Constr. Approx.

CONSTRUCTIVE APPROXIMATION

Rates of Convex Approximation in Non-Hilbert Spaces

Michael J. Donahue, Leonid Gurvits, Christian Darken, and Eduardo Sontag

Abstract.

This paper deals with sparse approximations by means of convex combinations of elements from a predetermined "basis" subset S of a function space. Specifically, the focus is on the *rate* at which the lowest achievable error can be reduced as larger subsets of S are allowed when constructing an approximant. The new results extend those given for Hilbert spaces by Jones and Barron, including in particular a computationally attractive incremental approximation scheme. Bounds are derived for broad classes of Banach spaces; in particular, for L_p spaces with $1 < p < \infty$, the $O(n^{-1/2})$ bounds of Barron and Jones are recovered when $p = 2$.

One motivation for the questions studied here arises from the area of "artificial neural networks," where the problem can be stated in terms of the growth in the number of "neurons" (the elements of S) needed in order to achieve a desired error rate. The focus on non-Hilbert spaces is due to the desire to understand approximation in the more "robust" (resistant to exemplar noise) L_p, $1 \leq p < 2$ norms.

The techniques used borrow from results regarding moduli of smoothness in functional analysis as well as from the theory of stochastic processes on function spaces.

1. Introduction

The subject of this paper concerns the problem of approximating elements of a Banach space X—typically presented as a space of functions—by means of finite linear combinations of elements from a predetermined subset S of X. In contrast to classical linear approximation techniques, where optimal approximation is desired and no penalty is imposed on the *number* of elements used, we are interested here in *sparse* approximants, that is to say, combinations that employ few elements. In particular, we are interested in understanding the rate at which the achievable error can be reduced as one increases the number allowed. Such questions are of obvious interest in areas such as signal representation, numerical analysis, and neural networks (see below).

AMS classification: 41A25, 46B09, 68T05

Key words and phrases: Banach space, convex combination, neural network, modulus of smoothness

Rather than arbitrary linear combinations $\sum_i a_i g_i$, with a_i's real and g_i's in S, it turns out to be easier to understand approximations in terms of combinations that are subject to a prescribed upper bound on the total coefficient sum $\sum_i |a_i|$. After normalizing S and replacing it by $S \cup -S$, one is led to studying approximations in terms of convex combinations. This is the focus of the current work.

To explain the known results and our new contributions, we first introduce some notation.

1.1. *Optimal Approximants*

Let X be a Banach space, with norm $\|\cdot\|$. Take any subset $S \subseteq X$. For each positive integer n, we let $\mathrm{lin}_n S$ consist of all sums $\sum_{i=1}^n a_i g_i$, with $g_1, \ldots, g_n$ in S and with arbitrary real numbers $a_1, \ldots, a_n$, while we let $\mathrm{co}_n S$ be the set of such sums with the constraint that all $a_i \in [0,1]$ and $\sum_i a_i = 1$. The distances from an element $f \in X$ to these spaces are denoted respectively by

$$\|\mathrm{lin}_n S - f\| := \inf \{\|h - f\|\,,\ h \in \mathrm{lin}_n S\}$$

and

$$\|\mathrm{co}_n S - f\| := \inf \{\|h - f\|\,,\ h \in \mathrm{co}_n S\}\ .$$

Of course, always $\|\mathrm{lin}_n S - f\| \le \|\mathrm{co}_n S - f\|$. For each subset $S \subseteq X$, $\mathrm{lin}\, S = \cup_n \mathrm{lin}_n S$ and $\mathrm{co}\, S = \cup_n \mathrm{co}_n S$ denote respectively the linear span and the convex hull of S. We use bars to denote closure in X; thus, for instance, $\overline{\mathrm{co} S}$ is the closed convex hull of S. Note that saying that $f \in \overline{\mathrm{lin}\, S}$ or $f \in \overline{\mathrm{co} S}$ is the same as saying that $\lim_{n\to+\infty} \|\mathrm{lin}_n S - f\| = 0$ and $\lim_{n\to+\infty} \|\mathrm{co}_n S - f\| = 0$ respectively; in this case, we say for short that f is (linearly or convexly) *approximable* by S. These distances as a function of n represent the convergence rates of the best approximants to the target function f. The study of such rates is standard in approximation theory (e.g., (Powell 1981)), but the questions addressed here are not among those classically considered.

Let ϕ be a positive function on the integers. We say that the space X admits a *(convex) approximation rate* $\phi(n)$ if for each bounded subset S of X and each $f \in \overline{\mathrm{co} S}$, $\|\mathrm{co}_n S - f\| = O(\phi(n))$. (The constant in this estimate is allowed to, and in general will, depend on S, typically through an upper bound on the norm of elements of S.) One could of course also define the analogous *linear* approximation rates; we do not do so because at this time we have no nontrivial results to report in that regard. (The implications of the restriction to convex approximates is examined in Appendix A.)

Jones (1992) and Barron (1991) showed that every Hilbert space admits an approximation rate $\phi(n) = 1/\sqrt{n}$. One of our objectives is the study of such rates for non-Hilbert spaces. To date the larger issue of convergence rates in more general Banach spaces and in the important subclass of L_p, $p \neq 2$, spaces has not been addressed. Barron (1992) showed that the same rate is obtained in the uniform norm, but only for approximation with respect to a particular class of sets S.

1.2. Incremental Approximants

Jones (1992) considered the procedure of constructing approximants to f incrementally, by forming a convex combination of the last approximant with a *single* new element of S; in this case, the convergence rate in L_2 is interestingly again $O(1/\sqrt{n})$. Incremental approximants are especially attractive from a computational point of view. In the neural network context, they correspond to adding one "neuron" at a time to decrease the residual error. We next define these concepts precisely.

Again let X be a Banach space with norm $\|\cdot\|$. Let $S \subseteq X$. An *incremental sequence* (for approximation in $\mathrm{co}\, S$) is any sequence $f_1, f_2, \ldots$ of elements of X so that $f_1 \in S$ and for each $n \geq 1$ there is some $g_n \in S$ so that $f_{n+1} \in \mathrm{co}\,(\{f_n, g_n\})$.

We say that an incremental sequence $f_1, f_2, \ldots$ is *greedy* (with respect to $f \in \overline{\mathrm{co} S}$) if

$$\|f_{n+1} - f\| \;=\; \inf\left\{\|h - f\| \mid h \in \mathrm{co}\,(\{f_n, g\})\,, g \in S\right\}, \quad n = 1, 2, \ldots .$$

The set S is generally not compact, so we cannot expect the infimum to be attained. Given a positive sequence $\epsilon = (\epsilon_1, \epsilon_2, \ldots)$ of allowed "slack" terms, we say that an incremental sequence $f_1, f_2, \ldots$ is ε*-greedy* (with respect to f) if

$$\|f_{n+1} - f\| \;<\; \inf\left\{\|h - f\| \mid h \in \mathrm{co}\,(\{f_n, g\})\,, g \in S\right\} + \varepsilon_n\,, \quad n = 1, 2, \ldots .$$

Let ϕ be a positive function on the integers. We say that S *has an incremental (convex) scheme with rate* $\phi(n)$ if there is an incremental schedule ε such that, for each f in $\overline{\mathrm{co} S}$ and each ε-greedy incremental sequence $f_1, f_2, \ldots$, it holds that

$$\|f_n - f\| \;=\; O(\phi(n))$$

as $n \to +\infty$. Finally, we say that the space X *admits incremental (convex) schemes with rate* $\phi(n)$ if every bounded subset S of X has an incremental scheme with rate $\phi(n)$.

The intuitive idea behind this definition is that at each stage we attempt to obtain the best approximant in the restricted subclass consisting of convex combinations $(1 - \lambda_n) f_n + \lambda_n g$, with λ_n in $[0, 1]$, g in S, and f_n being the previous approximant. It is also possible to select the sequence $\lambda_1, \lambda_2, \ldots$ beforehand. We say that an incremental sequence $f_1, f_2, \ldots$ is ϵ-greedy (with respect to f) with *convexity schedule* $\lambda_1, \lambda_2, \ldots$ if

$$\|f_{n+1} - f\| \;<\; \inf\left\{\|\left((1 - \lambda_n) f_n + \lambda_n g\right) - f\| \mid g \in S\right\} + \varepsilon_n\,, \quad n = 1, 2, \ldots .$$

One could also define the analogous linear incremental schemes, for which one does not require $\lambda_n \in [0, 1]$, but, as before, we only report results for the convex case.

Informally, from now on we refer to the rates for convex approximation as "optimal rates" and use the terminology "incremental rates" for the best possible rates for incremental schemes. For any incremental sequence, $f_n \in \mathrm{co}_n(S)$, so clearly optimal rates are always majorized by the corresponding incremental rates.

The main objective of this paper[1] is to analyze both optimal and incremental rates in broad classes of Banach spaces, specifically including L_p, $1 \leq p \leq \infty$. A summary of our rate bounds for the special case of the spaces L_p is given as Table 1. In general, we find that the worst-case rate of approximation in the "robust" L_p, $1 \leq p < 2$, norms is worse than that in L_2, unless some additional conditions are imposed on the set S.

Table 1. The order of the worst-case rate of approximation in L_p. "NO" means that the approximants do not converge in the worst case.

p	1	$(1,2)$	$[2,\infty)$	∞
optimal	1	$n^{-1+1/p}$	$n^{-1/2}$	1
incremental	NO	$n^{-1+1/p}$	$n^{-1/2}$	NO

1.3. Neural Nets

The problem is of general interest, but we were originally motivated by applications to "artificial neural networks." In that context the set S is typically of the form

$$S \ = \ \{g : \mathbb{R}^d \to \mathbb{R} \mid \exists\, a \in \mathbb{R}^d, b \in \mathbb{R}, \text{ s.t. } g(x) = \pm\sigma(a \cdot x + b)\},$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed function, called the *activation* or *response* function. Typically, σ is a smooth "sigmoidal" function such as the logistic function $(1 + e^{-x})^{-1}$, but it can be discontinuous, such as the Heaviside function (the characteristic function of $[0, \infty)$). The elements of $\mathrm{lin}_n S$ are called *single hidden layer neural networks* (with activation σ and a linear output layer) with *n hidden units*. For neural networks, then, the question that we investigate translates into the study of how the approximation error scales with the number of hidden units in the network.

Neural net approximation is a technique widely used in empirical studies. Mathematically, this is justified by the fact that, for each compact subset M of $\mathbb{R}^d$, restricting elements of S to M, one has that $\overline{\mathrm{lin}\, S} = C^0(M)$, that is, $\mathrm{lin}\, S$ is dense in the set of continuous functions under uniform convergence (and hence also in most other function spaces). This density result holds under extremely weak conditions on σ; being locally Riemann integrable and non-polynomial is enough. See for instance (Leshno et al., 1992).

Spaces L_p with p equal to or slightly greater than one are particularly important because of their usefulness for robust estimation (e.g., (Rey 1983)). In the particular context of regression with neural networks, Hanson (1988) presents experimental results showing the superiority of L_p $(p \ll 2)$ to L_2.

[1] A preliminary version of some of the results presented in this paper appeared as (Darken, Donahue, Gurvits and Sontag 1993).

1.4. Connections to Learning Theory

Of course, neural networks are closely associated with learning theory. Let us imagine that we are attempting to learn a target function that lies in the convex closure of a predetermined set of basis functions S. Our learned estimate of the target function will be represented as a convex combination of a subset of S. For each n in an increasing sequence of values of n, we optimize our choice of basis functions and their convex weighting over a sufficiently large data set (the size of which may depend on n). Let us assume that the problem is "learnable", e.g., that over the class of probability measures of interest on the domain of the functions in S, the difference between one's estimates of the error based on examples must converge to the true error uniformly over all possible approximants. Then the generalization error (expected loss over the true exemplar distribution) must go to zero at least as fast as the order of the upper bounds in this work. Thus our bounds represent a guarantee of how fast generalization error will decrease in the limiting case when exemplars are so cheap that we do not care how many we use during training.

Moreover, since for error functions that are L_p norms our bounds are tight, we can say something even stronger in this case. For L_p, there exists a set of basis functions and a function in their convex hull such that no matter *how many examples* are used in training, the error can decrease no faster than the bounds we have provided. Thus, our results exhibit a worst-case speed limitation for learning.

1.5. Contents of the Paper

It is a triviality that optimal approximants to approximable functions always converge. However, the rates of convergence depend critically upon the structure of the space. In some spaces, like L_1, there exist target functions for which the rate can be made arbitrarily slow (Sect. 2.1). In Banach spaces of (Rademacher) type t with $t > 1$, however, a rate bound of $O(n^{-1+1/t})$ is obtained (Sect. 2.2). For L_p spaces these results specialize to those of Table 1. Particular examples of L_p spaces are given to show that the orders given in our bounds cannot in general be sharpened (Sect. 2.3).

Section 3 studies incremental approximation. A particularly interesting aspect of these results is that the new element of S added to the incremental approximant is not required to be the best possible choice. Instead, the new element can meet a less stringent test (Theorem 3.5). Also, the convex combination of the elements included in the approximant is not optimized. Instead a simple average is used. (This is an example of a fixed convexity schedule, as defined in Sect. 1.2.) Thus, our incremental approximants are the simplest yet studied, simpler even than those of (Jones 1992). Nonetheless, the same worst-case order is obtained for these approximants on L_p, $1 < p < \infty$, as for the optimal approximant. In more general spaces, the incremental approximants may not even converge (Sect. 3.1). However, if the space has a modulus of smoothness of power type greater than one, or is of Rademacher type t, then rate bounds can be given (Sects. 3.2 and 3.3).

