# Entropy Numbers of Linear Function Classes

**Robert C. Williamson**
Department of Engineering
Australian National University
Canberra, ACT 0200, Australia

**Alex J. Smola**
Department of Engineering
Australian National University
Canberra, ACT 0200, Australia

**Bernhard Schölkopf**
Microsoft Research Limited
St. George House
1 Guildhall Street
Cambridge CB2 3NH, UK

## Abstract

This paper collects together a miscellany of results originally motivated by the analysis of the generalization performance of the "maximum-margin" algorithm due to Vapnik and others. The key feature of the paper is its operator-theoretic viewpoint. New bounds on covering numbers for classes related to Maximum Margin classes are derived *directly* without making use of a combinatorial dimension such as the VC-dimension. Specific contents of the paper include:

- a new and self-contained proof of Maurey's theorem and some generalizations with small explicit values of constants;
- bounds on the covering numbers of maximum margin classes suitable for the analysis of their generalization performance;
- the extension of such classes to those induced by balls in quasi-Banach spaces (such as $\ell_p$-norms with $0 < p < \infty$).
- extension of results on the covering numbers of convex hulls of basis functions to $p$-convex hulls ($0 < p \leq 1$);
- an appendix containing the tightest known bounds on the entropy numbers of the identity operator between $\ell_{p_1}^n$ and $\ell_{p_2}^n$ ($0 < p_1 < p_2 \leq \infty$).

## 1 Introduction

Linear classifiers have had a resurgence of interest in recent years because of the development of Support Vector machines [22, 24] which are based on Maximum Margin hyperplanes [25]. The generalization performance of support vector machines is becoming increasingly understood with an analysis of the covering numbers of the classes of functions they induce. Some of this analysis has made use of *entropy number* techniques.

In this paper we focus on the simple maximum margin case more closely and do not consider kernel mappings at all. The effect of the kernel used in support vector machines has been analysed using similar techniques in [32, 11] The classical maximum margin algorithm effectively works with the class of functions

$$\mathcal{F} := \left\{ \mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} \colon \|\mathbf{w}\|_{\ell_2^M} \leq 1,\ \|\mathbf{x}\|_{\ell_2^M} \leq 1 \right\}.$$

(The standard notation used here is defined precisely below in Definition 3 .) The focus of the present paper is to consider what happens when different norms are used in the definition of $\mathcal{F}$.

Apart from the purely mathematical interest in developing the connection between problems of determining covering numbers of function classes to those of determining entropy numbers of operators, the results in the paper indicate the effect to be expected by using different norms to define linear function classes in practical learning algorithms. There is a considerable body of work in the mistake-bounded framework for analysing learning algorithms for linear function classes exploring the effect of different norms. The present paper can be considered as a similar exercise in the statistical learning theory framework.

The following section collects all the definitions we need. All proofs in the paper are relegated to the appendix.

## 2 Definitions

We will make use of several notions from the theory of Banach spaces and a generalization of these called a quasi-Banach spaces. A nice general reference for Banach spaces is [33]; for quasi-Banach spaces see [8].

**Definition 1 (Banach space)** *A Banach space $(X, \|\cdot\|_X)$ is a complete normed linear space $X$ with a norm on $X$, i.e. a map $\|\cdot\|_X \colon X \to [0, \infty)$ that satisfies*

1. $\|\mathbf{x}\|_X = 0$ *if and only if* $\mathbf{x} = 0$*;*
2. $\|\lambda \mathbf{x}\|_X = |\lambda| \|\mathbf{x}\|_X$ *for scalars* $\lambda \in \mathbb{R}$ *and all* $\mathbf{x} \in X$*;*
3. $\|\mathbf{x} + \mathbf{y}\|_X \leq \|\mathbf{x}\|_X + \|\mathbf{y}\|_X$ *for all* $\mathbf{x}, \mathbf{y} \in X$.

**Definition 2 (Quasi-norm)** *A* quasi-norm *is a map like a norm which instead of satisfying the triangle inequality (3 above) satisfies*

**3′.** *There exists a constant $C$ such that for all $\mathbf{x}, \mathbf{y} \in X$, $\|\mathbf{x} + \mathbf{y}\|_X \leq C(\|\mathbf{x}\|_X + \|\mathbf{y}\|_X)$*

All of the spaces considered in this paper are real. A norm (quasi-norm) induces a *metric* (*quasi-metric*) via $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_X$. We will use $d$ to denote both the norm and the

induced metric, and use $(X, d)$ to denote the induced metric space. A *quasi-Banach space* is a complete quasi-normed linear space.

**Definition 3 ($\ell_p$ Norms)** *Suppose $\mathbf{x} \in \mathbb{R}^n$ with $n \in \mathbb{N}$ or $\mathbf{x} \in \mathbb{R}^{\mathbb{N}}$ (for infinite dimensional spaces). Then*

- *for $0 < p < \infty$, $\|\mathbf{x}\|_{\ell_p} := \|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ (provided the sum converges).*
- *for $p = \infty$, $\|\mathbf{x}\|_{\ell_\infty} := \|\mathbf{x}\|_\infty = \sup_i |x_i|$.*

*If $\dim \mathbf{x} = m$, we often write $\ell_p^m$ to explicitly indicate the dimension. The space $\ell_p^M$ is defined as $\ell_p^m := \{\mathbf{x} : \|\mathbf{x}\|_{\ell_p^m} < \infty\}$. Given $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \ell_p^M$, write $\mathbf{X}^m = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Suppose $\mathcal{F}$ is a class of functions defined on $\mathbb{R}^M$. The $\ell_p^m$ norm* with respect to $\mathbf{X}^m$ *of $f \in \mathcal{F}$ is defined as*

$$\|f\|_p^{\mathbf{X}^m} := \|(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m))\|_{\ell_p^m}. \tag{1}$$

For $1 \le p \le \infty$, $\|\cdot\|_{\ell_p^m}$ is a norm, and for $0 < p < 1$ it is a quasi-norm. Note that a different definition of the $\ell_p^m$ norm is used in some papers in learning theory, e.g. [28, 34]. A useful inequality in this context is the following (cf. e.g. [16, p.21]):

**Theorem 4 (Hölder's inequality)** *Suppose $p, q \ge 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$ and that $u \in \ell_p$ and $v \in \ell_q$. Then*

$$|u \cdot v| \le \|u\|_p \|v\|_q.$$

Suppose $(X, d)$ is a metric space and let $S \subset X$. We say $A_\epsilon \subset X$ is an $\epsilon$-cover for $S$ if for all $\mathbf{x} \in S$, there is an $a \in A_\epsilon$ such that $d(\mathbf{x}, a) < \epsilon$.

**Definition 5 (Covering and Entropy number)** *The* $\epsilon$-covering number *of $S$, denoted by $\mathcal{N}(\epsilon, S, d)$, is the size of the smallest $\epsilon$-cover of $S$. The $n$th* entropy number *of a set $S \subset X$ is defined by*

$$\epsilon_n(S) = \epsilon_n(S, d) := \inf\{\epsilon > 0 : \mathcal{N}(\epsilon, S, d) \le n\} \tag{2}$$

*Given a class of functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, the* uniform covering number *or $\epsilon$-growth function is*

$$\mathcal{N}^m(\epsilon, \mathcal{F}) := \sup_{\mathbf{X}^m \in \mathcal{X}^m} \mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_\infty^{\mathbf{X}^m}). \tag{3}$$

Covering numbers are of considerable interest to learning theory because generalization bounds can be stated in terms of them [1, 30].

**Definition 6 (Operator Norm)** *Denote by $X = (X, d)$ a (quasi)-normed space, $U_d$ is the (closed) unit ball: $U_d := \{\mathbf{x} \in X : \|\mathbf{x}\|_X \le 1\}$. Suppose $X$ and $Y$ are (quasi)-Banach spaces and $T$ is a linear operator mapping from $X$ to $Y$. Then the* operator norm *of $T$ is defined by*

$$\|T\| := \sup\{\|T\mathbf{x}\|_Y : \mathbf{x} \in U_X\}. \tag{4}$$

*and $T$ is called* bounded *if $\|T\| < \infty$.*

We denote by $\mathfrak{L}(X, Y)$ the set of all bounded linear operators from $X$ to $Y$.

**Definition 7 (Entropy numbers of operators)** *Suppose $T \in \mathfrak{L}(X, Y)$. The* entropy numbers of the operator $T$ *are defined by*

$$\epsilon_n(T) = \epsilon_n(T, X) := \epsilon_n(T(U_X)). \tag{5}$$

*The* dyadic entropy numbers $e_n(T)$ *are defined by*

$$e_n(T) = e_n(T, d) := \epsilon_{2^{n-1}}(T) \text{ for } i \in \mathbb{N}. \tag{6}$$

The main reference for entropy numbers of operators is [7]. Many of the properties shown there for Banach spaces $X, Y$ actually carry over to quasi-Banach spaces — see e.g. [8]. The *factorization theorem for entropy numbers* is of considerable use:

**Lemma 8 (Edmunds and Triebel [8, p.7])** *Let $A, B, C$ be quasi-Banach spaces and let $S, T \in \mathfrak{L}(A, B)$ and $R \in \mathfrak{L}(B, C)$. Then*

1. $\|T\| \ge e_1(T) \ge e_2(T) \ge \cdots \ge 0$.
2. $\forall k, l \in \mathbb{N}$, $e_{k+l-1}(RS) \le e_k(R) e_l(S)$.

Finite rank operators have exponentially decaying entropy numbers:

**Lemma 9 (Carl and Stephani, [7, p.14,21])** *Denote by $A, B$ Banach spaces, let $T \in \mathfrak{L}(A, B)$ and suppose that $\operatorname{rank}(T) = m$. Then for all $k \in \mathbb{N}$,*

$$e_k(T) \le 4\|T\| 2^{-(k-1)/m}. \tag{7}$$

In fact this bound is tight to within a constant factor of 4.

