Agnostic Learning Nonconvex Function Classes

Shahar Mendelson and Robert C. Williamson

Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200
Australia
shahar.mendelson@anu.edu.au, Bob.Williamson@anu.edu.au

Abstract. We consider the sample complexity of agnostic learning with respect to squared loss. It is known that if the function class F used for learning is convex then one can obtain better sample complexity bounds than usual. It was also previously claimed that there is a lower bound that showed there was an essential gap in the rate. In this paper we show that the lower bound argument is flawed and that one can get "fast" sample complexity bounds for nonconvex F. The new bounds depend on the detailed geometry of F, in particular the distance in a certain sense of the target's conditional expectation from the set of nonuniqueness points of the class F.

1 Introduction

The agnostic learning model [5] is a generalization of the PAC learning model that does not presume the target function lies within the space of functions (hypotheses) used for learning. There are now a number of results concerning the sample complexity of agnostic learning, especially with respect to the squared loss functional. In particular, in [8] it was shown that if ε is the required accuracy, then the sample complexity (ignoring log factors and the confidence terms) of agnostic learning from a closed class of functions F with squared loss is $O(d/\varepsilon)$ if F is convex, where d is an appropriate complexity parameter (e.g. the empirical metric entropy of the class). This result was extended and improved in [9].

It was claimed in [8] that if F is not convex, there exists a lower bound of $\Omega(1/\varepsilon^2)$ on the sample complexity. Thus, whether or not F is convex seemed important for the sample complexity of agnostic learning with squared loss.

However, these are deceptive results. The claimed lower bound relies on a random construction and the fact that for nonconvex F, one can always find a target "function" (actually a target conditional expectation) f^* which has two best approximations in the class F. Unfortunately, as we show here, the random construction is wrong. It *is* the case though that sample complexity of agnostic learning does depend on the closeness of f^* to a point with a nonunique best approximation. In this article we will develop some *nonuniform* results which hold for "most" target conditional expectations in the agnostic learning scenario from a nonconvex class F and obtain sharper sample complexity upper bounds.

The proof we present here is based on recently developed methods which can be used for complexity estimates. It was shown in [9] that the complexity of a learning problem can be governed by two properties. The first is the Rademacher complexity of the class, which is a parameter that indicates "how large" the class is (see [10, 1]). The other property is the ability to control the mean square value of each loss function using its expectation. We will show that indeed the mean square value can be bounded in terms of the expectation as long as as one knows the distance of the target from the set of points which have more than a unique best approximation in the class.

In the next section we present some basic definitions, notation, and some general complexity estimates. Then, we present our nonuniform upper bound. Finally, we briefly present the proof on the lower bound claimed in [8], show where the argument fails, and prove that the claim itself can not be true.

2 Definitions, Notation and Background Results

If $(\mathcal{X}, d)$ is a metric space, and $U \subseteq \mathcal{X}$, then for $\varepsilon > 0$, we say that $C \subseteq \mathcal{X}$ is an *ε-cover* of U with respect to d if for all $v \in U$, there exists $w \in C$ such that $d(v, w) \leq \varepsilon$. The *ε-covering number* with respect to d, $N(\varepsilon, U, d)$, is the cardinality of the smallest ε-cover of U with respect to d. If the metric d is obvious, we will simply say ε-cover etc.

The closed ball centered at c of radius r is denoted by $B(c, r) := \{x \in \mathcal{X} \colon \|x - c\| \leq r\}$. Its boundary is $\partial B(c, r) := \{x \in \mathcal{X} \colon \|x - c\| = r\}$. If $x \in \mathcal{X}$, and $A \subset \mathcal{X}$, let the distance between A and x be defined as $d_A(x) := \inf\{d(x, a) \colon a \in A\}$. The *metric projection* of x onto A is $P_A(x) := \{a \in A \colon \|x - a\| = d_A(x)\}$. Hence, elements of $P_A(x)$ are all *best approximations* of x in A.

Denote by $L_\infty(\mathcal{X})$ the space of bounded functions on $\mathcal{X}$ with respect to the sup norm and set $B\big(L_\infty(\mathcal{X})\big)$ to be its unit ball. Let μ be a probability measure on $\mathcal{X}$ and put $L_2(\mu)$ to be the Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ with the norm endowed by the inner product $\langle f, g\rangle = \int f(x)g(x)\mathrm{d}\mu(x)$. Let $\mathcal{Y} \subset [-1, 1]$, and set F to be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$, and thus a subset of $L_2(\mu)$. Assumptions we will make throughout are that F is a closed subset of $L_2(\mu)$ and that it satisfies a measurability condition called "admissibility" (see [3, 4, 14]) for details.

Definition 1. *Let $F \subset L_2(\mu)$. A point $f \in L_2(\mu)$ is said to be a nup point (nonunique projection) of F with respect to (w.r.t.) $L_2(\mu)$ if it has two or more best approximations in F with respect to the $L_2(\mu)$ norm. Define*

$$\mathrm{nup}(F, \mu) := \{f \in L_2(\mu) : f \text{ is a nup point of } F \text{ w.r.t. } L_2(\mu)\}.$$

It is possible to show that in order to solve the agnostic learning problem of approximating a random variable Y with values in $\mathcal{Y}$ by elements in F, it suffices to learn the function $f^* = \mathbb{E}(Y|X = x)$. Indeed, for every $f \in F$,

$$\begin{aligned}\mathbb{E}\big(f(X) - Y\big)^2 &= \mathbb{E}\big(\mathbb{E}(Y|X) - f(X)\big)^2 + \mathbb{E}\big(\mathbb{E}(Y|X) - Y\big)^2 \\ &= \mathbb{E}\big(f^*(X) - f(X)\big)^2 + \mathbb{E}\big(f^*(X) - Y\big)^2.\end{aligned}$$

Thus, a minimizer of the distance between $f(X)$ and Y will depend only on finding a minimizer for $\mathbb{E}\big(f^*(X)-f(X)\big)^2$, that is, solving the function learning problem of approximating f^* by members of F with respect to the $L_2(\mu)$ norm.