Both optimal and incremental convergence rates may be improved if S has special structure (Sect. 4). In particular, we provide some analysis of the situation where S is a finite-VC dimension set of indicator functions and the sup norm is to be used, which is a common setting for neural network approximation.

2. Optimal Approximants

In this section we study rates of convergence for optimal convex approximates. To illustrate the fact that the issue is nontrivial, we begin by identifying a class of spaces for which the best possible rate $\phi(n)$ is constant, that is to say, no nontrivial rate is possible (Theorem 2.3). This class includes infinite dimensional L_1 and L_∞ (or $C(X)$) spaces.

In Theorem 2.5 we then study general bounds valid for spaces of (Rademacher) type t. It is well-known that L_p spaces with $1 \leq p < \infty$ are of type $\min\{p, 2\}$ (Ledoux and Talagrand 1991); on this basis an explicit specialization to L_p is given in Corollary 2.6.

We then close this section with explicit examples showing that the obtained bounds are tight.

2.1. Examples of Spaces Where No Rate Bound is Possible

In some spaces, the worst-case rate of convergence of optimal approximants can be shown to be arbitrarily slow.

Lemma 2.1. *Let (a_n) be a positive, convex ($a_n + a_{n+2} \geq 2a_{n+1}$) sequence converging to 0. Define $a_0 = 2a_1$ and $b_n = a_{n-1} - a_n$. Let $S = \{a_0 e_k\}$, where $\{e_k\}$ is the canonical basis in l_1, and consider $f = (b_n)$ as an element of l_1. Then $f \in \overline{\mathrm{co} S}$ and*

$$\|\mathrm{lin}_N S - f\| = a_N \qquad \text{for all } N. \tag{2.1}$$

Proof. Note that $\sum_{n=1}^{\infty} b_n/a_0 = 1$, so clearly $f \in \overline{\mathrm{co} S}$. By convexity (b_n) is a non-increasing sequence, so

$$\|\mathrm{lin}_N S - f\| = \sum_{i=N+1}^{\infty} b_i = \sum_{i=N+1}^{\infty} a_{i-1} - a_i = a_N. \tag{2.2}$$

□

Consider next the space l_∞. Let ϵ_k be an enumeration of all $\{-1, 0, 1\}$-valued sequences that are eventually constant, i.e., $\epsilon_k(n) \in \{-1, 0, 1\}$ for all $n \in \mathbb{N}$, and for each k there exists an N such that $\epsilon_k(n) = \epsilon_k(N)$ for all $n > N$. For each n, let $g_n \in l_\infty$ be the sequence $g_n(k) = \epsilon_k(n)$, and define the map $T : l_1 \to l_\infty$ by $T(e_n) = g_n$. The reader may check that T is an isometric embedding. Therefore T carries the example of Lemma 2.1 into l_∞.

What happens in c_0, the space of all sequences converging to 0? We will now construct a projection from $T(l_1)$ into c_0 that will retain the desired convergence rate. We will need, however, the extra restriction that the sequence (a_n) be strictly convex, i.e., that $a_n + a_{n+2} > 2a_{n+1}$.

Let $b_n = a_{n-1} - a_n$ as before, and define the auxiliary sequences

$$\underline{c}_N = \min\{n \in \mathbb{N} \mid b_n < a_N\}$$
$$\bar{c}_N = \min\{n \in \mathbb{N} \mid n > N+1 \text{ and } a_n < b_N - b_{N+1}\}.$$

The sequence $\underline{c}$ is well defined because $a_N > 0$ for all N and $b_n \downarrow 0$. Similarly, $\bar{c}$ is well defined since $a_n \downarrow 0$ and by strict convexity $b_N - b_{N+1} > 0$. Note that $\underline{c}_N \leq N+1 < N+2 \leq \bar{c}_N$. Moreover, $\underline{c}_N$ (and hence $\bar{c}_N$) goes to infinity with N since $b_n > 0$ for each n while $a_N \downarrow 0$.

Next define for each $N \in \mathbb{N}$,

$$A_N = \{k \in \mathbb{N} \mid \epsilon_k(n) = 0 \text{ for } n < \underline{c}_N \text{ and } \epsilon_k(n) = \epsilon_k(\bar{c}_N + 1) \text{ for } n > \bar{c}_N\},$$

and define for convenience the single element set $A_0 = \{k \mid \epsilon_k = \delta_k\}$, where $\delta_k(n)$ is the sequence that is 1 for $n = k$ and 0 otherwise. Then let $A = \cup_N A_N$.

Let P be the projection that sends an element h of l_∞ to the sequence $P(h)(k) = h(k)$ if $k \in A$ and $P(h)(k) = 0$ otherwise. Notice that if $\epsilon_k(n) \neq 0$, then $k \notin A_N$ for all N such that $n < \underline{c}_N$. Since $\underline{c}_N \to \infty$, if follows that there exists for each n only finitely many k's in A such that $\epsilon_k(n) \neq 0$. (Each A_N is a finite set.) Therefore $P(g_n) = P \circ T(e_n) \in c_0$ for each n, i.e., $P \circ T : l_1 \to c_0$.

It remains to show that

$$\|P \circ T(f) - \text{lin}_N P \circ T(S)\| = a_N.$$

Let us introduce the notation $\tilde{h}$ for $P \circ T(h)$, $h \in l_1$, and similarly $\tilde{S}$ for $P \circ T(S)$. It is clear that $\|\tilde{f} - \text{lin}_N \tilde{S}\| \leq a_N$, since T is an isometry and $\|P\| = 1$. To examine the bound from below, let $\tilde{f}_N = \sum_{n=1}^N d_n \tilde{e}_{m_n}$ be an arbitrary element of $\text{lin}_N \tilde{S}$, where $\{m_1, m_2, \ldots, m_N\}$ is a sampling of $\mathbb{N}$ of size N. We aim to produce a $k_0 \in A$ such that $\tilde{f}_N(k_0) = 0$ and $\tilde{f}(k_0) \geq a_N$, since then

$$\|\tilde{f} - \tilde{f}_N\| = \sup_{k \in A} |\tilde{f}(k) - \tilde{f}_N(k)| \geq |\tilde{f}(k_0) - \tilde{f}_N(k_0)| \geq a_N.$$

Let $n_0 = \min(\mathbb{N} \setminus \{m_1, m_2, \ldots, m_N\})$. If $n_0 < \underline{c}_N$, select k_0 such that $\epsilon_{k_0} = \delta_{n_0}$. If $n_0 = N+1$ (which is the largest possible value for n_0), select k_0 such that $\epsilon_{k_0}(n) = 0$ for $n \leq N$ and $= 1$ otherwise. It follows from $\underline{c}_N \leq N+1$ that $k_0 \in A$ in either event, and clearly $\tilde{f}_N(k_0) = 0$. Moreover, in the first case $\tilde{f}(k_0) = b_{n_0} \geq a_N$ by the definition of $\underline{c}_N$, while in the second case, $\tilde{f}(k_0) = \sum_{n=N+1}^\infty b_n = a_N$.

Lastly, consider the case $\underline{c}_N \leq n_0 \leq N$. Select k_0 so that $\epsilon_{k_0}(n) = 0$ if $n \in \{m_1, m_2, \ldots, m_N\}$ or if $n > \bar{c}_N$, and $\epsilon_{k_0}(n) = 1$ otherwise. This sequence is guaranteed to be in A_N, and $\tilde{f}_N(k_0) = 0$. Moreover,

$$\tilde{f}(k_0) = \sum_{n=\underline{c}_N}^{\bar{c}_N} b_n \epsilon_{k_0}(n) \geq b_N + \sum_{n=N+2}^{\bar{c}_N} b_n.$$

The inequality holds because $\epsilon_{k_0}(n) = 1$ for at least one $n \leq N$, and (b_n) is a decreasing sequence. It then follows from the definition of $\bar{c}_N$ that

$$b_N + \sum_{n=N+2}^{\bar{c}_N} b_n = b_N + a_{N+1} - a_{\bar{c}_N} > a_N,$$

so $\|\tilde{f} - \tilde{f}_N\| \geq a_N$, completing the proof.

Lemma 2.2. *Let (a_n) be a positive, strictly convex $(a_n + a_{n+2} > 2a_{n+1})$ sequence converging to 0. Then there exists a bounded set $S \subset c_0$ (with $\|g\| \leq 2a_1$ for all $g \in S$) and $f \in \overline{\mathrm{co}S}$ such that*

$$\|f - \mathrm{lin}_N S\| = a_N \qquad \text{for all } N \in \mathbb{N}.$$

An alternate method of proof is to replace the projection P in the discussion above with a map $T' : l_\infty \to c_0$ defined by $T'(h)(k) = \delta_k h(k)$, where $\delta_k \downarrow 0$ is carefully chosen (as a function of (a_n)) to preserve the inequality $\|T' \circ T(f) - \mathrm{lin}_N T' \circ T(S)\| \geq a_N$. The details are left to the reader.

In either method, the constructed base set $S \subset c_0$ depends on the rate sequence (a_n). It is interesting to compare this with the situation in l_1 and l_∞, where the set S is universal, i.e., independent of (a_n). (Though the limit function $f \in \overline{\mathrm{co}S}$ does vary with (a_n).)

The preceding discussion showing the absence of a rate bound in l_∞ relied upon an isometric embedding of l_1 into l_∞. The same argument can be used if we have only an isomorphic embedding, i.e., a bounded linear map with bounded inverse. Combining this with the two preceding lemmas gives

Theorem 2.3. *Let X be a Banach space with a subspace isomorphic to either l_1 or c_0. Then for any positive sequence (a_n) converging to 0, it is possible to construct a bounded set S and $f \in \overline{\mathrm{co}S}$ such that*

$$\|\mathrm{co}_N S - f\| \geq \|\mathrm{lin}_N S - f\| \geq a_N. \tag{2.3}$$

Proof. If (a_n) is not convex, then replace it with a strictly convex sequence $(\bar{a}_n)$ converging to zero such that $\bar{a}_n \geq a_n$ for all n. This is a well-known construction, for example (Stromberg 1981, p. 515). The result then follows from Lemmas 2.1 and 2.2. □

If a Banach space contains a subspace isomorphic to l_1 or c_0, then it is not reflexive. The converse is not true, though there are a variety of results that do imply the existence of subspaces isomorphic to l_1 or c_0, and such results combine with Theorem 2.3 to identify spaces with arbitrarily slow convergence of the type discussed. For example:

Theorem 2.4. *[(Lindenstrauss and Tzafriri 1979, p. 35)] A Banach lattice X is reflexive if and only if no subspace of X is isomorphic to l_1 or c_0.*

This theorem implies that examples of arbitrarily slow convergence exist in $L_1[0,1]$ and $C[0,1]$, although in these cases it is also easy to construct the embeddings directly (Darken et al. 1993). Related results can be found in (Bessaga and Pelczynski 1958), (Rosenthal 1974), (Rosenthal 1994), and (Gowers preprint).

In Section 3 we will show that a condition somewhat stronger than reflexivity (uniform smoothness) guarantees convergence of incremental approximates. In the next section, however, we show that a condition unrelated to reflexivity (the Rademacher type property) can guarantee convergence of optimal approximants and even give bounds on convergence rates.

2.2. Bounds for Type t Spaces

We recall some basic definitions first.

A *Rademacher sequence* $(\epsilon_i)_{i=1}^n$ is a finite sequence of independent zero mean random variables taking values from $\{-1,+1\}$. Given any Banach space X, any Rademacher sequence $(\epsilon_i)_{i=1}^n$, and any fixed finite sequence $(f_i)_{i=1}^n$ of elements of X, we can view in a natural manner the expression $\sum_{i=1}^n \epsilon_i f_i$ as a random variable taking values in X. With this understanding, the space X is of *(Rademacher) type t (with constant C)* if for each Rademacher sequence (ϵ_i) and each sequence (f_i) it holds that

$$E\left\|\sum \epsilon_i f_i\right\|^t \;\leq\; C\sum \|f_i\|^t . \tag{2.4}$$

It is interesting to note that t can never be greater than 2, that Hilbert spaces are of type 2, and that there exists nonreflexive Banach spaces of type 2 (James 1978).

Theorem 2.5. *Let X be a Banach space of type t, where $1 \leq t \leq 2$. Pick $S \subset X$, $f \in \overline{\mathrm{co}(S)}$, and $K > 0$ such that $\forall g \in S$, $\|g - f\| \leq K$. Then for all n,*

$$\|\mathrm{co}_n S - f\| \leq \frac{KC^{1/t}}{n^{1-1/t}}, \tag{2.5}$$

where C is a constant depending on X but independent of n.

