How Powerful Is Continuous Nonlinear Information for Linear Problems?

B. KACEWICZ

Institute of Informatics, University of Warsaw, 00-901 Warsaw, Poland

AND

G. W. WASILKOWSKI*

Department of Computer Science, Columbia University, New York, New York 10027

Received January 22, 1986

There are many papers dealing with the approximate solution of linear problems where only partial information is available. Two types of information have been considered: linear and discontinuous nonlinear. In particular, we know that discontinuous nonlinear information is far more powerful than linear information. In this paper we study continuous nonlinear information for linear problems, and we prove that:

— it is no more powerful than linear information in the worst case setting,
— it is much more powerful than linear information in the average case setting.


1. INTRODUCTION

To explain the setting of the paper consider the following approximation problem. Suppose that one wants to approximate a real function $f$ from a given class $F$. The function $f$ is not known. Instead, one knows information $N(f)$ provided by an information operator $N: F \rightarrow \mathbb{R}^n$ for some finite number $n$. A typical example of $N$ is given by $N(f) = [f(t_1), \ldots, f(t_n)]$. Note that such an $N$ is a linear operator. One can think about more general information operators including nonlinear ones such as $N(f) = [L_1(f), \ldots, L_n(f)]$ with nonlinear functions $L_i$. In many cases, $N$

*Supported by the National Science Foundation under Grant DCR-82-14322 and by the Advanced Research Projects Agency under Contract N00039-82-C-0427. Part of this work was done while the author was a member of the Mathematical Sciences Research Institute, Berkeley.
is partial since there are infinitely many functions which share the same information. This means that $N$ causes an intrinsic uncertainty which cannot be reduced no matter how $N(f)$ is combined to approximate $f$. The uncertainty is defined as the minimal error of any algorithm that uses $N$.

How is the error of an algorithm defined? It depends on the setting one is interested in. In this paper we consider two settings: worst and average case. In the first setting, the error of an algorithm is defined by its worst performance in the class $F$. In the second setting the error is defined by the average performance with respect to some probability measure defined on $F$. In either setting we are interested in information with the intrinsic uncertainty as small as possible.

There are many results on the minimal uncertainty for linear information operators whose range is $\mathbb{R}^a$ (see, e.g., Traub, Wasiłkowski, and Woźniakowski, 1984; Traub and Woźniakowski, 1980; Wasiłkowski and Woźniakowski, 1986). For some problems this minimal uncertainty may be very large. For nonlinear information, the situation is different. It is known (see Traub and Woźniakowski, 1980, Theorem 3.1, p. 153) that even for $n = 1$ there always exists a discontinuous nonlinear information operator whose uncertainty is arbitrarily small. This means that there is a big gap between uncertainty caused by linear and discontinuous nonlinear information. There is an obvious difficulty in computing discontinuous nonlinear information. Therefore, it seems natural to restrict the class of information operators to continuous ones. By continuous $N$ we mean that $N$ restricted to an arbitrary finite-dimensional space is a continuous operator. We study the following question:

Is continuous nonlinear information more powerful than linear information?

(1.1)

Continuous nonlinear information has richer structure than linear information. One might hope, therefore, that the intrinsic uncertainty caused by nonlinear information is smaller than that of linear information. This is the case for some nonlinear problems, such as zero-finding problems. For some other problems, such as nonlinear ODE, continuous nonlinear information is not more powerful than linear information, as shown in Kacewicz (1983).

In this paper we study (1.1) for the approximate solution of linear problems. We prove that the answer to this question depends on the setting. In the worst case setting, continuous nonlinear information is not more powerful than linear information. That is, the uncertainty caused by arbitrary continuous nonlinear information with range $\mathbb{R}^a$ is not smaller than the uncertainty caused by certain linear information with the same range $\mathbb{R}^a$. In the average case setting, continuous nonlinear information is in general much more powerful than linear information. Indeed, we prove that one can solve
a linear problem with arbitrarily small average error using only one evaluation of a continuous nonlinear functional.

2. Worst Case Setting

In this section we prove that in the worst case setting continuous information is not more powerful than the linear one. We begin with basic definitions (for more detailed discussion we refer to Traub and Woźniakowski, 1980).

