Some Local Measures of Complexity of Convex Hulls and Generalization Bounds

Olivier Bousquet[1], Vladimir Koltchinskii[2]*, and Dmitriy Panchenko[2]

[1] Centre de Mathématiques Appliquées
Ecole Polytechnique
91128 Palaiseau, FRANCE
bousquet@cmapx.polytechnique.fr
[2] Department of Mathematics and Statistics
The University of New Mexico
Albuquerque, NM 87131-1141, U.S.A.
vlad@math.unm.edu,panchenk@math.unm.edu

Abstract. We investigate measures of complexity of function classes based on continuity moduli of Gaussian and Rademacher processes. For Gaussian processes, we obtain bounds on the continuity modulus on the convex hull of a function class in terms of the same quantity for the class itself. We also obtain new bounds on generalization error in terms of localized Rademacher complexities. This allows us to prove new results about generalization performance for convex hulls in terms of characteristics of the base class. As a byproduct, we obtain a simple proof of some of the known bounds on the entropy of convex hulls.

1 Introduction

Convex hulls of function classes have become of great interest in Machine Learning since the introduction of AdaBoost and other methods of combining classifiers. Working with convex combinations of simple functions allows to enhance the approximation properties of the learning algorithm with a small increase of computational cost. However, since the convex hull of the class is typically much larger than the class itself, the learning complexity gets also increased. It is thus of importance to assess the complexity of a convex hull in terms of the complexity of the base class.

The most commonly used measure of complexity of convex hulls is based on covering numbers (or metric entropies). The first bound on the entropy of the convex hull of a set in a Hilbert space was obtained by Dudley [9] and later refined by Ball and Pajor [1] and a different proof was given independently by van der Vaart and Wellner [21]. These authors considered the case of polynomial growth of the covering numbers of the base class. Sharp bounds in the case of exponential growth of the covering numbers of the base class as well as extension of previously konwn results to the case of Banach spaces were obtained later [7, 19, 16, 11, 8].

* Partially supported by NSA Grant MDA904-99-1-0031

In Machine Learning, however, the quantities of primary importance for determining the generalization performance are not the entropies themselves but rather the so-called Gaussian or Rademacher complexities of function classes [3]. It turns out that there is a direct relationship between such quantities measured on the convex hull and measured on the class itself as pointed out in [14].

More recently, it has been proven that it is possible to obtain refined generalization bounds by considering localized Gaussian or Rademacher complexities of the function classes [13, 2]. These quantities are closely related to continuity moduli of the corresponding stochastic processes.

Our main purpose in this paper is to study such quantities defined on convex hulls of function classes. In section 2, we provide a bound on the continuity modulus of a Gaussian process on the convex hull of a class in terms of the continuity modulus on the class itself.

Then, in section 3, we combine this result with some new bounds on the generalization error in function learning problems based on localized Rademacher complexities. This allows us to bound the generalization error in a convex hull in terms of characteristics of the base class.

Finally, we use the bounds on continuity moduli on convex hulls to give very simple proofs of some previously known results on the entropy of such classes.

2 Continuity Modulus on Convex Hulls

Let $\mathcal{F}$ be a subset of a Hilbert space $\mathcal{H}$ and W denote an isonormal Gaussian process defined on $\mathcal{H}$, that is a collection $(W(h))_{h\in\mathcal{H}}$ of Gaussian random variables indexed by $\mathcal{H}$ such that

$$\forall h \in \mathcal{H},\ \mathbb{E}\left[W(h)\right] = 0 \quad \text{and} \quad \forall h, h' \in \mathcal{H},\ \mathbb{E}\left[W(h)W(h')\right] = \langle h, h' \rangle_{\mathcal{H}} \ .$$

We define the modulus of continuity of the process W as

$$\omega(\mathcal{F}, \delta) := \omega_{\mathcal{H}}(\mathcal{F}, \delta) = \mathbb{E}\left[\sup_{\substack{f,g\in\mathcal{F} \\ \|f-g\|\leq\delta}} |W(f) - W(g)|\right] .$$

Since we are in a Hilbert space, we use the natural metric induced by the inner product $d(f,g) = \|f-g\| = \sqrt{\langle f-g, f-g\rangle}$ to define balls in $\mathcal{H}$. Let $\mathcal{F}_\varepsilon$ denote a minimal ε-net of $\mathcal{F}$, i.e. a subset of $\mathcal{F}$ of minimal cardinality such that $\mathcal{F}$ is contained in the union of the balls of radius ε with centers in $\mathcal{F}_\varepsilon$. Let $\mathcal{F}^\varepsilon$ denote a maximal ε-separated subset of $\mathcal{F}$, i.e. a subset of $\mathcal{F}$ of maximal cardinality such that the distance between any two points in this subset is larger than or equal to ε. The ε-covering number of $\mathcal{F}$ is then defined as

$$N(\mathcal{F}, \varepsilon) := N_{\mathcal{H}}(\mathcal{F}, \varepsilon, d) = |\mathcal{F}_\varepsilon| \, ,$$

and the ε-entropy is defined as $H(\mathcal{F}, \varepsilon) = \log N(\mathcal{F}, \varepsilon)$ (this quantity is usually referred to as the metric entropy of $\mathcal{F}$).

