On the Existence of Weak Learners and Applications to Boosting*

Shie Mannor[†]and Ron Meir[‡]

Department of Electrical Engineering

Technion, Haifa 32000

Israel

Abstract

We consider the existence of a weak learner for boosting learning algorithms. A weak learner for binary classification problems is required to achieve a weighted empirical error on the training set which is bounded from above by $1/2 - \gamma$, $\gamma > 0$, for any distribution on the data set. Moreover, in order that the weak learner be useful in terms of generalization, γ must be sufficiently far from zero. While the existence of weak learners is essential to the success of boosting algorithms, a proof of their existence has been hitherto lacking. In this work we provide a proof that, under appropriate conditions, weak learners actually exist. In particular, we show that under certain natural conditions on the data set, a linear classifier is a weak learner. Our results can be directly applied to generalization error bounds for boosting, leading to closed-form bounds. We also provide a procedure for dynamically determining the number of boosting iterations required to achieve low generalization error. Finally, using the results of Kearns and Mansour, our results may also be applied to top-bottom decision trees with oblique splits.

Key words: Boosting, Weak learner, Geometric discrepancy

Running head: On the existence of weak learners

1   Introduction

One of the most exciting developments in the fields of Machine Learning and Pattern Recognition in recent years has been the development of the boosting approach to learning [Sch90, FS96]. Boosting, similarly to other ensemble based methods, such as Bagging [Bre96] and Mixture of Experts [JJ94], has been shown to be quite effective in the construction of successful classifiers. In addition to many impressive empirical demonstrations of the utility of the procedure, there has also been recently a great deal of theoretical work providing guaranteed performance bounds (e.g., [SFBL98]). It turns out that a key ingredient in the success of the boosting approach is its ability to yield classifiers that achieve a large margin, implying that the decision boundary defining the classifier, while separating the training points, is able to retain as large a distance as possible from them. The notion of margin also plays an essential role in the theory of support vector machines [Vap95].

In order to derive useful generalization bounds for boosting-type algorithms, it has turned out to be essential to be able to guarantee the existence of a so-called *weak learner*. In boosting, the final composite classifier is formed by a convex combination of base classifiers. In order for the procedure to be effective, these base classifiers, also known as weak hypotheses or weak learners, are required to obey the so-called weak-learning assumption, defined in Section 2 below. While many base classifiers have been used in practice, it has not been clear up to this point under what conditions weak learners actually exist. This situation is somewhat disturbing, as the existence of a weak learner is essential to the theory (and practice) of boosting.

In this work we prove that linear classifiers are indeed weak learners. Clearly, any other system based on linear classifiers, such as neural networks and decision trees with oblique splits are also weak learners. Moreover, the proof technique used suggests a randomized algorithm, which achieves the desired result. In order to establish our results, a simple assumption needs to be made concerning the data. As we show in Section 4, some regularity assumption is necessary, as no *effective* weak learner can exist for arbitrary data sets.

The proof method employed in this paper is based on the theory of geometric discrepancy, a sub-field of combinatorial geometry, which deals with irregularities of distributions. The reader is referred to the excellent text-book by Matoušek [Mat99] for a general introduction to this field.

The remainder of the paper is organized as follows. Section 2 recalls the boosting framework and motivates the importance of finding weak classifiers. In Section 3 we introduce some basic definitions and key results from the field of geometric discrepancy. The main result regarding the existence of a weak linear learner is presented in Section 4, and the application to error bounds for boosting is provided as a corollary of the main result. We close in Section 5 by drawing some conclusions and mentioning some open questions. Some of the more technical proofs have been relegated to appendices.

It should be commented that no attempt has been made to optimize the constants appearing in the lower bounds. Finally, we note that we use the term 'classifier' and 'hypothesis' interchangeably throughout this work.

2   Boosting and Weak Learners

Boosting is a general meta-procedure for constructing a strong classifier from a set of weak classifiers [Sch90]. Each weak classifier is constructed based on a re-weighted version of the data set. Although there exist many versions of boosting to-date (e.g., [FS96, JFT00, LMF00, SS99]), most of them fall into the general framework depicted in figure 1, adapted from [SS99].

We begin our discussion by recalling the notion of the Vapnik-Chervonenkis (VC) dimension.

Definition 1 *Let H be a class of functions from $\mathcal{X}$ to $\{-1, +1\}$, and let X be a set of m points from $\mathcal{X}$. We say that H shatters X if H can compute all 2^m dichotomies on x. The VC-dimension of H is the size of the largest shattered subset of $\mathcal{X}$.*

Since the final classifier constructed within the boosting framework is a convex com-

---

A generalized boosting algorithm

Given: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y} \subseteq \mathbb{R}$
Initialize: $P_1(i) = 1/m$
For $t = 1, 2, \dots, T$:

- Train a weak learner using the distribution P_t.

- Get a weak hypothesis $h_t : \mathcal{X} \mapsto \mathcal{Y}$.

- Update the distribution based on the performance of h_t, namely

$$P_{t+1} \Leftarrow P_t .$$

Output:

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \bigg/ \sum_{t=1}^{T} \alpha_t ,$$

where the non-negative parameters α_t are determined by the specific boosting algorithm used.

Figure 1: *A generalized boosting algorithm*

bination of weak classifiers, it is of interest to investigate error bounds for convex combinations of classifiers. In this context we recall the elegant results derived by Schapire *et al.* [SFBL98]. Let H be a class of binary classifiers of VC-dimension d_H, and denote by $co(H)$ the convex hull of H,

$$co(H) = \left\{ f \ : \ f(x) = \sum_i \alpha_i h_i(x), \ h_i \in H, \ \alpha_i \geq 0, \ \sum_i \alpha_i = 1 \right\} .$$

Given a sample $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, $x_i \in \mathcal{X}$ and $y_i \in \{-1, +1\}$, of m examples drawn independently at random from a probability distribution D over $\mathcal{X} \times \{-1, +1\}$, Schapire et al. [SFBL98] show that for $m > d_H$, with probability at least $1 - \delta$, for *every* $f \in co(H)$ and $\theta > 0$,

$$\mathbf{P}_D[Yf(X) \leq 0] \leq \mathbf{P}_S[Yf(X) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \left(\frac{d_H(\log(m/d_H))^2}{\theta^2} + \log\frac{1}{\delta} \right)^{1/2} \right), \quad (1)$$

where the *margin-error* $\mathbf{P}_S[Yf(X) \leq \theta]$ denotes the fraction of training points for which $y_i f(x_i) \leq \theta$, namely

$$\mathbf{P}_S[Yf(X) \leq \theta] = \frac{1}{m} \sum_{i=1}^{m} I(y_i f(x_i) \leq \theta),$$

where $I(E)$ is the indicator function for the event E. Note that the term $\mathbf{P}_D[Yf(X) \leq 0]$ is simply the probability of misclassification of the classifier $h(x) = \text{sgn}(f(x))$.