Let $F_1$ and $F_2$ be real separable Hilbert spaces with norms $\| \cdot \|_1$ and $\| \cdot \|_2$, respectively. Let $S$ be a continuous linear operator,

$$S: F_1 \to F_2.$$ 

We call $S$ a solution operator. We wish to construct an element $\alpha$, $\alpha = \alpha(f) \in F_2$, which approximates $Sf$ with error $\| Sf - \alpha \|_2$ as small as possible. The information about $f$ is provided by $N(f)$, where $N$, called the information operator (or information), is a mapping

$$N: F_1 \to \mathbb{R}^n. \quad (2.1)$$

The number $n$ is called the cardinality of $N$, $\text{card}(N) = n$. Knowing $N(f)$ we construct an approximation $\alpha$ by an algorithm $\phi$,

$$\alpha := \phi(N(f)).$$

Here by an algorithm that uses $N$ we mean any mapping

$$\phi: N(F_1) \to F_2.$$ 

In the worst case setting, the error of the algorithm $\phi$ is defined by

$$e^\ast(\phi, N) = \sup_{\| f \|_1 = 1} \| Sf - \phi(N(f)) \|_2. \quad (2.2)$$

The intrinsic uncertainty caused by $N$ is measured by the minimal error among all algorithms that use $N$,

$$r^\ast(N) = \inf\{e^\ast(\phi, N) : \phi \text{ uses } N\}. \quad (2.3)$$

Because of its geometrical interpretation, $r^\ast(N)$ is called the radius of $N$ (see Traub and Woźniakowski, 1980).

In this paper we consider the following classes of information operators:
CONTINUOUS NONLINEAR INFORMATION

\[ \Psi_n = \{ N : \text{card}(N) \leq n \}, \]
\[ C_n = \{ N \in \Psi_n : N \text{ is continuous} \}, \]
\[ L_n = \{ N \in \Psi_n : N \text{ is linear} \}. \]

Here, by continuity of \( N \) we mean that \( N \) restricted to any finite-dimensional space is a continuous operator. Therefore, \( N \in L_n \) implies that \( N \in C_n \). Since \( L_n \subseteq C_n \subseteq \Psi_n \),

\[ \inf_{N \in \Psi_n} r^w(N) \leq \inf_{N \in C_n} r^w(N) \leq \inf_{N \in L_n} r^w(N). \quad (2.4) \]

It is known that \( \inf_{N \in \Psi_n} r^w(N) = 0 \) (see Traub and Woźniakowski, 1980, Theorem 3.1, p. 153). This means that we can solve the problem with arbitrarily small error using information with cardinality one. This information is discontinuous. For the class \( L_n \), we know that the sequence \( \inf_{N \in L_n} r^w(N) \) need not converge to zero, and if it does it might converge arbitrarily slowly (see Traub and Woźniakowski, 1980, Theorem 2.1, p. 86). Therefore there is a big difference between the intrinsic uncertainty caused by nonlinear discontinuous information and linear information. Is continuous nonlinear information more powerful than linear? In Theorem 2.1 we prove that this is not the case. We also characterize the minimal radius among all linear information operators.

To state Theorem 2.1 we need the following definitions. Let

\[ K = \mathbb{S} \ast \mathbb{S} : F_1 \to F_1. \quad (2.5) \]

Then \( K \) is continuous, symmetric, and nonnegative definite. Let \( \text{sp}(K) \) denote the spectrum of \( K \). For such an operator \( K \) the spectrum is nonempty and \( \text{sp}(K) = p(K) \cup c(K) \). Here \( p(K) \) is the point spectrum of \( K \) and \( c(K) \) is the continuous spectrum of \( K \). That is, \( p(K) \) is the set of all eigenvalues of \( K \) (an eigenvalue of multiplicity \( k \) is counted \( k \)-times) and \( c(K) \) is the set of all positive numbers \( x \) for which \( (K - xI)^{-1} \) is well defined on a dense subspace of \( F_1 \) and is unbounded. See Dunford and Schwartz (1963, p. 907). Obviously, if \( \lambda \in \text{sp}(K) \) then \( \lambda \geq 0 \).

Let

\[ R_n = \inf \{ \sup_{\lambda \in \text{sp}(K) - B} \lambda : B \subseteq \text{sp}(K) \text{ and } B \text{ has at most } n \text{ elements} \}, \quad (2.6) \]

with the convention that \( \sup \emptyset = 0 \). Then \( R_n \) is the \((n + 1)st\) maximal element from the spectrum of \( K \), if such an element exists. Otherwise, \( R_n \) is the maximal attraction point from \( \text{sp}(K) \) and in this case \( R_k = R_n \), \( \forall k \geq n \).

