψ be any Orlicz function satisfying Δ_2. Then the functions of bounded ψ-variation on an interval form an algebra, and there is a $K < \infty$ with*

$$v_\psi(gh) \le K(v_\psi(\|g\|_\infty h) + v_\psi(\|h\|_\infty g)),$$

for any functions g and h. For $\psi(u) \equiv u^p, 1 < p < \infty$, we have

$$\|gh\|_{(p)} \le \|h\|_\infty \|g\|_{(p)} + \|g\|_\infty \|h\|_{(p)},$$

so that $\|gh\|_{[p]} \le \|g\|_{[p]}\|h\|_{[p]}$, for any two functions g and h.

PROOF. For any points x and y, we have

$$\begin{aligned}|(gh)(y) - (gh)(x)| &= |h(y)(g(y) - g(x)) + g(x)(h(y) - h(x))| \\ &\le \|h\|_\infty |g(y) - g(x)| + \|g\|_\infty h(y) - h(x)|.\end{aligned}$$

Since the Δ_2 condition implies quasi-subadditivity, ψ can be distributed over the sum, with a constant K. Then, taking sums of ψ of such increments over nonoverlapping intervals, the result for v_ψ follows.

For $\psi(u) \equiv u^p, 1 < p < \infty$, the result follows instead from Minkowski's inequality. □

Now apply Lemma 5.1 to $g(t) = M^{(n)}(t)$ and $h(t) = 1 - t$, and set $\alpha_n(t) := n^{1/2}(F_n(t) - t) \equiv (gh)(t)$. For any function f on an interval $[a, b]$ and Orlicz function ψ, denote the ψ-variation of f on $[a, b]$ by $v_\psi(f)_{[a,b]}$. Then

$$v_\psi(f)_{[0,1]} \le v_\psi(f)_{[0,1/2]} + v_\psi(f)_{[1/2,1]} + \psi((\sup - \inf)_{[0,1]} f),$$

and $v_{\psi_1}(\alpha_n)_{[0,1/2]}$ is uniformly bounded in probability. Symmetrically, the same is true for $[\frac{1}{2}, 1]$. Thus $v_{\psi_1}(\alpha_n)$ are uniformly bounded in probability. For $r > 2, |u|^r = o(\psi_1(u))$ as $u \to 0$. It follows that, for each $r > 2, \|\alpha_n\|_{[r]}$ are also uniformly bounded in probability, completing the alternate proof of Dudley [(1992a), Corollary 3.8].

I have a proof that $v_{\psi_1}(F_n - F) = O_p(n^{-3/2})$ and, consequently, that in the situation of Theorem 2.6, $\|R_{g_n}\|_2 = O_p((\log \log n)^{1/2}/n^{3/4})$. The proof is omitted since this is a little weaker than Theorem 2.6, which in any case is essentially a known fact.

Lepingle (1976) also treated r-variation of martingales.

Acknowledgment. I thank Gilles Pisier for pointing out the possibility of martingale proofs of the boundedness in probability of $\|\alpha_n\|_{[p]}, p > 2$. He has told me of another proof which uses the notion of "Type 2," as mentioned in Dudley (1992a) and references there.

REFERENCES

APPELL, J. and ZABREJKO, P. P. (1990). *Nonlinear Superposition Operators.* Cambridge Univ. Press.

BAHADUR, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 577–580.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univ. Press.

BROKATE, M. and COLONIUS, F. (1990). Linearizing equations with state-dependent delays. *Appl. Math. Optim.* **21** 45–52.

CSÖRGŐ, M. and RÉVÉSZ, P. (1978). Strong approximations of the quantile process. *Ann. Statist.* **6** 882–894.

DEHEUVELS, P. and MASON, D. M. (1990). Bahadur–Kiefer–type processes. *Ann. Probab.* **18** 669–697.

DOLLARD, J. D. and FRIEDMAN, C. N. (1979). *Product Integration with Applications to Differential Equations.* Addison-Wesley, Reading, MA.

DUDLEY, R. M. (1989). *Real Analysis and Probability.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

DUDLEY, R. M. (1991). Differentiability of the composition and inverse operators for regulated and a.e. continuous functions. Unpublished manuscript.

DUDLEY, R. M. (1992a). Fréchet differentiability, p-variation and uniform Donsker classes. *Ann. Probab.* **20** 1968–1982.

DUDLEY, R. M. (1992b). Empirical processes: p-variation for $p \leq 2$ and the quantile–quantile and $\int F\, dG$ operators. Unpublished manuscript.

DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642–669.

ESTY, W., GILLETTE, R., HAMILTON, M. and TAYLOR, D. (1985). Asymptotic distribution theory of statistical functionals. *Ann. Inst. Statist. Math.* **37** 109–129.

FERNHOLZ, L. T. (1983). *Von Mises Calculus for Statistical Functionals. Lecture Notes in Statist.* **19**. Springer, New York.

FILIPPOVA, A. (1961). Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory Probab. Appl.* **7** 24–57.

FREEDMAN, M. A. (1983). Operators of p-variation and the evolution representation problem. *Trans. Amer. Math. Soc.* **279** 95–112.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1). *Scand. J. Statist.* **16** 97–128.

GILL, R. D. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18** 1501–1555.

GRAY, A. (1975). Differentiation of composites with respect to a parameter. *J. Austral. Math. Soc. Ser. A* **19** 121–128.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

HOLT, R. J. (1986). Computation of gamma and beta tail probabilities. Technical report, Dept. Mathematics, MIT.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

KENDALL, M. G. and STUART, A. (1977). *The Advanced Theory of Statistics* **1**. *Distribution Theory*, 4th ed. Macmillan, New York.

KIEFER, J. (1967). On Bahadur's representation of sample quantiles. *Ann. Math. Statist.* **38** 1323–1342.

KIEFER, J. (1970). Deviations between the sample quantile process and the sample df. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 299–319. Cambridge Univ. Press.

KRABBE, G. L. (1961). Integration with respect to operator-valued functions. *Bull. Amer. Math. Soc.* **67** 214–218.

KRASNOSEL'SKII, M.A. and RUTICKII, YA. B. (1961). *Convex Functions and Orlicz Spaces*. Noordhoff, Groningen. (Translated by L. F. Boron.)

LEPINGLE, D. (1976). La variation d'ordre p des semi-martingales. *Z. Wahrsch. Verw. Gebiete* **36** 295–316.

MOLENAAR, W. (1970). *Approximations to the Poisson, Binomial and Hypergeometric Distribution Functions*. Math. Centrum, Amsterdam.

MONROE, I. (1972). On embedding right continuous martingales in Brownian motion. *Ann. Math. Statist.* **43** 1293–1311.

MONROE, I. (1976). Almost sure convergence of the quadratic variation of martingales: A counterexample. *Ann. Probab.* **4** 133–138.

MUSIELAK, J. and ORLICZ, W. (1959). On generalized variations (I). *Studia Math.* **18** 11–41.

PRATT, J. W. (1968). A normal approximation for binomial, F, beta, and other common, related tail probabilities II. *J. Amer. Statist. Assoc.* **63** 1457–1483.

REEDS, J. A., III (1976). On the definition of von Mises functionals. Ph.D. dissertation, Dept. Statistics, Harvard Univ.

RÉNYI, A. (1953). On the theory of order statistics. *Acta Math. Acad Sci. Hungar.* **4** 191–227.

SEBASTIÃO E SILVA, J. (1956). Le calcul différentiel et intégral dans les espaces localement convexes, réels ou complexes I, II. *Rend. Accad. Lincei Sci. Fis. Mat.* (8) **20** 743–750; **21** 40–46.

SHORACK, G. R. (1982). Kiefer's theorem via the Hungarian construction. *Z. Wahrsch. Verw. Gebiete* **61** 369–373.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

TAYLOR, S. J. (1972). Exact asymptotic estimates of Brownian path variation. *Duke Math. J.* **39** 219–241.

VAN DER VAART, A. (1991). Efficiency and Hadamard differentiability. *Scand. J. Statist.* **18** 63–75.

VAN DER VAART, A., and WELLNER, J. (1994). *Weak Convergence and Empirical Processes*. IMS, Hayward, CA. To appear.

VERVAAT, W. (1972). Functional central limit theorems for processes with positive drift and their inverses. *Z. Wahrsch. Verw. Gebiete* **12** 245–253.

VON MISES, R. (1936). Les lois de probabilité pour les fonctions statistiques. *Ann. Inst. H. Poincaré* **6** 185–212.

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

WHITT, W. (1980). Some useful functions for functional limit theorems. *Math. Oper. Res.* **5** 67–85.

WONG, W. H. and SEVERINI, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *Ann. Statist.* **19** 603–632.

YOUNG, L. C. (1936). An inequality of the Hölder type, connected with Stieltjes integration. *Acta Math.* **67** 251–282.

DEPARTMENT OF MATHEMATICS
ROOM 2-245
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139