One operator that we will make repeated use of is the *identity* operator If $A$ and $B$ are (quasi)-Banach spaces,

$$\mathrm{id} : A \to B, \text{ defined by } \mathrm{id} : x \mapsto x. \tag{8}$$

This seemingly trivial operator is of interest when $A \ne B$ because of the definition of the operator norm. As we shall see below the entropy numbers of $\mathrm{id} : \ell_{p_1}^n \to \ell_{p_2}^n$ play a central role in many of the results we develop. In the following we will use $c$ to denote a positive constant and log is logarithm base 2.

## 3 Entropy Numbers of Linear Function Classes

Since we are mainly concerned about the capacity of linear functions of the form

$$f(\mathbf{x}) := \mathbf{w} \cdot \mathbf{x} \text{ with } \|\mathbf{x}\|_A \le c_{\mathbf{x}} \text{ and } \|\mathbf{w}\|_B \le c_{\mathbf{w}} \tag{9}$$

we will consider bounds on entropy numbers of linear operators. (Here $\|\cdot\|_A$ and $\|\cdot\|_B$ are norms or quasi-norms.) This will allow us to deal with function classes derived from (9). In particular we will analyze the class of functions

$$\mathcal{F}_{p,q}^M := \left\{\mathbf{x} \mapsto \mathbf{w} \cdot \mathbf{x} : \|\mathbf{w}\|_{\ell_p^M} \le 1,\ \|\mathbf{x}\|_{\ell_q^M} \le 1\right\} \tag{10}$$

where $M \in \mathbb{N}$, $p, q \ge 0$, and $\frac{1}{p} + \frac{1}{q} \ge 1$. More specifically we will look at the evaluation of $\mathcal{F}_{p,q}^M$ on an $m$–sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \ell_q^M$ and the entropy numbers of the evaluation map in terms of the $\ell_\infty^m$ metric. Formally, we will study the entropy numbers of the operator $S_{\mathbf{X}^m}$ defined as

$$\begin{array}{rccc} S_{\mathbf{X}^m} & : & \ell_p^M & \to & \ell_\infty^m \\ S_{\mathbf{X}^m} & : & \mathbf{w} & \mapsto & (\mathbf{w} \cdot \mathbf{x}_1, \ldots, \mathbf{w} \cdot \mathbf{x}_m) = \mathbf{w} \cdot \mathbf{X}^m. \end{array}$$

The connection between $S_{\mathbf{X}^m}$ and $\mathcal{F}_{p,q}^M$ is given in Lemma 11 below. Since one cannot expect that all problems can be cast into (10) without proper rescaling we will be interested in constraints on $\mathbf{w}$ and $\mathbf{x}_i$ such that $|\mathbf{w} \cdot \mathbf{x}_i| \leq 1$ (and rescale later).

**Lemma 10 (Product bounds from Hölder's Inequality)** *Suppose $p, q \geq 0$ with $\frac{1}{p}+\frac{1}{q} \geq 1$. Furthermore suppose $M \in \mathbb{N}$, $\mathbf{w} \in \ell_p^M$, $\mathbf{x} \in \ell_q^M$ with $\|\mathbf{w}\|_{\ell_p^M} \leq c_{\mathbf{w}}$ and $\|\mathbf{x}\|_{\ell_q^M} \leq c_{\mathbf{x}}$. Then*

$$\sup_{\mathbf{x}_1,\dots,\mathbf{x}_m \in c_{\mathbf{x}} U_{\ell_q^M}} \|S_{\mathbf{X}^m}\| \leq c_{\mathbf{w}} c_{\mathbf{x}}. \tag{11}$$

In order to avoid tedious notation we will assume that $c_{\mathbf{w}} = c_{\mathbf{x}} = 1$. This is no major restriction since the general results follow simply by rescaling $\epsilon$ by $c_{\mathbf{w}} c_{\mathbf{x}}$.

Our interest in the operator $S_{\mathbf{X}^m}$ and its entropy numbers is explained by the following lemma which connects $e_k(S_{\mathbf{X}^m})$ to the uniform covering numbers of $\mathcal{F}_{p,q}^M$.

**Lemma 11 (Entropy and Covering Numbers)** *Let $k \in \mathbb{N}$. If for all $\mathbf{X}^m \in (U_{\ell_q^M})^m$, $e_k(S_{\mathbf{X}^m} : \ell_p^M \to \ell_\infty^m) \leq \epsilon$, then*

$$\log \mathcal{N}^{\mathbf{X}^m}(\epsilon, \mathcal{F}_{p,q}^M) \leq k - 1. \tag{12}$$

### 3.1 The Maurey-Carl Theorem

In this section we present a special case of the famous Maurey-Carl theorem. The proof (in the appendix) presented provides a (small) explicit constant.

The result is not only of fundamental importance in statistical learning theory — it is of central importance in pure mathematics. Carl and Pajor [6] prove the Maurey theorem via the "Little Grothendieck theorem" which is related to Grothendieck's "fundamental theorem of the metric theory of tensor products". Furthermore the Little Grothendieck theorem can be proved in terms of Maurey's theorem (and thus they are formally equivalent). See [7, pages 254–267] for details. The version proved by Carl [3] (following Maurey's proof) uses a characterization of Banach spaces in terms of their Rademacher type. The latter is defined as follows.

**Definition 12 (Rademacher type of Banach spaces)** *A Banach space $X$ is of Rademacher type $p$, $1 \leq p \leq 2$ if there is constant $\tau > 0$ such that for every finite sequence $\{x_1, \dots, x_n\} \subset X$ we have*

$$\int_0^1 \left\| \sum_{i=1}^n r_i(t) x_i \right\| dt \leq \tau \left( \sum_{i=1}^n \|x_i\|^p \right)^{1/p}. \tag{13}$$

*Here $r_i(t) = \operatorname{sgn} \sin(2^i \pi t)$ is the $i$th Rademacher function on $[0,1]$. The Rademacher type $p$ constant $\tau_p(X)$ is the smallest constant $\tau$ satisfying (13).*

**Theorem 13 (Maurey-Carl)** *Let $X$ be a Banach space of Rademacher type $p$, $1 < p \leq 2$. Let $m \in \mathbb{N}$ and let $S \in \mathfrak{L}(\ell_1^m, X)$. Then there exists a constant $c$ such that for all $k \in \mathbb{N}$, $k \leq m$*

$$e_k(S) \leq c \tau_p(X) \|S\| \left( k^{-1} \log \left( \frac{m}{k} + 1 \right) \right)^{1-\frac{1}{p}}. \tag{14}$$

It is of interest to determine an explicit value for the constant $c$. Carl [3] proved that for $S : \ell_1^m \to E$,

$$\epsilon_{\binom{2m+k-1}{k}}(S) \leq 4 \tau_p(E) k^{-1+1/p} \|S\| \tag{15}$$

This leads to the following straight-forward corollary:

**Corollary 14 (Small Constants for Maurey's theorem)** *If $X$ is a Hilbert space, (14) holds with $p = 2$, $\tau_p(E) = 1$, and $c \leq 4.4377$.*

The dual version of Theorem 13, i.e. bounds on $e_k(X \to \ell_\infty^m)$ has identical formal structure as Theorem 13. We only state the Hilbert space case here.

**Theorem 15 (Dual Version of the Maurey-Carl Theorem)** *Suppose $H$ is a Hilbert space, $m \in \mathbb{N}$ and $T$ a linear operator $T : H \to \ell_\infty^m$. Then*

$$e_k(T : H \to \ell_\infty^m) \leq c \|T\| \left( k^{-1} \log \left( \frac{m}{k} + 1 \right) \right)^{1/2}, \tag{16}$$

*where $c = 102.88$.*

As explained in Appendix B, we suspect that a smaller value of $c$ is possible (we conjecture 1.86). Theorem 15 can be improved by taking advantage of operators with low rank via Lemma 9.

**Lemma 16 (Improved Dual Maurey-Carl Theorem)** *Let $H$ be a Hilbert space, and suppose $m, k \in \mathbb{N}$ with $m \geq 4$ and $T$ is a linear operator $T : H \to \ell_\infty^m$. Then*

$$e_k(T : H \to \ell_\infty^m) \leq \|T\| \begin{cases} 1 & \text{if } k \leq \log m \\ c \left( k^{-1} \log \left( \frac{m}{k} + 1 \right) \right)^{1/2} & \text{if } \log m \leq k \leq m \\ 8cm^{-\frac{1}{2}} 2^{-k/m} & \text{if } m < k \end{cases} \tag{17}$$

*where $c = 102.88$.*

## 4 Dimensionality and Sample Size

Lemma 16 already indicated that the size $m$ of the sample generating the evaluation operator $S_{\mathbf{X}^m}$ plays a crucial role in the scaling behaviour of entropy numbers. The dimensionality $M$ of $X$ (i.e. the dimension of the samples $\mathbf{x}_i$) also comes into play. This guides our analysis in the present section. In section 4.1 we will deal with the case where $M = m$, section 4.2 deals with the situation where $M$ is polynomial in $m$.

Depending on the setting of the learning problem we need bounds for the entropy numbers of the identity map between $\ell_p^m$ and $\ell_q^m$. We will use such bounds repeatedly below. We have collected together a number of bounds on this in appendix C.

### 4.1 Dimensionality of $X$ and Sample Size $m$ are Equal

We begin with the simplest case — $X$ is a finite dimensional Hilbert space of dimensionality $M = m$, hence we will be dealing with $\mathcal{F}_{2,2}^m$. Lemma 16 applies. It is instructive to restate this result in terms of covering numbers.