It is helpful to compare (1) to the classic VC bounds [VC71], which *do not* take into account the fact that f is real-valued and that it is a convex combination of functions from H. Let $d_{co(H)}$ denote the VC-dimension of $co(H)$. Then one finds (e.g., [AB99]) that with probability at least $1 - \delta$, for *every* $f \in co(H)$ and $m > d_{co(H)}$,

$$\mathbf{P}_D[Yf(X) \leq 0] \leq \mathbf{P}_S[Yf(X) \leq 0] + O\left(\frac{1}{\sqrt{m}} \left(d_{co(H)} \log(m/d_{co(H)}) + \log(1/\delta) \right)^{1/2} \right)$$

$$(2)$$

Note that the first term in (2) is simply the fraction of misclassified points in the data set. Comparing (1) and (2) we note two major differences. The former bound contains the extra parameter θ, which allows one to fine tune the bound in order to achieve better performance. However, since $\mathbf{P}_S[Yf(X) \leq 0] \leq \mathbf{P}_S[Yf(X) \leq \theta]$ the first term in (2) is always *smaller* that the corresponding term in (1). However, observe that the second term in (1) may be significantly smaller than the corresponding term in (2), since it contains the VC-dimension of H rather than the VC-dimension of $co(H)$, which may be *significantly* larger (e.g., Section 14.4 in [AB99]). It should be clear that there is a trade-off between the two terms appearing on the r.h.s. of (1). While the first term is monotonically increasing with θ, the second term decreases monotonically. This

situation is very much reminiscent of the idea of structural risk minimization, suggested by Vapnik [Vap82] in the context of hierarchical hypothesis classes.

In order to obtain a useful bound, one needs to be able to estimate the first term in (1). In order to do so, we formally introduce the notion of a *weak hypothesis*.

Definition 2 *Let* $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$, *be a set of* m *examples. We say that a hypothesis* h γ-*satisfies the weak hypothesis assumption if, for any probability distribution* P *on the* m *points* S, *the weighted empirical error of* h *is smaller than* $1/2 - \gamma$, *namely*

$$\sum_{i=1}^{m} P_i I(h(x_i) \neq y_i) \leq \frac{1}{2} - \gamma. \tag{3}$$

There are two important facts to note concerning this definition. First, it is required to hold for *any* distribution P, and second, one only demands that the error be smaller than $1/2 - \gamma$. Since an error of $1/2$ is always achievable[1], this is a very weak assumption if γ is small.

Schapire *et al.* [SFBL98] have recently provided a bound on the empirical error of the composite classifier formed by the AdaBoost algorithm [FS96] run for T steps. In particular, denote the error of each weak learner by ϵ_t, $t = 1, \dots, T$, then Theorem 5 in [SFBL98] shows that

$$\mathbf{P}_S[Y f(X) \leq \theta] \leq \prod_{t=1}^{T} \sqrt{4\epsilon_t^{1-\theta}(1 - \epsilon_t)^{1+\theta}} \tag{4}$$

$$\leq \left((1 - 2\gamma)^{1-\theta}(1 + 2\gamma)^{1+\theta}\right)^{T/2} \qquad (\gamma > \theta),$$

where the second inequality holds if each weak classifier attains an error smaller than $1/2 - \gamma$, $\gamma > \theta$. This bound decreases to zero exponentially fast if $\gamma > \theta$. In other words, *if* a sufficiently large value of γ can be guaranteed, then the first term in (1) converges to zero. However, if γ (and thus θ) behaves like $m^{-\beta}$ for some $\beta > 0$, the rate of convergence in the second term in (1) will increase, possibly leading to worse bounds than those available by using standard VC results (2). What is needed then is a characterization of conditions under which the achievable θ does not decrease too rapidly with m. For example, if $\gamma = \Omega(1/\log m)$ this will only contribute a logarithmic factor to the complexity penalty, and not change the rate which will remain $O(1/\sqrt{m})$, up to logarithmic factors. On the other hand, we have shown in [MM00] that a value of γ of order $1/m$ can always be achieved; however, such a small value is useless from the point of view of the bound (1). We introduce the notion of an *effective* weak learner, which characterizes a weak learner for which γ is sufficiently large to guarantee that the second

---

[1]To see this, pick a random classifier h. If the error is smaller or equal to $1/2$, we're done, if not the error of the classifier $-h$ is smaller than $1/2$.

term in (1) converges to zero as the sample size increases. In the following definition we use the notation $f(m) = \omega(g(m))$ to indicate that $f(m)$ becomes arbitrarily large relative to $g(m)$ when m approaches infinity.

Definition 3 *A weak hypothesis h is* effective *if it satisfies (3) with* $\gamma > \omega \left(\log m / \sqrt{m} \right)$.

In this work we consider conditions under which an effective weak learner exists for one of the simplest possible classifiers, namely the linear classifier.

3   Background from Geometric Discrepancy

We introduce the key definitions and results which will enable us to show in Section 4 that an *effective* weak learner exists. The results of this section essentially summarize the work of Alexander [Ale90, Ale91, Ale94], setting the nomenclature and drawing the connections to classification and the notion of weak learners. We start with several definitions. Let h denote a hyperplane

$$h = \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x} + b = 0 \right\}.$$

The corresponding (open) half-spaces h^+ and h^- are given, respectively, by $h^+ = \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x} + b > 0 \right\}$ and $h^- = \left\{ \mathbf{x} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x} + b < 0 \right\}$. We denote by H a generic half-space. The linear classifier associated with the hyperplane h will be denoted by ξ_h, namely