Proof. $\forall\epsilon > 0, \exists n_\epsilon$, $\alpha_1, \ldots, \alpha_{n_\epsilon} \in \mathbb{R}^+$, and $f_1, \ldots, f_{n_\epsilon} \in S$ such that

$$\sum_{i=0}^{n_\epsilon} \alpha_i = 1, \qquad \sum_{i=0}^{n_\epsilon} \alpha_i f_i + \Delta = f,$$

and $\|\Delta\| < \epsilon$. Take ξ_j to be a sequence of independent random variables on X taking value f_i with probability α_i. Then for any $\beta \in (0,1)$,

$$\begin{aligned} &E\left\|f - \frac{1}{n}\sum_{j=1}^n \xi_j\right\|^t \\ &= \frac{1}{n^t} E\left\|\sum_{j=1}^n (f - \xi_j - \Delta) + n\Delta\right\|^t \end{aligned} \tag{2.6}$$

$$
\begin{aligned}
&\leq \frac{1}{n^t} E \left((1-\beta) \frac{\left\| \sum_{j=1}^n (f - \xi_j - \Delta) \right\|}{1-\beta} + \beta n \frac{\|\Delta\|}{\beta} \right)^t \\
&\leq \frac{1}{n^t} \left[(1-\beta) E \left(\frac{\left\| \sum_{j=1}^n (f - \xi_j - \Delta) \right\|}{1-\beta} \right)^t + \beta \left(\frac{n \|\Delta\|}{\beta} \right)^t \right] \\
&= \frac{1}{n^t (1-\beta)^{t-1}} E \left\| \sum_{j=1}^n (f - \xi_j - \Delta) \right\|^t + \frac{1}{\beta^{t-1}} \|\Delta\|^t ,
\end{aligned}
$$

which follows because $\phi(x) = x^t$ is a convex function for $1 \leq t \leq 2$. Since the range of ξ_j has finitely many values and the space is type t, by (Ledoux and Talagrand 1991, Prop. 9.11, p. 248) it follows that:

$$
E \left\| \sum_{j=1}^n (f - \xi_j - \Delta) \right\|^t \leq C \sum_{j=1}^n E \|f - \xi_j - \Delta\|^t . \tag{2.7}
$$

On the other hand, we have:

$$
\begin{aligned}
E\|f - \xi_1 - \Delta\|^t &= \sum_{i=1}^{n_\epsilon} \alpha_i \|f - f_i - \Delta\|^t \\
&\leq \sum_{i=1}^{n_\epsilon} \alpha_i \left(\|f - f_i\| + \|\Delta\| \right)^t \\
&< \sum_{i=1}^{n_\epsilon} \alpha_i (K + \epsilon)^t \\
&= (K + \epsilon)^t.
\end{aligned} \tag{2.8}
$$

Without loss of generality, assume $0 < \epsilon < 1$ and take $\beta = \epsilon$. Then combining (2.6), (2.7), and (2.8),

$$
E \left\| f - \frac{1}{n} \sum_{j=1}^n \xi_j \right\|^t < \frac{C(K+\epsilon)^t}{n^{t-1}(1-\epsilon)^{t-1}} + \epsilon.
$$

We conclude that for some realization of the ξ_j (labeled g_j) the inequality must hold, i.e.,

$$
\left\| f - \frac{1}{n} \sum_{j=1}^n g_j \right\|^t < \frac{C(K+\epsilon)^t}{n^{t-1}(1-\epsilon)^{t-1}} + \epsilon.
$$

Taking the infimum with respect to all $\epsilon > 0$ proves the theorem. $\square$

We now give a specialization to L_p, $1 \leq p < \infty$. These spaces are of type $t = \min\{p, 2\}$. From (Haagerup 1982) we find that the best value for C in (2.5) is 1 if $1 \leq p \leq 2$ and $\sqrt{2}\,[\Gamma((p+1)/2)/\sqrt{\pi}]^{1/p}$ if $2 < p < \infty$. One may use Stirling's formula to get an asymptotic formula for the latter expression.

Corollary 2.6. *Let X be an L_p space with $1 \leq p < \infty$. Suppose $S \subset X$, $f \in \overline{\mathrm{co}S}$, and $K > 0$ such that $\forall g \in S$, $\|g - f\| \leq K$. Then for all n,*

$$\|\mathrm{co}_n S - f\| \leq \frac{KC_p}{n^{1-1/t}}, \tag{2.9}$$

where $t = \min\{p, 2\}$, and $C_p = 1$ if $1 \leq p \leq 2$ and $C_p = \sqrt{2}\,[\Gamma((p+1)/2)/\sqrt{\pi}]^{1/p}$ for $2 < p < \infty$. For large p, $C_p \sim \sqrt{p/e}$.

2.3. Tightness of Rate Bounds

In this section we show that order of the rate bound given for L_p in (2.9) is tight. That is, we give specific examples of L_p spaces and subsets S with target functions $f \in \overline{\mathrm{co}S}$ where optimal approximants converge with the specified order.

Theorem 2.7. *There exists $S \subseteq l_p$, $1 < p < \infty$, and $f \in \overline{\mathrm{co}S}$ such that $\|\mathrm{co}_n S - f\|_p = Kn^{-1+1/p}$ where $K = \sup_{g \in S} \|f - g\|_p$.*

Proof. Let S consist of the elements of the canonical basis for l_p. Then $f := 0$ is in the closed convex hull of S and $\sup_{g \in S} \|f - g\|_p = 1$. Let f_n be an element of $\mathrm{co}_n S$ that is to approximate 0. So f_n is of the form $f_n = \sum_{k=1}^n a_k g_{n_k}$, where each g_{n_k} is an element of S, and the a_k are non-negative and sum to 1. Without loss of generality, we may assume the g_{n_k} are distinct, since otherwise we would be effectively working in $\mathrm{co}_m S$ with $m < n$. Then

$$\|f_n - 0\|^p = \sum_{k=1}^n a_k^p.$$

It is easy to see that the error is minimized by taking the a_k's all equal, namely $\forall k$, $a_k = 1/n$ (Jensen). Therefore

$$\|f_n - 0\| \geq n^{(1-p)/p} = n^{-1+1/p}.$$

□

Next we show that the $O(n^{-1/2})$ bound for L_p, $2 < p < \infty$, is tight (see Corollary 2.6). We borrow from computer science the notation $\psi(n) = \Omega(\phi(n))$ to mean that there is a constant $C > 0$ so that $\psi(n)/\phi(n) \geq C$ for all n large enough. For the purposes of the next result, we say a space L_p is *admissible* if there exists a Rademacher sequence definable on it; for instance, $L_p(0, 1)$ with the usual Lebesgue measure is one such space.

Proposition 2.8. *For any admissible L_p, $2 < p < \infty$, there exists a subset S and an $f \in \overline{\mathrm{co}S}$ such that $\|\mathrm{co}_n S - f\|_p = \Omega(n^{-1/2})$.*

Proof. Let ξ_i, i.i.d. on {-1,+1}, be a Rademacher sequence. Define $S = \{\xi_i\}$, which is a subset of the unit ball in L_p. Using the upper bound of Khintchine's Inequality (Khintchine 1923; Ledoux and Talagrand 1991, Lemma 4.1, p. 91), one can show that $0 \in \overline{\mathrm{co}S}$. Suppose the best approximation of 0 by a convex sum of n elements of S is $\sum_{i=1}^{n} \alpha_i \xi_{k(i)}$. Then by the lower bound of Khintchine's Inequality,

$$\left\| \sum_{i=1}^{n} \alpha_i \xi_{k(i)} \right\|_{L_p} \geq A_p \sqrt{\sum_i \alpha_i^2} = A_p \|(\alpha_i)\|_{l_2}.$$

But we have already given an example in l_2 (in the proof of Theorem 2.7) for which the last term is $\Omega(n^{-1/2})$. □

3. Incremental Approximants

We now start the study of incremental approximation schemes. Unlike the situation with optimal approximation, incremental approximations are not guaranteed to even converge. In general, the convergence of incremental schemes appears to be intimately tied to the concept of norm smoothness. In Theorem 3.1 we show that smoothness is equivalent to at least a monotonic decrease of the error, and then in Theorem 3.4 it is proved that uniform smoothness is a sufficient condition to guarantee convergence. (It is possible to construct a smooth space with an ϵ-greedy sequence that does not converge—Appendix D. However, if an ϵ-greedy sequence converges, then it can only converge to the desired target function—Corollary 3.2.)

In Sects. 3.2 and 3.3 we study upper bounds on the rate of convergence for spaces with modulus of smoothness of power type greater than 1 and for spaces of (Rademacher) type t, $1 < t \leq 2$. The L_p spaces, $1 < p < \infty$, are examples of spaces with modulus of smoothness of power type $t = \min(p, 2)$ (see Appendix B), which themselves sit inside the more general class of spaces of type t. (See, for example, (Lindenstrauss and Tzafriri 1979, p. 78; Deville et al. 1993, p. 166), while (James 1978) shows that the containment is strict. See also (Figiel and Pisier 1974).) The upper bounds obtained for incremental approximation error in the power type spaces agree with the bounds for optimal approximation error obtained in Sect. 2.2 (albeit with a slightly larger constant of proportionality), which were shown to be the best possible in Sect. 2.3. Therefore, little is lost by using incremental approximates instead of optimal approximates, at least in worst-case settings. The incremental convergence bounds obtained in Sect. 3.3 for type-t spaces are only slightly weaker than the optimal approximation error bounds obtained in Sect. 2.2

3.1. Convergence of Greedy Approximants

The first remark is that for some spaces there may not exist any nondecreasing rate whatsoever. In the terminology given in the introduction, it may be the case that there are greedy incremental sequences for f for which $\|f_n - f\| \not\to 0$. This will happen in particular if there are a set S and two elements $f \neq f_n \in \mathrm{co}S$ so that for each $g \in S$ and each $h \in \mathrm{co}\,(\{f_n, g\})$ different from f_n, $\|h - f\| > \|f_n - f\|$; in that case, the successive minimizations result in the sequence $f_n, f_n, \ldots$, which doesn't converge to f. Geometrically, convergence of incremental approximants can fail to occur if the unit ball for the norm has a sharp corner. This is illustrated by the example in Fig. 1, for the plane $\mathbb{R}^2$ under the L^1 norm. In order to use the intuition gained from this example, we need a number of standard but perhaps less-known notions from functional analysis.

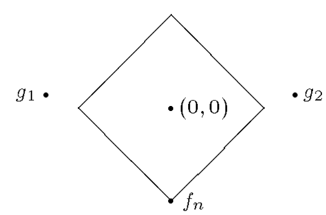

Fig. 1. Incremental approximants may fail to converge in some spaces. Consider approximating $f = (0,0)$ by linear combinations of elements from $\{f_n, g_1, g_2\}$ according to the $L_1(\mathbb{R}^2)$ metric. The best approximant by one point is f_n. The rotated square is the contour of the norm about f on which f_n lies. The best approximant by a linear combination of f_n and g_1 or g_2 is once again f_n. Thus even though f is in the convex hull of $\{f_n, g_1, g_2\}$, incremental approximants fail to converge or even to decrease monotonically.

If X is a Banach space and $f \neq 0$ is an element of X, a *peak functional* for f is a bounded linear operator, that is, an element $F \in X^*$, that has norm $= 1$ and satisfies $F(f) = \|f\|$. (The existence for each $f \neq 0$ of at least one peak functional is guaranteed by the Hahn-Banach Theorem.) Geometrically, one may think of the null space of $F - \|f\|$ as a hyperplane tangent at f to the ball centered at the origin of radius $\|f\|$. (For Hilbert spaces, there is a unique peak functional for each f, which can be identified with $(1/\|f\|)f$ acting by inner products.) The space X is said to be *smooth* if for each f there is a unique peak functional. (Roughly, this means that balls in X have no "corners.") The *modulus of smoothness* of any Banach space X is the function $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined by

$$\rho(r) := \frac{1}{2}\left(\sup_{\|f\|=\|g\|=1} \{\|f + rg\| + \|f - rg\|\} - 2\right)$$

Note that, by sub-additivity of norms, always $\rho(r) \leq r$. For Hilbert spaces, one has $\rho(r) = \sqrt{1 + r^2} - 1$. A Banach space is said to be *uniformly smooth* if $\rho(r) = o(r)$ as $r \to 0$; in particular, Hilbert spaces are uniformly smooth, but so are L_p spaces with $1 < p < \infty$, as is reviewed in Appendix B. (Here one has an upper bound of the type $\rho(r) < cr^t$, for some $t > 1$, which implies uniform smoothness.) As a remark, we point out that uniform smooth spaces are in particular smooth. Moreover, it can be shown that uniformly smooth spaces are also reflexive, though the converse implication does not hold.

The next result implies that greedy approximants always result in monotonically decreasing error if and only if the space is smooth. In particular, if the space is not smooth, then one can not expect greedy (or even ϵ-greedy) incremental sequences to converge. (Since one may always consider the translate $S - \{f\}$ as well as a translation by f of all elements in a greedy sequence, no generality is lost in taking $f = 0$.)