**Theorem 17 (Covering Numbers for $\mathcal{F}^m_{2,2}$)** *There exists constants $c, c' > 0$ such that for all $n \in \mathbb{N}$, and all $\epsilon > 0$,*

$$\log \mathcal{N}^m(\epsilon, \mathcal{F}^m_{2,2}) \leq \begin{cases} c\frac{\log m}{\epsilon^2} & \text{if } \epsilon > \frac{c'}{\sqrt{m}} \\ \max(1, -cm\log(m\epsilon^2)) & \text{if } \epsilon \leq \frac{c'}{\sqrt{m}}. \end{cases} \tag{18}$$

It is interesting to note the analogy with the Sauer-Vapnik-Chervonenkis lemma [31, 20, 1] which shows that the growth function has two regimes. We will now develop generalizations of the above result for $\mathcal{F}^m_{p,q}$ with $(p,q) \neq (2,2)$. An existing result in this direction is

**Lemma 18 (Carl [3, p. 94])** *Let $S \in \mathfrak{L}(\ell^m_p, \ell^m_\infty)$, $m \in \mathbb{N}$ and let $1 < p \leq 2$. Then there exists a constant $c = c(p)$ such that for all $k \in \mathbb{N}$,*

$$e_k(S) \leq c\,\|S\| \left(k^{-1}\log\left(1 + \frac{m}{k}\right)\right)^{\frac{1}{2}}. \tag{19}$$

(For $k > m$ one can get a better bound along the lines of Lemma 16.) This leads to the following theorem:

**Theorem 19 (Slack in $\mathcal{F}^m_{p,q}$)** *Let $p > 0$, $q \geq 2$, and $\frac{1}{p} + \frac{1}{q} > 1$. Then there exist constants $c, c'$ such that with $\beta := \frac{1}{p} + \frac{1}{q} - \frac{1}{2}$ we have $e_k(\mathcal{F}_{p,q}) \leq c\left(k^{-1}\log\left(\frac{m}{k}+1\right)\right)^{\beta}$ and $\log \mathcal{N}^m(\epsilon, \mathcal{F}^m_{p,q}) \leq c' \log m\, \epsilon^{-1/\beta}$.*

### 4.2 Dimensionality of $X$ is Polynomial in the Sample Size $m$

Now we will consider $\mathcal{N}^m(\epsilon, \mathcal{F}^M_{p,q})$ when $M > m$. We will derive results that are useful when $M$ is polynomial in $m$. With $S_{\mathbf{X}^m} \colon \ell^M_p \to \ell^m_\infty$ defined as before, we proceed to bound $e_k(S_{\mathbf{X}^m})$.

**Lemma 20 (Slack in $\mathcal{F}^M_{p,q}$)** *Let $0 < p \leq 2$, $q \geq 1$, $\frac{1}{p} + \frac{1}{q} \geq 1$, and $M, m \in \mathbb{N}$. Then there exists a constant $c' \geq 0$ such that with*

$$\begin{aligned} & e_{2k+1}(\mathcal{F}^M_{p,q}, \ell^m_\infty) \\ \leq\ & e_k(\mathrm{id}\colon \ell^M_p \to \ell^M_2)\, e_k(\tilde{S}_{\mathbf{X}^m} \colon \ell^M_2 \to \ell^m_\infty) && (20) \\ \leq\ & c'\left(k^{-1}\log\left(\tfrac{M}{k}+1\right)\right)^{\frac{1}{p}-\frac{1}{2}} \left(k^{-1}\log\left(\tfrac{m}{k}+1\right)\right)^{\frac{1}{2}} && (21) \end{aligned}$$

Consider now the situation that $p = 1$ and $q = 2$. From Lemma 20,

$$\begin{aligned} & e_{2k+1}(\mathcal{F}^M_{1,2}, \ell^m_\infty) \\ \leq\ & c'\left(k^{-1}\log\left(\frac{M}{k}+1\right)\right)^{\frac{1}{2}} \left(k^{-1}\log\left(\frac{m}{k}+1\right)\right)^{\frac{1}{2}} \\ =\ & c' k^{-1} \log^{1/2}\left(\frac{M}{k}+1\right) \log^{1/2}\left(\frac{m}{k}+1\right). \end{aligned}$$

Thus regression with $\mathcal{F}^M_{1,2}$ has a sample complexity of $m(\epsilon) \sim \frac{1}{\epsilon^{\tau+1}} \log^{1/2} M$ ignoring $\log(m)$ factors. Interestingly Zhang [34] has developed a result going in the other direction: he makes use of mistake bounds to determine similar covering numbers, whereas our covering number bounds (computed directly) recover the general form of the mistake bounds when turned into batch learning results. (Note, too, that Zhang uses a normalised definition of $\|\cdot\|_{\ell^m_p}$ and so care needs to be taken in comparing his results to ours.)

## 5 Covering Numbers of $\mathrm{co}_p(F)$

Adaboost [10] is an algorithm related to the variants on the maximum margin algorithm considered in this paper. It outputs an hypothesis which is the convex hull of the set of weak learners and its generalization performance can be expressed in terms of the covering number of the convex hull of weak learners at a scale related to the margin achieved [21]. In this section we consider what effect there is on the covering numbers of the class used (and hence the generalization bounds) when the $p$-convex hull is used (with $p \in (0,1)$). Variants of Adaboost can be developed which use the $p$-convex hull and experimental results indicate that $p$ affects the generalization in a manner consistent with what is suggested by the theory below [2]. The argument below is inspired by the results in [4, 5].

We are interested in $\mathcal{N}(\epsilon, \mathrm{co}_p(F), \ell^m_\infty)$ when $F$ is a subset of a Hilbert space $H$ and $\mathrm{co}_p(F)$ denotes the $p$-convex hull of $F$ (see Definition 21). Recalling the definition of $\|f\|^{\mathbf{X}^m}_\infty$, let $\|f\|^{\mathbf{X}^m}_q := (\sum_{i=1}^m |f(x_i)|^q)^{1/q}$. For all $q > 0$, we have $\|f\|^{\mathbf{X}^m}_\infty \leq \|f\|^{\mathbf{X}^m}_q$ and thus $\mathcal{N}(\epsilon, F, \ell^m_\infty) \leq \mathcal{N}(\epsilon, F, \ell^m_2)$. Since $\|\cdot\|^{\mathbf{X}^m}_2$ induces a Hilbert space we will now bound $\mathcal{N}(\epsilon, \mathrm{co}_p(F), \ell^m_\infty)$ in terms of covering numbers with respect to the Hilbert space norm $\|\cdot\|^{\mathbf{X}^m}_2$. Since the results will hold regardless of $\mathbf{X}^m$ in fact we will bound the uniform covering number $\mathcal{N}^m(\epsilon, \mathrm{co}_p(F))$.

**Definition 21 ($p$-Convex Hull)** *Suppose $p > 0$, and $F$ is a set. Then the $p$-convex hull of $F$ (strictly speaking the $p$-absolutely convex hull) is defined by [13, chapter 6]*

$$\mathrm{co}_p(F) = \bigcup_{n\in\mathbb{N}} \left\{ \sum_{i=1}^n \alpha_i f_i \colon f_i \in F,\, \alpha_i \in \mathbb{R},\, \sum_{i=1}^n |\alpha_i|^p \leq 1 \right\}$$

As an example, consider the $p$-convex hull of the set of Heaviside functions on $[0,1]$. It is well known that the convex hull ($p = 1$) is the set of functions of bounded variation. In Appendix D we explore the analogous situation for $0 < p < 1$.

**Lemma 22** *Let $0 < p \leq 1$, let $A$ be a set and let $U_\epsilon(A)$ be an $\epsilon$-cover of $A$. Then $\mathrm{co}_p(U_\epsilon(A))$ is an $\epsilon$-cover of $\mathrm{co}_p(A)$.*

**Lemma 23** *Let $0 < p \leq 1$, $\delta > 0$, $\epsilon_1, \epsilon_2 > 0$ and $\epsilon_1 + \epsilon_2 = \delta$. Then $\mathcal{N}(\delta, \mathrm{co}_p(A)) \leq \mathcal{N}(\epsilon_1, \mathrm{co}_p(U_{\epsilon_2}(A)))$.*

**Lemma 24** *Let $A = \{a_1, \ldots, a_n\} \subset H$ where $H$ is a Hilbert space. Define the linear operator $S : \ell^n_p \to H$ such that $S\mathrm{e}_i = a_i$, $i = 1, \ldots, n$, where $\mathrm{e}_i$ is the canonical basis of $\ell^n_p$ (e.g. $\mathrm{e}_1 = (1, 0, \ldots, 0)$, $\mathrm{e}_2 = (0, 1, \ldots, 0)$, etc.). Then $\mathrm{co}_p(A) = S(U_{\ell^n_p})$.*

**Theorem 25 (Covering Numbers of $p$-Convex Hull)** *Let $H$ be a Hilbert space, let $F \subset H$ be compact, and let $B := \sup_{h\in F} \|h\|$. Then $\log \mathcal{N}(\delta, \mathrm{co}_p(F)) \leq$*

$$\min_{\epsilon_2 \in (0,\delta)} \left\{ k \colon cB\left(k^{-1}\log\left(1 + \tfrac{\mathcal{N}(\epsilon_2, F)}{k}\right)\right)^{\frac{1}{p}-\frac{1}{2}} = \delta - \epsilon_2 \right\}.$$

Suppose $\mathcal{N}(\epsilon, F) \sim \left(\frac{1}{\epsilon}\right)^d$ for some $d \in \mathbb{N}$. We can determine the rate of growth of $\log \mathcal{N}(\delta, \mathrm{co}_p(F))$ as follows. Neglecting the $k$ inside the log in (45) we can explicitly solve the

equation. Numerical evidence suggests that the dependence of the value of $k$ in (45) is not very strong, so we choose $\epsilon_2 = \delta/2$. Then a simple approximate calculation yields:

**Corollary 26** *Suppose $F \subset H$ is such that for some $d \in \mathbb{N}$, $\mathcal{N}(\epsilon, F) \sim \left(\frac{1}{\epsilon}\right)^d$. Then for $0 < p \leq 1$,*

$$\log \mathcal{N}(\delta, \mathrm{co}_p(F)) \sim c(p) d \left(\frac{1}{\delta}\right)^{\frac{2p}{2-p}} \log 1/\delta.$$

For $p = 1$ this is $O((1/\delta)^2 \log(1/\delta))$ whereas we know from [4] that the rate should be $O((1/\delta)^{\frac{2d}{d+2}})$. Of course for large $d$, the difference is negligible (asymptotically in $1/\delta$).