$$\xi_h(\mathbf{x}) = \text{sgn} \left(\mathbf{w} \cdot \mathbf{x} + b \right)$$

An essential ingredient of the proof of our main result is the notion of a *motion invariant measure* on the space of hyperplanes in $\mathbb{R}^d$. We follow Section 6.4 in [Mat99] in describing this measure. First, we ignore the set of hyperplanes passing through the origin, which has measure zero. Second, note that any hyperplane h is *uniquely* determined by r, the distance from the origin and Ω, the solid d-dimensional angle describing the inclination of the shortest line from the origin to the hyperplane (see Figure 2 for the case $d = 2$, where Ω is simply the polar angle ϕ, $0 \leq \phi \leq 2\pi$). Let χ be the bijective map from the set of hyperplanes to the d-dimensional Euclidean space (r, Ω). If L is a set of hyperplanes, we define the measure $\mu(L)$ as the Lebesgue measure of the set $\chi(L)$ in the (r, Ω) space, where r and Ω are interpreted as cartesian coordinates. For example, if the set L is supported in the unit ball in $\mathbb{R}^2$, then we have the set $\{(r, \phi), 0 \leq r \leq 1, 0 \leq \phi \leq 2\pi\}$. It is easy to see that the measure μ is *motion-invariant*, i.e., if L' arises from L by rigid motion (i.e., without changing the relative positions of the lines) then $\mu(L') = \mu(L)$. In fact, it turns out that, up to scaling, this is the only motion-invariant measure on hyperplanes (see Section 6.4 in [Mat99]).
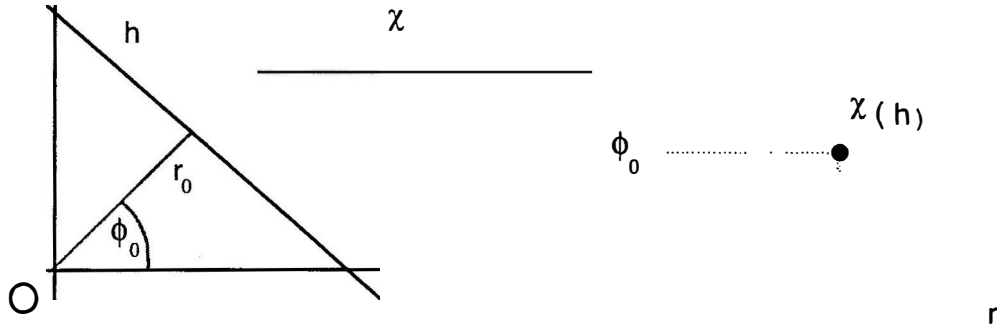
Figure 2: Mapping from hyperplanes to spherical coordinates (based on [Mat99])

Given a set of points $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ in $\mathbb{R}^d$, and a signed atomic measure ν. The geometric discrepancy $D_s(\nu)$ is given by

$$D_s(\nu) \stackrel{\triangle}{=} \sup_H \{\nu(H)\} = \sup_H \left\{ \sum_{\mathbf{x}_i \in X \cap H} \nu(\mathbf{x}_i) \right\},$$

where the supremum is over all halfspaces.

Intuitively, the discrepancy relates to the maximal measure of points in X, which belong to a half-space. Consider, for example, the case of an even number of points, such that $\nu_i = 1$, $i = 1, 2, \dots, m/2$, and $\nu_i = -1$ for $i = m/2 + 1, \dots, m$, where $\nu_i = \nu(\mathbf{x}_i)$. Then, $D_s(\nu)$ is the maximal difference between the sizes of positive and negative sets belonging to any halfspace. In the sequel we will be interested in finding a *lower* bound on $D_s(\nu)$. The main motivation for the introduction of $D_s(\nu)$ is its relationship to the weighted empirical error defined in (3). Let $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$, be a labeled sample of m points weighted according to $\{P_1, \dots, P_m\}$. Set

$$\nu_i \stackrel{\triangle}{=} y_i P_i \qquad (i = 1, 2, \dots, m), \tag{5}$$

be a signed measure over the points X. We will observe in Section 4 that when $\sum_i \nu_i = 0$,

$$\boxed{\inf_\xi \left\{ \sum_{i=1}^m P_i I(\xi(\mathbf{x}_i) \neq y_i) \right\} = \frac{1}{2} - D_s(\nu) \qquad (\nu_i = y_i P_i)} \tag{6}$$

where the infimum is over all linear classifiers ξ.

Furthermore, it follows from Lemma 4.1 that if a lower bound on $D_s(\nu)$ is available for any measure ν such that $\nu(X) = \sum_{i=1}^m \nu_i = 0$, then a lower bound can be obtained

for any signed measure ν. For this purpose we define a set of measures obeying the condition,

$$\Psi(\mathbb{R}^d) = \left\{ \nu : \nu \text{ is a signed measure such that } |\nu|(\mathbb{R}^d) < \infty \text{ and } \nu(\mathbb{R}^d) = 0 \right\}.$$

Let K be the support of ν. Then for any $\nu \in \Psi(\mathbb{R}^d)$, $\nu(K) = 0$. In this work K will be a finite set, $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$. Furthermore, since P is a normalized probability distribution, clearly we must also demand that $\sum_{i=1}^m |\nu_i| = 1$. Thus, we make the following assumption throughout this section.

Assumption 3.1 *The signed measure ν obeys the two conditions*

$$\sum_{i=1}^m |\nu_i| = 1 \qquad \sum_{i=1}^m \nu_i = 0.$$

A crucial quantity used in the sequel is the function

$$I(\nu) \overset{\triangle}{=} \sum_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \nu_i \nu_j \qquad \left(\sum_i \nu_i = 0 \right),$$

which is a weighted measure of dispersion of the points $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$. This function turns out to be very useful for our purposes since it is connected to the discrepancy, through the following result.

Lemma 3.1 ([Ale94], Lemma 2) *Let ν belong to $\Psi(\mathbb{R}^d)$, and denote by μ the unique motion invariant measure over hyperplanes. Then*

$$-I(\nu) = \int \nu(h^+)^2 d\mu(h) \leq D_s(\nu)^2 \mu(\mathbf{H}) \tag{7}$$

where $\mathbf{H}$ is the set of planes that cut the convex hull of the support of ν.

Two important observations follow from (7).

1. The function $I(\nu)$ is negative. This is in itself a surprising and non-trivial result, following from the classic work of [Sch37] on isometric embeddings in Hilbert space. The reader is referred to Section 6.7 in [Mat99] for a more modern proof of this fact using the notion of positive definite functions.

2. The resulting inequality

$$D_s(\nu) \geq \sqrt{\frac{-I(\nu)}{\mu(\mathbf{H})}}$$

immediately yields a lower bound on $D_s(\nu)$ if a lower bound for $-I(\nu)$ and an *upper* bound for $\mu(\mathbf{H})$ are available. An exact value for $\mu(\mathbf{H})$ is provided by Alexander for hyper-cubes and hyper-spheres [Ale94], and can be immediately used in our case. For example, from Lemma 3 in [Ale94] we find that for the case of a ball of radius r in $\mathbb{R}^d$,

$$\mu(\mathbf{H}) = 2r(d-1)\frac{O_{d-1}}{O_{d-2}} = 2\pi r(d-1)\frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}$$

where $O_k = 2\pi^{(k+1)/2}\left(\Gamma((k+1)/2)\right)^{-1}$ is the volume of the unit ball. (We comment that Alexander's Lemma discusses the unit ball. However, the extension to any ball is straightforward using results from Section 13.6 in [San76]). Since, without loss of generality, we may always assume that the set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is contained in a ball of radius $\max_{i,j}\|\mathbf{x}_i - \mathbf{x}_j\|/2$, we need not concern ourselves any further with the factor $\mu(\mathbf{H})$.