Theorem 3.1. *Let X be Banach space. Then:*

1. *Assume that X is smooth, and pick any $S \subset X$ so that $0 \in \overline{\mathrm{co} S}$. Then for each nonzero $f \in X$ there is some $g \in S$ and some $\tilde{f} \in \mathrm{co}(\{f, g\})$ different from f so that $\|\tilde{f}\| < \|f\|$.*
2. *Conversely, if X is not smooth, then there exist an $S \subset X$ so that $0 \in \overline{\mathrm{co} S}$ and an $f \in S$ so that, for every $g \in S$ and every $\tilde{f} \in \mathrm{co}(\{f, g\})$ different from f, $\|\tilde{f}\| > \|f\|$.*

Proof. Assume that X is smooth, and let S and f be as in the statement. Let F be the (unique) peak functional for f. There must be some $g \in S$ for which

$$F(g) < \|f\|/2\,, \tag{3.10}$$

since otherwise $\{h \in X \mid F(h) = \|f\|/2\}$ would be a hyperplane separating $\mathrm{co}(S)$ from $0 \in \overline{\mathrm{co} S}$. Define

$$f_\lambda = (1 - \lambda) f + \lambda g \qquad \text{for } \lambda \in [0, 1].$$

We wish to show that $\|f_\lambda\| < \|f\|$ for some $\lambda \in (0, 1]$, as this will establish the first part of the Theorem. For this, consider the peak functional F_λ for f_λ. Note that

$$\lim_{\lambda \downarrow 0} F_\lambda(f) = \lim_{\lambda \downarrow 0} F_\lambda(f_\lambda) + F_\lambda(f - f_\lambda) = \lim_{\lambda \downarrow 0} \|f_\lambda\| = \|f\|. \tag{3.11}$$

The unit ball in X^* is weak-$*$ compact (Alaoglu), so the net $(F_\lambda)_{\lambda \in (0,1)}$ (where $\lambda \downarrow 0$) has a convergent subnet, say $F_{\lambda_\alpha} \to F^*$, with $\|F^*\| \leq 1$. In particular, $F_{\lambda_\alpha}(f) \to F^*(f)$, so it follows from (3.11) that $F^*(f) = \|f\|$, and hence F^* is a peak functional for f. But X is smooth, so $F^* = F$. Thus there exists $\lambda_0 \in (0, 1]$ such that $|F_{\lambda_0}(g) - F(g)| < \|f\|/2$, which combines with (3.10) to give $F_{\lambda_0}(g) < \|f\|$. Therefore

$$\|f_{\lambda_0}\| = F_{\lambda_0}(f_{\lambda_0}) = (1 - \lambda_0) F_{\lambda_0}(f) + \lambda_0 F_{\lambda_0}(g) < \|f\|,$$

since $\lambda_0 > 0$. This proves the first assertion.

We now prove the converse. Since X is not smooth, there is some unit vector f with two distinct peak functionals F and F'. Since $F \neq F'$, there is some $h \in X$ so that $F(h) \neq F'(h)$. Let

$$g := h - \left(\frac{F(h) + F'(h)}{2} \right) f\,.$$

Note that $F(g) + F'(g) = 0$ and $F'(g) \neq 0$; scaling g, we may assume that $F'(g) = 2$. Consider now the set $S = \{f, g_1, g_2\}$, where $g_1 = g$ and $g_2 = -g$; note that $0 \in \operatorname{co} S$. This provides the needed counterexample, since

$$\|(1-\lambda)f + \lambda g_1\| \geq F'((1-\lambda)f + \lambda g) = 1 + \lambda > 1 = \|f\|$$

and

$$\|(1-\lambda)f + \lambda g_2\| \geq F((1-\lambda)f - \lambda g) = 1 + \lambda > 1 = \|f\|$$

for each $\lambda > 0$. □

It is interesting to remark that, for the set S built in the last part of the proof, even elements in the affine span of f and g_1 (or of f and g_2) have norm > 1 if distinct from f, that is, the inequalities hold in fact for all $\lambda \neq 0$. (For $\lambda < 0$, interchange F and F' in the last pair of equations.)

It is an easy consequence of the first part of Theorem 3.1 that greedy incremental approximates in a smooth Banach space can converge only to the target function:

Corollary 3.2. *Let X be a smooth Banach space with $S \subset X$. Let $f \in \overline{\operatorname{co} S}$ and suppose $f_1, f_2, \ldots$ is an incremental ϵ-greedy sequence with respect to f, where the schedule $\epsilon_1, \epsilon_2, \ldots$ converges to 0. If the sequence (f_n) converges, then it converges to f.*

Proof. Without loss of generality, we may assume that $f = 0$. Suppose $\lim_n f_n = f_\infty \neq 0$. Then by the first part of Theorem 3.1, there exists $g \in S$ and $\lambda \in [0,1]$ such that $\|(1-\lambda)f_\infty + \lambda g\| < \|f_\infty\|$. Define $\delta = \|f_\infty\| - \|(1-\lambda)f_\infty + \lambda g\|$, and choose N large enough so $\|f_\infty - f_n\| < \delta/3$ and $\epsilon_n < \delta/3$ for all $n > N$. Fix $n > N$. Then $\|(1-\lambda)f_n + \lambda g\| < \|f_\infty\| - 2\delta/3$, but (f_n) is ϵ-greedy, which implies $\|f_{n+1}\| < \|f_\infty\| - \delta/3$. But this is impossible since by choice of N, $\|f_\infty - f_{n+1}\| < \delta/3$. Therefore the limit $f_\infty = 0$, as desired. □

It is possible, however, to have an ϵ-greedy sequence that fails to converge. See Appendix D. This situation is avoided if X is uniformly smooth, as we shall see below. But first we need a technical lemma that captures the geometric properties of smoothness necessary to obtain stepwise estimates of convergence. This lemma is used not only in Theorem 3.4, but also throughout Sect. 3.2.

Lemma 3.3. *Let X be a Banach space with modulus of smoothness $\rho(u)$, and let $S \subset X$. Assume that $0 \in \overline{\operatorname{co}(S)}$ and let $f \neq 0$ be an element of $\operatorname{co}(S)$. Let F be a peak functional for f. Then*

$$(3.12)\quad \|(1-\lambda)f + \lambda g\| \leq (1-\lambda)\left[1 + 2\rho\left(\frac{\lambda\|g\|}{(1-\lambda)\|f\|}\right)\right]\|f\| + \lambda F(g),$$

for all $0 \leq \lambda < 1$ and all $g \in S$. Furthermore, for any $\epsilon > 0$, there exists a $g \in S$ such that $F(g) < \epsilon$.

Proof. Pick any $0 \leq \lambda < 1$ and $g \in S$. If $g = 0$ then (3.12) is trivially satisfied, so assume $g \neq 0$.

Define $h = f + ug/\|g\|$ and $h^* = f - ug/\|g\|$ for $u \geq 0$. Then from the definition of the modulus of smoothness we have

$$\|h\| + \|h^*\| \leq 2\|f\| \left[1 + \rho(u/\|f\|)\right].$$

But

$$\|h^*\| \geq F\left(f - \frac{u}{\|g\|} g\right) = \|f\| - uF(g)/\|g\|.$$

Therefore

$$\|h\| \leq \|f\| \left(1 + 2\rho(u/\|f\|)\right) + uF(g)/\|g\|. \tag{3.13}$$

If we set $u = \lambda\|g\|/(1-\lambda)$, we get

$$(1-\lambda)f + \lambda g \; = (1-\lambda)h,$$

which combines with (3.13) to prove (3.12).

Finally, given $\epsilon > 0$, suppose there is no $g \in S$ such that $F(g) < \epsilon$. Then the affine hyperplane $\{h \in X \mid F(h) = \epsilon/2\}$ would separate S from 0, contradicting $0 \in \overline{\mathrm{co}(S)}$. $\square$

Theorem 3.4. *Let X be a uniformly smooth Banach space. Let S be a bounded subset of X and let $f \in \overline{\mathrm{co}(S)}$ be given, and let (ϵ_n) be an incremental schedule with $\sum_{k=1}^{\infty} \epsilon_k < \infty$. Then any ϵ-greedy (with respect to f) incremental sequence $(f_n) \subset \mathrm{co}\, S$ converges to f.*

Proof. Pick $K \geq \sup_{g\in S} \|f - g\|$, and let (f_n) be an ϵ-greedy incremental sequence. Define

$$a_n := \|f_n - f\|.$$

We want to show that $a_n \to 0$. To this end, let $a_\infty = \liminf_{n\to\infty} a_n$. Since (f_n) is ϵ-greedy, $a_{n+1} \leq a_n + \epsilon_n$ and $a_{n+m} \leq a_n + \sum_{k=n}^{m-1} \epsilon_k$. But $\sum_{k=n}^{\infty} \epsilon_k \to 0$ as $n \to \infty$, so in fact $a_\infty = \lim_{n\to\infty} a_n$. Suppose $a_\infty > 0$. Then from the definition of ϵ-greedy and (3.12), it follows that

$$a_{n+1} \leq \inf_{\lambda,g} \left\{(1-\lambda)\left[1 + 2\rho\left(\frac{\lambda K}{(1-\lambda)a_n}\right)\right] a_n + \lambda F_n(g - f)\right\} + \epsilon_n,$$

where ρ is the modulus of smoothness for X, F_n is the peak functional for $f_n - f$, and the infimum is taken over all $0 \leq \lambda < 1$ and $g \in S$. (In relation to Lemma 3.3, everything here is translated by $-f$.) The modulus of smoothness is a non-decreasing function, so certainly

$$\rho\left(\frac{\lambda K}{(1-\lambda)a_n}\right) \leq \rho\left(\frac{2\lambda K}{(1-\lambda)a_\infty}\right)$$

for large enough n. Using this and taking the limit as $n \to \infty$ in the preceding inequality yields

$$a_\infty \leq a_\infty \left\{ 1 + \inf_\lambda \lambda \left[\frac{4K}{a_\infty} \frac{\rho\left(u(\lambda)\right)}{u(\lambda)} - 1 \right] \right\},$$

where $u(\lambda) := (2\lambda K)/[(1-\lambda)a_\infty]$. But $u(\lambda) \to 0$ as $\lambda \to 0$, so by uniform smoothness $\rho(u)/u \to 0$ as $\lambda \to 0$. Therefore the quantity in the infimum is negative, and a contradiction is reached. Thus, a_∞ must be zero. □

The stepwise selection of $\lambda = \lambda_n$ in the above proof apparently depends upon the modulus of smoothness $\rho(u)$. We shall see in the next section that if we have a non-trivial power type estimate on the modulus of smoothness then it suffices to use $\lambda_n = 1/(n+1)$, i.e., f_{n+1} becomes a simple average of certain elements $g_0, g_1, \ldots, g_n$, chosen to satisfy a particular inequality (see (3.15) below).

3.2. Spaces with Modulus of Smoothness of Power Type Greater than One

We now give rate bounds for incremental approximates that hold for all Banach spaces with modulus of smoothness of power type greater than one (Theorems 3.5 and 3.7). Keep in mind that $\rho(u) \leq \gamma u^t$ with $t > 1$ is a sufficient condition for X to be uniformly smooth, and holds in particular for L_p-spaces if $1 < p < \infty$. (See Appendix B.)

Theorem 3.5. *Let X be a uniformly smooth Banach space having modulus of smoothness $\rho(u) \leq \gamma u^t$, with $t > 1$. Let S be a bounded subset of X and let $f \in \overline{\mathrm{co}(S)}$ be given. Select $K > 0$ such that $\|f - g\| \leq K$ for all $g \in S$, and fix $\epsilon > 0$. If the sequences $(f_n) \subset \mathrm{co}(S)$ and $(g_n) \subset S$ are chosen recursively such that*

$$f_1 \in S \tag{3.14}$$

$$F_n(g_n - f) \leq \frac{2\gamma}{n^{t-1}\|f_n - f\|^{t-1}}\left((K+\epsilon)^t - K^t\right) := \delta_n \tag{3.15}$$

$$f_{n+1} = \left(\frac{n}{n+1}\right) f_n + \left(\frac{1}{n+1}\right) g_n, \tag{3.16}$$

(where F_n is the peak functional for $f_n - f$; we terminate the procedure if $f_n = f$), then

$$\|f_n - f\| \leq \frac{(2\gamma t)^{1/t}(K+\epsilon)}{n^{1-1/t}} \left[1 + \frac{(t-1)\log_2 n}{2tn} \right]^{1/t}. \tag{3.17}$$

Recall that (3.15) can always be obtained, since otherwise $\{h \in X \mid F_n(h - f) = \delta_n/2\}$ would be a hyperplane separating S from $f \in \overline{\mathrm{co}(S)}$.

Proof. Replacing S with $S - f$ allows us to assume without loss of generality that $f = 0$ and $\|g\| \leq K$ for all $g \in S$.

Applying Lemma 3.3 with $g = g_n$ and $\lambda = 1/(n+1)$ yields

$$\begin{aligned} \|f_{n+1}\| &\leq \frac{n\|f_n\|}{n+1}\left[1 + 2\rho\left(\frac{\|g_n\|}{n\|f_n\|}\right)\right] + \frac{\delta_n}{n+1} \\ &\leq \frac{n\|f_n\|}{n+1}\left[1 + \left(\frac{(2\gamma)^{1/t}(K+\epsilon)}{n\|f_n\|}\right)^t\right]. \end{aligned} \tag{3.18}$$

If we set

$$a_n := \frac{n\|f_n\|}{(2\gamma)^{1/t}(K+\epsilon)}$$

into the previous inequality we obtain

$$a_{n+1} \leq a_n(1 + 1/a_n^t).$$

Comparing to Lemma C.1, we see that this is half of (C.46). We can get the rest by applying the triangle inequality to (3.16) and making use of Lemma B.3 (B.38), i.e.,

$$a_{n+1} \leq a_n + 1/(2\gamma)^{1/t} < a_n + 3/2.$$

So to employ Lemma C.1, it remains only to show that (C.47) holds for $n = 2$ or for $n = 1$ with $a_1 \geq 1$. Note first that

$$a_1 = \frac{\|f_1\|}{(2\gamma)^{1/t}(K+\epsilon)} < \frac{1}{(2\gamma)^{1/t}} < t^t$$

by Lemma B.3 (B.41). So (3.17) holds for $n = 1$, and if $a_1 \geq 1$ then we can apply Lemma C.1 immediately. Otherwise $a_1 < 1$, in which case

$$a_2 \leq a_1 + \frac{1}{(2\gamma)^{1/t}} < \left(\frac{5t-1}{2}\right)^{1/t},$$

by Lemma B.4, and so (C.47) holds for $n = 2$.