**Theorem 2.1.** Continuous nonlinear information is not more powerful than linear information, i.e.,
\[
\inf_{N \in \mathcal{C}_n} r^w(N) = \inf_{N \in \mathcal{L}_n} r^w(N) = \sqrt{R_n}.
\]

**Proof.** We first prove that

\[
\inf_{N \in \mathcal{L}_n} r^w(N) \leq \sqrt{R_n}. \tag{2.7}
\]

Observe that there exists an index \(k, k \in \{0, 1, \ldots, n\}\), and \(k\) eigenvalues \(\lambda_1, \ldots, \lambda_k \in p(K)\) such that \(\lambda_i \geq R_n\) for \(i = 1, \ldots, k\) and

\[
R_n = \sup\{\lambda \in \text{sp}(K) : \lambda \notin \{\lambda_1, \ldots, \lambda_k\}\}.
\]

If \(k = 0\), take \(N^* = 0\). Otherwise, if \(k \geq 1\), take \(N^*(\cdot) = [\langle \cdot, \xi_1 \rangle, \ldots, \langle \cdot, \xi_k \rangle]\), where \(\xi_i\)'s are orthonormalized eigenelements of \(K\) corresponding to eigenvalues \(\lambda_i\). Obviously, \(N^*\) is linear and of cardinality \(k \leq n\). From Traub and Woźniakowski (1980, p. 76), we know that

\[
r^w(N^*)^2 = \sup\{\langle Kf, f \rangle : \|f\|_1 = 1\text{ and } \langle f, \xi_i \rangle = 0\text{ for } i = 1, \ldots, k\}.
\]

Finally, let \(X^\perp\) be the orthogonal complement to \(\text{lin}\{\xi_1, \ldots, \xi_k\}\), and let \(K'\) be the operator \(K\) restricted to \(X^\perp\). Obviously, the last supremum is equal to \(\sup_{\lambda \in \text{sp}(K')\lambda}\). Since \(\text{sp}(K') = \text{sp}(K) - \{\lambda_1, \ldots, \lambda_k\}\), we have that

\[
r^w(N^*) = \sqrt{R_n}.
\]

This proves (2.7).

We now prove

\[
\inf_{N \in \mathcal{L}_n} r^w(N) \geq \sqrt{R_n}. \tag{2.8}
\]

This together with (2.4) and (2.7) will complete the proof. We need the following

**Lemma 2.1.** For every integer \(k\) and positive \(\delta\), there exist elements \(g_1, \ldots, g_k \in F_i\) such that

\[
\langle g_i, g_j \rangle = \delta_{i,j}, \quad \forall i, j \in \{1, \ldots, k\}, \tag{2.9}
\]

and

\[
\langle Kg_i, g_j \rangle = \begin{cases} 
R_{k-1} - \delta & \text{if } i = j \\
0 & \text{otherwise}.
\end{cases} \tag{2.10}
\]

**Proof.** (Induction on \(k\)). For \(k = 1\), \(R_0 = \sup_{\|f\|_1 = 1} \langle Kf, f \rangle\). Hence such \(g_1\) exists. Therefore suppose that (2.9) and (2.10) hold for some number \(k\). We prove that they hold for \(k + 1\).

Observe that if \(K\) has \(k + 1\) dominating eigenvalues \(\lambda_1, \ldots, \lambda_{k+1}\), i.e., \(\lambda_i \in p(K)\) and \(\lambda_1 \geq \cdots \geq \lambda_{k+1}\) with

\[
\lambda_{k+1} = \sup\{\lambda : \lambda \in \text{sp}(K) - \{\lambda_1, \ldots, \lambda_k\}\},
\]

and
then $R_k = \lambda_{k+1}$. Hence the corresponding orthogonalized eigenelements $g_1, \ldots, g_{k+1}$ satisfy (2.9) and (2.10). A similar situation holds if the point spectrum $p(K)$ has an attraction point $\lambda$ for which $\lambda = R_k$. Therefore we can assume that the value $R_k$ is attained by an element from the continuous spectrum, i.e.,

$$R_k = \sup_{\lambda \in c(K)} \lambda.$$ 

Due to the inductive assumption, there exist $g_1, \ldots, g_k$ satisfying (2.9) and (2.10). Let $X_k = \text{lin}\{g_1, Kg_1, \ldots, g_k, Kg_k\}$. Since $R_k \in c(K)$ (recall that the continuous spectrum is closed), the operator $K - R_k I$ is one-to-one. Let $Y = (K - R_k I)^{-1}(X)$ be the preimage of $X$ under $K - R_k I$. Then $Y$ is a finite-dimensional linear subspace of $F_1$. Let $Y^\perp$ be the orthogonal complement of $Y$. From the definition of continuous spectrum we know that there exists a sequence $\{h_m\}_{m=1}^\infty$ of elements from $F_1$ such that $\|h_m\|_1 = 1$ and