In the remainder, K will denote a non-negative constant. Its value may change from one line to another.

2.1 Main Result

Our main result relates the continuity modulus of an isonormal Gaussian process defined on the convex hull of a set $\mathcal{F}$ to the continuity modulus of the same process on this set.

Theorem 1. *We have for all* $\delta \geq 0$

$$\omega(conv(\mathcal{F}), \delta) \leq \inf_{\varepsilon} \left(2\omega(\mathcal{F}, \varepsilon) + \delta \sqrt{N(\mathcal{F}, \varepsilon)} \right) .$$

Proof. Let $\varepsilon > 0$, and let L be the linear span of $\mathcal{F}_\varepsilon$. We denote by Π_L the orthogonal projection on L. We have for all $f \in \mathcal{F}$,

$$f = \Pi_L(f) + \Pi_{L^\perp}(f) ,$$

so that

$$\omega(conv(\mathcal{F}), \delta) \leq \mathbb{E}\left[\sup_{\substack{f,g \in conv(\mathcal{F}) \\ \|f-g\| \leq \delta}} |W(\Pi_L f) - W(\Pi_L g)| \right] + \mathbb{E}\left[\sup_{\substack{f,g \in conv(\mathcal{F}) \\ \|f-g\| \leq \delta}} |W(\Pi_{L^\perp} f) - W(\Pi_{L^\perp} g)| \right] .$$

Now since for any orthogonal projection Π, $\|\Pi(f) - \Pi(g)\| \leq \|f - g\|$ we have

$$\omega(conv(\mathcal{F}), \delta) \leq \omega(\Pi_L conv(\mathcal{F}), \delta) + \omega(\Pi_{L^\perp} conv(\mathcal{F}), \delta) . \tag{1}$$

We will upper bound both terms in the right hand side separately. For the first term, since $\Pi_L(conv(\mathcal{F})) \subset L$, we have

$$\omega(\Pi_L conv(\mathcal{F}), \delta) \leq \omega(L, \delta) ,$$

and by linearity of W and the fact that L is a vector space,

$$\omega(L, \delta) = \mathbb{E}\left[\sup_{\substack{f \in L \\ \|f\| \leq \delta}} |W(f)| \right] \leq \delta \mathbb{E}\left[\sup_{\substack{\|y\|_{\mathbb{R}^d} \leq 1 \\ y \in \mathbb{R}^d}} \langle Z, y \rangle \right] ,$$

where Z is a standard normal vector in $\mathbb{R}^d$ (with $d = \dim L$ and $\|\cdot\|_{\mathbb{R}^d}$ the euclidean norm in $\mathbb{R}^d$). This gives

$$\omega(L, \delta) \leq \delta \mathbb{E}\left[\|Z\|_{\mathbb{R}^d} \right] \leq \delta \sqrt{\mathbb{E}\left[\|Z\|^2_{\mathbb{R}^d} \right]} \leq \delta \sqrt{d} .$$

Since L is the linear span of $\mathcal{F}_\varepsilon$ which is a finite set of size $N(\mathcal{F}, \varepsilon)$ we have $d \leq N(\mathcal{F}, \varepsilon)$ so that

$$\omega(L, \delta) \leq \delta \sqrt{N(\mathcal{F}, \varepsilon)} .$$

Now let's consider the second term of (1). We crudely upper bound it as follows

$$\omega(\Pi_{L^\perp} conv(\mathcal{F}), \delta) \leq 2\mathbb{E}\left[\sup_{f \in conv(\mathcal{F})} |W(\Pi_{L^\perp} f)|\right] .$$

Since $\Pi_{L^\perp}$ is linear, the supremum is attained at extreme points of the convex hull which are elements of $\mathcal{F}$, that is

$$\omega(\Pi_{L^\perp} conv(\mathcal{F}), \delta) \leq 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} |W(\Pi_{L^\perp} f)|\right] .$$

Now for each $f \in \mathcal{F}$, let g be the closest point to f in $\mathcal{F}_\varepsilon$. By definition we have $\|f - g\| \leq \varepsilon$ and $g \in L \cap \mathcal{F}$ so that $\Pi_{L^\perp} g = 0$ and thus

$$\omega(\Pi_{L^\perp} conv(\mathcal{F}), \delta) \leq 2\mathbb{E}\left[\sup_{\substack{f,g \in \mathcal{F} \\ \|f-g\| \leq \varepsilon}} |W(\Pi_{L^\perp} f) - W(\Pi_{L^\perp} g)|\right] .$$

Now since $\Pi_{L^\perp}$ is a contraction, using Slepian's lemma (see [15], Theorem 3.15 page 78) we get

$$\omega(\Pi_{L^\perp} conv(\mathcal{F}), \delta) \leq 2\mathbb{E}\left[\sup_{\substack{f,g \in \mathcal{F} \\ \|f-g\| \leq \varepsilon}} |W(f) - W(g)|\right] = 2\omega(\mathcal{F}, \varepsilon) .$$

This concludes the proof. □

Note that Theorem 1 allows us to give a positive answer to a question raised by Giné and Zinn in [12]. Indeed, we can prove that the convex hull of a uniformly Donsker class is uniformly Donsker. Due to lack of space, we do not give the details here.