It follows in either case that

$$a_n \leq \left(tn + \frac{t-1}{2}\log_2 n\right)^{1/t} \qquad \text{for all } n \geq 1.$$

Rewriting in terms of f_n proves the theorem. □

Recall that in L_p spaces, the modulus of smoothness is of power order t, where $t = \min(p, 2)$. The next corollary follows immediately from the preceding theorem and Lemma B.1.

Corollary 3.6. *Let S be a bounded subset of L_p, $1 < p < \infty$, with $f \in \overline{\mathrm{co}(S)}$ given. Define $q = p/(p-1)$ and select $K > 0$ such that $\|f-g\| \leq K$ for all $g \in S$. Then for each $\epsilon > 0$, there exists a sequence $(g_n) \subset S$ such that the sequence $(f_n) \subset \mathrm{co}(S)$ defined by*

$$f_1 = g_1 \qquad f_{n+1} = nf_n/(n+1) + g_n/(n+1)$$

satisfies

$$\|f - f_n\| \leq \frac{2^{1/p}(K+\epsilon)}{n^{1/q}} \left[1 + \frac{(p-1)\log_2 n}{n}\right]^{1/p}$$

if $1 < p \leq 2$ and

$$\|f - f_n\| \leq \frac{(2p-2)^{1/2}(K+\epsilon)}{n^{1/2}} \left[1 + \frac{\log_2 n}{n}\right]^{1/2}$$

if $2 \leq p < \infty$.

We now interpret Theorem 3.5 in terms of ϵ-greedy sequences. Let f, S, and X be as in that theorem, and let (f_n) be an ϵ-greedy sequence with respect to f, which as before we can assume to be 0. Then

$$\begin{aligned} \|f_{n+1}\| &\leq \inf_{\lambda,g} \left\{ (1-\lambda)\left[1 + 2\gamma\left(\frac{\lambda\|g\|}{(1-\lambda)\|f_n\|}\right)^t\right]\|f_n\| + \lambda F_n(g) \right\} + \epsilon_n \\ &\leq \inf_g \left\{ \frac{n\|f_n\|}{n+1}\left[1 + 2\gamma\left(\frac{\|g\|}{n\|f_n\|}\right)^t\right] + F_n(g)/(n+1) \right\} + \epsilon_n, \end{aligned}$$

by Lemma 3.3. The outside inequality holds also if (f_n) is only ϵ-greedy with respect to the convexity schedule $\lambda_n = 1/(n+1)$. Now given that the modulus of convexity satisfies $\rho(u) \leq \gamma u^t$ $(t > 1)$, fix $\epsilon > 0$, and select an incremental schedule (ϵ_n) satisfying $\epsilon_n \leq \epsilon\gamma/n^t$. Then using the fact that $\|g\| \leq K$ for all $g \in S$ and that there exists $g \in S$ with $F_n(g)$ smaller than any preassigned positive value, we get

$$\|f_{n+1}\| \leq \frac{n\|f_n\|}{n+1}\left[1 + 2\gamma\left(\frac{K}{n\|f_n\|}\right)^t\right] + \epsilon\gamma/n^t.$$

Recalling the definition of δ_n in (3.15), we see that $\epsilon\gamma/n^t \leq \delta_n/(n+1)$, and a comparison with (3.18) shows that the bound obtained in Theorem 3.5 also holds for the ϵ-greedy sequence (f_n) as well. This proves

Theorem 3.7. *Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^t$, with $t > 1$. Then X admits incremental convex schemes with rate $1/n^{1-1/t}$. Moreover, if the incremental schedule (ϵ_n) satisfies $\epsilon_n \leq \epsilon\gamma/n^t$ where ϵ is any fixed positive value, and if (f_n) is either ϵ-greedy or ϵ-greedy with convexity schedule $\lambda_n = 1/(n+1)$, then the error to the target function at step $n+1$ is bounded above by (3.17).*

The specialization of this result to L_p spaces, analogous to Corollary 3.6, is straightforward and is left to the reader.

Remark. The only non-constructive step in the proof of Theorem 3.5 is the determination of $g \in S$ such that $F_n(g-f) \leq \delta_n$, where F_n is the peak functional for $f_n - f$. In L_p spaces, F_n can be associated with the function in L_q ($q = p/(p-1)$) defined by

$$h_n(x) := \operatorname{sign}(f_n - f(x))|f_n - f(x)|^{p-1}/\|f_n - f\|_p^{p-1},$$

so

$$F_n(g-f) = \int h_n(x)\,(g(x) - f(x))\;dx.$$

This means that to satisfy (3.15), one must find $g \in S$ such that

$$\int h_n(x)g(x)\,dx \leq \delta_n + \int h_n(x)f(x)\,dx.$$

(We should point out that $\int h_n(x)f(x)\,dx$ is likely to be negative.)

The details of finding such a g will depend upon the specific class S under consideration. As an example, in the context of neural networks, suppose S consists of those functions $g(x)$ having the form $\pm\sigma(a \cdot x + b)$, where $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, σ is a fixed activation function, and $x \in \mathbb{R}^d$ is allowed to vary over a subset $\Omega \subset \mathbb{R}^d$. Then we are left with finding an a and b such that

$$\left|\int_\Omega h_n(x)\sigma(a \cdot x + b)\,dx\right| \geq -\int_\Omega h_n(x)f(x)\,dx - \delta_n. \tag{3.19}$$

Actually, the condition $f \in \overline{\mathrm{co}S}$ implies the existence of an a and b such that the left hand side of (3.19) is at least as large as $-\int_\Omega h_n(x)f(x)\,dx$, so this may be viewed as a maximization problem. We do not need to find the global maximum, however, but only a value satisfying the weaker condition (3.19).

3.3. Rate Bounds in Type t Spaces

We turn our attention now to determining rate bounds for incremental approximants in Rademacher type t spaces with $1 < t \leq 2$ (Corollary 3.13). The constants in the bounds are implicit. Furthermore, the rate bounds for the case of L_p spaces (not given explicitly) are slightly weaker than those established in the previous section.

Banach and Saks (1930) showed that if a sequence $g_1, g_2, \ldots$ is weakly convergent to f in $L_p(0,1)$, $1 < p < \infty$, then there is a subsequence $g_{k_1}, g_{k_2}, \ldots$ that is Cesàro summable in the norm topology to f, i.e., $\|f - \sum_{i=1}^n g_{k_i}/n\| \to 0$. This result was extended to uniformly convex spaces by Kakutani (1938). It is interesting to compare the proof in (Kakutani 1938) to the proof of Theorem 3.4, or of the corresponding Hilbert space result in (Jones 1992). The latter proofs use $f \in \overline{\mathrm{co}S}$ to insure the existence of elements g_k to which the modulus of smoothness inequality can be gainfully employed. The Kakutani result is proved by using the weak convergence of the g_k's to find elements g_{k_i} to which the modulus of convexity inequality can be usefully applied. We now give a variant of the Banach-Saks property that holds in Banach spaces of (Rademacher) type $t > 1$.

Definition 3.8 Generalized Banach-Saks Property (GBS).
A Banach space X has the GBS property if for each bounded set S and each $f \in \overline{\mathrm{co}S}$, there exists a sequence $g_1, g_2, \ldots$ in S such that $\|f - \sum_{k=1}^{n} g_k/n\| \to 0$. If ϕ is a given function on $\mathbb{N}$, we say that X has the GBS(ϕ) property if for each f and set S as above one can always find some sequence satisfying the convergence rate $\|f - \sum_{k=1}^{n} g_k/n\| = O(\phi(n))$.

A probabilistic proof of the GBS(ϕ) property for arbitrary Banach spaces of type $t > 1$ is given below. We will make use of the following basic property of type t spaces: If a Banach space X is of type t, then for any independent mean zero random variables $\xi_i \in X$ taking finitely many values, $E(\|\xi_1 + \ldots + \xi_n\|^t) \leq C\sum_{i=1}^{n} E\|\xi_i\|^t$ (Ledoux and Talagrand 1991, p. 248). We also need the following result from Ledoux and Talagrand (1991, Theorem 6.20, p. 171).

Theorem 3.9. *Suppose that ξ_i are independent mean zero random variables in X, $E\|\xi_i\|^N \leq \infty$, $1 \leq i \leq n$, $N > 1$, $N \in \mathbb{N}$. Then there is a universal constant K such that*

$$(3.20)\quad E\left\|\sum_{i=1}^{n}\xi_i\right\|^N \leq \left(K\frac{N}{\log N}\left[E\left\|\sum_{i=1}^{n}\xi_i\right\| + \left(E\max_{1\leq i\leq n}\|\xi_i\|^N\right)^{\frac{1}{N}}\right]\right)^N .$$

The following corollary plays a crucial role in our argument.

Corollary 3.10. *Suppose that $\|\xi_i\| \leq M$, and Banach space X is of type $t > 1$. Then*

$$(3.21)\qquad E\left\|\sum_{i=1}^{n}\xi_i\right\|^N \leq A_N n^{\frac{N}{t}},$$

where A_n is a constant depending on N, M, and X.

Proof.

$$\begin{aligned} E\left\|\sum_{i=1}^{n}\xi_i\right\|^N &\leq \left(K\frac{N}{\log N}\left[E\left(\left\|\sum\xi_i\right\|\right) + M\right]\right)^N \\ &\leq \left(K\frac{N}{\log N}\left[\left(E\left\|\sum\xi_i\right\|^t\right)^{\frac{1}{t}} + M\right]\right)^N \\ &\leq \left(K\frac{N}{\log N}\left[C^{\frac{1}{t}}\left(\sum_{i=1}^{n}E\|\xi_i\|^t\right)^{\frac{1}{t}} + M\right]\right)^N \end{aligned}$$

$$\leq \left(K \frac{N}{\log N}\right)^N \left(C^{\frac{1}{t}} M n^{\frac{1}{t}} + M\right)^N$$
$$\leq A_N n^{\frac{N}{t}}.$$

□

Below we follow a standard construction using the Borel-Cantelli Lemma.

Theorem 3.11. *Let us consider a sequence ξ_i of independent, bounded, zero mean random variables in a Banach space X of type $t > 1$. Then with probability one,*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\| = o(n^{\frac{1}{t}-1+\epsilon}) \tag{3.22}$$

for any $\epsilon > 0$.

Proof. By the Chebyshev inequality, $\forall N \in \mathbb{N}$,

$$P\left\{ \left\| \sum_{i=1}^{n} \xi_i \right\| \geq \delta \right\} \leq E\left(\left\| \sum_{i=1}^{n} \xi_i \right\|^N \right) \delta^{-N}$$
$$\leq A_N n^{\frac{N}{t}} \delta^{-N}.$$

The second inequality follows from Corollary 3.10. Set $\delta = a n^{\epsilon + 1/t}$, where $a > 0$ and $\epsilon > 0$. Then the right hand side of the last inequality becomes $A_N a^{-N} n^{-\epsilon N}$. For sufficiently large N, $N\epsilon > 1$ and

$$\sum_{n \geq 1} \frac{A_N}{a^N n^{N\epsilon}} < \infty.$$

Thus by Borel-Cantelli,

$$\forall a > 0, \quad P\left\{ \exists (n_k), n_k \stackrel{k\to\infty}{\longrightarrow} \infty \text{ s.t. } \left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \xi_i \right\| \geq a(n_k)^{\frac{1}{t}-1+\epsilon} \right\} = 0.$$

Since the union of countably-many zero-measure sets also has zero measure, it follows that for any (a_l) converging to 0,

$$P\left\{ \exists l,\ \exists (n_k),\ n_k \stackrel{k\to\infty}{\longrightarrow} \infty \text{ s.t. } \left\| \frac{1}{n_k} \sum_{i=1}^{n_k} \xi_i \right\| \geq a_l(n_k)^{\frac{1}{t}-1+\epsilon} \right\} = 0,$$

which implies that

$$P\left\{ \lim_{n\to\infty} \frac{\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\|}{n^{\frac{1}{t}-1+\epsilon}} = 0 \right\} = 1.$$

□

We are now ready to investigate the GBS(ϕ) property. Suppose that $f \in \overline{\mathrm{co}S} \subset X$. Then for any $\epsilon > 0$ there exists a $k(\epsilon) < \infty$ and $g_i^\epsilon \in S$, $\alpha_i^\epsilon > 0$ for $1 \le i \le k(\epsilon)$ with $\sum_{i=1}^{k(\epsilon)} \alpha_i^\epsilon = 1$ such that

$$\left\| f - \sum_{i=1}^{k(\epsilon)} \alpha_i^\epsilon g_i^\epsilon \right\| < \epsilon.$$

Let $\tilde{f}_\epsilon := \sum_{i=1}^{k(\epsilon)} \alpha_i^\epsilon g_i^\epsilon$, and consider a positive sequence (ϵ_n) such that

$$\frac{1}{n} \sum_{j=1}^{n} \epsilon_j \le n^{\frac{1}{t}-1}.$$

Also select a sequence of independent random variables ξ_j such that $P\{\xi_j = g_i^{\epsilon_j}\} = \alpha_i^{\epsilon_j}$ for each i, $1 \le i \le k(\epsilon_j)$. Then

$$\begin{aligned} \left\| \frac{1}{n} \sum_{j=1}^{n} \xi_j - f \right\| &= \left\| \frac{1}{n} \sum_{j=1}^{n} (\xi_j - \tilde{f}_{\epsilon_j}) - \frac{1}{n} \sum_{j=1}^{n} (f - \tilde{f}_{\epsilon_j}) \right\| \\ &\le \left\| \frac{1}{n} \sum_{j=1}^{n} \eta_j \right\| + n^{\frac{1}{t}-1}. \end{aligned}$$