## 6 Conclusions

We have computed covering numbers for a range of variants on the maximum margin algorithm. In doing so we made explicit use of an operator theoretic viewpoint already used fruitfully in analysing the effect of the kernel in SV machines. We also analysed the covering numbers of $p$-convex hulls of simple classes of functions.

We have seen how the now classical results for maximum margin hyperplanes can be generalized to function classes induced by different norms. The scaling behaviour of the resulting covering number bounds gives some insight into how related algorithms will perform in terms of their generalization performance. In other work [32, 11, 26] we have explored the effect of the kernel used in support vector machines for instance. In that case the eigenvalues of the kernel play a key role. In all this work the viewpoint that he function class is the image under the multiple evaluation map (considered as a linear operator) of a ball induced by a norm has been used.

Gurvits [12] asked (effectively) what can learning theory do for the geometric theory of Banach spaces. It seems the assistance flows more readily in the other direction. Perhaps the one contribution learning theory has made is pointing out an interesting research direction [27] by giving an answer to Pietsch's implicit question where he said of entropy numbers [19, p.311] that "at present we do not know any application in the real world." Now at least there is one!

**Acknowledgements** We would like to thank detailed comments and much help from Bernd Carl, Anja Westerhoff and Ingo Steinwart. This work was supported by the Australian Research Council. AS was supported by the DFG, grant SM 62/1-1. Parts of this work were done while BS was visiting the Australian National University.

## References

[1] M. Anthony and P. Bartlett. *A Theory of Learning in Artificial Neural Networks*. Cambridge University Press, 1999.

[2] J. Barnes. Capacity control in boosting using a $p$-convex hull. Master's thesis, Department of Engineering, Australian National University, 1999.

[3] B. Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Ann. de l'Institut Fourier*, 35(3):79–118, 1985.

[4] B. Carl. Metric entropy of convex hulls in hilbert spaces. *Bulletin of the London Mathematical Society*, 29:452–458, 1997.

[5] B. Carl, I. Kyrezi, and A. Pajor. Metric entropy of convex hulls in Banach spaces. *Proceedings of the London Mathematical Society*, 1999. to appear.

[6] B. Carl and A. Pajor. Gelfand numbers of operators with values in a Hilbert space. *Inventiones Mathematicae*, 94:479–504, 1988.

[7] B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators*. Cambridge University Press, Cambridge, UK, 1990.

[8] D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.

[9] D.E. Edmunds and H. Triebel. Entropy numbers and approximation numbers in function spaces. *Proceedings of the London Mathematical Society*, 58:137–152, 1989.

[10] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. 13th International Conference on Machine Learning*, pages 148–146. Morgan Kaufmann, 1996.

[11] Y. G. Guo, P. L. Bartlett, J. Shawe-Taylor, and R. C. Williamson. Covering numbers for support vector machines. In *Proceedings of COLT99*, 1999.

[12] L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In M. Li and A. Maruoka, editors, *Algorithmic Learning Theory ALT-97*, LNAI-1316, pages 352–363, Berlin, 1997. Springer.

[13] H. Jarchow. *Locally Convex Spaces*. B.G. Teubner, 1981.

[14] H. König. *Eigenvalue Distribution of Compact Operators*. Birkhäuser, Basel, 1986.

[15] M. Laczkovich and D. Preiss. $\alpha$-Variation and transformation into $c^n$ functions. *Indiana University Mathematics Journal*, 34(2):405–424, 1985.

[16] I.J. Maddox. *Elements of Functional Analysis*. Cambridge University Press, Cambridge, 1970.

[17] A.M. Olevskiĭ. Homeomorphisms of the circle, modifications of functions, and Fourier series. *American Mathematical Society Translations (2)*, 147:51–64, 1990.

[18] A. Pietsch. *Operator ideals*. North-Holland, Amsterdam, 1980.

[19] A. Pietsch. *Eigenvalues and s-Numbers*. Cambridge University Press, Cambridge, 1987.

[20] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory*, 13:145–147, 1972.

[21] R. Schapire, Y. Freund, P. L. Bartlett, and W. Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 1998.

[22] B. Schölkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999.

[23] C. Schütt. Entropy numbers of diagonal operators between symmatric Banach spaces. *Journal of Approximation Theory*, 40:121–128, 1984.

[24] J. Shawe-Taylor and N. Cristianini. Margin distribution and soft margin. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 349 – 358, Cambridge, MA, 2000. MIT Press.

[25] A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, 2000.

[26] A.J. Smola, A. Elisseeff, B. Schölkopf, and R.C. Williamson. Entropy numbers for convex combinations and MLPs. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 369 – 387, Cambridge, MA, 2000. MIT Press.

[27] I. Steinwart. Some estimates for the entropy numbers of convex hulls with finitely many extreme points. Technical report, University of Jena, 1999.

[28] M. Talagrand. The Glivenko–Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9(2):371–384, 1996.

[29] H. Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland, Amsterdam, 1978.

[30] V. Vapnik. *Statistical Learning Theory*. Wiley, N.Y., 1998.

[31] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280, 1971.

[32] R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report 19, NeuroCOLT, http://www.neurocolt.com, 1998. Accepted for publication in IEEE Transactions on Information Theory.

[33] P. Wojtaszczyk. *Banach Spaces for Analysts*. Cambridge University Press, 1991.

[34] T. Zhang. Analysis of regularised linear functions for classification problems. IBM Research Report RC-21572, 1999.

# A Proofs

**Proof (Lemma 10)** If $p \leq 1$ choose $p' = 1$ and $q' = \infty$, otherwise simply set $p' = p$ and $q' = (1 - p^{-1})^{-1}$. Since $\frac{1}{p} + \frac{1}{q} \geq 1$ we can always find two positive numbers $p', q' \geq 1$ such that $p' \leq q'$ and $\frac{1}{p'} + \frac{1}{q'} = 1$. Note that by convexity $\|\xi\|_r \leq \|\xi\|_{r'}$ for $0 \leq r \leq r'$ and therefore $\ell_r^M \subseteq \ell_{r'}^M$. Thus we may apply Hölder's inequality with $p'$ and $q'$ to obtain

$$\sup_{1 \leq i \leq m} |\mathbf{x}_i \cdot \mathbf{w}| \leq \sup_{1 \leq i \leq m} \|\mathbf{x}_i\|_{p'} \|\mathbf{w}\|_{q'} \leq c_{\mathbf{x}} c_{\mathbf{w}} \tag{22}$$

which proves (11). ∎

**Proof (Lemma 11)** Fix $\mathbf{X}^m$. We have

$$e_k(S_{\mathbf{X}^m}) = \epsilon_{2^{k-1}}(S_{\mathbf{X}^m}) = \epsilon(S_{\mathbf{X}^m}(U_{\ell_p^M})) \leq \epsilon \tag{23}$$

where the last equality follows from the definition of $S_{\mathbf{X}^m}$. Since $\mathbf{X}^m \in (U_{\ell_q^M})^m$, we have $S_{\mathbf{X}^m}(U_{\ell_p^M}) = \mathcal{F}_{p,q}^M$. Thus $\epsilon_{2^{k-1}}(\mathcal{F}_{p,q}^M) = \epsilon_{2^{k-1}}(S_{\mathbf{X}^m}) \leq \epsilon$ and thus $\mathcal{N}(\epsilon, \mathcal{F}_{p,q}^M, \ell_\infty^{\mathbf{X}^m}) \leq 2^{k-1}$. Since $\mathbf{X}^m \in (U_{\ell_q^M})^m$ was arbitrary, we conclude that $\log_2 \mathcal{N}^{\mathbf{X}^m}(\epsilon, \mathcal{F}_{p,q}^M) \leq k - 1$. ∎

**Proof (Corollary 14)** The fact that $p = 2$ and $\tau_p(E) = 1$ follows from the construction of Hilbert spaces [33]. What remains to be shown is the bound on $c$. Specializing (15) for Hilbert spaces, rewriting $l(m,k) := \log \binom{2m+k-1}{k} + 1$ (and conversely $k(l,m)$) we obtain

$$e_l(s) \leq 4k(l,m)^{-1/2} \|S\|. \tag{24}$$

The next step is to find a (simple) function $\tilde{k}(l,m)$ such that $4k(l,m)^{-1/2}\|S\| \leq 4c\tilde{k}(l,m)^{-1/2}\|S\|$ or equivalently $c^2 \geq \tilde{k}(l,m)/k(l,m)$ for all $1 \leq k \leq m$. We choose $\tilde{k}(l,m) = l/\log(\frac{m}{l} + 1)$. Next we have to bound $c^2$. Set

$$\rho(k,m) := \frac{\tilde{k}(l(m,k),m)}{k} = \frac{l(m,k)}{\log(\frac{m}{l(m,k)} + 1)k}. \tag{25}$$

One can readily check that for any $m \in \mathbb{N}$, $\rho(k,m)$ attains its maximum value at $k = m$. Furthermore $m \mapsto \rho(m,m)$ is monotonically non-decreasing. Thus taking the limit (note that $\binom{3m-1}{m} = \frac{2}{3}\binom{3m}{m}$ and that $\log := \log_2$) we take

$$\lim_{m \to \infty} \rho(m,m) < 6.16615 \tag{26}$$

Resubstitution yields $c < 4\sqrt{6.16615} < 9.9327$ which completes the proof.[1] ∎

**Proof (Lemma 16)** The first bound (for $k \leq \log m$) follows immediately from the definition of entropy numbers by $e_k(T) \leq \|T\|$. The second line of (17) is a direct consequence of Theorem 15. All that remains is the third line: for $k > m$ we factorise $T$ as $T = T \circ \mathrm{id} : \ell_2^m \to \ell_2^m$ and subsequently

$$e_k(T) \leq e_m(T) e_{k-m+1}(\mathrm{id} : \ell_2^m \to \ell_2^m). \tag{27}$$

[1] A (marginally) tighter version of the theorem could be stated by using $\rho$ directly to bound $e_k \leq 4\sqrt{\rho(m,m)}\left(k^{-1}\log(\frac{n}{k}+1)\right)^{-1/2}\|S\|$.