One final ingredient needed in order to prove our main result is the construction of an auxiliary measure in $\mathbb{R}^{d+1}$. The main idea here, due to Alexander [Ale90], is to view $\mathbb{R}^d$ as a hyper-plane in $\mathbb{R}^{d+1}$, and construct a measure in $\mathbb{R}^{d+1}$ for which a lower bound on $D_s(\nu)$ can be established. We briefly describe the construction and properties of the new measure, referring the reader to [Ale90, Ale91, Ale94] for the full details. Let Φ be an atomic measure over $\mathbb{R}$, concentrated on the finite set $R = \{r_1, r_2, \ldots, r_n\}$. For $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, consider the set $X \times R \subset \mathbb{R}^{d+1}$, and define the convolution measure $\nu \star \Phi$ over $X \times R$ by

$$(\nu \star \Phi)(\mathbf{x}_i, r_k) \triangleq \nu(\mathbf{x}_i)\Phi(r_k) = \nu_i\phi_k.$$

From Corollary 3 in [Ale91] we conclude that

$$-I(\nu \star \Phi) \leq -\|\Phi\|_1^2 I(\nu),$$

where $\|\Phi\|_1 = \sum_{k=1}^{\ell} |\phi_k|$. Finally, let $\mathbf{q}_{ik} = (\mathbf{x}_i^T, r_k)^T \in \mathbb{R}^{d+1}$ be a point in $X \times R$. From Lemma 9 in [Ale91] we have that

$$-I(\nu \star \Phi) = -\sum_{i \neq j}\sum J_\Phi(\mathbf{x}_i, \mathbf{x}_j)\nu_i\nu_j - I(\Phi)\sum \nu_i^2, \qquad (8)$$

where

$$J_\Phi(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{\ell}\sum_{l=1}^{\ell} \|\mathbf{q}_{ik} - \mathbf{q}_{jl}\|_2 \phi_k\phi_l \qquad (9)$$

It is important to observe that since $\|\mathbf{q}_{ik} - \mathbf{q}_{jl}\|_2^2 = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 + |r_k - r_l|^2$, the function $J_\Phi(\mathbf{x}_i, \mathbf{x}_j)$ depends on $\mathbf{x}_i$ and $\mathbf{x}_j$ only through $d = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. In Section 4 we use (8) and (9) as a starting point for our proof. We quote two essential features of the function $J_\Phi(\mathbf{x}_i, \mathbf{x}_j)$.

Lemma 3.2 ([Ale91], Theorem 4 and Lemma 7) *Let Φ be an atomic measure concentrated on the n points $\{r_1, \cdots, r_n\}$ in $\mathbb{R}$, and such that $\sum_{i=1}^{n} \Phi(r_i) = 0$. Assume further that $d = \|\mathbf{x} - \mathbf{x}'\|_2$ exceeds δ, the diameter of the support of Φ. Then*

$$J_\Phi(\mathbf{x}, \mathbf{x}') = d \sum_{k=1}^{\infty} c_k I^{2k}(\Phi) d^{-2k},$$

where c_k is defined by $(1 + x^2)^{1/2} = \sum_{k=0}^{\infty} c_k x^{2k}$, and

$$I^\ell(\Phi) \stackrel{\triangle}{} \sum_{i=1}^{n} \sum_{j=1}^{n} |r_i - r_j|^\ell \Phi(r_i) \Phi(r_j). \tag{10}$$

Moreover,

$$I^\ell(\Phi) \le \|\Phi\|_1^2 \delta^\ell. \tag{11}$$

Lemma 3.3 ([Ale91], Theorem 6) *Let Φ be an atomic measure concentrated on the n points $\{r_1, \cdots, r_n\}$ in $\mathbb{R}$, and such that $\sum_{i=1}^{n} \Phi(r_i) = 0$. Then $-J_\Phi(\mathbf{x}, \mathbf{x}')$ is a strictly decreasing positive function of $d = \|\mathbf{x} - \mathbf{x}'\|_2$.*

4   On the existence of weak learners

With the basic mathematical tools of Section 3 in hand, we proceed now to the proof of the existence of an effective weak learner based on hyperplanar decision boundaries. A few words are in order concerning the construction of linear classifiers for arbitrary data sets, which are, of course, not linearly separable in general. When the points are linearly separable, it is known that there are effective (i.e., polynomial time) procedures for determining the separating hyperplane. For example, linear programming achieves the desired goal. When the points are *not* linearly separable, it has been proved that the problem of finding the linear classifier with the minimal number of errors is NP-hard (see, for example, Section 24.2 in [AB99]). Recently, it was shown in [BBD99] that the problem is NP-hard even if only an approximate solution is required. Note, however, that our problem is somewhat different, as we are interested only in finding, for any distribution on the data points, a 'weak' solution, i.e., one that is always slightly better than the trivial error of $1/2$. Our main focus here is on the *existence* of such a linear classifier, since this will establish the existence of a weak learner. We proceed now to the mathematical analysis of the problem.

Let

$$\xi(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x} + b),$$

be the classification according to the hyper-plane h defined by $h = \{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b = 0\}$.
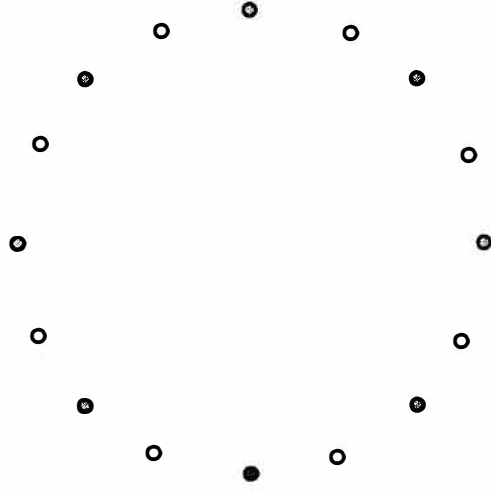
Figure 3: *Sixteen alternating points on a sphere. No line can achieve an error lower than* $1/2 - 1/8 = 3/8$

Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$, be a sample of m points. We need to establish an upper bound on the error

$$\epsilon(P, \xi) \triangleq \sum_{i=1}^{m} P_i I(y_i \neq \xi(\mathbf{x}_i)) \tag{12}$$

where $P_i \geq 0$ and $\sum_i P_i = 1$. Observe that $\epsilon(P, \xi)$ may also be expressed as $\sum_{i=1}^{m} P_i[1 - y_i \xi(\mathbf{x}_i)]/2$.