2.2 Examples

As an application of Theorem 1, we will derive bounds on the continuity modulus of convex hulls of classes for which we know the rate of growth of the metric entropy.

Let's first recall a well-known relationship between the modulus of continuity of a Gaussian process defined on a class and the metric entropy of that class[1]. By Dudley's entropy bound (see [15], Theorem 11.17, page 321) we have

$$\omega(\mathcal{F}, \varepsilon) \leq K \int_0^\varepsilon H^{1/2}(\mathcal{F}, u)\, du .$$

[1] with respect to the natural metric associated to the process, which in the case of the isonormal Gaussian process is simply the metric induced by the inner product.

It is well known (and easy to check by inspection of the proof of the above result) that when the class is δ-separated, the integral can start at δ, so that for a maximal δ-separated subset $\mathcal{F}^\delta$ of $\mathcal{F}$ we have

$$\omega(\mathcal{F}^\delta, \varepsilon) \leq K \int_\delta^\varepsilon H^{1/2}(\mathcal{F}^\delta, u)\, du\,,$$

for all $\varepsilon > \delta$.

We first consider the case when the entropy of the base class grows logarithmically (e.g. classes of functions with finite Vapnik-Chervonenkis dimension V).

Example 1. If for all $\varepsilon > 0$,

$$N(\mathcal{F}, \varepsilon) \leq K\varepsilon^{-V}\,,$$

then for all $\delta > 0$,

$$\omega(conv(\mathcal{F}), \delta) \leq K\delta^{2/(2+V)} \log^{V/(2+V)} \delta^{-1}\,.$$

Proof. We have from Theorem 1,

$$\begin{aligned}\omega(conv(\mathcal{F}), \delta) &\leq \inf_\varepsilon \left(K \int_0^\varepsilon \log^{1/2} u^{-1} du + \delta\varepsilon^{-V/2} \right) \\ &\leq \inf_\varepsilon \left(K\varepsilon \log^{1/2} \varepsilon^{-1} + \delta\varepsilon^{-V/2} \right)\,.\end{aligned}$$

Choosing

$$\varepsilon = \delta^{2V/(2+V)} \log^{2V/(2+V)} \delta^{-1}\,,$$

we obtain for $\delta \leq 1$,

$$\omega(conv(\mathcal{F}), \delta) \leq K\delta^{2/(2+V)} \log^{V/(2+V)} \delta^{-1}\,.$$

□

Although the main term in the above bound is correct, we obtain a superfluous logarithm. This logarithm can be removed if one uses directly the entropy integral in combination with results on the entropy of the convex hull of such classes [1, 21, 19]. At the moment of this writing, we do not know a simple proof of this fact that does not rely upon the bounds on the entropy of convex hulls.

Now we consider the case when the entropy of the base class has polynomial growth. In this case, we shall distinguish several situations: when the exponent is larger than 2, the class is no longer pre-Gaussian which means that the continuity modulus is unbounded. However, it is possible to study the continuity modulus of a restricted class. Here we consider the convex hull of a δ-separated subset of the base class, for which the continuity modulus is bounded when computed at a scale proportional to δ.

Example 2. If for all $\varepsilon > 0$,

$$H(\mathcal{F}, \varepsilon) \leq K\varepsilon^{-V} ,$$

then for all $\delta > 0$, for $0 < V < 2$,

$$\omega(conv(\mathcal{F}), \delta) \leq K \log^{1/2-1/V} \delta^{-1} ,$$

for $V = 2$,

$$\omega(conv(\mathcal{F}^{\delta/4}), \delta) \leq K \log \delta^{-1} ,$$

and for $V > 2$,

$$\omega(conv(\mathcal{F}^{\delta/4}), \delta) \leq K\delta^{1-V/2} .$$

Proof. We have from Theorem 1, for $\varepsilon > \delta/4$,

$$\omega(conv(\mathcal{F}^{\delta/4}), \delta) \leq \inf_{\varepsilon} \left(K \int_{\delta/4}^{\varepsilon} u^{-V/2} du + \delta \exp(K\varepsilon^{-V}/2) \right) .$$

For $0 < V < 2$, this gives

$$\omega(conv(\mathcal{F}), \delta) \leq \inf_{\varepsilon} \left(K\varepsilon^{(2-V)/2} + \delta \exp(K\varepsilon^{-V}/2) \right) .$$

Choosing

$$\varepsilon = K^{1/V} \log^{-1/V} \delta^{-1} ,$$

we obtain for δ small enough

$$\omega(conv(\mathcal{F}), \delta) \leq K \log^{(V-2)/2V} \delta^{-1} .$$

For $V = 2$, we get

$$\omega(conv(\mathcal{F}^{\delta/4}), \delta) \leq \inf_{\varepsilon} \left(K \log \frac{4\varepsilon}{\delta} + \delta \exp(K\varepsilon^{-2}/2) \right) .$$