Assume that we have fixed the target f^*. We denote by f_a its best approximation in F with respect to the given $L_2(\mu)$ norm. For any function $f \in F$, let the squared loss function associated with f^* and f be

$$g_{f,f^*} : x \mapsto (f(x)-f^*(x))^2 - (f_a(x)-f^*(x))^2,$$

and set $\mathcal{L}(f^*) = \mathcal{L} := \{g_{f,f^*} : f \in F\}$.

Interestingly, although a "randomly chosen" $f^* \in L_2(\mu)$ is unlikely[1] to be in $\operatorname{nup}(F,\mu)$, as we shall see below, $\operatorname{nup}(F,\mu)$ nevertheless controls the sample complexity of learning f^* for all $f^* \in L_2(\mu) \setminus F$.

Definition 2. *For any set $\{x_1,\dots,x_n\} \subset \mathcal{X}$, let μ_n be the empirical measure supported on the set; i.e. $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Given a class of functions F, a random variable Y taking values in $\mathcal{Y}$, and parameters $0<\varepsilon,\delta<1$ let $C_F(\varepsilon,\delta,Y)$ be the smallest integer such that for any probability measure μ*

$$\Pr\{\exists g_{f,f^*} \in \mathcal{L}(f^*) : \mathbb{E}_{\mu_n} g_{f,f^*} < \varepsilon, \mathbb{E}_\mu g_{f,f^*} \geq 2\varepsilon\} < \delta, \tag{1}$$

where $f^* = \mathbb{E}(Y|X=x)$.

The quantity $C_F(\varepsilon,\delta,Y)$ is known as the *sample complexity of learning a target Y with the function class F*. The definition means that if one draws a sample of size greater than $C_F(\varepsilon,\delta,Y)$ then with probability greater than $1-\delta$, if one "almost minimizes" the empirical loss (less than ε) then the expected loss will not be greater than 2ε. Typically, the sample complexity of a class is defined as the "worst" sample complexity when going over all possible selections of Y.

Recent results have yielded good estimates on the probability of the set in (1). These estimates are based on the Rademacher averages as a way of measuring the complexity of a class of functions. The averages are better suited to proving sample complexity results than classical techniques using the union bound over an ε-cover, mainly because of the "functional Bennett inequality" due to Talagrand [13].

Definition 3. *Let μ be a probability measure on $\mathcal{X}$ and suppose F is a class of uniformly bounded functions. For every integer n, set*

$$R_n(F) := \mathbb{E}_\mu \mathbb{E}_\varepsilon \frac{1}{\sqrt{n}} \sup_{f\in F} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|$$

where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ and $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables.

[1] Since Hilbert spaces are uniformly convex it follows from a theorem of Stechkin [12] (see [16, page 9]) that $L_2(\mu) \setminus \operatorname{nup}(F,\mu)$ is a countable intersection of open dense sets. This implies that if one puts a reasonable probability measure ν on $L_2(\mu)$, then $\nu(\{f \in L_2(\mu) : f \notin \operatorname{nup}(F,\mu)\}) = 1$.

Various relationships between Rademacher averages and classical measures of complexity are shown in [10, 11]. It turns out that the best sample complexity bounds to date are in terms of *local* Rademacher averages. Before presenting these bounds, we require the next definition.

Definition 4. *We say that $F \subset L_2(\mu)$ is star-shaped with centre f if for every $g \in F$, the interval $[f,g] = \{tf + (1-t)g : 0 \leq t \leq 1\} \subset F$. Given F and f, let*

$$\operatorname{star}(F,f) := \bigcup_{g \in F} [f,g].$$

Theorem 1. *Let $F \subset B\big(L_\infty(\mathcal{X})\big)$, fix some f^* bounded by 1 and set $\mathcal{L}(f^*)$ to be the squared loss class associated with F and f^*. Assume that there is a constant B such that for every $g \in \mathcal{L}(f^*)$, $\mathbb{E}g^2 \leq B\mathbb{E}g$.*

Let $\mathcal{G} := \operatorname{star}(\mathcal{L}, 0)$ and for every $\varepsilon > 0$ set $\mathcal{G}_\varepsilon = \mathcal{G} \cap \{h : \mathbb{E}h^2 \leq \varepsilon\}$. Then for every $0 < \varepsilon, \delta < 1$,

$$\Pr\left\{\exists g \in \mathcal{L}, \mathbb{E}_{\mu_n} g \leq \varepsilon/2, \mathbb{E}g \geq \varepsilon\right\} \leq \delta$$

provided that

$$n \geq C \max\left\{\frac{R_n^2(\mathcal{G}_\varepsilon)}{\varepsilon^2}, \frac{B\log\frac{2}{\delta}}{\varepsilon}\right\},$$

where C is an absolute constant.

Using this result one can determine an upper bound on the sample complexity in various cases. The one we present here is a bound in terms of the metric entropy of the class.

Theorem 2 ([11]). *Let Y be a random variable on $\mathcal{Y}$ and put $f^* = \mathbb{E}(Y|X=x)$. Let F,$\mathcal{L}$, $\mathcal{G}$ and B be as in theorem 1.*

1. *If there are $\gamma, p, d \geq 1$ such that for every $\varepsilon > 0$,*

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, F, L_2(\mu_n)) < d\log^p\left(\frac{\gamma}{\varepsilon}\right),$$

then for every $0 < \varepsilon, \delta < 1$,

$$C(\varepsilon, \delta, Y) \leq \frac{C_{p,\gamma}}{\varepsilon} \max\Big\{d\log^p\frac{1}{\varepsilon}, B\log\frac{2}{\delta}\Big\},$$

where $C_{p,\gamma}$ depends only on p and γ.

2. *If there are $0 < p < 2$ and $\gamma \geq 1$ such that for every $\varepsilon > 0$,*

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, F, L_2(\mu_n)) < \gamma\varepsilon^{-p}$$

then

$$C(\varepsilon, \delta, Y) \leq C_{p,\gamma} \max\Big\{\big(\frac{1}{\varepsilon}\big)^{1+\frac{p}{2}}, B\log\frac{2}{\delta}\Big\},$$

where $C_{p,\gamma}$ depends only on p and γ.

From this result it follows that if the original class F is "small enough", one can establish good generalization bounds, if, of course, the mean-square value of each member of the loss class can be uniformly controlled by its expectation. This is trivially the case in the proper learning scenario, since each loss function is nonnegative. It was known to be true if F is convex in the squared loss case [8] and was later extended in the more general case of p-loss classes for $2 \leq p < \infty$ [9].