By Lemma 9,

$$e_{m-k+1}(\mathrm{id}: \ell_2^m \to \ell_2^m) \le 4 \cdot 2^{-(k-m)/m} \le 8 \cdot 2^{-k/m}.$$

Theorem 15 tells us $e_m(T) \le c\|T\| m^{-1/2}$. Substituting, we are done. ∎

**Proof (Theorem 17)** We distinguish between the two last bounds of Lemma 16. For $2 \le k \le m$ we can bound

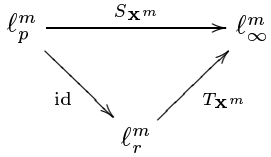
$$\begin{aligned} \epsilon_k(S_{\mathbf{X}^m} : \ell_2^m \to \ell_\infty^m) &\le c\left(k^{-1}\log\left(\frac{m}{k}+1\right)\right)^{1/2} \\ &\le c\left(k^{-1}\log m\right)^{1/2} \end{aligned} \tag{28}$$

and hence $k \le c^2\epsilon^{-2}\log m$. Application of Lemma 11 proves the first inequality of (18)[2]. For the second inequality simply note that the third line of Lemma 16 states (for $k \ge m$)

$$\begin{aligned} &\epsilon_k(S_{\mathbf{X}^m} : \ell_2^m \to \ell_\infty^m) \le c\, 2^{-\frac{k}{m}} m^{-\frac{1}{2}} \\ \Leftrightarrow \quad &k \le -\frac{1}{2} m \log(m\epsilon^2) + c \end{aligned}$$

Rewriting the conditions on $k$ in terms of $\epsilon$ and collecting all remaining terms into the constants $c$ and $c'$ completes the proof. ∎

**Proof (Theorem 19)** As before we observe that $\epsilon_k(\mathcal{F}_{p,q}^m) = \epsilon_k(S_{\mathbf{X}^m})$ and we will factorise the operator $S_{\mathbf{X}^m} : \ell_p^m \to \ell_\infty^m$ as in the diagram (with $\frac{1}{r} = 1 - \frac{1}{q}$):




(29)

The idea is that the identity operator uses up the "slack" between $p$ and $r$ implicit in the constraint $\frac{1}{p} + \frac{1}{q} \ge 1$. As will be seen, a smaller value of $\beta$ is achieved for larger values of $r$. But $r$ can be no larger than $(1 - \frac{1}{q})^{-1}$ in order for us to be able to use Lemma 10. If equality is achieved in the $p, q$ constraint, then id maps $\ell_p^m$ to $\ell_p^m$ and of course nothing (additional) is gained.

We will show that $\|T_{\mathbf{X}^m}\| \le 1$ and then use Lemma 8 to bound $e_k(S_{\mathbf{X}^m})$. The operator $T_{\mathbf{X}^m}$ is identical to $S_{\mathbf{X}^m}$ except its domain is $\ell_r^m$. Thus $T_{\mathbf{X}^m} : \mathbf{w} \mapsto (\mathbf{w}\cdot\mathbf{x}_1, \ldots, \mathbf{w}\cdot\mathbf{x}_m)$. By Lemma 10, since $\|\mathbf{x}_i\|_{\ell_q^m} \le 1$, and $\frac{1}{r} + \frac{1}{q} \ge 1$, we have

$$\begin{aligned} \|T_{\mathbf{X}^m}\| &= \sup_{\mathbf{w}\in U_{\ell_r^m}} \|T_{\mathbf{X}^m}\mathbf{w}\|_{\ell_\infty^m} \\ &\le \max_{i=1,\ldots,m} \sup_{\mathbf{w}\in U_{\ell_r^m}} \|\mathbf{w}\cdot\mathbf{x}_i\| \le 1. \end{aligned} \tag{30}$$

By Lemma 8

$$\begin{aligned} e_{k-1}(S_{\mathbf{X}^m}) &= e_{k-1}(T_{\mathbf{X}^m} \circ \mathrm{id}: \ell_p^m \to \ell_r^m) \\ &\le e_{k/2}(\mathrm{id}: \ell_p^m \to \ell_r^m) \cdot e_{k/2}(T_{\mathbf{X}^m}) \end{aligned} \tag{31}$$

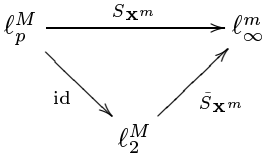
[2]A slightly more involved and longer argument gives a better bound: there are constants $c_1, c_2, c_3 > 0$ such that $\log \mathcal{N}(\epsilon, \mathcal{F}_{2,2}^m, \ell_\infty^m) \le \frac{c_1 \log(c_2\epsilon^2 m)}{\epsilon^2}$ for $\epsilon \ge c_3/\sqrt{m}$.

Hence using Lemmas 11, 31 and 18 we obtain

$$\begin{aligned} e_k(S_{\mathbf{X}^m}) &\le c\left(\frac{2}{k}\log\left(\frac{2m}{k}+1\right)\right)^{\frac{1}{p}-\frac{1}{r}} && (32)\\ &\quad \|T_{\mathbf{X}^m}\|\left(\frac{2}{k}\log\left(\frac{2m}{k}+1\right)\right)^{\frac{1}{2}} && (33)\\ &\le c\left(\frac{2}{k}\log\left(\frac{2m}{k}+1\right)\right)^{\frac{1}{p}+\frac{1}{q}-1} && (34)\\ &\quad \|T_{\mathbf{X}^m}\|\left(\frac{2}{k}\log\left(\frac{2m}{k}+1\right)\right)^{\frac{1}{2}} && (35)\\ &\le c'\|T_{\mathbf{X}^m}\|\left(k^{-1}\log\left(\frac{m}{k}+1\right)\right)^{\beta}. && (36) \end{aligned}$$

Here we used the product inequality for entropy numbers in (33), (35) holds since $p^{-1} - r^{-1} \le p^{-1} + q^{-1} - 1$, and moreover there exists a constant $\tilde{c}$ such that $\left(\frac{2}{k}\log\left(\frac{2m}{k}+1\right)\right) \le \tilde{c}\left(k^{-1}\log\left(\frac{m}{k}+1\right)\right)$ for $k \ge 1$. Solving for $k$ then immediately gives the bound on the covering numbers. ∎

**Proof (Lemma 20)** Similar to Theorem 17 consider the factorization




(37)

Exploiting the factorization we can bound

$$\begin{aligned} &e_{2k+1}(S_{\mathbf{X}^m} : \ell_p^M \to \ell_\infty^m) \\ &\le e_k(\mathrm{id}: \ell_p^M \to \ell_2^M) e_k(\tilde{S}_{\mathbf{X}^m}: \ell_2^M \to \ell_\infty^m) \end{aligned}$$

which proves (20). Next we have to bound the individual terms separately. For the first factor Lemma 31 can be applied. The dual version of the Maurey-Carl theorem applies to the second term. What remains to be shown is that $\|\tilde{S}_{\mathbf{X}^m}\| \le 1$. This, however, follows immediately from Lemma 10 in analogy to the previous section. ∎

**Proof (Lemma 22)** Let $f \in \mathrm{co}_p(A)$ be arbitrary, $f = \sum_{i=1}^N \alpha_i a_i$, $\sum_{i=1}^N |\alpha_i|^p \le 1$, where $N$ may be infinite. For $i = 1, \ldots, N$, let $\hat{a}_i \in U_\epsilon(A)$ be such that $\|\hat{a}_i - a_i\| \le \epsilon$. (Such $\hat{a}_i$ exist by definition of $U_\epsilon(A)$.) Let $\hat{f} := \sum_{i=1}^N \alpha_i \hat{a}_i$. We have

$$\begin{aligned} \|f - \hat{f}\| &= \left\|\sum_{i=1}^N \alpha_i a_i - \sum_{i=1}^N \alpha_i \hat{a}_i\right\| \\ &\le \sum_{i=1}^N |\alpha_i| \sup_j \|a_j - \hat{a}_j\| \\ &\le \sum_{i=1}^N |\alpha_i|\epsilon = \epsilon \end{aligned} \tag{38}$$

∎

**Proof (Lemma 23)** Suppose $V_{\epsilon_1}$ is an $\epsilon_1$-cover of $\mathrm{co}_p(U_{\epsilon_2}(A))$. From Lemma 22, for any $f \in \mathrm{co}_p(A)$, there exists a $\hat{f} \in \mathrm{co}_p(U_{\epsilon_2}(A))$ such that $\|f - \hat{f}\| \leq \epsilon_2$. But for any $\hat{f} \in \mathrm{co}_p(U_{\epsilon_2}(A))$, there exists $\hat{\hat{f}} \in V_{\epsilon_1}$ such that $\|\hat{f} - \hat{\hat{f}}\| \leq \epsilon_1$. By the triangle inequality $\|f - \hat{\hat{f}}\| \leq \|f - \hat{f}\| + \|\hat{f} - \hat{\hat{f}}\| \leq \epsilon_1 + \epsilon_2 = \delta$. Thus $V_{\epsilon_1}$ is an $\delta$-cover of $\mathrm{co}_p(A)$. ∎

**Proof (Lemma 24)** Let $\mathbf{x} = (x_1, \ldots, x_n) \in U_{\ell_p^n}$. Write $\mathbf{x} = \sum_{i=1}^n x_i \mathrm{e}_i$. Then $S\mathbf{x} = S(\sum_{i=1}^n x_i \mathrm{e}_i) = \sum_{i=1}^n x_i S\mathrm{e}_i = \sum_{i=1}^n x_i a_i$. Since $\mathbf{x} \in U_{\ell_p^n}$, $\sum_{i=1}^n |x_i|^p \leq 1$. Thus $S(U_{\ell_p^n}) \subseteq \mathrm{co}_p(A)$. Likewise, for any $f \in \mathrm{co}_p(A)$, $f = \sum_{i=1}^n x_i a_i$, with $\mathbf{x} = (x_1, \ldots, x_n) \in U_{\ell_p^n}$. Thus $\mathrm{co}_p(A) \subseteq S(U_{\ell_p^n})$. ∎