Taking $\varepsilon = 1/4$ we get for δ small enough

$$\omega(conv(\mathcal{F}^{\delta/4}), \delta) \leq K \log \delta^{-1} .$$

For $V > 2$, we get

$$\omega(conv(\mathcal{F}^{\delta/4}), \delta) \leq \inf_{\varepsilon} \left(K\delta^{(2-V)/2} - \varepsilon^{(2-V)/2} + \delta \exp(K\varepsilon^{-2}/2) \right) .$$

Taking $\varepsilon \to \infty$, we obtain

$$\omega(conv(\mathcal{F}^{\delta/4}), \delta) \leq K\delta^{(2-V)/2} .$$

□

3 Generalization Error Bounds

3.1 Results

We begin this section with a general bound that relates the error of the function minimizing the empirical risk to a local measure of complexity of the class which is the same in spirit as the bound in [13].

Let $(S, \mathcal{A})$ be a measurable space and let $X_1, \dots, X_n$ be n i.i.d. random variables in this space with common distribution P. P_n will denote the empirical measure based on the sample

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} .$$

In what follows, we choose[2] $\mathcal{H} = L_2(P_n)$ and we are using the notations of Section 2.

We consider a class $\mathcal{F}$ of measurable functions defined on S with values in $[0, 1]$. We assume in what follows that $\mathcal{F}$ also satisfies standard measurability conditions used in the theory of empirical processes as in [10, 21].

We define

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) ,$$

and let ψ_n be an increasing concave (possibly data-dependent random) function with $\psi_n(0) = 0$ such that

$$\mathbb{E}_\varepsilon \left[\sup_{P_n f \le r} |R_n(f)| \right] \le \psi_n(\sqrt{r}), \ \forall r \ge 0 .$$

Let $\hat{r}_n$ be the largest solution of the equation

$$r = \psi_n(\sqrt{r}) . \tag{2}$$

The solution $\hat{r}_n$ of (2) gives what is usually called zero error rate for the class $\mathcal{F}$[13], i.e. the bound for Pf given that $P_n f = 0$.

The bounds we obtain below are data-dependent and they do not require any structural assumptions on the class (such as VC conditions or entropy conditions). Note that $\hat{r}_n$ is determined only by the restriction of the class $\mathcal{F}$ to the sample $(X_1, \dots, X_n)$.

Theorem 2. *If ψ_n is a non-decreasing concave function and $\psi_n(0) = 0$ then there exists $K > 0$ such that with probability at least $1 - e^{-t}$ for all $f \in \mathcal{F}$*

$$Pf \le K \left(P_n f + \hat{r}_n + \frac{t + \log\log n}{n} \right) . \tag{3}$$

[2] This choice implies that we work with the random metric induced by the training data. In other words, we measure the distance between two functions f and g by $d(f, g) = (\sum (f(X_i) - g(X_i))^2)^{1/2}$.

It is most common to estimate the expectation of Rademacher processes via an entropy integral (Theorem 2.2.4 in [21]):

$$\mathbb{E}_\varepsilon \left[\sup_{P_n f \leq \delta} |R_n(f)| \right] \leq \frac{4\sqrt{3}}{\sqrt{n}} \int_0^{\sqrt{\delta}/2} H^{1/2}(\mathcal{F}, u) du \,,$$

which means one can choose $\psi_n(\delta)$ as the right hand side of the above bound. This approach was used for instance in [13].

If the (empirical L_2) covering numbers of the base class grow polynomially, e.g. the base class is a Vapnik-Chervonenkis class of VC dimension $V > 0$, then

$$N(\mathcal{G}, \varepsilon) \leq K \varepsilon^{-V} \,,$$

and thus using results in [19]

$$\log N(conv(\mathcal{G}), \varepsilon) \leq K \varepsilon^{-2V/(2+V)} \,,$$

so that the entropy integral is upper bounded by $K n^{-1/2} \delta^{1/(2+V)}$ and thus we obtain $\hat{r}_n$ of the order of

$$n^{-\frac{1}{2}\frac{2+V}{1+V}} \,.$$

If the metric entropy is polynomial with exponent $0 < V < 2$, the same reasoning gives $\hat{r}_n$ of the order of

$$n^{-\frac{1}{2}} \log^{1/2-1/V} n \,.$$

These results are optimal in the sense that there are classes with such entropy growth for which they cannot be improved. However, our goal is to avoid using entropies as measures of the complexity of the classes and to rather use localized Rademacher or Gaussian complexities.

We will thus apply the bound of Theorem 2 to the function learning problem in the convex hull of a given class.

Let $\mathcal{G}$ be a class of measurable functions from S into $[0, 1]$. Let $g_0 \in conv(\mathcal{G})$ be an unknown target function. The goal is to learn g_0 based on the data $(X_1, g_0(X_1)), \ldots, (X_n, g_0(X_n))$. We introduce $\hat{g}_n$ defined as

$$\hat{g}_n := \arg \min_{g \in conv(\mathcal{G})} P_n |g - g_0| \,,$$

which in principle can be computed from the data.