Our aim is to investigate this condition and to see what assumptions must be imposed on f^* to ensure such a uniform control of the mean square value in terms of the expectation.

3 Nonuniform Agnostic Learnability of Nonconvex Classes

We will now study agnostic learning using nonconvex hypothesis classes. The key observation is that whilst in the absence of convexity one can not control $\mathbb{E}[g_{f,f^*}^2]$ in terms of $\mathbb{E}[g_{f,f^*}]$ *uniformly* in f^*, one can control it *nonuniformly* in f^* by exploiting the geometry of F. The main result is corollary 1.

The following result is a generalization of [7, lemma 14] (cf. [6, lemma A.12]).

Lemma 1. *Let F be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$. Put $\alpha \in [0,1)$, set $f^* \in L_2(\mu)$ and suppose f^* has range contained in $[0,1]$. If for every $f \in F$*

$$\langle f_a - f^*, f_a - f \rangle \leq \frac{\alpha}{2} \|f_a - f\|^2, \tag{2}$$

then for every $g_{f,f^*} \in \mathcal{L}(f^*)$,

$$\mathbb{E}[g_{f,f^*}^2] \leq \frac{16}{1-\alpha} \mathbb{E}[g_{f,f^*}].$$

Proof. For the sake of simplicity, we denote each loss function by g_f. Observe that

$$\begin{aligned}
\mathbb{E}[g_f^2] &= \mathbb{E}[((f^*(X) - f(X))^2 - (f^*(X) - f_a(X))^2)^2] \\
&= \mathbb{E}[((2f^*(X) - f(X) - f_a(X))(f_a(X) - f(X)))^2] \\
&\leq 16\mathbb{E}[(f(X) - f_a(X))^2] \\
&= 16\|f_a - f\|^2. \qquad (3)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}[g_f] &= \mathbb{E}[(f^*(X) - f(X))^2 - (f^*(X) - f_a(X))^2] \\
&= \mathbb{E}[(f^*(X) - f_a(X))^2 + (f_a(X) - f(X))^2 \\
&\qquad +2(f^*(X) - f_a(X))(f_a(X) - f(X)) - (f^*(X) - f_a(X))^2] \\
&= \mathbb{E}[(f_a(X) - f(X))^2 + 2(f^*(X) - f_a(X))(f_a(X) - f(X))] \\
&= \mathbb{E}[(f_a(X) - f(X))^2] + 2\mathbb{E}[(f^*(X) - f_a(X))(f_a(X) - f(X))] \\
&= \|f_a - f\|^2 + 2\langle f^* - f_a, f_a - f\rangle \\
&= \|f_a - f\|^2 - 2\langle f_a - f^*, f_a - f\rangle \\
&\geq \|f_a - f\|^2 - \alpha\|f_a - f\|^2 \\
&= (1-\alpha)\|f_a - f\|^2 \\
&= \frac{1-\alpha}{16}\mathbb{E}[g_f^2].
\end{aligned}$$

∎

Lemma 2. *Fix $f^* \in L_2(\mu)$. Then, $f \in L_2(\mu)$ satisfies (2) if and only if f is not contained in*

$$B^{(\alpha)} := B\left(\frac{1}{\alpha}(f^* - f_a) + f_a, \frac{1}{\alpha}\|f^* - f_a\|\right),$$

which is the closed ball in $L_2(\mu)$ centered at $\frac{1}{\alpha}(f^* - f_a) + f_a$ with radius $\frac{1}{\alpha}\|f^* - f_a\|$.

Proof. Note that $\langle f^* - f_a, f - f_a\rangle \leq \frac{\alpha}{2}\|f_a - f\|^2$ if and only if $\|f_a - f\|^2 - \frac{2}{\alpha}\langle f^* - f_a, f - f_a\rangle \geq 0$. Clearly, the latter is equivalent to

$$\begin{aligned}
\langle f_a - f, f_a - f\rangle + \langle\frac{1}{\alpha}(f^* - f_a), \frac{1}{\alpha}(f^* - f_a)\rangle - \langle\frac{1}{\alpha}(f^* - f_a), \frac{1}{\alpha}(f^* - f_a)\rangle \\
+ \frac{2}{\alpha}\langle f^* - f_a, f_a - f\rangle \geq 0.
\end{aligned}$$

Thus,

$$\langle f_a - f + \frac{1}{\alpha}(f^* - f_a), f_a - f + \frac{1}{\alpha}(f^* - f_a)\rangle \geq \langle\frac{1}{\alpha}(f^* - f_a), \frac{1}{\alpha}(f^* - f_a)\rangle. \quad (4)$$

Clearly, f satisfies (4) if and only if $\|f - (f_a + \frac{1}{\alpha}(f^* - f_a))\| \geq \frac{1}{\alpha}\|f^* - f_a\|$; hence it belongs to the region outside of $B^{(\alpha)}$. ∎

In the limit as $\alpha \to 0$, ∂B_α approaches a hyperplane. Then by the unique supporting hyperplane characterization of convex sets [15, theorem 4.1] this implies F is convex.

We will use lemma 2 as indicated in figure 1. The key factor in bounding B is the closeness of f^* to $\mathrm{nup}(F, \mu)$ in a particular sense. Suppose $f^* \in L_2(\mu) \setminus (F \cup \mathrm{nup}(F, \mu))$, and let

$$r_{F,\mu}(f^*) := \inf\{\|f - P_F(f^*)\| : f \in \{\lambda(f^* - P_F(f^*)) : \lambda > 0\} \cap \mathrm{nup}(F, \mu)\}.$$

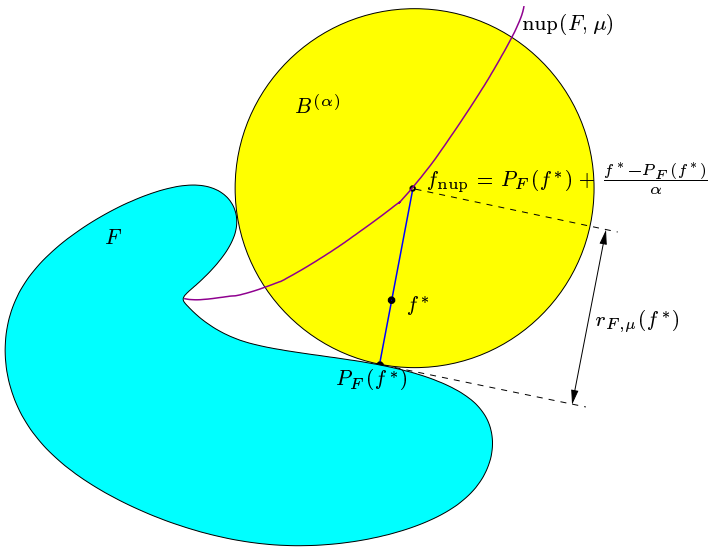

Fig. 1. Illustration of lemma 2 and the definition of $r_{F,\mu}(f^*)$.