**Proof (Theorem 25)** Let $n := \mathcal{N}(\epsilon_2, F)$ and let $\epsilon_{k,n,\epsilon_2} := e_k(\mathrm{co}_p(U_{\epsilon_2}(F))$. Thus we have $\log \mathcal{N}(\epsilon_{k,n,\epsilon_2}, U_{\epsilon_2}(F)) \leq k$ and by Lemma 23, if

$$\epsilon_{k,n,\epsilon_2} + \epsilon_2 \leq \delta, \tag{39}$$

then $\log \mathcal{N}(\delta, \mathrm{co}_p(F)) \leq k$. Thus

$$\log \mathcal{N}(\delta, \mathrm{co}_p(F)) \leq \min_{\epsilon_2 \in (0,\delta)} \{k \colon (39) \text{ holds}\}. \tag{40}$$

By Lemma 24, $\epsilon_{k,n,\epsilon_2} = e_k(S)$. In order to compute $e_k(S)$ we factorise $S$ as follows

$$\begin{array}{ccc} \ell_{p_1}^m & \xrightarrow{\quad S \quad} & H \\ & {\scriptstyle \mathrm{id}} \searrow \quad \nearrow {\scriptstyle \tilde{S}} & \\ & \ell_1^m & \end{array} \tag{41}$$

Recalling the definition of $S$, and observing that $\mathrm{e}_i \in U_{\ell_1^m}$ for $i = 1, \ldots, m$, we obtain

$$\|\tilde{S}\| = \max_{i=1,\ldots,n} \|\tilde{S}\mathrm{e}_i\|_{\ell_2^m} = \max_{i=1,\ldots,m} \|a_i\|_{\ell_2^m} \leq B \tag{42}$$

By 8,

$$e_{k+l-1}(S) \leq e_k(\mathrm{id} \colon \ell_p^m \to \ell_1^m) e_l(\tilde{S} \colon \ell_1^m \to H). \tag{43}$$

Choosing $k = l$ and using Corollary 14 and Lemma 31 we obtain

$$e_{2k}(S) \leq c(p) 9.93269\, B \left( \frac{\log(\frac{m}{k} + 1)}{k} \right)^{\frac{1}{p} - \frac{1}{2}} \tag{44}$$

Furthermore we have that $\|S\| \leq \max_{i=1,\ldots,n} \|a_i\| \leq B$. Thus we obtain

$$\epsilon_{k,n,\epsilon_2} \leq cB \left( \frac{\log(1 + \frac{n}{k})}{k} \right)^{\frac{1}{p} - \frac{1}{2}} \tag{45}$$

Combining (40) and (45) concludes the proof. ∎

# B Maurey's Theorem

In this appendix we provide a proof of Maurey's theorem for operators $S \colon H \to \ell_\infty^m$ which gives an explicit value for the constant. This is considerably more work, and we get a correspondingly poorer estimate, than in the dual case of theorem 13 for operators $S \colon \ell_1^m \to H$. The argument of this section is due to Professor Bernd Carl.

**Theorem 27**

$$e_l(T : \ell_2^m \to \ell_\infty^m) \leq 102.88 \|T\| \left( \frac{\log(\frac{m}{l} + 1)}{l} \right)^{1/2}. \tag{46}$$

We compute the entropy of an operator $T : \ell_2^m \to \ell_\infty^m$ by factorizing it as

$$\begin{array}{ccc} \ell_2^m & \xrightarrow{\quad T \quad} & \ell_\infty^m \\ & {\scriptstyle T} \searrow \quad \nearrow {\scriptstyle \mathrm{id}} & \\ & \ell_p^m & \end{array} \tag{47}$$

where $p \leq 2 < \infty$ is a free parameter that we will optimize over at the end. The two factors will be dealt with by using the following two propositions, respectively.

**Definition 28** *$\ell(T)$ denotes the average*

$$\ell(T) = \int_{\mathbb{R}^m} \|T\mathbf{x}\|_E d\gamma(\mathbf{x}), \tag{48}$$

*of $\|T\mathbf{x}\|_E$ over the $m$-dimensional Gaussian measure*

$$\gamma(A) := \frac{1}{(2\pi)^{m/2}} \int_A e^{-\frac{1}{2}\|\mathbf{x}\|_2^2} d\mathbf{x}. \tag{49}$$

**Proposition 29 (Pajor, Talagrand)**

$$k^{1/2} e_k(T) \leq 4\sqrt{2}\, \ell(T) \tag{50}$$

We exploit the fact that the entropy numbers of the identity operator from $\ell_p^m$ to $\ell_\infty^m$ are known. They take the form:

**Proposition 30 (Schütt)** *For $0 < p < \infty$, there exists a constant $B$ such that for all $k \in \mathbb{N}$,*

$$e_k(\mathrm{id} : \ell_p^m \to \ell_\infty^m) \leq B \left( \frac{\log(m/k + 1)}{k} \right)^{\frac{1}{p}} \tag{51}$$

Here and below log is to base 2. Schütt [23] did not provide an explicit value for $B$. We compute one below in Lemma 32.

In order to use Proposition 29 to bound $e_k(T : \ell_2^m \to \ell_p^m)$, we need to upper bound $\ell(T)$. Using $x_i$, $i = 1, \ldots, m$, to denote the coordinates of $\mathbf{x}$ in the orthonormal basis $\{\mathrm{e}_i \colon i = 1, \ldots, n\}$, we have

$$\begin{aligned} \ell(T) &= \int_{\mathbb{R}^m} \|T\mathbf{x}\|_p d\gamma(\mathbf{x}) \\ &\leq \left( \int_{\mathbb{R}^m} \|T\mathbf{x}\|_p^p d\gamma(\mathbf{x}) \right)^{\frac{1}{p}} \\ &= \left( \int_{\mathbb{R}^m} \sum_{k=1}^m \left| \sum_{i=1}^m x_i \langle T\mathrm{e}_i, \mathrm{e}_k \rangle \right|^p d\gamma(\mathbf{x}) \right)^{\frac{1}{p}} \\ &= \left( \sum_{k=1}^m \int_{\mathbb{R}^m} \left| \sum_{i=1}^m x_i \langle T\mathrm{e}_i, \mathrm{e}_k \rangle \right|^p d\gamma(\mathbf{x}) \right)^{\frac{1}{p}}. \end{aligned}$$

The Khinchin inequality [33] states that

$$\left(\int_{\mathbb{R}^m}\left|\sum_{i=1}^m x_i\alpha_i\right|^p d\gamma(x_1,\ldots,x_m)\right)^{1/p} \le c_p\left(\sum_{i=1}^m|\alpha_i|^2\right)^{1/2}$$

where

$$c_p = 2\left(\frac{\Gamma((1+p)/2)}{\Gamma(1/2)}\right)^{1/p} \le \sqrt{p}.$$

Hence

$$\begin{aligned}
\ell(T) &\le \sum_{k=1}^m\left(\sqrt{p}^p\left(\sum_{i=1}^m|\langle T\mathrm{e}_i,\mathrm{e}_k\rangle|^2\right)^{p/2}\right)^{1/p}\\
&= \sqrt{p}\left(\sum_{k=1}^m\left(\sum_{i=1}^m|\langle T\mathrm{e}_i,\mathrm{e}_k\rangle|^2\right)^{p/2}\right)^{1/p}\\
&\le \sqrt{p}\left(\sum_{k=1}^m\|T'\mathrm{e}_k\|_2^p\right)^{1/p}\\
&\le \sqrt{p}m^{1/p}\sup_{1\le k\le m}\|T'\mathrm{e}_k\|_2\\
&= \sqrt{p}m^{1/p}\|T'\colon\ell_1^m\to\ell_2^m\|\\
&= \sqrt{p}m^{1/p}\|T\colon\ell_2^m\to\ell_\infty^m\|
\end{aligned}$$

By Proposition 29, we get

$$e_k(T:\ell_2^m\to\ell_p^m) \le k^{-\frac12}4\sqrt{2}\sqrt{p}m^{1/p}\|T:\ell_2^m\to\ell_\infty^m\|.$$

Next, we combine the obtained bound with Schütt's result, using

$$e_{2k-1}(T\colon\ell_2^m\to\ell_\infty^m) \le e_k(T\colon\ell_2^m\to\ell_p^m)\,e_k(\mathrm{id}\colon\ell_p^m\to\ell_\infty^m),$$

to obtain $e_{2k-1}(T:\ell_2^m\to\ell_\infty^m) \le$

$$k^{-\frac12}4\sqrt{2}\sqrt{p}m^{1/p}\|T:\ell_2^m\to\ell_\infty^m\|\;B\left(\frac{\log(m/k+1)}{k}\right)^{\frac1p}.$$

Now choose $p = 2\log(\frac{m}{k}+1)$. Observe $p\ge 2$ for $k\le m$. Thus

$$\begin{aligned}
e_{2k-1}(T:\ell_2^m\to\ell_\infty^m) &\le k^{-1/2}4\sqrt{2}\sqrt{2}(\log(\tfrac{m}{k}+1))^{1/2}\\
&\quad\times\|T\|m^{\frac{1}{2\log(\frac{m}{k}+1)}}B\left(\frac{\log(\frac{m}{k}+1)}{k}\right)^{1/p}\\
&= 8\;B\|T\|\left(\frac{\log(\frac{m}{k}+1)}{k}\right)^{1/2}\Lambda
\end{aligned}$$

where

$$\begin{aligned}
\Lambda &= m^{\frac{1}{2\log(\frac{m}{k}+1)}}\left(\frac{\log(\frac{m}{k}+1)}{k}\right)^{\frac{1}{2\log(\frac{m}{k}+1)}}\\
&= x^{\frac{1}{2\log(x+1)}}(\log(x+1))^{\frac{1}{2\log(x+1)}}
\end{aligned}$$

with $x = m/k$. Note $k\le m \Rightarrow x\ge 1$. Let $\tau = \log(x+1)$, so $x = 2^\tau - 1$. Then $\tau\ge 0$ and

$$\Lambda = \Lambda(t) = (2^\tau-1)^{\frac{1}{2\tau}}\tau^{\frac{1}{2\tau}}.$$

One can check that $\lim_{\tau\to\infty}\Lambda(\tau) = \sqrt{2}$ and $\Lambda(\tau)$ has a unique maximum for $\tau\in[0,\infty)$. Computing $\frac{\partial\Lambda(\tau)}{\partial\tau}$, setting it equal to zero and solving numerically, one finds the maximum occurs for $\tau = 3.66661119696101\ldots$ at which point $\Lambda(\tau) = 1.66956682\ldots =: C$. Thus

$$e_{2k-1}(T:\ell_2^m\to\ell_\infty^m) \le 8CB\|T\|\left(\frac{\log(\frac{m}{k}+1)}{k}\right)^{1/2} \tag{52}$$

and all that remains is to bound $B$, which we now do.