Here $\eta_j = \xi_j - \tilde{f}_{\epsilon_j}$, so $E\eta_j = 0$. Applying Theorem 3.11 immediately yields:

Theorem 3.12. *Any Banach space of type t, $1 < t \le 2$, has the GBS($n^{\frac{1}{t}-1+\epsilon}$) property for all $\epsilon > 0$.*

This can be restated as:

Corollary 3.13. *Let X be a Banach space of type t, $1 < t \le 2$, and let S be a bounded subset of X with $f \in \overline{\mathrm{co}S}$. Then for all $\epsilon > 0$, there exists a sequence $(g_i) \subset S$ such that the incremental sequence*

$$f_n = \frac{1}{n} \sum_{i=1}^{n} g_i = \frac{n-1}{n} f_{n-1} + \frac{1}{n} g_n \tag{3.23}$$

satisfies

$$\|f - f_n\| = o(n^{\frac{1}{t}-1+\epsilon}). \tag{3.24}$$

Remark. The GBS(ϕ) property guarantees for $f \in \overline{\mathrm{co}S}$ and S bounded the existence of $(g_n) \subset S$ such that $\|f - \sum_{i=1}^{n} g_i/n\| \to 0$. It is not true in general that one may pick all the g_n distinct. Consider for example the Hilbert space ℓ_2 (which is of type 2, i.e., GBS($1/\sqrt{n}$)) and $S = (e_n)$, where (e_n) is an orthonormal basis. Then $\overline{\mathrm{co} S} = \sum_{i \geq 0} \alpha_i e_i$ where $\alpha_i \geq 0$ and $\sum \alpha_i = 1$. But if $e_{n_k} \neq e_{n_l}$, $(n_k \neq n_l)$ then

$$\frac{1}{N} \sum_{i=1}^{N} e_{n_i} \xrightarrow{\|\cdot\|} 0.$$

However, it is possible to give necessary and sufficient conditions for the existence of a sequence of distinct elements $(g_i) \subset S$ as above:

Theorem 3.14. *Let X be a Banach space of type t, $1 < t \leq 2$, $S \subset X$, $f \in \overline{\mathrm{co} S}$. There exists a sequence $(g_i) \subset S$ where $\forall i \neq j$, $g_i \neq g_j$ such that $\| \sum g_i / n - f \| \to 0$ if and only if for all finite $K \subset S$, $f \in \overline{\mathrm{co}(S \setminus K)}$.*

Proof. That $f \in \overline{\mathrm{co}(S \setminus K)}$ for all finite K is sufficient follows from the discussion preceding Theorem 3.12. With this condition holding, we are free to choose the $g_i^{\epsilon_j}$ to be distinct for all i and j.

Necessity follows from considering a single finite-cardinality set $K \subset S$ such that $f \notin \overline{\mathrm{co}(S \setminus K)}$. Assume there exists a sequence of distinct elements $(g_i) \subset S$ such that $\|f - \sum_{i=1}^{n} g_i / n\| \to 0$. We will construct a sequence in $\mathrm{co}(S \setminus K)$ converging to f, which is a contradiction.

Since $|K| < \infty$ there must be an $r > 0$ such that $\forall n > r$, $g_n \in S \setminus K$. Let $s > r$. Then

$$f_s := \frac{1}{s} \sum_{i=1}^{s} g_i = \frac{1}{s} \sum_{i=1}^{r} g_i + \frac{1}{s} \sum_{i=r+1}^{s} g_i.$$

As $s \to \infty$, the first sum tends to zero, so the second must tend to f. Therefore the sequence

$$\frac{1}{s-r} \sum_{i=r+1}^{s} g_i = \frac{s}{s-r} \frac{1}{s} \sum_{i=r+1}^{s} g_i$$

must also tend to f. Each element of this sequence is in $\mathrm{co}(S \setminus K)$. □

4. Additional Assumptions on S

In the previous sections we have studied the convergence of convex approximates from arbitrary bounded sets S as a function of the space X. However, it is possible to improve the convergence behavior for a given space by imposing additional assumptions on S. For example, suppose $X = L_p(\Omega)$ with $1 \leq p < 2$ and $m(\Omega) < \infty$. If there exists $h \in L_2(\Omega)$ such that $|g(x)| < h(x)$ for all $x \in \Omega$ and $g \in S$, then $S \subset L_2(\Omega)$ as well, and the convex closure of S is the same in both $L_p(\Omega)$ and $L_2(\Omega)$. Working in $L_2(\Omega)$, we can find for each $f \in \overline{\mathrm{co} S}$ an incremental sequence f_n converging to f with rate $O(1/\sqrt{n})$. This convergence rate (with a different constant) then holds in $L_p(\Omega)$ as well, though in the latter space the sequence might not be ϵ-greedy, especially in view of Theorem 3.1.

At the other end of the L_p spectrum, recall from Theorem 2.3 that there can be no general rate bound in L_∞. However, by restricting S to classes of indicator functions, we can recover L_2-like convergence in L_∞. This approach, common in the artificial neural network community, is especially interesting in pattern classification applications, and the connections with *Vapnik-Chervonenkis (VC) dimension*—first discovered by Barron in (Barron 1992)—are especially intriguing.

In the following we adopt the convention of identifying an index (indicator) function with the set it indexes (i.e., its support), where the distinction is determined by context.

Definition 4.1. *Let F be a class of indicator functions on a set X. Let W be the set of all $Y \subseteq X$ such that for all $Z \subseteq Y$ there exists an $f \in F$ such that $f \cap Y = Z$. The "VC (Vapnik-Chervonenkis) dimension of F" is $\sup_{Y \in W} |Y|$.*

Our bounds (Theorem 4.5) are slightly weaker than those given by (Barron 1992) (ours are "big O" as compared to "little O"). However, the proof method here seems worthy of note, especially as it makes use of more basic results than the central limit theorem relied upon in (Barron 1992).

We also explore the implications of good convergence rate bounds on the VC dimension of the corresponding set S. Good convergence rate bounds for S do *not* imply that S has a finite VC dimension (Theorem 4.6), i.e., the converse of Theorem 4.5 is false. However, if any nontrivial rate bound holds *uniformly* for all subsets of S, its VC dimension must be finite (Theorem 4.7).

Definition 4.2. *Let F be a class of indicator functions on a set X. Its dual F' is defined as the class $\{\mathrm{ev}_x, x \in X\}$ of indicator functions on F, where*

$$\mathrm{ev}_x(f) = f(x) \ \ \forall f \in F .$$

Thus, for each element of X there is a unique element of F', named ev_x. Note that it may be the case that $\mathrm{ev}_x = \mathrm{ev}_y$ for $x \neq y$. One thinks of ev_x as the "evaluation at x operator" for elements of F. Note that ev_x contains the members of F which contain x.

Definition 4.3. *Let VC be the operator on classes of indicator functions that measures the Vapnik-Chervonenkis (VC) dimension. If F is a class of indicator functions, we define the co-VC dimension by*

$$coVC(F) := VC(F'). \tag{4.25}$$

The next lemma is a consequence of (Ledoux and Talagrand 1991, Theorem 14.15, p. 418).

Lemma 4.4. *Denote by $M(\Omega, \Sigma)$ the Banach space of all bounded signed measures μ on (Ω, Σ) equipped with the norm $\|\mu\|_M := |\mu|(\Omega) \equiv \mu^+(\Omega) + \mu^-(\Omega)$. Let $C \subset \Sigma$; consider the operator $j : M(\Omega, \Sigma) \to l_\infty(C)$ defined by $j(\mu) = (\mu(c))_{c \in C}$. Then $VC(C) < \infty$ if and only if there exists a constant K such that for any Rademacher sequence γ_i and all finite sequences $(\mu_i) \subset M(\Omega, \Sigma)$,*

$$E \left\| \sum_i \gamma_i j(\mu_i) \right\|_{\sup} \leq K \left(\sum_i \|\mu_i\|_M^2 \right)^{1/2} . \tag{4.26}$$

If such a K exists, $K = K'\sqrt{VC(C)}$, where K' is a universal constant.

Theorem 4.5. *Let F be a class of indicator functions on a set X. Then $F \subset l_\infty(X)$. Let $coVC(F) = d < \infty$. Then for any $h \in \overline{\mathrm{co}F}$, $\|\mathrm{co}_n F - h\|_{\sup} \leq K(d/n)^{1/2}$, where K is a universal constant.*

Proof. Since $h \in \overline{\mathrm{co}F}$, $\forall\epsilon > 0$, $\exists k_\epsilon \in \mathrm{co}F$ s.t. $\|h - k_\epsilon\|_{\sup} < \epsilon$. We will neglect the ϵ and write merely k where convenient. For some n_ϵ, $k_\epsilon = \sum_{i=1}^{n_\epsilon} \alpha_i f_i$, where $\sum_{i=1}^{n_\epsilon} \alpha_i = 1$, $a_i > 0$, and $f_i \in F$. Let $\Sigma_k := \sigma(F' \cup \cup_{i=1}^{n_\epsilon}\{\{f_i\}\})$. $M(F, \Sigma_k)$ is a Banach space of bounded measures on F. The norm of $\mu \in M(F, \Sigma_k)$ is $\|\mu\|_M = |\mu|(F) = \mu^+(F) + \mu^-(F)$. Define $m_i \in M(F, \Sigma_k)$ to be the probabilistic point-mass measure with support on f_i, i.e., m_i assigns measure 1 to sets containing f_i and 0 otherwise. Define $\mu_k := \sum_{i=1}^{n} \alpha_i m_i$. Let ξ_l be finite-valued i.i.d. random variables taking value m_i with probability α_i. Define $j : M(F, \Sigma_k) \to l_\infty(X)$ by $j(\mu) := (\mu(\mathrm{ev}_x))_{x\in X}$. Note that j is a linear operator. By the triangle inequality,

$$E_\xi \left\| j\left(\sum_{i=1}^{n} \xi_l/n\right) - h\right\|_{\sup} \leq E_\xi \left\| j\left(\sum_l (\xi_l - \mu_k)\right)\right\|_{\sup} /n + \|h-k\|_{\sup}.$$

By a simple inequality (Ledoux and Talagrand, 1991, Lemma 6.3, p. 152),

$$E_\xi \left\|\sum_l j\left(\xi_l - \mu_k\right)\right\|_{\sup} \leq 2E_\xi E_\gamma \left\|\sum_l \gamma_l j(\xi_l - \mu_k)\right\|_{\sup},$$

where γ_l are i.i.d. random variables taking values +1 and -1 with equal probability. Then by Lemma 4.4,

$$E_\gamma \left\|\sum_l \gamma_l j(\xi_l - \mu_k)\right\|_{\sup} \leq K\sqrt{d}\sqrt{\sum_l \|\xi_l - \mu_k\|_M^2}.$$

Since ξ_l and μ_k are both probability measures, $\|\xi_l - \mu_k\|_M \leq \|\xi_l\|_M + \|\mu_k\|_M = 2$. Combining the above, we have

$$E_\xi \left\| j\left(\sum_{i=1}^{n} \xi_l/n\right) - h\right\|_{\sup} \leq \frac{4K\sqrt{d}}{\sqrt{n}} + \epsilon.$$

Since the inequality is true for all $\epsilon > 0$ it remains true for $\epsilon = 0$. Since for some realization of the ξ_l the inequality must still hold, the theorem is proven. □

The mere fact of the existence of a convergence rate bound for a set of indicator functions does not imply that the set has finite VC dimension. We give an example to make this point.

Theorem 4.6. *Let X be a measure space with an infinite number of measurable sets. Let S denote the set of indicator functions of all measurable sets. Clearly the VC dimension of S is infinite. However, if $f \in \overline{\mathrm{co}S}$, then f can be approximated by an element of $co_n S$ with error less than $1/n$ (in the uniform metric).*

Proof. If $f \in \overline{\mathrm{co}S}$, then clearly $0 \leq f(x) \leq 1$ for all $x \in X$. Fix n and define

$$A_k = f^{-1}\left([(k-1)/n, 1]\right)$$

for $k = 1, 2, \ldots, n$. (Note that some A_k may be empty.) Let g_k be the indicator function for A_k. Each g_k is in S since f must be a measurable function. The function

$$f_n = \sum_{k=1}^{n} (1/n) g_k$$

is in $\mathrm{co}_n S$ and satisfies

$$0 \leq f_n(x) - f(x) < 1/n$$

for all $x \in X$. (Moreover, this shows that $\overline{\mathrm{co}S}$ equals the set of measurable functions with range in $[0, 1]$.) □

However, if it is the case that one given convergence rate bound holds *for all finite subsets* of a set of indicator functions, then the VC dimension of this set is finite.