$$\lim_{m \to \infty} \|(K - R_k I)^{-1}h_m\|_1 = \infty. \quad (2.11)$$

Without loss of generality we assume that $h_m \in Y^\perp, \forall m$. Let

$$f_m = \frac{(K - R_k I)^{-1}h_m}{\|(K - R_k I)^{-1}h_m\|_1}.$$ 

Then $\|f_m\|_1 = 1$. Furthermore, $f_m$ are orthogonal to $X$, which follows directly from the fact that $h_m$ are orthogonal to $Y$ and that $(K - R_k I)^{-1}$ is self-adjoint. Note that

$$(k - R_k I)f_m = \frac{h_m}{\|(K - R_k I)^{-1}h_m\|_1} \xrightarrow{m \to \infty} 0. \quad (2.12)$$

Hence $\langle Kf_m, f_m \rangle - R_k \langle f_m, f_m \rangle = \langle Kf_m, f_m \rangle - R_k$ tends to zero. Thus, letting $g_{k+1} = f_m$ for sufficiently large $m$, we have an element which satisfies (2.9) and (2.10). This completes the proof of Lemma 2.1.

We are ready to prove (2.8). Let $g_1, \ldots, g_{n+1}$ be as in Lemma 2.1 with $k = n + 1$ and arbitrary $\delta$. Let $J$ be the unit sphere in $\text{lin}\{g_1, \ldots, g_{n+1}\}$,

$$J = \left\{ f = \sum_{i=1}^{n+1} a_i g_i : \sum_{i=1}^{n+1} a_i^2 = 1 \right\}.$$ 

Take arbitrary information $N \in C_n$. Since $N$ restricted to $J$ is continuous, the Borsuk–Ulam theorem (see Kuratowski, 1968, Vol. II, p. 477) implies that there exists an element
such that $N(f^*) = N(-f^*)$. Since $f^*$ and $-f^*$ share the same information, $\phi(N(f^*)) = \phi(N(-f^*))$ for any algorithm $\phi$ that uses $N$. Hence

$$r^*(N)^2 \geq \inf_{x \in F_2} \max\{\|Sf^* - x\|_2^2, \|S(-f^*) - x\|_2^2\}$$

$$= \|Sf^*\|_2^2$$

$$= \langle Kf^*, f^* \rangle$$

$$= \sum_{i=1}^{n+1} (a_i^*)^2 \langle Kg_i, g_i \rangle$$

$$\geq R_n - \delta.$$

Since $N$ and $\delta$ are arbitrary, (2.8) follows. This completes the proof of Theorem 2.1.

Remark 2.1. Theorem 2.1 can be slightly extended by assuming continuity of $N$ only on finite-dimensional spheres. Another extension is due to the definition of the algorithm error. For the sake of simplicity we have chosen a very simple definition of $e^*(\phi, N)$. Theorem 2.1 is true, however, for errors defined in a more general way. For example, it holds if

$$e^*(\phi, N) = \sup_{f \in F_1} \|Sf - \phi(N(f))\|_2 \rho(\|Tf\|_3),$$

where $T: F_1 \rightarrow F_3$ is continuous and linear, $F_3$ is a separable Hilbert space, and $\rho: R_+ \rightarrow R_+$ is an arbitrary function. For a discussion of this kind of error see Traub, Wasilkowski, and Woźniakowski (1983, App. E, p. 139).

Remark 2.2. It is proven in Traub and Woźniakowski (1980, Theorem 5.3, p. 40) that the minimal radius of linear information is equal to the square root of the $(n + 1)$st maximal eigenvalue $\lambda_{n+1}$ of $K$, whenever $K$ is compact. For compact $K$, $\text{sp}(K) \subset \rho(K) \cup \{0\}$, which implies that $R_n = \lambda_{n+1}$. This gives a correspondence between Theorem 2.1 and Theorem 5.3 from Traub and Woźniakowski (1980) for compact operators $K$.