From (66) $B\le 2(3.70789)^{1/p}$. But our choice of $p\ge 2$ and hence $B\le 2(3.70789)^{1/2} = 3.851176\ldots$. Substituting this value of $B$ into (52) along with the numerical value of $C$ we get

$$e_{2k-1}(T\colon\ell_2^m\to\ell_\infty^m) \le 51.44\|T\|\left(\frac{\log(\frac{m}{k}+1)}{k}\right)^{1/2}.$$

Noting that $e_{2k}(T\colon\ell_2^m\to\ell_\infty^m)\le e_{2k-1}(T\colon\ell_2^m\to\ell_\infty^m)$, we obtain a statement valid for even as well as odd numbers. To infer a bound on $e_l$, $l\in\mathbb{N}$, we set $l = 2k$, hence $k = l/2$. Then

$$e_l(T\colon\ell_2^m\to\ell_\infty^m) \le 51.44\|T\|\left(\frac{2\log(\frac{2m}{l}+1)}{l}\right)^{1/2}.$$

Now for $\log(m)\le l\le m$ we have $\frac{\log(\frac{2m}{l}+1)}{l} \le \frac{\log(\frac{2m}{l}+2)}{l} = \frac{\log(\frac{m}{l}+1)+1}{l} \le \frac{2\log(\frac{m}{l}+1)}{l}$. Thus for $\log(m)\le l\le m$,

$$e_l(T:\ell_2^m\to\ell_\infty^m) \le 102.88\|T\|\left(\frac{\log(\frac{m}{l}+1)}{l}\right)^{1/2}. \tag{53}$$

**Conjecture on Best Value of Maurey Constant**

A value of 102.88 is not particularly satisfying. We believe in fact it is quite loose. Our reasoning is as follows. Consider all operators $T\colon\ell_2^m\to\ell_\infty^m$ such that $\|T\| = 1$. By definition $e_1(T) = 1$. Observe that the unit ball $U_{\ell_2^m}$ is the ellipsoid of maximum volume contained inside $U_{\ell_\infty^m}$. Since

$$\mathrm{vol}(T(U_{\ell_2^m})) = \lim_{\epsilon\to 0}\mathcal{N}(\epsilon, T(U_{\ell_2^m}))\mathrm{vol}(\epsilon U_{\ell_\infty^m})$$

where $\mathcal{N}(\epsilon,S)$ is the covering number of $S$ we have that

$$\mathrm{vol}(T(U_{\ell_2^m})) = \lim_{n\to\infty} n\mathrm{vol}(\epsilon_n(T(U_{\ell_2^m}))U_{\ell_\infty^m}).$$

But $\mathrm{vol}(T(U_{\ell_2^m}))$ is maximized over all $T$ such that $\|T\| = 1$ by choosing $T = \mathrm{id}$. Thus we have for $\|T\| = 1$

$$e_1(T)\le e_1(\mathrm{id}) \quad\text{and}\quad \lim_{n\to\infty}e_n(T)\le\lim_{n\to\infty}e_n(\mathrm{id}).$$

We conjecture that for all $n\in\mathbb{N}$ and all $T$ with $\|T\| = 1$

$$e_n(T:\ell_2^m\to\ell_\infty^m) \le e_n(\mathrm{id}:\ell_2^m\to\ell_\infty^m)$$

If this were true, by (55) the Maurey constant would be 1.86.

## C Bounds on $e_k(\mathrm{id}\colon\ell_{p_1}^n\to\ell_{p_2}^n)$

We now determine bounds on $e_{k,p_1,p_2}^m := e_k(\mathrm{id}_{p_1,p_2}^m)$. These have been given in [29, 4.10.3], [18, page 172], [14, 3.c.8], [23], and [9, p.141], (see also [8, page 101]). All but the last two references only considered $p_1\ge 1$. The most recent contribution by Edmunds and Triebel [9, page 141], [8, page 101] subsumes all of the others and is summarized in the Lemma below. For $p_1\ge 1$ by an argument of Schütt [23] it is asymptotically optimal in $n$ and $k$.

**Lemma 31** *Let $0 < p_1 \leq p_2 \leq \infty$. Then $e_k(\mathrm{id}\colon \ell_{p_1}^n \to \ell_{p_2}^n) \leq$*

$$c \begin{cases} 1 & \text{if } 1 \leq k \leq \log n \\ (k^{-1}\log(1+\frac{n}{k}))^{1/p_1 - 1/p_2} & \text{if } \log n \leq k \leq n \\ 2^{-k/n} n^{1/p_2 - 1/p_1} & \text{if } k \geq n \end{cases}$$

*for $k \in \mathbb{N}$ where $c$ is a positive constant independent of $M$ and $k$ depends on $p_1$ and $p_2$.*

We *can* find a nice (small) explicit value for $c$ for the case when $p_1 = 1$ which is of interest in its own right. We proceed with that case now.

### C.1 When $p_1 \geq 1$

For $p_1 \geq 1$ by an argument of Schütt [23] it is asymptotically optimal in $m$ and $k$. Since we need an explicit value of the constant we provide the explicit proof below.

**Lemma 32** *For all $m \in \mathbb{N}$, and all $l \leq m$,*

$$e_{l+1}(\mathrm{id} : \ell_1^n \to \ell_\infty^m) \leq \frac{3.707893773\log(\frac{m}{l}+1)}{l} \quad (54)$$

$$e_{l+1}(\mathrm{id} : \ell_2^m \to \ell_\infty^m) \leq 1.86\left(\frac{\log(\frac{m}{l}+1)}{l}\right)^{1/2}. \quad (55)$$

**Proof** Let $B(\epsilon, k)$ denote a $\ell_\infty$ $\epsilon$-ball in $k$ dimensions. Let $p > 0$. For a given number of dimensions $k$, we determine the smallest $\epsilon$ so that $U_p^k$ can be covered by

$$S_p^k := \bigcup_{i=1}^{k} \bigcup_{j=\lfloor -1/2\epsilon \rfloor}^{\lceil 1/2\epsilon \rceil} B(\epsilon, k) + 2j\epsilon \mathrm{e}_i \quad (56)$$

where $\mathrm{e}_i$ is the $i$th canonical basis vector. For $x \in U_p^k$ we have $\sum_{i=1}^k x_i^p \leq 1 \Rightarrow k x_1^p \leq 1 \Rightarrow x_1 \leq k^{-1/p}$. Setting $x_1 = \epsilon$ (the radius of $B(\epsilon, k)$) gives $\epsilon^p \leq 1/k \Rightarrow k \leq \epsilon^{-p}$. We will now set $k = \epsilon^{-p}$. Along each of the $k$ axes of $U_p^k$ we have used $(\lceil 1/2\epsilon \rceil - \lfloor -1/2\epsilon \rfloor - 1 + 1)$ cubes. We have added and subtracted one here so we do not multiply count the box that will live at the very center of $U_p^k$. Therefore along all $k$ axes, we have

$$\begin{aligned} & k(2\lceil 1/2\epsilon \rceil - 1) + 1 \\ = \; & k(2\lceil 1/2\epsilon \rceil - 1) + 1 \\ \leq \; & k(2(1/2\epsilon + 1) - 1) + 1 \\ = \; & k(1/\epsilon + 1) + 1. \end{aligned}$$

Thus $|S_p^k| \leq k(1/\epsilon + 1) + 1$. Observe that by construction of $k$, any point in $(U_p^m \setminus U_p^k)$ which is not covered by $S$ must lie within a $(m-k)$-dimension $\epsilon$-ball on one of the $(m-k)$ principal axes of $U_p^m$ not contained in $U_p^k$. Thus by separately covering all possible choices of $k$ axes in the manner above, we cover $U_p^m$. Since there are $\binom{m}{k}$ ways to choose the $k$ axes, we have that

$$\lambda_m^p(k) := (k(k^{1/p} + 1) + 1)\binom{m}{k} \quad (57)$$

$\epsilon$-$\ell_\infty^m$-balls cover $U_p^m$, where $\epsilon = \frac{1}{k^{1/p}}$. In other words

$$\epsilon_{\lambda_m^p(k)}(\mathrm{id}_{p,\infty}^m) \leq \frac{1}{k^{1/p}}. \quad (58)$$

for $k \leq m$ where $\mathrm{id}_{p,\infty}^m = \mathrm{id}\colon \ell_p^m \to \ell_\infty^m$. Now set

$$\psi_m^p(k) := \log \lambda_m^p(k). \quad (59)$$

Thus

$$e_{\psi_m^p(k)+1}(\mathrm{id}_{p,\infty}^m) \leq \frac{1}{k^{1/p}} \quad (60)$$

Let $l = \psi_m^p(k)$ and so $e_{l+1} \leq \frac{1}{k^{1/p}}$. Suppose $h_m(\cdot)$ is a function such $h_m(\psi_m^p(k)) \leq k$ for $k \leq m$. Then $\frac{1}{(h_m(l))^{1/p}} \geq \frac{1}{k^{1/p}}$ and so

$$e_{l+1} \leq \frac{1}{(h_m^p(l))^{1/p}} \text{ for } l \leq m. \quad (61)$$

Choose $\tilde{h}_m(l) = l/\log(\frac{m}{l}+1)$. Thus we need to show

$$\rho(m,k,p) := \frac{\tilde{h}_m^p(\psi_n^p(k))}{k} \leq c_p \quad (62)$$

for some constants $c_p$ depending on $p$. We will then set $h_m^p(l) = \tilde{h}_m(l)/c_p$. Numerical calculations indicate that $\rho(m,k,1)$ achieves a maximum value of $c_1 := 3.707893773$ at $(k,m) = (3,12)$. Similarly, we can numerically determine that $\rho(m,k,2)$ achieves a unique maximum value of $c_2 = 3.459446772$ at $(k,m) = (2,8)$. Thus for $p \in \{1,2\}$, $h_m(l) = \tilde{h}_m(l)/c_p$ will do. We then have from (61) that for $l \leq m$,

$$\begin{aligned} e_{l+1}(\mathrm{id}_{2,\infty}^m) &\leq \left(\frac{c_2}{\tilde{h}_m(l)}\right)^{1/2} \\ &\leq \left(3.45944677 l^{-1}\log\left(\frac{m}{l}+1\right)\right)^{1/2} \\ &\leq 1.85996\left(l^{-1}\log\left(\frac{m}{l}+1\right)\right)^{1/2} \end{aligned}$$ ∎

The following interpolation lemma follows immediately from [18, p.173].