Theorem 4.7. *Let F be a set of indicator functions on a set X and $C \subseteq F$. Let $\alpha(C, r)$ be the worst-case rate of approximation of a member of the closed convex hull of C by r elements of C. Assume that there is some function h so that $h(r) \to 0$ as $r \to \infty$ such that $\alpha(C, r) \leq h(r)$ for all finite $C \subset F$. Then $VC(F) < \infty$.*

Proof. We argue by contradiction. Assume that $VC(F) = \infty$. Then also $coVC(F) = \infty$. Thus, for each integer n there are elements $x_1, x_2, \ldots, x_{2^n} \in X$ and functions $f_1, \ldots, f_n \in F$ such that $(f_1, \ldots, f_n)$ takes all 2^n possible values on these points. Define $C_n := \{f_1, \ldots, f_n\}$. Consider approximating $f := \sum_{i=1}^{n} f_i/n \in \mathrm{co}C_n$ by r elements of C_n. By the symmetry of f, we can without loss of generality write the approximant as $g = \sum_{i=1}^{r} \alpha_i f_i$. Then

$$\begin{aligned}
\|g - f\|_{\sup} &= \sup_X |g(x) - f(x)| \\
&= \sup_X \left| \sum_{i=1}^{r} \left(\alpha_i - \frac{1}{n}\right) f_i(x) - \sum_{i=r+1}^{n} \frac{1}{n} f_i(x) \right| \\
&= \frac{1}{2} \sup_X \left| \sum_{i=1}^{r} \left(\alpha_i - \frac{1}{n}\right) [2f_i(x) - 1] - \sum_{i=r+1}^{n} \frac{1}{n} [2f_i(x) - 1] \right| \\
&= \frac{1}{2} \left(\sum_{i=1}^{r} \left|\alpha_i - \frac{1}{n}\right| + \sum_{i=r+1}^{n} \frac{1}{n} \right)
\end{aligned}$$

since there is an element $x \in X$ for which $2f_i(x) - 1$ is of the same sign as $\alpha_i - 1/n$ for $1 \leq i \leq r$ and is negative one for $i > r$. Thus

$$\|g - f\|_{\sup} \geq \frac{1}{2} \frac{n-r}{n} = \frac{1}{2}\left(1 - \frac{r}{n}\right). \tag{4.27}$$

Since the bounding function for the error, h, converges to zero, there exists $q > 0$ such that $h(q) < 1/4$. But the worst-case error approximating elements of C_{2q} with q functions from this set is greater than $[1-q/(2q)]/2 = 1/4$, a contradiction. This establishes the claim. $\square$

Acknowledgments

This work was partially completed while Michael Donahue and Eduardo Sontag were visiting Siemens Corporate Research. Michael Donahue is currently with the National Institute of Standards and Technology, Gaithersburg, MD 20899.

A. Implications of the Convexity Assumption

Constraining f to lie within the convex closure of S (instead of the linear span) effectively reduces the size of the set of approximable functions. Which functions are being left out? For the case of functions on a real interval under the sup norm where S is the set of Heavisides, there is a tidy answer.

Define $F_v :=$

$$\{f : [a,b] \to \mathbb{R} \mid \mathrm{Var}(f) \le v,\ \forall t \in [a,b],\ |f(t)| \le v\}. \tag{A.28}$$

Let H be the set of Heavisides on $\mathbb{R}$, i.e., $H = \{h : \mathbb{R} \to \mathbb{R} \mid \exists\theta, h = I[\theta,\infty)\}$. Define

$$\begin{aligned} H_v := \{f : \mathbb{R} \to \mathbb{R} \mid \exists h \in H \text{ s.t. } f = vh \text{ or } f = -vh \\ \text{or } f = vh' \text{ or } f = -vh' \\ \text{where } \forall t \in [a,b],\ h'(t) = h(-t)\}. \end{aligned} \tag{A.29}$$

Clearly $H_v \subset F_v$, and a straightforward argument shows that in fact $F_v = \overline{\mathrm{co}(H_v)}_{\sup}$. By an elementary argument (analogous to the proof of Theorem 4.6), one can also show that if $f \in F_v$, then $\|\mathrm{co}_n H_v - f\|_{\sup} = O(1/n)$.

At this point, it is natural to ask what is the class of functions that can be uniformly approximated by neural nets with Heaviside activations, that is, what is the closure of the *linear span* (not the convex hull) of the maps from H_v (which is the same for all v of course). This is a classical question; see for instance (Dieudonné 1960, VII.6): the closure is the set of all regulated functions, that is, the set of functions $f : [a,b] \to R$ for which $\lim_{x \to x_0^-} f(x)$ and $\lim_{x \to x_0^+} f(x)$ exist for all $x_0 \in [a,b)$ and $x_0 \in (a,b]$, respectively. Thus, by constraining target functions to the convex closure of H_v instead of the span, we are losing the ability to approximate those regulated functions that are not of bounded variation.

In the multivariable case, that is, $f : K \to \mathbb{R}$ with K a compact subset of $\mathbb{R}^n$, the situation is far less clear. If f has "bounded variation with respect to half-spaces" (i.e., is in the convex hull of the set of all half spaces (Barron 1992)), and in particular if f admits a Fourier representation

$$f(x) = \int_{\mathbb{R}^n} e^{i\langle\omega,x\rangle} \tilde{f}(\omega)\, d\omega$$

with

$$\int_{\mathbb{R}^n} \|\omega\| |\tilde{f}(\omega)| d\omega < \infty,$$

then by (Barron 1992) there are approximations with rate $O(1/\sqrt{n})$ (since f is in the convex hull of the Heavisides). But the precise analog of regulated functions is less obvious. One might expect that piecewise constant functions can be uniformly approximated, for instance, at least if defined on polyhedral partitions, but this is false.

For a counterexample, let f be the characteristic function of the square $[-1,1]^2$ in $\mathbb{R}^2$, and let K be, for instance, a disc of radius 2 centered on $(0,0)$. Then it is impossible to approximate f to within error $1/8$ by a one hidden-layer neural net, that is, a linear combination of terms of the form $H(\langle w, x\rangle + b)$, with $w \in \mathbb{R}^2$ and $b \in \mathbb{R}$. (Constant terms on K can be included, without loss of generality, by choosing b appropriately.) This is proved as follows. If there would exist a function g of this form, that approximates to within $1/8$, then close to the boundary of the disc its values are in the range $(-1/8, 1/8)$, and near the center of the square it has values $> 7/8$. Moreover, everywhere the values of g are in $(-1/8, 1/8) \bigcup (7/8, +\infty)$. Now, the function $4g - (1/2)$ is again in the same span, and it now has values in $(-1, 0)$ and $(3, +\infty)$ in the same regions. This contradicts (Sontag 1992, Prop. 3.8). (Of course, one can also prove this directly.)

B. Properties of the Modulus of Smoothness

In this section we collect some inequalities related to power type estimates of the modulus of smoothness. In particular, the first lemma shows that L_p spaces are of power type $t = \min(p, 2)$. (See Corollary 3.6 and Theorem 3.7.)

Lemma B.1. *If $X = L_p$, $1 \le p < \infty$, then the modulus of smoothness $\rho(u)$ satisfies*

$$\rho(u) \le \begin{cases} u^p/p & \text{if } 1 \le p \le 2 \\ (p-1)u^2/2 & \text{if } 2 \le p < \infty \end{cases} \tag{B.30}$$

for all $u \ge 0$.

Proof. From the definition of the modulus of smoothness we have

$$\begin{aligned} 2(\rho(u)+1) &= \sup\{\|f+g\| + \|f-g\| \mid \|f\| = 1, \|g\| = u\} \\ &\le 2^{1/q} \sup\{(\|f+g\|^p + \|f-g\|^p)^{1/p} \mid \|f\| = 1, \|g\| = u\}, \end{aligned} \tag{B.31}$$

where $q = p/(p-1)$. The inequality follows from the concavity of the function $t \to t^{1/p}$. Next we make use of some inequalities given by Hanner (1956):

$$\begin{aligned} (\|f\| + \|g\|)^p + |\, \|f\| - \|g\| \,|^p &\le \|f+g\|^p + \|f-g\|^p \\ &\le 2(\|f\|^p + \|g\|^p), \end{aligned} \tag{B.32}$$

for $1 < p \leq 2$. The inequalities hold in the reverse sense if $2 \leq p < \infty$. (The second inequality above is actually due to Clarkson (1936).) Combining this with (B.31) yields

$$\rho(u) \leq \begin{cases} (1+u^p)^{1/p} - 1 & \text{if } 1 \leq p \leq 2 \\ \left[\dfrac{(1+u)^p + |1-u|^p}{2}\right]^{1/p} - 1 & \text{if } 2 \leq p < \infty \end{cases} \tag{B.33}$$

This result is cited in (Lindenstrauss 1963). It is possible to use the methods in (Hanner 1956) to show that the above bounds on $\rho(u)$ are tight.

For $1 < p \leq 2$, (B.30) now follows immediately. For $p \geq 2$ and $0 \leq u \leq 1$ a Taylor series expansion shows

$$\left(\frac{(1+u)^p + (1-u)^p}{2}\right)^{1/p} \leq \frac{p-1}{2}u^2 + 1, \tag{B.34}$$

which proves (B.30) for $2 \leq p < \infty$ and $0 \leq u \leq 1$. For $u > 1$, divide by u and use (B.34) again:

$$\begin{aligned} \left(\frac{(u+1)^p + (u-1)^p}{2}\right)^{1/p} &= u\left(\frac{(1+1/u)^p + (1-1/u)^p}{2}\right)^{1/p} \\ &\leq u + \frac{p-1}{2u} \\ &\leq \frac{p-1}{2}u^2 + 1, \end{aligned}$$

for $2 \leq p < \infty$ and $u > 1$. □

For completeness we note the following result, the proof of which is left to the reader.

Theorem B.2. *Let X be a Banach space. The modulus of smoothness for X satisfies*

$$\rho(u) \leq u \qquad \textit{for all } u \geq 0. \tag{B.35}$$

Furthermore, if X is L_1 or L_∞ with dimension at least 2, then

$$\rho(u) = u \qquad \textit{for all } u \geq 0. \tag{B.36}$$

The following technical lemma uses an inequality of Lindenstrauss to provide several lower bounds on γ as a function of t for spaces with modulus of smoothness $\rho(u) \leq \gamma u^t$. These are needed in Theorem 3.5.

Lemma B.3. *Let X be a Banach space with modulus of smoothness $\rho(u)$ satisfying $\rho(u) \leq \gamma u^t$ for all $u \geq 0$. Then $1 \leq t \leq 2$ and*

$$\gamma t \geq \begin{cases} [(2-t)t]^{1-t/2}(t-1)^{t-1} & \textit{if } 1 < t < 2 \\ 1 & \textit{if } t = 1 \textit{ or } t = 2. \end{cases} \tag{B.37}$$

Moreover, for all $t \in [1,2]$,

$$\gamma \geq \sqrt{2} - 1, \tag{B.38}$$

$$\gamma \geq 3^{-t/2}, \tag{B.39}$$

$$\gamma \geq 2^{t-1}/5^{t/2}, \tag{B.40}$$

$$\gamma > \frac{1}{t} e^{-3/2e}. \tag{B.41}$$

Proof. Lindenstrauss (1963) gives

$$\sqrt{1+u^2} - 1 \leq \rho(u) \leq \gamma u^t \qquad \text{for all } u \geq 0. \tag{B.42}$$

Letting $u \to \infty$ shows that $t \geq 1$, and if $t = 1$ then $\gamma \geq 1$. Furthermore,

$$\gamma u^t \geq \sqrt{1+u^2} - 1 \geq \frac{u^2}{2+u^2},$$

so letting $u \downarrow 0$ proves that $t \leq 2$ and if $t = 2$ then $\gamma \geq 1/2$.

Therefore t satisfies $1 \leq t \leq 2$, and inequalities (B.37) through (B.41) hold if $t = 1$ or $t = 2$. Next assume $1 < t < 2$, and rewrite (B.42) as

$$\gamma \geq \frac{\sqrt{1+u^2} - 1}{u^t} \qquad \text{for all } u \geq 0.$$

In particular, this inequality holds if we replace u with $\sqrt{(2-t)t}/(t-1)$, which gives

$$\gamma \geq \frac{(2-t)^{1-t/2}(t-1)^{t-1}}{t^{t/2}}. \tag{B.43}$$

This completes the proof of (B.37).

Inequality (B.38) follows from (B.42) by setting $u = 1$. Using $u = \sqrt{3}$ provides (B.39), and (B.40) is obtained with $u = \sqrt{5}/2$.

To prove (B.41), place the inequality $x^x \geq e^{-1/e}$ into (B.43) to obtain

$$\gamma \geq t^{-t/2} e^{-1/2e} e^{-1/e}.$$

Then use $1 < t < 2$ to complete the proof. $\square$

These estimates are compared in Figure 2.

Lemma B.4. *Let* $\rho(u)$ *be the modulus of smoothness of a Banach space, and assume that* $\rho(u) \leq \gamma u^t$ *for all* $u \geq 0$ *(*$t \in [1,2]$ *is fixed). Then*

$$1 + \frac{1}{(2\gamma)^{1/t}} < \left(\frac{5t-1}{2}\right)^{1/t}. \tag{B.44}$$

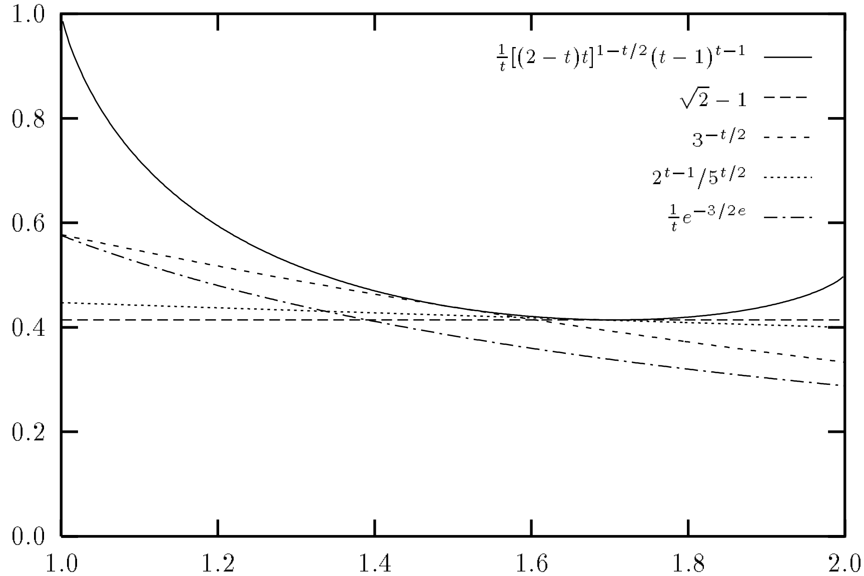

Fig. 2. Comparison of lower bound estimates on γ (from Lemma B.3) as a function of t.