First, we wish to show that unless some conditions are specified, there can be no effective linear classifiers. To show this consider the following example, which demonstrates that, for any classifier ξ, there exists a set of points such that $\epsilon(P, \xi) \geq 1/2 - \gamma$, where $\gamma \leq O(1/m)$.

Example 1 Consider the points $\mathbf{x}_1, \ldots, \mathbf{x}_{2m}$ on the boundary of a circle, and set $\nu_i = (-1)^i (2m)^{-1}$. Clearly $\sum_i \nu_i = 0$ and $\sum_i |\nu_i| = 1$. For example, in Figure 3 there are 8 grey (positive) points and 8 white (negative) points alternating on the circumference of a circle. From the symmetry of the problem, it it is clear that any line cutting the circle (and not passing through any of the points) is such that there is a difference of at most 1 between the number of points from either color on either side of the line. Clearly the error incurred by *any* hyperplane h is at least $1/2 - 1/m$, and cannot fulfill the condition required from an effective weak learner (recall Definition 3). $\qquad\square$

We introduce an auxiliary signed measure, defined by $\nu_i = P_i y_i$, as in (5). For any measure ν and classifier ξ let

$$\bar{D}(\nu, \xi) = \sum_{i=1}^{m} \nu_i \xi(\mathbf{x}_i) \qquad (\nu_i = y_i P_i),$$

and note that

$$\epsilon(P,\xi) = \frac{1 - \tilde{D}(\nu,\xi)}{2} \qquad (\nu_i = y_i P_i).$$

Let $I^{\pm}$ be the subsets of indices for which $y_i = \pm 1$. A probability distribution is said to be *symmetric*, with respect to a sample S, if

$$\sum_{i \in I^+} P_i = \sum_{i \in I^-} P_i, \qquad (13)$$

i.e., P assigns equal weights to positive and negative examples. We show in Lemma 4.1 that it suffices to consider symmetric distributions. Observe that the symmetry and normalization conditions on P are equivalent to the conditions $\sum_{i=1}^m \nu_i = 0$ and $\sum_{i=1}^m |\nu_i| = 1$, which constitute Assumption 3.1.

For any halfspace $h^+ = \{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} + b > 0\}$ and points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, let

$$D(\nu, h^+) \stackrel{\triangle}{=} \nu(h^+) = \sum_{\mathbf{x}_i \in X \cap h^+} \nu_i = \sum_{\mathbf{x}_i \in X} \left(\frac{1 + \xi_h(\mathbf{x}_i)}{2}\right) \nu_i,$$

where $\nu_i = \nu(\mathbf{x}_i)$.

Under the assumption that $\sum_{\mathbf{x}_i \in X} \nu_i = 0$, it follows that

$$\tilde{D}(\nu, \xi_h) = 2D(\nu, h^+) \qquad (\xi_h(\mathbf{x}) = 2I(\mathbf{x} \in h^+) - 1).$$

We conclude that

$$\epsilon(P, \xi_h) = \frac{1}{2} - D(\nu, h^+) \qquad (\xi_h(\mathbf{x}) = 2I(\mathbf{x} \in h^+) - 1 \ \& \ \nu_i = y_i P_i).$$

Therefore, in order to derive an an *upper* bound on $\inf_\xi \epsilon(P, \xi)$ it suffices to obtain a *lower* bound on the geometric discrepancy $D_s(\nu) = \sup_h D(\nu, h^+)$, as claimed in (6).

We now show that is suffices to consider the case $\sum_i \nu_i = 0$ (equivalently, $\sum_{i \in I^+} P_i = \sum_{i \in I^-} P_i$). The proof of the Lemma is given in the Appendix.

Lemma 4.1 *Let m arbitrary, but distinct, points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ in $\mathbb{R}^d$ be given. Let Ξ be a class of binary classifiers, containing the constant classifiers $\xi(\mathbf{x}) = +1$ and $\xi(\mathbf{x}) = -1$. Assume that for <u>any</u> symmetric probability distribution P, a classifier $\xi \in \Xi$ can be found such that*

$$\epsilon(P, \xi) \leq \frac{1}{2} - \gamma, \qquad (14)$$

with γ depending only on the points $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Then for any probability distribution (not necessarily symmetric) on X, there exists a classifier $\xi \in \Xi$ for which

$$\epsilon(P, \xi) \leq \frac{1}{2} - \frac{\gamma}{2}.$$

Note that Lemma 4.1 is more general than we need, as it applies to any class of classifiers Ξ, rather than the class of linear classifiers, which is our main concern.

4.1   Main Theorem

We are now ready to state our main result. We begin with the more general theorem, specializing to some specific cases in two corollaries. We begin by defining an auxiliary function that will result from the proof of Theorem 4.1.

$$d^{-1}F_n(d,\eta) \stackrel{\triangle}{=} -\binom{2n}{n} - 2\sum_{\ell=1}^{n}(-1)^{\ell}(1+\ell^2(\eta/4nd)^2)^{1/2}\,\frac{(2n)!}{(n-\ell)!(n+\ell)!} \qquad (15)$$

Before stating our main result we describe some of the properties of $F_n(d,\eta)$, as this function plays a central role in the bounds we derive.

Lemma 4.2 *The function $F_n(d,\eta)$ satisfies the following properties:*

1. *For fixed η, $F_n(d,\eta)$ is a positive and monotonically decreasing function of d.*

2. *$F_n(d,\eta) \approx ad + b\eta/d$ when η/d is large (relative to n).*

3. *When η/d is small,*

$$F_n(d,\eta) \approx \left[\left(\frac{1}{4n}\right)^{2n}\frac{1\cdot 1\cdot 3\cdots(2n-3)}{2\cdot 4\cdot 6\cdots(2n)}(2n)!\right]d\left(\frac{\eta}{d}\right)^{2n}.$$

The symbol '$\approx$' indicates that the additional terms are of smaller order.

The proof of Lemma 4.2 is presented in the appendix. In Figure 4 we plot $F_n(d,\eta)$ as a function of η for $d=2$ and $n=4$ and as a function of d for $\eta=1$ and $n=4$. The properties mentioned above are evident in this figure.

We are now ready to state our main result.