**Lemma 33** *Let $1 \leq p_1 \leq p_2 \leq \infty$. Then for all $k \in \mathbb{N}$*

$$e_k(\mathrm{id}_{1,p_2}^n) \leq 2e_k(\mathrm{id}_{1,\infty}^m)^{1-1/p_2} \quad (63)$$

$$e_k(\mathrm{id}_{p_1,\infty}^m) \leq 2e_k(\mathrm{id}_{1,\infty}^m)^{1/p_1} \quad (64)$$

$$e_k(\mathrm{id}_{p_1,p_2}^m) \leq 4e_k(\mathrm{id}_{1,\infty}^m)^{1/p_1 - 1/p_2}. \quad (65)$$

Combining this Lemma with (54) gives for $p \geq 1$,

$$e_k(\mathrm{id}_{p,\infty}^m) \leq 2e_k(\mathrm{id}_{1,\infty}^m)^{1/p} \leq 2\left(\frac{3.70789\log(\frac{m}{k}+1)}{k}\right)^{1/p} \quad (66)$$

### C.2 When $p_1 < 1$

Lemma 31 and the proof of Lemma 32 suggests a similar form of the result in Lemma 32 should be obtainable when $p_1 < 1$. However it turns out that for $p_1 < 1$ the value of constant obtained is quite unsatisfactory. This is because its value is dominated by the behaviour of $\rho(p_1,k,n)$ for very small $k$ and $n$ ($(k,n) = (3,3)$). For learning applications these are uninteresting values of $k$ and $n$. Furthermore for some applications we are actually interested in how $e_{k,p_1,\infty}^n$ behaves as a function of $p$ for *fixed* $n$. For example if we

wanted to use $p_1$ as a "capacity control knob." In that case it is necessary to determine the dependence of $c$ on $p_1$ explicitly. In doing so it turns out that in the case of $p_1 < 1$ one has to pay too high a price for the elegance of an expression of the form (54) and we end up being better served by an implicit formula, which nevertheless can be easily computed. This implicit formula *does* exhibit the expected behaviour in $p$ for fixed $n$.

Setting $j := k^{1/p}$ we have from (60) that

$$e_{\psi_n^p(j^p)+1} \leq \frac{1}{2j}. \tag{67}$$

For a given $k$ we can easily determine $e_k$ numerically: Let $j_0 := \{j\colon \psi_n^p(j^p) + 1 = k\}$. Such a $j_0$ is unique. Then $e_k \leq 1/2j_0$. Using this method one can plot $k = k(p,\epsilon) = \log \mathcal{N}(\epsilon, \mathcal{F}_{p,\infty})$ as a function of $p \leq 1$ for various $\epsilon$. As one would expect, the log covering numbers decrease with decreasing $p$.

## D $p$-Convex Hull of Heavisides

We take the following definitions from [15]. Suppose $I \subset \mathbb{R}$ and $f\colon I \to \mathbb{R}$. For $\alpha \geq 0$,

$$V_\alpha(F, I) := \sup \sum_{i=1}^{n} |f(b_i) - f(a_i)|^\alpha$$

where $\{[a_i, b_i]\}_{i=1}^n$ is an arbitrary finite system of non-overlapping intervals with $a_i, b_i \in I$ for $i = 1, \ldots, n$.

Suppose $f$ is continuous on $[a,b]$. Let $G$ be the union of all open subintervals on which $f$ is either strictly monotonic or constant. Then

$$K = K_f = [a,b] \setminus G$$

is the *set of points of varying monotonicity*. If $f$ is not continuous everywhere, we let $D$ denote the set of points of discontinuity and set $\bar{K} = K \cup D$.

If $\alpha \geq 0$, $f$ is said to be of bounded $\alpha$-variation and we write $f \in \mathrm{CBV}_\alpha$ if $V_\alpha(f, \bar{K}) < \infty$. If $V_\alpha(f, \bar{K}) < M$ we say $f \in \mathrm{CBV}_\alpha(M)$.

Let $\mathcal{H} = \{H(x-\theta)\colon \theta \in [a,b]\}$ where $H(x)$ is the Heaviside (step function).

**Lemma 34** *For any $N \in \mathbb{N}$, for $0 < \alpha \leq 1$, $\mathrm{co}_\alpha^N(\mathcal{H}) \subseteq \mathrm{CBV}_\alpha(1)$.*

**Proof** For any $f \in \mathrm{co}_\alpha^N(\mathcal{H})$, we can write $f$ as

$$f(x) = \sum_{i=1}^{N} \beta_i H(x - \theta_i)$$

where we will assume that $\theta_1 \leq \theta_2 \leq \cdots \theta_N$. Observe that $\bar{K}_f \subseteq \{\theta_i\colon i = 1, \ldots, N\}$. Thus $V_\alpha(f, \bar{K}_f)$

$$\begin{aligned} &\leq \sum_{k=2}^{N} \left| \sum_{i=1}^{N} \beta_i H(\theta_k - \theta_i) - \sum_{i=1}^{N} \beta_i H(\theta_{k-1} - \theta_i) \right|^\alpha \\ &= \sum_{k=2}^{N} \left| \sum_{i=1}^{k} \beta_i - \sum_{i=1}^{k-1} \beta_i \right|^\alpha = \sum_{k=2}^{N} |\beta_k|^\alpha \leq 1. \end{aligned}$$

∎

Since $\mathrm{co}_\alpha^N(\mathcal{H}) \subseteq \mathrm{CBV}_\alpha(1)$ for all $N \in \mathbb{N}$, we have for $0 < \alpha \leq 1$

$$\bigcup_{N=1}^{\infty} \mathrm{co}_\alpha^N(\mathcal{H}) \subseteq \mathrm{CBV}_\alpha(1).$$

As a way of illustrating the "size" of $\mathrm{CBV}_\alpha$, in a fashion that gives some additional intuition to our entropy number results determined directly, we will now compute the fat-shattering dimension of $\mathrm{CBV}_\alpha(1)$.

**Proposition 35** *Suppose $0 < \alpha \leq 1$ and $0 < \gamma < 1$. Then*

$$\mathrm{Fat}_{\mathrm{CBV}_\alpha}(\gamma) \leq \frac{3}{\gamma^\alpha}.$$

**Proof** Suppose $\{f_1, \ldots, f_{2^m}\} \subset \mathrm{CBV}_\alpha(1)$ $\gamma$-shatter the points $(x_1, \ldots, x_m) \subset [a,b]$ with respect to $(r_1, \ldots, r_m)$. We will show that $m \leq 3\gamma^{-\alpha}$. For $i = 1, \ldots, 2^m$, we have $K_{f_i} \supseteq \{x_1, \ldots, x_m\}$. There must always be a sign assignment $b_i \in \{-1, 1\}^m$ which is realized w.r.t. $(r_1, \ldots, r_m)$ by $f_i$ such that

$$|f_i(x_j) - f_i(x_{j-1})| \geq 2\gamma$$

for $j = 2, \ldots, m$. Thus

$$V_\alpha(f_I, \bar{K}_{f_i}) \geq \sum_{j=2}^{m} |f_i(x_j) - f_i(x_{j-1})|^\alpha \geq (m-1)(2\gamma)^\alpha.$$

But by hypothesis $V_\alpha(f_i, \bar{K}_{f_i}) \leq 1$ and so $(m-1)(2\gamma)^\alpha \leq 1$. Thus

$$m \leq 2^{-\alpha}\gamma^{-\alpha} + 1 \leq 3\gamma^{-\alpha}$$

for $m > 1$ and $\gamma \leq 1$. ∎

The smallness of $\mathrm{CBV}_\alpha$ is well illustrated by the following theorem from [15]. For $0 \leq s \leq 1$ define

$$\begin{aligned} \mathrm{Lip}(s) \quad = \quad & \{f : [a,b] \to \mathbb{R}\colon \exists K > 0, \forall x, y \in [a,b], \\ & |f(y) - f(x)| \leq K|x-y|^s\}. \end{aligned}$$

For $k \in \mathbb{N}$ and $k < s \leq k+1$,

$$\begin{aligned} \mathrm{Lip}(s) \quad = \quad & \{f\colon f \text{ is } k\text{-times differentiable and} \\ & f^{(k)} \in \mathrm{Lip}(s-k)\}. \end{aligned}$$

**Theorem 36 (Laczkovich and Preiss)** *Let $s, \alpha > 0$, $s = 1/\alpha$. Then $f \in \mathrm{CBV}_\alpha$ if and only if there exists a homeomorphism $\phi$ of $[a,b]$ into itself such that $f \circ \phi \in \mathrm{Lip}(s)$.*

(For related results see [17] and references therein.)