Observe that $r_{F,\mu}(f^*) = \|f_{\text{nup}} - P_F(f^*)\|$ where f_{nup} is the point in $\text{nup}(F,\mu)$ found by extending a ray from $P_F(f^*)$ through f^* until reaching $\text{nup}(F,\mu)$ (see Figure 1). Let

$$\alpha_{F,\mu}(f^*) := \frac{\|f^* - P_F(f^*)\|}{r_{F,\mu}(f^*)} = \frac{\|f^* - P_F(f^*)\|}{\|f_{\text{nup}} - P_F(f^*)\|}$$

and observe that $\alpha_{F,\mu}(f^*) \in [0,1]$ is the largest α such that $B_F^{(\alpha)}(f^*)$ only intersects F at $P_F(f^*)$.

Note that if F is convex then $\text{nup}(F,\mu)$ is the empty set; hence for all $f^* \in L_2(\mu)$, $r_{F,\mu}(f^*) = \infty$ and $\alpha_{F,\mu}(f^*) = 0$.

Combining theorem 2 with lemmas 1 and 2 leads to our main positive result:

Corollary 1. *Let $F \subset L_2(\mu)$ be a class of functions into $[0,1]$, set Y to be a random variable taking its values in $[0,1]$, and assume that $f^* = \mathbb{E}(Y|X) \notin \text{nup}(F,\mu)$. Assume further that there are constants $d, \gamma, p \geq 1$ such that for every empirical measure μ_n, $\log N\big(\varepsilon, F, L_2(\mu_n)\big) \leq d \log^p(\gamma/\varepsilon)$. Then, there exists a constant $C_{p,\gamma}$, which depends only on p and γ, such that for every $0 < \varepsilon, \delta < 1$,*

$$C_F(\varepsilon, \delta, Y) \leq \frac{C_{p,\gamma}}{\varepsilon} \max\left\{ d \log^p \frac{1}{\varepsilon},\ \frac{\log \frac{2}{\delta}}{1 - \alpha_{F,\mu}(f^*)} \right\}.$$

Note that this result is *non-uniform* in the target Y because some functions f^* are harder to learn than others. For all $f^* \in F^\alpha := \{f \in H : \alpha_{F,\mu}(f) \geq \alpha\}$, one obtains a uniform bound in terms of α. Figure 2 illustrates the boundaries of F^α

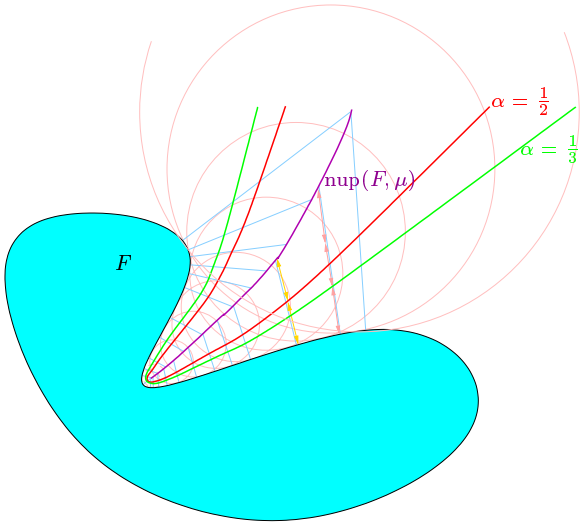

Fig. 2. Illustration of the sets F^α. The lines marked $\alpha = \frac{1}{3}$ and $\alpha = \frac{1}{2}$ are the boundaries of $F^{1/2}$ and $F^{1/3}$ respectively.

for a given F and two different values of α. If F is convex, then $\alpha_{F,\mu}(f^*) = 0$ always and one recovers a completely uniform result.

4 The Lower Bound

In this section we present the geometric construction which led to the claimed lower bound presented in [8]. We then show that the construction is wrong, and can not be corrected. In our discussion we shall use several geometric results, the first of which is the following standard lemma, whose proof we omit.

Lemma 3. *Let X be a Hilbert space and set $x \in X$ and $r > 0$. Put $B_1 = B(x, r)$ and let $y \in \partial B_1$. For any $0 < t < 1$ let $z_t = tx + (1-t)y$ and set $B_2 = B(z_t, \|z_t - y\|)$. Then $B_2 \subset B_1$ and $\partial B_1 \cap \partial B_2 = \{x\}$.*

Using lemma 3 it is possible to show that even if x has several best approximations in G, then any point on the interval connecting x and any one of the best approximations of x has a unique best approximation.

Corollary 2. *Let $x \in X$, set $y \in P_G(x)$ and for every $0 < t < 1$ let $z_t = tx + (1-t)y$. Then, $P_G(z_t) = y$.*

Proof. We begin by showing that $P_G(z_t) \subset P_G(x)$. To that end, note that $d_G(z_t) = \|z_t - y\| = (1-t)\|x - y\|$. Indeed, $d_G(z_t) \leq \|z_t - y\|$. If there is some $y' \in G$ such that $\|y' - z_t\| < (1-t)\|x - y\|$ then by the triangle inequality $\|y' - x\| \leq \|y' - z_t\| + \|z_t - x\| < \|x - y\|$, which is impossible. In fact, the