We introduce the function $\psi_n(\mathcal{G}, \delta)$ defined as

$$\psi_n(\mathcal{G}, \delta) := \sqrt{\frac{\pi}{2n}} \inf_{\varepsilon > 0} \left(\omega(\mathcal{G}, \varepsilon) + \delta \sqrt{N(\mathcal{G}, \varepsilon)} \right) \,.$$

Corollary 1. *Let $\hat{r}_n(\mathcal{G})$ be the largest solution of the equation*

$$r = \psi_n(\mathcal{G}, \sqrt{r}) \,.$$

Then there exists $K > 0$ such that for all $g_0 \in conv(\mathcal{G})$ the following inequality holds with probability at least $1 - e^{-t}$

$$P|\hat{g}_n - g_0| \leq K \left(\hat{r}_n(\mathcal{G}) + \frac{t + \log\log n}{n} \right) \,.$$

Proof. Let $\mathcal{F} = \{|g - g_0| : g \in conv(\mathcal{G})\}$. Note that $\psi_n(\mathcal{G}, \delta)$ is concave non-decreasing (as the infimum of linear functions) and $\psi_n(\mathcal{G}, 0) = 0$, it can thus be used in Theorem 2. We obtain (using bound (4.8) on page 97 of [15])

$$\begin{aligned}
\mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}\\ P_n f\le r}} |R_n(f)|\right] &\le \sqrt{\frac{\pi}{2n}}\mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}\\ P_n f\le r}} |W_{P_n}(f)|\right] \\
&\le \sqrt{\frac{\pi}{2n}}\mathbb{E}\left[\sup_{\substack{f\in\mathcal{F}\\ (P_n f^2)^{1/2}\le \sqrt{r}}} |W_{P_n}(f)|\right] \\
&\le \sqrt{\frac{\pi}{2n}}\omega(conv\mathcal{G}, \sqrt{r}) \le \psi_n(\mathcal{G}, \sqrt{r})\,,
\end{aligned}$$

where in the last step we used Theorem 1. To complete the proof, it is enough to notice that $P_n|\hat{g}_n - g_0| = 0$ (since $g_0 \in conv(\mathcal{G})$) and to use the bound of Theorem 2. □

A simple application of the above corollary in combination with the bounds of examples 1 and 2 give, for instance, the following rates. If the covering numbers of the base class grow polynomially, i.e.

$$N(\mathcal{G}, \varepsilon) \le K\varepsilon^{-V}\,,$$

we obtain $\hat{r}_n$ of the order of (up to logarithmic factors)

$$n^{-\frac{1}{2}\frac{2+V}{1+V}}\,.$$

If the random metric entropy is polynomial with exponent $0 < V < 2$, $\hat{r}_n$ is of the order of

$$n^{-\frac{1}{2}}\log^{1/2-1/V} n\,.$$

Notice that in both cases we obtain the same results as with a direct application of the entropy bound of the convex hull.

However the bound of corollary 1 contains only terms that can be computed from the data and that depend on the base class. This means that if one actually computes $\psi_n(\mathcal{G}, \delta)$ (without using an entropy upper bound), one can only obtain better results.

3.2 Proof of Theorem 2

STEP 1: Concentration. We define $\delta_k = 2^{-k}$ for $k \ge 0$, and consider a sequence of classes

$$\mathcal{F}_k = \{f \in \mathcal{F} : \delta_{k+1} < Pf \le \delta_k\}\,.$$

If we denote

$$R_k = \mathbb{E}_\varepsilon\left[\sup_{\mathcal{F}_k} |R_n(f)|\right]\,,$$

then the symmetrization inequality implies that

$$\mathbb{E}\left[\sup_{\mathcal{F}_k}|P_nf - Pf|\right] \leq 2\mathbb{E}\left[R_k\right] .$$

Note that for $f \in \mathcal{F}_k$, $P(f - Pf)^2 \leq Pf^2 \leq Pf \leq \delta_k$, so that Theorem 3 in [5] implies that with probability at least $1 - e^{-t}$ for all $f \in \mathcal{F}_k$

$$|P_nf - Pf| \leq 4\mathbb{E}\left[R_k\right] + \left(\frac{2\delta_k t}{n}\right)^{1/2} + \frac{4t}{3n} .$$

Now, Theorem 16 in [4] gives that with probability at least $1 - e^{-t}$

$$\mathbb{E}\left[R_k\right] \leq \frac{2t}{n} + 2R_k .$$

Therefore, with probability at least $1 - 2e^{-t}$ for all $f \in \mathcal{F}_k$

$$|P_nf - Pf| \leq 8R_k + \left(\frac{2\delta_k t}{n}\right)^{1/2} + \frac{10t}{n} . \tag{4}$$

STEP 2: Union Bound. Now we apply an union bound over $k = 0, 1, \ldots$. We define

$$l(\delta) = 2\log\left(\frac{\pi}{\sqrt{3}}\log_2\frac{2}{\delta}\right) .$$

We have

$$\sum_{k\geq 0} e^{-l(\delta_k)} = \frac{\pi^2}{3}\sum_{k\geq 0}\frac{1}{(k+1)^2} = \frac{1}{2} .$$

Therefore, replacing t by $t + l(\delta_k)$ in Inequality (4) and applying the union bound we get that with probability at least $1 - e^{-t}$ for all $k \geq 0$ and for all $f \in \mathcal{F}_k$

$$|P_nf - Pf| \leq 8R_k + \left(\frac{2\delta_k(t + l(\delta_k))}{n}\right)^{1/2} + \frac{10(t + l(\delta_k))}{n} . \tag{5}$$

Now we reason on the event where this inequality holds.