Theorem 4.1 *Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$ be a fixed sample, and let non-negative weights P_i be assigned to each point in S such that $\sum_{i=1}^{m} P_i = 1$. Denote by $X^{\pm}$ the points $\mathbf{x}_i$ for which $y_i = \pm 1$, $|X^{\pm}| = m_{\pm} = \alpha_{\pm} m$, and set*

$$\eta = \min_{i,j}\{\|\mathbf{x}_i - \mathbf{x}_j\|_2,\ \mathbf{x}_i \in X^{+},\ \mathbf{x}_j \in X^{-}\},$$
$$L = \max_{i,j}\|\mathbf{x}_i - \mathbf{x}_j\|_2.$$

Let the sets $X^{\pm}$ be partitioned into $K_{\pm}$ disjoint subsets denoted by $X_k^{\pm}$. Then for any P there exists a linear classifier ξ such that
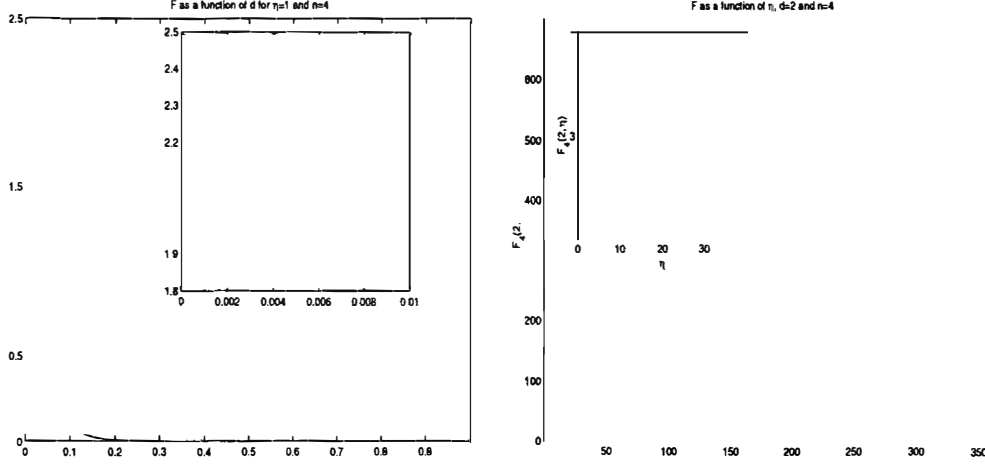
$$\epsilon(P,\xi) \le \frac{1}{2} - \gamma$$

Figure 4: *$F_4(d, \eta)$ as a function of d (for fixed η) and as a function of η (for fixed d)*

where

$$\gamma \geq \sqrt{-(C_d/4L)I(\nu)} \qquad (\nu_i = y_i P_i), \tag{16}$$

and where $C_d = \Gamma(d/2)/[\pi(d-1)\Gamma((d-1)/2)]$ depends only on the dimension d, and

$$-I(\nu) \geq 0.004\eta \left[\log_2(125\sqrt{\alpha_+\alpha_-}m)\right]^{-3/2} \sum_{i=1}^{m} \nu_i^2$$

$$+ \sum_{k=1}^{K_+} F_n(d_k^+, \eta)\sum\sum|\nu_i||\nu_j| + \sum_{k=1}^{K_-} F_n(d_k^-, \eta)\sum\sum|\nu_i||\nu_j|. \tag{17}$$

Here $d_k^\pm$ is the diameter of the k'th subset in the partition of $X^\pm$, and $n = \lceil (1/2)\log_2(125\sqrt{\alpha_+\alpha_-}m)\rceil$, i.e., $n = \Theta(\log m)$.

Remark 1 Observe that η measures the minimal distance between points belonging to the two oppositely labeled subsets. As can be expected, the lower bound improves for larger values of η. We refer to η as the *gap*, and observe that it is a quantity characterizing the geometric distribution of points. This should be contrasted with the parameter γ in Definition 2, which characterizes a classifier, rather than a set of points (the points themselves enter through (3)).

Remark 2 The statement of the theorem requires the partition of the sets $X^\pm$ into $K_\pm$ subsets. Clearly there are many such partitions, for values of $K_\pm$ between 1 and $m_\pm$. A careful inspection of the second and third terms in (17) implies that there is a trade-off to be considered. First, observe that if one of the subsets of $X^\pm$ is a singleton, the corresponding second or third term in (17) is absent. Next, assume that the number

of regions $K_\pm$, is large, and that each subset $X_k^\pm$ is composed of a relatively small set of nearby points. In this case, the number of summands is large, and the functions $F_n(d_k^\pm, \eta)$ are also large since $d_k^\pm$ are small and the function $F_n(d, \eta)$ is monotonically decreasing with d. On the other hand, the terms $\sum_{i \neq j} |\nu_i||\nu_j|$ are small, since they contain a small number of terms. In the extreme case where the size of the partition is equal to the number of points, the second and third terms in (17) are entirely absent. Consider now the other extreme situation, where $K_\pm = 1$, i.e., no partition of the points is constructed. In this case, the number of summands in the final two terms in (17) is reduced, and $F_n(d_k^\pm, \eta) \approx F_n(L, \eta)$ is also smaller. On the other hand, the terms $\sum_{i \neq j} |\nu_i||\nu_j|$ are large, since they contain a large number of terms. Thus, the optimal partition of the sets $X^\pm$ depends in each case on the specific geometry of the points, as well as on their assigned measures ν_i. For example, we show in example 2 below, a clear case where it is advantageous to choose $K_\pm = 2$ rather than $K_\pm = 1$.

Example 2 We give an example where a better bound may be obtained using a larger partition of the set X. Consider the XOR configuration of points in figure 5. Since the first term in the bound (17) is not affected by the partition we compare only the final two terms, for the case $K_\pm = 1$ and $K_\pm = 2$, where in the latter case the partition is into the separate four quadrants. Simple algebra leads to the following two bounds:

$$-I(\nu) \geq \frac{3}{16} F_n(D, \eta) \qquad (K_\pm = 1)$$

$$-I(\nu) \geq \frac{1}{16} F_n(d, \eta) \qquad (K_\pm = 2).$$

From the first property in Lemma 4.2, $F_n(d, \eta)$ is positive and strictly decreasing function of d. By picking d small enough and D large enough, the bound (17) can be made larger for the case $K_\pm = 2$. $\qquad\square$

In order to obtain a better understanding of the structure of the bounds, we consider the special case where all the weights are equal in magnitude, namely $|\nu_i| = 1/m$ for all i. We then have the following corollary.