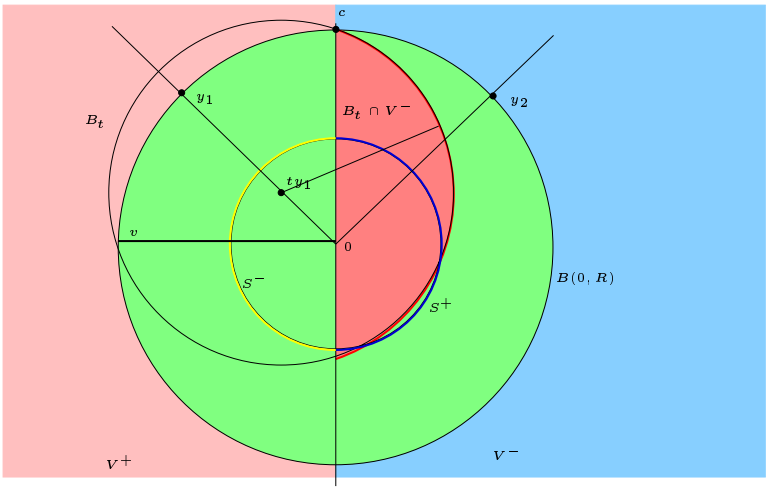

Fig. 3. Illustration for lemma 4

same argument shows that if $y' \in P_G(z_t)$ then $\|x - y'\| = d_G(x)$, implying that $y' \in P_G(x)$. Therefore,

$$P_G(z_t) \subset \partial B\big(x, d_G(x)\big) \cap \partial B\big(z_t, d_G(z_t)\big),$$

and thus, by lemma 3, $P_G(z_t)$ contains a single element — which has to be y. ∎

In the next two results we deal with the following scenario: let $R > 0$ and let $B = B(0,R)$. Fix any $y_1, y_2 \in \partial B$ which are linearly independent, and put $U \subset X$ to be the space spanned by y_1 and y_2. Let $c = R(y_1 + y_2)/\|y_1 + y_2\|$ and set $v \in U$ to be a unit vector orthogonal to c, such that $\langle v, y_1 \rangle > 0$. Denote $V_- = \{x : \langle x, v \rangle \leq 0\}$ and $S_- = V_- \cap S_X$, where S_X is the unit sphere $S_X = \{x \in X : \|x\| = 1\}$. In a similar fashion, let $V_+ = \{x : \langle x, v \rangle > 0\}$ and set $S_+ = V_+ \cap S_X$.

Lemma 4. *For every $0 < t < 1$, $d(ty_1, RS_-) = \|ty_1 - c\|$ and c is the unique minimum.*

The lemma has the following geometric interpretation: for every $0 < t < 1$, let B_t be the ball centered in ty_1 which passes through c. Then, by the lemma, any point in the intersection of B_t with V_- other than c is contained in the interior of $B(0,R)$. (See Figure 3.)

Proof. For any $Rs \in RS_-$,

$$\|ty_1 - Rs\|^2 = \|ty_1\|^2 + R^2 - 2tR\langle y_1, s \rangle,$$

hence, the minimum is attained for $s \in S_-$ which maximizes $\langle y_1, s \rangle$. Set $U = v^\perp$, and since $X = V \oplus U$ then

$$\langle y_1, s \rangle = \langle P_v y_1, P_v s \rangle + \langle P_U y_1, P_U s \rangle,$$

where P_U (resp. P_v) is the orthogonal projection on U (resp. v). Note that for every $s \in S_-$, $\langle P_v y_1, P_v s \rangle \leq 0$ and $\langle P_U y_1, P_U s \rangle \leq \|P_U y_1\|$. Thus, $\langle y_1, s \rangle \leq \|P_U y\|$ and the maximum is attained only by $s \in S_-$ such that $P_U s = P_U y_1 / \|P_U y_1\|$ and $P_v s = 0$. The only s which satisfies the criteria is $s = P_U y_1 / \|P_U y\|$, and $Rs = c$. ∎

Theorem 3. *Let $G \subset X$ be a compact set and $x \in \mathrm{nup}(G, \mu)$. Set $R = d_G(x)$, put $y_1, y_2 \in P_G x$ and set $c = R(y_1 + y_2)/\|y_1 + y_2\|$. For every $0 < t < 1$ let $z_t^1 = y_1 + tR(x - y_1)/\|x - y_1\|$, $z_t^2 = y_1 + tR(x - y_2)/\|x - y_2\|$ and $\varepsilon_t = \|z_t^1 - c\| - \|z_t^1 - y_1\|$. Then, if ρ satisfies that $d_G(z_t^1) \leq \rho < d_G(z_t^1) + \varepsilon_t$, and $g \in B(z_t^1, \rho) \cap G$ then $\|g - y_1\| < \|g - y_2\|$. Similarly, if $g \in B(z_t^2, \rho) \cap G$ then $\|g - y_2\| < \|g - y_1\|$.*

Proof. Clearly we may assume that $x = 0$, thus $z_t^1 = ty_1$ and $z_t^2 = ty_2$. Also, note that $\varepsilon_t = \|z_t^2 - c\| - \|z_t^2 - y_2\|$, hence the second part of the claim follows by an identical argument to the first part. By corollary 2, y_1 is the unique best approximation to z_t^1, hence $\varepsilon_t > 0$. If y_1 and y_2 are linearly dependent then $y_2 = -y_1$ and the result is immediate, thus, we may assume that y_1, y_2 are independent. Let U be the space spanned by $\{y_1, y_2\}$, and let $v \in U$ be a unit vector orthogonal to c such that $\langle v, y_1 \rangle > 0$. By lemma 4 and using the notation of that lemma,

$$B(ty_1, \rho) \cap V_- \subset \mathrm{int} B(ty_1, \|ty_1 - c\|) \cap V_- \subset \mathrm{int} B(0, R).$$

Thus $G \cap (B(ty_1, \rho) \cap V_-) \subset G \cap \mathrm{int} B(0, R) = \emptyset$, implying that $G \cap B(ty_1, \rho) \subset V_+$. Clearly, for every $g \in V_+$, $\|g - y_1\| < \|g - y_2\|$, as claimed. ∎

Remark 1. It is easy to see that there is a constant $C_{d,R}$ such that for every $0 < t < 1$, $\|z_t^i - c\| - \|z_t^i - y_i\| \geq C_{d,R} t$, where $R = d_G(x)$ and $d = \|y_1 - y_2\|$.