Remark 2.3. The proof of Theorem 2.1 uses the fact that $F_1$ and $F_2$ are separable Hilbert spaces. This theorem remains true for some problems defined on Banach spaces (for nonlinear ODE problems, see Kacewicz, 1983). For general linear problems defined on Banach spaces, it is an open question whether linear information is as powerful as continuous nonlinear information.
3. Average Case Setting

In this section we prove that continuous information is extremely powerful on the average. We begin with a brief discussion of the average case setting (for more details see Traub, Wasilkowski, and Woźniakowski, 1984; Wasilkowski and Woźniakowski, 1986).

Let $F_1$, $F_2$, and $S$ be as in Section 2. Also the definitions of information and of the algorithm remain the same. What differentiates the average case setting from the worst case one is the definition of the error of an algorithm. More precisely, in the average case setting the error of an algorithm $\phi$ is measured by the average performance of $\phi$,

$$e^a(\phi, N) = \int_{F_1} \|Sf - \phi(N(f))\| \mu(df). \quad (3.1)$$

The (average) radius of $N$ is then defined by

$$r^a(N) = \inf\{e^a(\phi, N) : \phi \text{ uses } N\}. \quad (3.2)$$

Here we assume that $\mu$ is a given probability measure defined on the $\sigma$-field $\mathcal{B}(F_1)$ of Borel sets of $F_1$ and

$$\int_{F_1} \|f\| \mu(df) < \infty. \quad (3.3)$$

**Theorem 3.1.**

$$\inf_{n \in C_1} r^a(N) = 0. \quad (3.4)$$

**Proof.** To prove (3.4) we show that for every positive $\delta$, there exists $N^* \in C_1$ with $r^a(N^*) \leq \delta$.

Let $X$ be an $m$-dimensional subspace of $F_1$ such that

$$\int_{F_1} \|(I - P_X)f\|_1 \mu(df) \leq \frac{\delta}{2\|S\|}. \quad (3.5)$$

Here $P_X$ is the orthogonal projection onto $X$. Obviously, such $X$ exists. Let $\xi_1, \ldots, \xi_m$ be an orthonormal basis of $X$ and let $\xi_{m+1}, \xi_{m+2}, \ldots$ be an orthonormal system of $X^\perp$, the orthogonal complement of $X$. Then for every $f \in F_1$, $f = x + x'$, where $x = P_X f = \sum_{i=1}^m \langle f, \xi_i \rangle \xi_i$ and $x^\perp = (I - P_X)f = \sum_{i=m+1}^\infty \langle f, \xi_i \rangle \xi_i$. We shall construct $N^*$ so that $N^*(f) = N^*(x + x^\perp) = N^*(x)$.

Let $\mu_X = \mu P_X^{-1}$ be the induced probability measure on $X$, i.e.,
\[
\mu_X(A) = \mu(P_X^{-1}(A)), \quad \forall \text{ measurable } A \subseteq X.
\]

Then for every \( \mu \)-integrable function \( G \),
\[
\int_{F_i} G(P_x f) \mu(df) = \int_X G(f) \mu_X(df).
\]

Since \((f, \xi_i)\) as a functional in \( f \in X \) is measurable for every \( i = 1, 2, \ldots, m \), there exists an index \( k \) such that
\[
\left| \langle f, \xi_i \rangle - \sum_{j=1}^k \langle f_j, \xi_i \rangle \chi_{A_j}(f) \right| \leq \frac{\delta}{4m\|S\|}, \quad \forall f \in X, \forall i \leq m, \tag{3.6}
\]
where \( A_1, \ldots, A_k \) are disjoint measurable subsets of \( X \), \( A_1 \cup \cdots \cup A_k = X \) and \( f_j \in A_j \). Here \( \chi_{A_j} \) denotes the characteristic function for the set \( A_j \). Note that (3.6) implies that for
\[
g(f) = \sum_{i=1}^m \sum_{j=1}^k \langle f_j, \xi_i \rangle \chi_{A_i}(f) \xi_i,
\]
we have
\[
\int_X \|f - g(f)\|_1 \mu_X(df) = \int_{F_i} \|P_x f - g(P_x f)\|_1 \mu(df) \leq \frac{\delta}{4\|S\|}. \tag{3.7}
\]
Note also that \( g(f) = f_j \) if \( f \in A_j \). We shall construct \( N^* \) and an algorithm \( \phi \) such that \( \phi(N^*(f)) \) is close to \( g(f) \).