STEP 3: Empirical Complexity. If we denote

$$U_k = \delta_k + 8R_k + \left(\frac{2\delta_k(t + l(\delta_k))}{n}\right)^{1/2} + \frac{10(t + l(\delta_k))}{n} ,$$

then on this event for any k and for all $f \in \mathcal{F}_k$, $P_nf \leq U_k$ so that

$$R_k \leq \mathbb{E}_\varepsilon\left[\sup_{P_nf\leq U_k}|R_n(f)|\right] \leq \psi_n(\sqrt{U_k}) ,$$

and thus

$$U_k \leq \delta_k + 8\psi_n(\sqrt{U_k}) + \left(\frac{2\delta_k(t+l(\delta_k))}{n}\right)^{1/2} + \frac{10(t+l(\delta_k))}{n}\,.$$

Notice that for $k \geq \log_2 n$, we have for $f \in \mathcal{F}_k$, $Pf \leq \delta_k \leq 1/n$ so that the result holds trivially for such functions. We thus have to prove the result for the other values of k. For $k \leq \log_2 n$, $\delta_k \geq 1/n$ so that we have

$$U_k \leq \delta_k + 8\psi_n(\sqrt{U_k}) + \sqrt{2\delta_k r_0} + 10r_0 \leq 8\psi_n(\sqrt{U_k}) + 2\delta_k + 11r_0\,,$$

with $r_0 = (t + K\log\log n)/n$ for some large enough K (we used the fact that $2\sqrt{ab} \leq a+b$).

STEP 4: Solving. Let's denote by $f(\sqrt{U_k})$ the right hand side of the above inequality. f is a non-decreasing concave function so that the above inequality implies that U_k is upper bounded by the largest solution x^* of $f(\sqrt{x}) = x$. Moreover, any non-negative real number z such that $f(\sqrt{z}) \leq z$ is an upper bound for x^* and thus for U_k. Let's prove that $z = K(\hat{r}_n + \delta_k + r_0)$ satisfies such a condition (for some large enough K).

Since ψ_n is concave and $\psi_n(0) = 0$, we have for $x > 0$, $\psi_n(\sqrt{Kx}) \leq \sqrt{K}\psi_n(\sqrt{x})$. Also, for $x > 0$, $\psi_n(\sqrt{\hat{r}_n + x}) \leq \hat{r}_n + x$ since $\hat{r}_n + x$ upper bounds the solution of $\psi_n(\sqrt{r}) = r$. We thus have

$$f(\sqrt{K(\hat{r_n} + \delta_k + r_0)}) \leq 8\sqrt{K}(\hat{r}_n + \delta_k + r_0) + 2\delta_k + 11r_0\,,$$

and this is less than $K(\hat{r}_n + \delta_k + r_0)$ for a large enough K.

We thus get $U_k \leq K(\hat{r}_n + \delta_k + r_0)$. Finally, (5) implies that for all $k \leq \log_2 n$ and $f \in \mathcal{F}_k$

$$Pf \leq P_n f + 8\psi_n(\sqrt{K(\hat{r}_n + \delta_k + r_0)}) + \sqrt{2\delta_k r_0} + 10r_0\,.$$

If $f \in \mathcal{F}_k$ then $\delta_k \leq 2Pf$, which proves

$$Pf \leq P_n f + 8\psi_n(\sqrt{K(\hat{r}_n + 2Pf + r_0)}) + \sqrt{4r_0 Pf} + 10r_0\,.$$

Now the same reasoning as above proves that there exists a constant K such that Pf is upper bounded by $K(P_n f + \hat{r}_n + r_0)$ which concludes the proof. □

4 Entropy of Convex Hulls

4.1 Relating Entropy With Continuity Modulus

By Sudakov's minoration (see [15], Theorem 3.18, page 80) we have

$$\sup_{\varepsilon>0} \varepsilon H^{1/2}(\mathcal{F}, \varepsilon) \leq K\mathbb{E}\left[\sup_{f\in\mathcal{F}} |W(f)|\right]\,.$$

Let $B(f,\delta)$ be the ball centered in f of radius δ. We define

$$H(\mathcal{F},\delta,\varepsilon) := \sup_{f\in\mathcal{F}} H(B(f,\delta)\cap\mathcal{F},\varepsilon)\,.$$

The following lemma relates the entropy of $\mathcal{F}$ with the modulus of continuity of the process W. This type of bound is well known (see e.g. [17]) but we give the proof for completeness.