Theorem 3 was the key idea behind the argument in [8] of the incorrect lower bound. What was claimed by the authors is the following:

Theorem 4 (False). *Let $G \subset L_2(\mu) \cap bB\big(L_\infty(\mu)\big)$ be compact and nonconvex. For any $\varepsilon > 0$ there exists a random variable W_ε such that $C_G(\varepsilon, \delta, W_\varepsilon) = \Omega(1/\varepsilon^2)$.*

Note that even this claim does not guarantee that there is a target concept for which the sample complexity is $O(\varepsilon^{-2})$. Rather, its actual claim is that it is impossible to obtain an upper bound which is better than the GC limit in the nonconvex case. The targets W_ε will be constructed in such a way that for smaller values of ε, the conditional expectation of W_ε is closer to $\mathrm{nup}(F, \mu)$. It does *not* (as incorrectly claimed in [8]) imply that $C_G(\varepsilon, \delta, Y) = \Omega(1/\varepsilon^2)$ for some random variable Y. As we will show in the sequel, even this weaker claim is wrong. In the next few lines we shall present the outline of the incorrect proof.

"*Proof*". Set $x \in \mathrm{nup}(G, \mu)$ and put $d = \mathrm{diam} P_G(x)$ and $R = d_G(x)$. Let $y_1, y_2 \in P_G(x)$ such that $\|y_1 - y_2\| = d$ and for every $0 \leq t \leq 1$ and $i = 1, 2$ let $z_t^i = ty_i + (1 - t)x$. Assume that y_1 and y_2 are linearly independent (the other case

follows from a similar argument) and let $U = \text{span}\{y_1, y_2\}$. As in theorem 3, set $c = R(y_1 + y_2)/\|y_1 + y_2\|$ and choose $v \in U$ to be a unit vector orthogonal to c such that $\langle v, y_1 \rangle > 0$. Since $v = (y_1 - y_2)/\|y_1 - y_2\|$ then $\|v\|_\infty \leq 2b/d$. Also, recall that by remark 1, there is a constant $C_{d,r}$ such that for $0 \leq t \leq 1$ and $i = 1, 2$, $\rho_t = \|z_t^i - c\| - \|z_t^i - y_i\| \geq C_{d,R} t$. For every $0 < \varepsilon < 1$ define the following $\{0,1\}$-valued random sequence: let ξ_i be i.i.d Bernoulli random variables such that

$$\Delta = \left| \mathbb{E}\xi_i - \frac{1}{2} \right| = \frac{\varepsilon}{2}.$$

By a result stated as lemma 3 in [8] and originally proved in [2], there is an absolute constant C such that one needs a sample of at least $C(\log(1/\delta)\varepsilon^{-2})$ points to identify whether $\Delta = \varepsilon/2$ or $\Delta = -\varepsilon/2$ with probability $1 - \delta$. The idea is to assume that there is an efficient algorithm for agnostic learning, and to use this algorithm to identify the value of Δ with high confidence.

Let $c_\varepsilon = (1 - \varepsilon)x + \varepsilon(y_1 + y_2)/2$, and for every integer n set

$$W_\varepsilon^n = \begin{cases} dv + c_\varepsilon & \text{if } \xi_n = 1, \\ -dv + c_\varepsilon & \text{if } \xi_n = 0. \end{cases}$$

Note that the $\mathbb{E}W_\varepsilon^n = dv/2 + c_\varepsilon = z_\varepsilon^1$ if $\Delta = \varepsilon/2$, and $\mathbb{E}W_\varepsilon = z_\varepsilon^2$ if $\Delta = -\varepsilon/2$. Also, for every $0 < \varepsilon < 1$, $\|W_\varepsilon^n\|_\infty < 3b$. By remark 1, there is a constant $C_{d,R}$ such that for every $0 < t < 1$, $\|z_t^i - c\| - \|z_t^i - y_i\| \geq C_{d,R} t$. Let $\varepsilon < 1/2$ and set $\varepsilon' = \frac{R}{2} C_{d,R} \varepsilon$. Assume that $h \in G$ is such that

$$\|z_\varepsilon^1 - h\|^2 - \|z_\varepsilon^1 - y_1\|^2 < \varepsilon'.$$

Since $\|z_\varepsilon^1 - h\| > \|z_\varepsilon^1 - y_1\|$ then

$$\|z_\varepsilon^1 - h\| - \|z_\varepsilon^1 - y_1\| < \frac{\varepsilon'}{\|z_\varepsilon^1 - h\| + \|z_\varepsilon^1 - y_1\|} \leq \frac{\varepsilon'}{2\|z_\varepsilon^1 - y\|} < \frac{C_{d,r}\varepsilon}{2}.$$

Therefore, $h \in \text{int}B(z_\varepsilon^1, \|z_\varepsilon^1 - c\|)$, implying that $\|h - y_1\| < \|h - y_2\|$.

Now, assume that one had an efficient algorithm to learn W_ε at the scale ε'. Since the conditional expectation of W_ε is either z_ε^1 or z_ε^2 depending on Δ, then the learning rule would produce a function h which satisfies either $\|z_\varepsilon^1 - h\|^2 - \|z_\varepsilon^1 - y_1\|^2 < \varepsilon'$ or $\|z_\varepsilon^2 - h\|^2 - \|z_\varepsilon^2 - y_1\|^2 < \varepsilon'$. We can identify which one it was according to the distances $\|h - y_1\|$ and $\|h - y_2\|$. We predict that $\Delta = \varepsilon/2$ if h is closer to y_1 and that $\Delta = -\varepsilon/2$ if h is closer to y_2.

By the lower bound on the sample complexity of estimating Δ, it was claimed that the agnostic sample complexity

$$C_G(\varepsilon', \delta, W_\varepsilon) \geq \frac{K}{\varepsilon'^2} \log \frac{1}{\delta},$$

where K is an absolute constant. To conclude, if this were the case, one could find constants K_1, K_2 (which depend only on R and on d) such that for every $0 < \varepsilon < 1/2$ there is a random variable Y, such that $\mathbb{E}(Y|X) = z_\varepsilon^1$ and

$$C_G(K_1\varepsilon, \delta, Y) \geq \frac{K_2}{\varepsilon^2} \log \frac{1}{\delta}.$$

The error in the proof is due to the different probability spaces used. The first is the space defined by the Bernoulli random variables, and the second is the one associated with the learning problem. The transition between the two algorithms corresponds to a map between the two probability spaces. The problem arises because this map does not have to be measure preserving. Hence, a "large" subset in one space can be mapped to a very small set in the other. ∎

Below we will show that this random construction can not be corrected.