Corollary 4.1 *Let the condition of Theorem 4.1 hold, with the additional constraint that $|\nu_i| = 1/m$ for all $i = 1, 2, \dots, m$. Denote the number of positively/negatively labeled points by $m_\pm = \alpha_\pm m$, $2 < m_+ < m - 2$. Then there exists a linear classifier ξ such that*

$$\epsilon(P, \xi) \leq \frac{1}{2} - \gamma$$

where

$$\gamma \geq \sqrt{C_d/2L} \left[\left(0.002\eta (\log_2(125\sqrt{\alpha_+\alpha_-}m))^{-3/2} m^{-1} \right)^{1/2} + \left([\alpha_+^2 + \alpha_-^2] F_n(L, \eta) \right)^{1/2} \right],$$
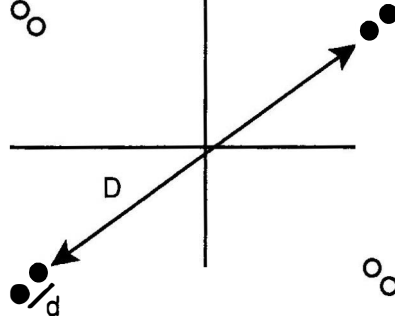
Figure 5: *8 points in a XOR position, D is the maximal distance between equally labeled points, and d is the distance between equally labeled points within a localized region.*

$C_d = \Gamma(d/2)/[4\pi(d-1)\Gamma((d-1)/2)]$ and $n = \lceil (1/2)\log_2(125\sqrt{\alpha_+\alpha_-}m)\rceil$. For small values of η/L we find that $F_n(L,\eta) \geq \Omega\left(m^{-2\log(2Le/\eta)}(\log m)^{-3/2}\right)$, implying that

$$\gamma \geq \Omega\left((\log m)^{-3/4}m^{-1/2}\right) + \Omega\left((\log m)^{-3/4}m^{-\log(\frac{2Le}{\eta})}\right).$$

Proof The claim follows by simple substitution of $|\nu_i| = 1/m$ and $K_\pm = 1$ in (17), i.e., using the trivial partition of the sets $X^\pm$. As in Example 2 above, we note that $F_n(d,\eta)$ is a positive and monotonically decreasing function of $d = \|\mathbf{x} - \mathbf{x}'\|_2$. Thus, since $d_k^\pm \leq L$, we may replace $F_n(d_k^\pm,\eta)$ by $F_n(L,\eta)$ to get a somewhat weaker lower bound. The lower bound on $F_n(L,\eta)$ follows by using the results of Lemma 4.2 to show that $F_n(L,\eta) \geq \Omega((\eta/n)^{2n}a_n(2n)!)$, where a_n is the $(n+1)$-th coefficient in the expansion of $\sqrt{1+x}$, and noting that $a_n \geq (n+3)^{-2}$. Finally, we use the inequality $\sqrt{a+b} \geq (\sqrt{a} + \sqrt{b})/\sqrt{2}$, which follows from Jensen's inequality and the concavity of the square root function. □

Finally, we comment on a hypothesis class which is widely used in many of the applications of boosting, namely the class of *stumps*. This class consists of linear classifiers for which the decision boundary is parallel to one of the axes. We show that under the conditions of Theorem 4.1 stumps cannot achieve an error lower than $1/2$, while this is clearly possible using general linear classifiers.

Lemma 4.3 *Let Ξ be the class of axis parallel linear classifiers (stumps). Then there exists a set of points S, obeying the conditions of Theorem 4.1, for which $\epsilon(P,\xi) = 1/2$ for any $\xi \in \Xi$.*

Proof The result follows by simply considering the set of points in figure 6, composed of a XOR configuration of equally weighted points. It is clear by symmetry that no axis-
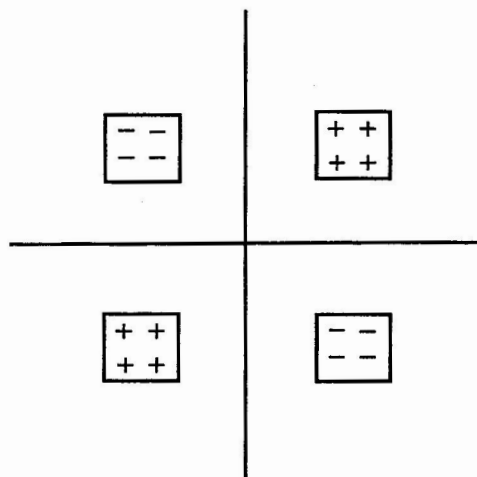
Figure 6: *A configuration of points that cannot be classified by stumps with an error smaller than 1/2.*

parallel line can achieve an error lower than 1/2, for any choice of the distance between positive and negative points.   □

Finally, we comment of the necessity of a positive value for η. One can easily show that in the special case where the hyperplanes are homogeneous, i.e., constrained to pass through the origin, the positivity of η is essential. To see this, consider a spherically symmetric two-dimensional configuration of points, such that all points in the first and third quadrants are positively labeled, while points in the second and fourth quadrants are negatively labeled. Moreover, assume that η, the distance between the closest points from the two classes is vanishingly small. Then it is easy to see by symmetry that no homogeneous hyperplane can achieve an error smaller than 1/2. We do not know at this point whether the positivity of η is necessary in the case of general hyperplanes.

4.2   Application to boosting

We now apply Theorem 4.1 to boosting algorithms, and in particular to the AdaBoost algorithm. The basic claims are derived by combining the generalization error bound (1) and the bound (17). First, note that the variable θ in (1) is a free parameter, which may be tuned at will. We wish to consider situations where the bound converges to zero as m increases, and thus need to guarantee that both terms in (1) vanish for large m. Consider first the margin error $\mathbf{P}_S[Yf(X) \leq \theta]$. In view of (4) and the comments following it, it suffices that $\theta < \gamma$, where each weak learner achieves an error which is smaller than $1/2 - \gamma$. Retaining only the first term of the lower bound (17), we observe that it behaves like $\Omega\left(\eta(\log m)^{-3/2}\sum_{i=1}^{m}\nu_i^2\right)$. The worst possible situation occurs when all the weights are equal, namely $|\nu_i| = 1/m$ for all i, in which case the lower bound on γ is (see (16)) $\Omega\left(\left[\eta(\log m)^{-3/2}m^{-1}\right]^{1/2}\right)$. In this case, upon substitution in (1), we

obtain that the second (complexity penalty) term does not decay to zero, due to the fact that the gap parameter γ decays to zero too quickly. However, it is well known that in most successful applications of boosting, only a small number of weights, say s, retain non-negligible values, while most weights shrink (exponentially fast) to zero. In this situation, keeping in mind the normalization condition $\sum_{i=1}^{m} |\nu_i| = 1$, the term $\sum_{i=1}^{m} \nu_i^2$ is in fact of order $\zeta = 1/s$, instead of order $1/m$. As long as $s = o(m)$, the existence on an effective weak learner is established. In this situation we may in fact obtain much better rates of convergence, as attested by the following theorem.