Lemma 1. *Assume $\mathcal{F}$ is of diameter 1. For all integer k we have*

$$H^{1/2}(\mathcal{F},2^{-k}) \le K\sum_{i=0}^{k} 2^i\omega(\mathcal{F},2^{1-i})\,.$$

This can also be written

$$H^{1/2}(\mathcal{F},\delta) \le K\int_{\delta}^{1} u^{-2}\omega(\mathcal{F},u)\,du\,.$$

Proof. We have

$$\begin{aligned}\omega(\mathcal{F},\delta) &= \mathbb{E}\left[\sup_{\substack{f,g\in\mathcal{F}\\ \|f-g\|\le\delta}} |W(f)-W(g)|\right]\\ &\ge \sup_{f\in\mathcal{F}}\mathbb{E}\left[\sup_{g\in B(f,\delta)\cap\mathcal{F}} |W(f)-W(g)|\right]\\ &\ge \sup_{f\in\mathcal{F}}\sup_{\varepsilon>0}\varepsilon H^{1/2}(B(f,\delta)\cap\mathcal{F},\varepsilon)\,,\end{aligned}$$

so that we obtain

$$\frac{\delta}{2}H^{1/2}(\mathcal{F},\delta,\frac{\delta}{2}) \le K\omega(\mathcal{F},\delta)\,.$$

Notice that we can construct a 2^{-k} covering of $\mathcal{F}$ by covering $\mathcal{F}$ by $N(\mathcal{F},1)$ balls of radius 1 and then covering the intersection of each of these balls with $\mathcal{F}$ with $N(B(f,1)\cap\mathcal{F},1/2)$ balls of radius $1/2$ and so on. We thus have

$$N(\mathcal{F},2^{-k}) \le \prod_{i=0}^{k}\sup_{f\in\mathcal{F}} N(B(f,2^{1-i})\cap\mathcal{F},2^{-i})\,.$$

Hence

$$H(\mathcal{F},2^{-k}) \le \sum_{i=0}^{k} H(\mathcal{F},2^{1-i},2^{-i})\,.$$

We thus have

$$H^{1/2}(\mathcal{F},2^{-k}) \le \sum_{i=0}^{k} H^{1/2}(\mathcal{F},2^{1-i},2^{-i}) \le K\sum_{i=0}^{k} 2^i\omega(\mathcal{F},2^{1-i})\,,$$

which concludes the proof. □

Next we present a modification of the previous lemma that can be applied to δ-separated subsets.

Lemma 2. *Assume $\mathcal{F}$ is of diameter 1. For all integer k we have*

$$H^{1/2}(\mathcal{F}, 2^{-k}) \leq K \sum_{i=0}^{k} 2^i \omega(\mathcal{F}^{2^{-i-1}}, 2^{2-i}) .$$

Proof. Notice that for $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}^{\delta/4}$ such that

$$B(f, \delta) \cap \mathcal{F} \subset B(f', \delta + \delta/4) \cap \mathcal{F} .$$

Moreover, since a maximal δ-separated set is a δ-net,

$$N(\mathcal{F}, \delta) \leq \left| N^\delta \right| = N(\mathcal{F}^\delta, \delta/2) ,$$

since for a δ-separated set A we have $N(A, \delta/2) = |A|$.

Let's prove that we have for any γ,

$$\left| (B(f, \gamma) \cup \mathcal{F})^{\delta/2} \right| \leq \left| B(f, \gamma + \delta/4) \cup \mathcal{F}^{\delta/4} \right| .$$

Indeed, since the points in $\mathcal{F}^{\delta/4}$ form a $\delta/4$ cover of $\mathcal{F}$, all the points in $(B(f, \gamma) \cup \mathcal{F})^{\delta/2}$ are at distance less than $\delta/4$ of one and only one point of $\mathcal{F}^{\delta/4}$ (the unicity comes from the fact that they are $\delta/2$ separated). We can thus establish an injection from points in $(B(f, \gamma) \cup \mathcal{F})^{\delta/2}$ to corresponding points in $\mathcal{F}^{\delta/4}$ and the image of this injection is included in $B(f, \gamma + \delta/4)$ since the image points are within distance $\delta/4$ of points in $B(f, \gamma)$.

Now we obtain

$$N((B(f', \delta + \delta/4) \cup \mathcal{F})^{\delta/2}, \delta/4) \leq N(B(f', 3\delta/2) \cup \mathcal{F}^{\delta/4}, \delta/8) .$$

We thus have

$$\begin{aligned} N(B(f, \delta) \cup \mathcal{F}, \delta/2) &\leq N(B(f', \delta + \delta/4) \cup \mathcal{F}, \delta/2) \\ &\leq N((B(f', \delta + \delta/4) \cup \mathcal{F})^{\delta/2}, \delta/4) \\ &\leq N(B(f', 3\delta/2) \cup \mathcal{F}^{\delta/4}, \delta/8) . \end{aligned}$$

This gives

$$\begin{aligned} \sup_{f \in \mathcal{F}} N(B(f, \delta) \cap \mathcal{F}, \delta/2) &\leq \sup_{f \in \mathcal{F}^{\delta/4}} N(B(f, 3\delta/2) \cap \mathcal{F}^{\delta/4}, \delta/8) \\ &= N(\mathcal{F}^{\delta/4}, 3\delta/2, \delta/8) . \end{aligned}$$