Definition 5. *Let G be a class of functions on a probability space (X, Σ, μ), and let f be a (deterministic) function in $L_\infty(\mathcal{X})$. For every integer n and every $\varepsilon > 0$, let $M_{\varepsilon,n}(f)$ be the set of n-tuples $\{x_1, ..., x_n\}$ which satisfy that for every $x \in M_{\varepsilon,n}(f)$ the learning rule assigns to the sample $s_n = \big(x_i, f(x_i)\big)$ a function $L_{s_n} \in G$ such that*

$$\|f - L_{s_n}\|^2 - \inf_{g \in G} \|f - g\|^2 < \varepsilon.$$

Clearly, for a class to be learnable, the probability $\mu^n\big(M_{\varepsilon,n}(f)\big)$ must tend to 1 as n tends to infinity for every $f \in L_\infty(\mathcal{X})$. Thus, we can have a "deterministic" version of the sample complexity:

Definition 6. *For a class of functions G on a probability space $(\mathcal{X}, \Sigma, \mu)$ and a (deterministic) function f let $D_f(\varepsilon, \delta)$ be the smallest integer n_0 such that for every $n \geq n_0$, $\mu^n\big(M_{\varepsilon,n}(f)\big) \geq 1 - \delta$.*

If X is distributed according to μ, let $C_f(\varepsilon, \delta)$ be defined as

$$C_f(\varepsilon, \delta) = \sup_Y C(\varepsilon, \delta, Y), \tag{5}$$

where the supremum is taken with respect to all real random variables Y, such that $\mathbb{E}(Y|X) = f$.

Moreover, if one has an efficient learning rule L for f, it depends only on $\{x_1, ..., x_n\}$. Thus, if $\{x_1, ..., x_n\} \in M_{\varepsilon,n}(f)$, then for every $\{y_1, ..., y_n\}$ L will serve as a learning rule for (x_i, y_i). Hence, for every Y such that $\mathbb{E}(Y|X) = f$ and every $0 < \varepsilon, \delta < 1$,

$$C(\varepsilon, \delta, Y) \leq D_f(\varepsilon, \delta), \tag{6}$$

implying that

$$C_f(\varepsilon, \delta) \leq D_f(\varepsilon, \delta).$$

The next step in our construction is to observe that if $f \in \mathrm{nup}(G, \mu)$ and if $g \in P_G(f)$, then the deterministic sample complexity of the function $z_t = (1-t)f + tg$ increases in some sense for $t \in (0, 1)$. In other words, as long as one is restricted to the interval (f, g), it is easier to learn as one moves closer to f.

Lemma 5. *Let $\varepsilon > 0$ and $0 < t_1 < t_2 < 1$. Then, there is a constant C^* which depends only on t_2 such that if $h \in G$ satisfies that $\|h - z_{t_2}\|^2 - \|g - z_{t_2}\|^2 < \varepsilon$ then $\|h - z_{t_1}\|^2 - \|g - z_{t_1}\|^2 < C^*\varepsilon$, where $g \in P_G(f)$. In fact, if $t_2 \leq 1/2$ then one can take $C^* = 2$.*

Proof. By the definition of h, it follows that

$$\|h - z_{t_2}\| - \|g - z_{t_2}\| < \frac{\varepsilon}{\|h - z_{t_2}\| + \|g - z_{t_2}\|} \equiv \rho.$$

Note that for every $\rho > 0$,

$$B\big(z_{t_2}, \|g - z_{t_2}\| + \rho\big) \subset B\big(z_{t_1}, \|g - z_{t_1}\| + \rho\big).$$

Indeed, if $\|u - z_{t_2}\| < \|g - z_{t_2}\| + \rho$, then

$$\begin{aligned}\|u - z_{t_1}\| &\le \|z_{t_1} - z_{t_2}\| + \|u - z_{t_2}\| < \|z_{t_1} - z_{t_2}\| + \|g - z_{t_2}\| + \rho \\ &= \|z_{t_1} - g\| + \rho.\end{aligned}$$

Clearly, $h \in B\big(z_{t_2}, \|g - z_{t_2}\| + \rho\big)$ and thus $h \in B\big(z_{t_1}, \|g - z_{t_1}\| + \rho\big)$. Therefore,

$$\|h - z_{t_1}\|^2 - \|g - z_{t_1}\|^2 \le \rho(\|h - z_{t_1}\| + \|g - z_{t_1}\|) = \varepsilon \frac{\|h - z_{t_1}\| + \|g - z_{t_1}\|}{\|h - z_{t_2}\| + \|g - z_{t_2}\|} = (*)$$

Since $h \in G$ and g is the best approximation of z_{t_2} in G then

$$\|h - z_{t_2}\| \ge \|g - z_{t_2}\| = (1 - t_2)\,\|f - g\|.$$

Thus, by the triangle inequality and the observation above,

$$\begin{aligned}(*) &\le \varepsilon \frac{\|h - z_{t_2}\| + \|z_{t_1} - z_{t_2}\| + \|g - z_{t_1}\|}{\|h - z_{t_2}\| + \|g - z_{t_2}\|} \le \varepsilon\Big(1 + \frac{(t_2 - t_1)\,\|f - g\|}{\|h - z_{t_2}\| + \|g - z_{t_2}\|}\Big) \\ &\le \Big(1 + \frac{t_2}{1 - t_2}\Big)\varepsilon.\end{aligned}$$

∎

Corollary 3. *Let f, g and z_t be as in lemma 5. Fix some $0 < t^* < 1$ and let C^* be the constant appearing in the lemma for $t_2 = t^*$. Then, for every $\varepsilon > 0$, every integer n and every $0 < t < t^*$,*

$$M_{\varepsilon,n}(z_{t^*}) \subset M_{C^*\varepsilon,n}(z_t),$$

and in particular, for every $0 < \varepsilon, \delta < 1$ and every $0 < t < 1/2$

$$D_{z_{1/2}}(\varepsilon, \delta) \ge D_{z_t}(2\varepsilon, \delta).$$

Proof. Fix an integer n and $\varepsilon > 0$ and assume that L is a learning rule for z_{t^*}. If $s_n \in M_{\varepsilon,n}(z_{t^*})$ then $L_{s_n} \in G$ and

$$\|L_{s_n} - z_{t^*}\|^2 - \|g - z_{t^*}\|^2 < \varepsilon.$$

By lemma 5, for every $0 < t < t^*$,

$$\|L_{s_n} - z_t\|^2 - \|g - z_t\|^2 < C^*\varepsilon,$$

and thus $s_n \in M_{C^*\varepsilon,n}(z_t)$, by using the learning rule L. The second part of the claim follows immediately from the first. ∎

Now we can prove that it is impossible to obtain a lower bound in a similar way to the one stated in theorem 4. Indeed, we show that if the random construction where to be true, there must have been a function which is not on $\mathrm{nup}(G,\mu)$ for which the sample complexity is $\Omega(1/\varepsilon^2)$.