Proof. Define $h(t)$ to be the right-hand side of (B.44). Then the derivative of $h(t)$ has the same sign as

$$\frac{5t-1}{2}\left(1-\log\left(\frac{5t-1}{2}\right)\right)+\frac{1}{2}. \tag{B.45}$$

But (B.45) is decreasing in t for $t > 3/5$, so $h'(t) = 0$ for at most one point $t_0 \in [1, 2]$ (in fact $t_0 \approx 1.4724$), which would be a local maximum for h. In particular, for any subinterval $[t_1, t_2] \subset [1, 2]$, it must be that

$$h(t) \geq \min\left(h(t_1), h(t_2)\right) \qquad \text{for all } t \in [t_1, t_2].$$

For our purposes we divide the interval $[1, 2]$ into two subintervals: $[1, 5/4]$ and $[5/4, 2]$. Evaluating h at the endpoints yields

$$\begin{aligned} h(1) &= 2 \\ h(5/4) &= (21/8)^{4/5} > 2.16 \\ h(2) &= 3/\sqrt{2} > 2.12. \end{aligned}$$

Therefore $h(t) \geq 2$ for $t \in [1, 5/4]$, and $h(t) \geq 2.12$ for $t \in [5/4, 2]$.

To obtain the corresponding upper bounds on the left-hand side of (B.44), apply (B.39) for $t \in [1, 5/4]$, and (B.40) for $t \in [5/4, 2]$. $\square$

C. A Sequence Inequality

The following lemma may be viewed as pertaining to a discrete analogue of the differential equation $y' = y^{1-t}$, the general solution of which is $y(x) = (tx+C)^{1/t}$. It can be used to provide bounds on the convergence rate of sequences having incremental changes compatible with the estimates from Lemma 3.3, with which it is combined to produce Theorems 3.5 and 3.7.

Lemma C.1. *Let (a_n) be a nonnegative sequence satisfying*

$$a_{n+1} \leq a_n + \min\left(\frac{3}{2}, \frac{1}{a_n^{t-1}}\right) \tag{C.46}$$

for all n, where $1 \leq t \leq 2$. If

$$a_n \leq \left(tn + \frac{t-1}{2}\log_2 n\right)^{1/t} \tag{C.47}$$

is satisfied for some $n \geq 2$, or for $n = 1$ with $a_1 \geq 1$, then it is satisfied for all $n' > n$ as well.

Proof. Let us take $b_n := tn + (1/2)(t-1)\log_2 n$, assume that $a_n \leq b_n^{1/t}$, and proceed by induction, i.e., show that $a_{n+1} \leq b_{n+1}^{1/t}$.

Suppose first that $a_n < 1$ and $n \geq 2$. Then

$$b_{n+1}^{1/t} \geq b_3^{1/t} \geq (6 + (1/2)\log_2 3)^{1/2} > 5/2.$$

The second inequality holds because the function $t \mapsto (3t + (1/2)(t-1)\log_2 3)^{1/t}$ is decreasing in t for $t \in [1,2]$, and so attains its minimum at $t = 2$. But then

$$a_{n+1} \leq a_n + 3/2 < 5/2 < b_{n+1}^{1/t},$$

as desired.

Alternatively, suppose that $a_n \geq 1$. Then since the function $x \mapsto x(1 + 1/x^t)$ is nondecreasing for $x \geq 1$, we have

$$\begin{aligned} a_{n+1} &\leq a_n(1 + 1/a_n^t) \ \leq\ b_n^{1/t}(1 + 1/b_n) \\ &\leq b_n^{1/t}\left[1 + \frac{t}{b_n} + \frac{t(t-1)}{2b_n^2}\right]^{1/t} \\ &\leq \left[t(n+1) + \frac{t-1}{2}\log_2(n+1) + \frac{t-1}{2n} - \frac{t-1}{2}\log_2(1+1/n)\right]^{1/t}. \end{aligned}$$

The third inequality above follows from a Taylor series expansion of $(1+1/b_n)^t$. Combining this with the definition of b_{n+1} and the fact that $\log_2(1+1/n) \geq 1/n$ for $n \geq 1$ yields

$$a_{n+1} \leq b_{n+1}^{1/t},$$

concluding the proof. □

D. An Example of a Smooth Banach Space with Non-Converging ϵ-Greedy Sequences

In Sect. 3.1 it was shown that ϵ-greedy sequences in uniformly smooth spaces always converge (provided $\sum \epsilon_k < \infty$), and that smoothness is necessary and sufficient for monotonically decreasing error in incremental approximants. We now construct an example showing that simple smoothness is insufficient to guarantee convergence of ϵ-greedy sequences.

Let $a = (a(n))$ be a sequence of real numbers. Define the sequence of functions (F_n) (the *norm sequence*) from $\mathbb{R}^{\mathbb{N}}$ to $\mathbb{R}^+ \cup \{0\}$ recursively by

$$F_1(a) = |a(1)|$$
$$F_n(a) = [(F_{n-1}(a))^{p_n} + |a(n)|^{p_n}]^{1/p_n} ,$$

where (p_n) is a fixed sequence (called the *norm power sequence*) with $1 \le p_n < \infty$ for all n.

Note that for each a, $F_n(a)$ is nondecreasing with n. Define

$$X_{(p_n)} := \left\{ a \in \mathbb{R}^{\mathbb{N}} \mid \sup_n F_n(a) < \infty \right\}, \tag{D.48}$$

and for $a \in X_{(p_n)}$

$$\|a\| := F(a) := \lim_{n\to\infty} F_n(a). \tag{D.49}$$

The reader may verify that $X_{(p_n)}$ equipped with the norm (D.49) is a Banach space. (This space is similar to the modular sequence spaces studied in (Woo 1973).) If (p_n) is bounded and $p_n > 1$ for all n, then it can be shown that $X_{(p_n)}$ is smooth. Let us use the notation $e_n \in X_{(p_n)}$ to denote the canonical basis element $e_n(k) = \delta_n(k)$. (Note that $\|e_n\| = 1$ for all n independent of (p_n).)

Proposition D.1. *Let (γ_n) be a strictly decreasing sequence converging to 1. Then there exists a norm power sequence (p_n) that is non-increasing, converges to 1, and has $p_n > 1$ for all n, such that the bounded set $S = \{-\gamma_1 e_1\} \cup \{\gamma_n e_n\}_{n\in\mathbb{N}}$ in $X_{(p_n)}$ admits for each incremental schedule (ϵ_n) an incremental sequence $(a_n) \subset \mathrm{co}S$ that is ϵ-greedy with respect to 0 but does not converge. (Note that $0 \in \mathrm{co}S$.)*

Proof. $X_{(p_n)}$ is determined by its power sequence (p_n). Choose $p_1 > 1$ arbitrarily, and recursively select $p_n \in (1, p_{n-1}]$ to satisfy

$$\inf_{0\le\lambda\le1} [(1-\lambda)\gamma_{n-1}]^{p_n} + (\lambda\gamma_n)^{p_n} > 1. \tag{D.50}$$

This can always be done because the inequality holds for $p_n = 1$ and so by continuity in p_n holds also for all p_n sufficiently close to 1.

Now let (ϵ_k) be a fixed incremental schedule. We build an ϵ-greedy (with respect to 0) sequence $(a_k) \subset S \subset \mathrm{co}S$ as follows. Let $a_1 = \gamma_{n_1} e_{n_1}$ be any ϵ_1-greedy element of S with $n_1 > 1$ (i.e., $\gamma_{n_1} < \min\{1 + \epsilon_1, \gamma_1\}$). Assuming that

$a_{k-1} = \gamma_{n_{k-1}} e_{n_{k-1}}$ with $n_{k-1} > 1$, we will show that we can pick $a_k = \gamma_{n_k} e_{n_k}$ with $n_k > n_{k-1}$ to be ϵ_k-greedy. Indeed, suppose that

$$b_k = (1-\lambda)a_{k-1} + \lambda g_k$$

is an ϵ_k-greedy step, where $g_k \in S$. It follows from (D.50) and the monotonicity of γ_n that $\|b_k\| > 1$, so we can pick $n_k > n_{k-1}$ such that

$$\|b_k\| > \gamma_{n_k} = \|\gamma_{n_k} e_{n_k}\|.$$

Therefore, taking $a_k = \gamma_{n_k} e_{n_k}$ yields an ϵ_k-greedy increment.

We define the sequence (a_k) recursively in this manner to complete the construction. □

References

Banach, S., and S. Saks (1930) Sur la convergence forte dans les champs L^P, *Studia Math.*, **2**:51–57.

Barron, A. R. (1991) Approximation and estimation bounds for artificial neural networks, in *Proc. Fourth Annual Workshop on Computational Learning Theory*, Morgan Kaufmann, 243–249.

Barron, A. R. (1992) Neural net approximation, in *Proc. of the Seventh Yale Workshop on Adaptive and Learning Systems*, 69–72.

Bessaga, C., and A. Pelczynski (1958) A generalization of results of R. C. James concerning absolute bases in Banach spaces, *Studia Math.*, **17**:151–164.

Clarkson, J. A. (1936) Uniformly convex spaces, *Trans. Amer. Math. Soc.*, **40**:396–414.

Darken, C., M. Donahue, L. Gurvits, and E. Sontag (1993) Rate of approximation results motivated by robust neural network learning, in *Proc. of the Sixth Annual ACM Conference on Computational Learning Theory*, The Association for Computing Machinery, New York, N.Y. 303–309.

Deville, R., G. Godefroy, and V. Zizler (1993) *Smoothness and Renormings in Banach Spaces.* Wiley, N.Y..

Dieudonné, J. (1960) *Foundations of Modern Analysis.* Academic Press, N.Y..

Figiel, T. and G. Pisier (1974) Séries aléatoires dans les espaces uniformément convexes ou uniformément lisses, *C. R. Acad. Sci. Paris*, **279**, Série A:611–614.

Gowers, W. T., A Banach space not containing c_0, l_1, or a reflexive subspace. Preprint.

Haagerup, U. (1982) The best constants in the Khintchine inequality, *Studia Math.*, **70**:231–283.

Hanner, O. (1956) On the uniform convexity of L^p and ℓ^p, *Arkiv för Matematik*, **3**:239–244.

Hanson, S. J. and D. J. Burr (1988) Minkowski-r back-propagation: learning in connectionist models with non-Euclidean error signals, in *Neural Information Processing Systems*, American Institute of Physics, New York, 348.

James, R. C. (1978) Nonreflexive spaces of type 2, *Israel J. Math.*, **30** :1–13.

Jones, L. K. (1992) A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Annals of Statistics*, **20**:608–613.

Kakutani, S. (1938) Weak convergence in uniformly convex spaces, *Tôhoku Math. J.*, **45**:188–193.

Khintchine, J. (1923) Über die diadischen Brüche. *Math. Z.*, **18**:109-116.

Ledoux, M. and M. Talagrand (1991) *Probability in Banach Space.* Springer-Verlag, Berlin.

Leshno, M., V. Lin, A. Pinkus, and S. Schocken (1992) Multilayer feedforward networks with a non-polynomial activation function can approximate any function. Preprint. Hebrew University.

Lindenstrauss, J. (1963) On the modulus of smoothness and divergent series in Banach spaces, *The Michigan Math. J.*, **10**:241–252.

Lindenstrauss, J. and L. Tzafriri (1979) *Classical Banach Spaces II: Function Spaces.* Springer-Verlag, Berlin.

Powell, M. J. D. (1981) *Approximation theory and methods.* Cambridge University Press, Cambridge.

Rey, W. J. (1983) *Introduction to Robust and Quasi-Robust Statistical Methods.* Springer-Verlag, Berlin.

Rosenthal, H. (1974) A characterization of Banach spaces containing l_1, *Proc. Nat. Acad. Sci. (USA)*, **71**:2411–2413.

Rosenthal, H. (1994) A subsequence principle characterizing Banach spaces containing c_0, *Bull. Amer. Math. Soc.*, **30**:227–233.

Sontag, E. D. (1992) Feedback stabilization using two-hidden-layer nets, *IEEE Trans. Neural Networks*, **3**:981–990.

Stromberg, K. R. (1981) *An Introduction to Classical Real Analysis.* Wadsworth, New York.

Woo, J. Y. T. (1973) On modular sequence spaces, *Studia Math.*, **48**:271–289.

Michael J. Donahue
Institute for Mathematics and its Applications
University of Minnesota
Minneapolis, MN 55455

Leonid Gurvits
NEC Research Institute
4 Independence Way
Princeton, NJ 08540

Christian Darken
Learning Systems Department
Siemens Corporate Research, Inc.
755 College Road East
Princeton, NJ 08540

Eduardo Sontag
Department of Mathematics
Rutgers University
New Brunswick, NJ 08903

This article was processed using the LATEX macro package with CAP style