Hence

$$H(\mathcal{F}, \delta, \delta/2) \leq H(\mathcal{F}^{\delta/4}, 3\delta/2, \delta/8) .$$

By the same argument as in previous Lemma we obtain

$$\frac{\delta}{8} H^{1/2}(\mathcal{F}^{\delta/4}, 3\delta/2, \delta/8) \leq K \omega(\mathcal{F}^{\delta/4}, 3\delta/2) .$$

□

4.2 Applications

Example 3. If for all $\varepsilon > 0$,

$$N(\mathcal{F}, \varepsilon) \leq \varepsilon^{-V} ,$$

then for all $\varepsilon > 0$,

$$H(conv(\mathcal{F}), \varepsilon) \leq \varepsilon^{-2V/(2+V)} \log^{2V/(2+V)} \varepsilon^{-1} .$$

Proof. Recall from Example 1 that

$$\omega(conv(\mathcal{F}), \delta) \leq K \delta^{2/(2+V)} \log^{V/(2+V)} \delta^{-1} .$$

Now, using Lemma 1 we get

$$\begin{aligned} H^{1/2}(conv(\mathcal{F}), 2^{-k}) &\leq K \sum_{i=0}^{k} 2^i 2^{2(1-i)/(2+V)} (i-1)^{V/(2+V)} \\ &= K \sum_{i=0}^{k} (2^{V/(2+V)})^i (i-1)^{V/(2+V)} . \end{aligned}$$

We check that in the above sum, the i-th term is always larger than twice the $i-1$-th term (for $i \geq 2$) so that we can upper bound the sum by the last term,

$$H^{1/2}(\mathcal{F}, 2^{-k}) \leq K (2^{V/(2+V)})^k (k-1)^{V/(2+V)} ,$$

hence, using $\varepsilon = 2^{-k}$, we get the result. □

Note that the result we obtain contains an extra logarithmic factor compared to the optimal bound [21, 19].

Example 4. If for all $\varepsilon > 0$,

$$H(\mathcal{F}, \varepsilon) \leq \varepsilon^{-V} ,$$

then for all $\varepsilon > 0$, for $0 < V < 2$,

$$H(conv(\mathcal{F}), \varepsilon) \leq \varepsilon^{-2} \log^{1-V/2} \varepsilon^{-1} ,$$

for $V = 2$,

$$H(conv(\mathcal{F}), \varepsilon) \leq \varepsilon^{-2} \log^{2} \varepsilon^{-1} ,$$

and for $V > 2$,

$$H(conv(\mathcal{F}), \varepsilon) \leq \varepsilon^{-V} .$$

Proof. The proof is similar to the previous one. □

In this example, all the bounds are known to be sharp [7, 11].

References

1. K. Ball and A. Pajor. The entropy of convex bodies with "few" extreme points. *London MAth. Soc. Lectrure Note Ser.* 158, pages 25–32, 1990.
2. P. Bartlett, O. Bousquet and S. Mendelson. Localized Rademacher Complexity. *Preprint*, 2002.
3. P. Bartlett, S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. In *Proceeding of the 14th Annual Conference on Computational Learning Theory*, Srpinger, 2001.
4. S. Boucheron, G. Lugosi and P. Massart. Concentration inequalities using the entropy method. *Preprint*, 2002.
5. O. Bousquet. A Bennett concentration inequality and its application to empirical processes. *C. R. Acad. Sci. Paris, Ser. I* 334, pages 495-500, 2002.
6. B. Carl. Metric entropy of convex hulls in Hilbert spaces. *Bulletin of the London Mathematical Society*, 29, pages 452-458, 1997.
7. B. Carl, I. Kyrezi and A. Pajor. Metric entropy of convex hulls in Banach spaces. *Journal of the London Mathematical Society*, 2001.
8. J. Creutzig and I. Steinwart. Metric entropy of convex hulls in type p spaces – the critical case. 2001.
9. R. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 15, pages 1306-1326, 1987.
10. R. Dudley. Uniform central limit theorems. Cambridge University Press, 2000.
11. F. Gao. Metric entropy of convex hulls. *Israel Journal of Mathematics*, 123, pages 359-364, 2001.
12. E. Giné and J. Zinn. Gaussian characterization of uniform Donsker classes of functions. *Annals of Probability*, 19, pages 758-782, 1991.
13. V. I. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, Eds. E.Gine, D.Mason and J.Wellner, pp. 443-459, 2000.
14. V. I. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1), 2002.
15. M. Ledoux and M. Talagrand Probability in Banach spaces. Springer-Verlag, 1991.
16. W. Li and W. Linde. Metric entropy of convex hulls in Hilbert spaces. *Preprint*, 2001.
17. M. Lifshits. Gaussian random functions. Kluwer, 1995.
18. P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX, pages 245-303, 2000.
19. S. Mendelson. On the size of convex hulls of small sets. *Preprint*, 2001.
20. S. Mendelson. Improving the sample complexity using global data. *Preprint*, 2001.
21. A. van der Vaart and J. Wellner. Weak convergence and empirical processes with applications to statistics. John Wiley & Sons, New York, 1996.