Note that in the claim of theorem 4 there is no assumption on the "size" of G. In particular, if the construction is correct, its assertion should hold for classes for which $\log N\big(\varepsilon, G, L_2(\mu)\big) = O(\varepsilon^{-p})$ for $p < 2$. We will show that this creates a contradiction, since the lower bound will exceed the upper bound proved in corollary 1.

Proof. Assume that the assertion of theorem 4 is true. Using its notation, there are constants K_1 and K_2 such that for every $0 < t < 1/2$ there is a random variable Y_t such that $\mathbb{E}(Y_t|X) = z_t^1$ and

$$D_{z_t^1}(K_1 t, \delta) \geq C_{z_t^1}(K_1 t, \delta) \geq C_G(K_1 t, \delta, Y_t) \geq \frac{K_2}{t^2} \log \frac{1}{\delta}, \tag{7}$$

where the first inequality follows from (6). By corollary 3, for every $\varepsilon > 0$ and every $0 < t < 1/2$, $D_{z_{1/2}^1}(\varepsilon, \delta) \geq D_{z_t^1}(2\varepsilon, \delta)$. Fix $0 < t < 1/2$ and select $\varepsilon = K_1 t/2$. By (7),

$$D_{z_{1/2}^1}\big(K_1 t/2, \delta\big) \geq D_{z_t^1}(K_1 t, \delta) \geq \frac{K_2}{t^2} \log \frac{1}{\delta}.$$

Renaming the variables it follows that there are absolute constants K and K^* such that for every $\varepsilon > 0$,

$$D_{z_{1/2}^1}(K^* \varepsilon, \delta) \geq \frac{K}{\varepsilon^2} \log \frac{1}{\delta}. \tag{8}$$

On the other hand, by corollary 1, if G is a class such that $\log N\big(\varepsilon, G, L_2(\mu_n)\big) = O(\varepsilon^{-p})$ for $p < 2$ for every empirical measure μ_n, then for any $h \notin \mathrm{nup}(G,\mu)$ there is a constant $C(h,G)$ such that for every $0 < \varepsilon, \delta < 1$,

$$D_h(\varepsilon, \delta) \leq \frac{C}{\varepsilon^{1+\frac{p}{2}}} \log \frac{1}{\delta}.$$

This creates an impossible situation, since by construction, for every $0 < t < 1$, $z_t^1 \notin \mathrm{nup}(G,\mu)$. ∎

5 Conclusion

We have shown that a previously presented lower bound on the sample complexity of agnostically learning nonconvex classes of functions F with squared loss is false. We have also presented an improved upper bound on that sample complexity in terms of the geometry of F and $\mathrm{nup}(F)$. The interesting thing about the upper bound is that it is intrinsically nonuniform — functions closer to $\mathrm{nup}(F,\mu)$ are harder to learn.

Acknowledgement This work was supported by the Australian Research Council.

References

1. P.L. Bartlett, O. Bousquet, S. Mendelson, "Localized Rademacher averages", submitted to COLT2002.
2. S. Ben-David and M. Lindenbaum, "Learning Distributions by their Density Levels — A Paradigm for Learning without a Teacher," in *Computational Learning Theory — EUROCOLT'95*, pages 53-68 (1995).
3. R.M. Dudley, *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics 63, Cambridge University Press 1999.
4. David Haussler, "Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications," *Information and Computation*, **100**, 78–150 (1992).
5. Michael J. Kearns, Robert E. Schapire and Linda M. Sellie, "Toward Efficient Agnostic Learning," pages 341–352 in *Proceedings of the 5th Annual Workshop on Computational Learning Theory,* ACM press, New York, 1992.
6. Wee Sun Lee, *Agnostic Learning and Single Hidden Layer Neural Networks*, Ph.D. Thesis, Australian National University, 1996.
7. Wee Sun Lee, Peter L. Bartlett and Robert C. Williamson, "Efficient Agnostic Learning of Neural Networks with Bounded Fan-in," *IEEE Transactions on Information Theory*, **42**(6), 2118–2132 (1996).
8. Wee Sun Lee, Peter L. Bartlett and Robert C. Williamson, "The Importance of Convexity in Learning with Squared Loss" *IEEE Transactions on Information Theory* **44**(5), 1974–1980, 1998 (earlier version in *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 140–146, 1996.)
9. S. Mendelson, "Improving the sample complexity using global data," *IEEE transactions on Information Theory*, to appear. http://axiom.anu.edu.au/∼shahar
10. S. Mendelson "Rademacher averages and phase transitions in Glivenko-Cantelli classes" *IEEE transactions on Information Theory*, **48**(1), 251–263, (2002).
11. S. Mendelson "A few remarks on Statistical Learning Theory", preprint. http://axiom.anu.edu.au/∼shahar
12. S.B. Stechkin, "Approximation Properties of Sets in Normed Linear Spaces," *Revue de mathematiques pures et appliquees*, **8**, 5–18, (1963) [in Russian].
13. M. Talagrand, "Sharper bounds for Gaussian and empirical processes", *Annals of Probability*, **22**(1), 28–76, (1994).
14. Aad W. van der Vaart and Jon A. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York, 1996.
15. Frederick A. Valentine, *Convex Sets*, McGraw-Hill, San Francisco, 1964.
16. L.P. Vlasov, "Approximative Properties of Sets in Normed Linear Spaces," *Russian Mathematical Surveys*, **28**(6), 1–66, (1973).