Without loss of generality we assume that \( A_j \) are rectangles, i.e.,
\[
A_j = \{f \in X : a_{i,j} \leq \langle f, \xi_i \rangle < b_{i,j}, i = 1, \ldots, m\}
\]
for some \( a_{i,j}, b_{i,j} \in \mathbb{R} \cup \{-\infty, +\infty\} \). For \( 0 < \gamma < 1 \), let \( A_j(\gamma) = \{f : a_{i,j} \leq \langle f, \xi_i \rangle \leq b_{i,j}(1 - \gamma \text{ sgn}(b_{i,j})), i = 1, \ldots, m\} \). Then \( A_j(\gamma) \) are closed and disjoint. Therefore there exists a continuous functional \( L_\gamma \), defined on \( F_i \) such that
\[
L_\gamma(f) = j \text{ if } P_x f \in A_j(\gamma), \quad \text{and} \quad |L_\gamma(f)| \leq k, \forall f \in F_i.
\]
\[
\forall j = 1, \ldots, k.
\]
Note that \( L_\gamma P_x = L_\gamma \). Take a function \( g_\gamma : \mathbb{R} \to X, g_\gamma(y) = f_j \) if \( y = j \) for \( j = 1, \ldots, k \), and \( g_\gamma(y) = 0 \) otherwise. Note that \( g_\gamma(L_\gamma(f)) = g(f) \) if \( P_x f \in \bigcup_{j=1}^k A_j(\gamma) \). Furthermore, \( g_\gamma(L_\gamma(f)) \in \{f_1, \ldots, f_k\}, \forall f \in F_i \). Thus, letting \( B(\gamma) = \bigcup_{j=1}^k (A_j - A_j(\gamma)) \), we have
\[ \int_{F_1} \| g(P_x f) - g_y(L_y(f)) \|_1 \mu(df) = \int_{F_1} \| g(f) - g_y(L_y(f)) \|_1 \mu_x(df) \]

\[ = \int_{B(\gamma)} \| g(f) - g_y(L_y(f)) \|_1 \mu(df) \leq \mu_x(B(\gamma)) \max_{i,j} \| f_i - f_j \|. \]

Since \( \mu_x(B(\gamma)) \) tends to zero with \( \gamma \), there exists \( \gamma^* \) such that
\[ \int_{F_1} \| g(f) - g_{\gamma^*}(L_{\gamma^*}(f)) \|_1 \mu_x(df) \leq \frac{\delta}{4\|S\|}. \]

This and (3.7) imply that
\[ \int_{F_1} \| P_x f - g_{\gamma^*}(L_{\gamma^*}(f)) \|_1 \mu(df) \leq \frac{\delta}{2\|S\|}. \quad (3.8) \]

Now take \( N^*(f) = L_{\gamma^*}(f) \) and \( \phi(N^*(f)) = S(g_{\gamma^*}(L_{\gamma^*}(f))). \) Then
\[ e^\sigma(\phi, N^*) = \int_{F_1} \| Sf - S(g_{\gamma^*}(L_{\gamma^*}(f))) \|_1 \mu(df) \]
\[ \leq \| S \| \int_{F_1} \| P_x f - g_{\gamma^*}(L_{\gamma^*}(f)) + (I - P_x)f \|_1 \mu(df) \]
\[ \leq \| S \| \int_{F_1} \| P_x f - g_{\gamma^*}(L_{\gamma^*}(f)) \|_1 + \| (I - P_x)f \|_1 \| \mu(df). \]

Hence due to (3.5) and (3.8) we have that
\[ r^\sigma(N^*) \leq e^\sigma(\phi, N^*) \leq \delta, \]

which completes the proof of Theorem 3.1. ■

Remark 3.1. We now comment on extensions of Theorem 3.1. Note that in the proof we never used the fact that \( F_2 \) is a Hilbert space. Thus, the theorem holds for any linear normed space \( F_2 \). Furthermore, the only place where the assumption that \( F_1 \) is a separable Hilbert was used was to deduce (3.5). This means that Theorem 3.1 is valid under the following assumptions:

(i) \( F_1 \) and \( F_2 \) are normed linear spaces,

(ii) \( \| f \|_1 \) is \( \mu \)-integrable,

(iii) there exists a sequence of finite-dimensional projections \( P_m \) such that
\[ \lim_{m \to \infty} \int_{F_1} \| (I - P_m)f \|_1 \mu(df) = 0. \]
The result is also valid when the error of an algorithm is defined differently than in (3.1). For instance, Theorem 3.1 holds if the error is defined by

\[ e^u(\phi, N) = \sqrt{p \int f | Sf - \phi(N(f)) \|_p \mu(df)}, \]

for \( p \geq 1 \), with (ii) replaced by the assumption that \( \|f\|_p \) is \( \mu \)-integrable.

REFERENCES