Theorem 4.2 *Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{\pm 1\}$ be a sample of m points drawn independently at random from a distribution D over $\mathbb{R}^d \times \{-1, +1\}$, and such that the minimal distance between differently labeled points in S is η. Assume that the AdaBoost algorithm has been run for T steps, and that there exists an integer $T_0 \leq T$, such that the probabilities P_i assigned to the points in S obey $\sum_{i=1}^{m} P_i^2 \geq \zeta$ for $T > T_0$. Then for sufficiently large m, with probability at least $1 - \delta$,*

$$\mathbf{P}_D[Y f(X)] \leq c \exp \left\{ -\frac{2\eta(1-\mu)(T-T_0)}{\zeta(\log m)^{3/2}} \right\} + O\left(\frac{1}{\sqrt{m}} \left(\frac{d_H}{\eta\zeta\mu^2} (\log(m/d_H))^{7/2} + \log\frac{1}{\delta} \right)^{1/2} \right),$$

where the constant c depends on T_0, and $\mu \in (0, 1)$ is a free parameter.

Proof First, observe that the generalization bound (1) applies to all $f \in \text{co}(H)$, and in particular to the f obtained via the boosting algorithm. Second, under the conditions of the theorem, we know that there exists a constant a and a positive number $\gamma_0 = \left[a\zeta\eta(\log m)^{-3/2} \right]^{1/2}$ such that each weak learner achieves an error smaller than $1/2 - \gamma_0$ for $T > T_0$. Set $\theta = \mu\gamma_0$, $0 < \mu < 1$. Then from the work of [SFBL98], see (4), we find that the empirical error of the classifier f obeys

$$\mathbf{P}_S[Y f(X) \leq \theta] \leq \prod_{t=1}^{T_0} \sqrt{4\epsilon_t^{1-\theta}(1-\epsilon_t)^{1+\theta}}$$
$$\times \exp \left\{ \frac{T-T_0}{2} [(1 - \mu\gamma_0)\ln(1 - 2\gamma_o) + (1 + \mu\gamma_0)\ln(1 + 2\gamma_0)] \right\},$$

where we have split the product (4) into $t \leq T_0$ and $T > T_0$. For large values of m, i.e., small values of γ_0, we expand the logarithm, substitute for γ_0 as a function of m, and obtain the desired claim for the first term in (1). The second term is obtained by straightforward substitution of $\theta = \mu\gamma_0$ in the second term of (1). $\qquad\square$

Remark 3 The bound of Theorem 4.2 is particularly interesting as it relates the number of boosting iterations to the sample size and to the properties of the data (through η). In particular, as long as the gap parameter η is strictly positive (independently of m), and

as long as T increases faster than $(\log m)^{3/2}$, we are guaranteed that the classification error will converge to zero. For example, Let $T = \Theta((\log m)^{5/2})$. Then we find that the bound behaves like

$$\mathbf{P}_D[Yf(X)] \leq O\left(\frac{1}{m^{2\eta(1-\mu)/\zeta}}\right) + O\left(\left(\frac{d_H}{\zeta\eta\mu^2}\right)^{1/2}\left(\frac{(\log(m/d_H))^{7/2}}{m}\right)^{1/2}\right),$$

which decays to zero as a function of m at a rate depending on ζ and on the gap parameter η. It is important to observe that an appropriate value of T is determined *dynamically* here, and not preset in advance, since the condition that $\sum_i P_i^2 \geq \zeta$ is only known after the algorithm is run.

It should be clear that Theorem 4.2 and Remark 3 may be used in order to select an appropriate value for the number of boosting iterations needed to guarantee that the total error bound approaches zero. As far as we aware there has been no systematic method to-date to address this problem.

Finally, we comment that Kearns and Mansour [KM96] have shown that several widely used algorithms for the construction of top-down decision trees, are in fact boosting algorithms. Considering decision trees formed by oblique splits at each node, we conclude that our results establish bounds on the generalization error for decision trees as well.

5   Conclusions

We have considered the existence of a weak learner in boosting algorithms. The existence of such a learner is a crucial requirement for the success of boosting algorithms, but there has been no general proof to-date that such a learner exists. In this work we have shown that an effective linear weak learner exists, and have provided bounds on its performance. It turns out that the only condition needed for the existence of a weak linear learner is that the positively/negativey labeled points be separated by some nonzero gap. Combining our results with the standard generalization bounds for boosting, we are able to establish complete bounds on the performance of boosting algorithms. Since the existence of a weak learner has been established for linear classifiers, the results clearly hold for any more complex classifiers for which linear classifiers are a sub-class. Two notable examples are neural networks and decision trees with oblique splits. We also argued that our bounds can be directly used in order to determine the number of boosting iteration needed in order to guarantee convergence of the expected error to zero. Whether the bounds are sufficiently tight in order for this procedure to be practically useful remains to be seen.

It is important to stress that the only condition needed in order to guarantee the existence of a weak linear classifier is a finite gap, i.e., a positive distance between points belonging to the two classes. In practical pattern recognition applications it may be hard

to guarantee a sizable gap as the sample size increases. A challenging open problem here would be to investigate under what conditions a weak learner may be shown to exist without this assumption. We have argued that, at least for homogeneous hyperplanes, this requirement is necessary, while the question for general hyperplanes remains open. A promising approach to this issue is based on the work of Rogers [Rog94], which extends the results of Alexander [Ale90, Ale91] to piecewise uniform measures.

It is interesting to comment that the main result of this paper required some rather advanced tools from the field of combinatorial geometry, specifically the sub-field of geometric discrepancy. It would seem extremely difficult to establish the existence of effective weak learners using elementary techniques. An immediate question that arises here relates to the establishment of weak learnability for other types of classifiers, which are *not* based on hyperplanes. This issue is currently under investigation.

We have not touched in this paper upon the issue of algorithmic design. The construction of the motion-invariant measure in Section 3 immediately suggests a stochastic Monte-Carlo type algorithm for this purpose. It would also be interesting to see whether deterministic algorithms exist, for which the property of weak learnability can be established.

Appenix

A   Proof of Lemma 4.1

Proof Since

$$\epsilon(P, \xi) = \frac{1}{2} - \frac{1}{2} \sum_{i=1}^{m} \nu_i \xi(x_i),$$

it suffices to prove that if for every symmetric ν

$$\sum_{i=1}^{m} \nu_i \xi(x_i) > 2\epsilon \qquad \left(\sum_i \nu_i = 0 \right),$$

holds for some ξ, then for any (not necessarily symmetric) distribution ν there exists a classifier $\xi \in \Xi$ such that

$$\sum_{i=1}^{m} \nu_i \xi(x_i) > \epsilon.$$

Let ν be an arbitrary measure on X. If $\left| \sum_{i=1}^{m} \nu_i \right| > \epsilon$ then one can take ξ to be the constant classifier $\xi(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{m} \nu_i)$ and the result follows. We therefore can assume, without loss of generality, that

$$\left| \sum_{i=1}^{m} \nu_i \right| \leq \epsilon.$$

Let us construct the following signed measure on X. Denote by I^+ the set of points with positive measure ν_i and I^- the set of points with negative measure, and let

$$\nu_i' = \begin{cases} \frac{1}{2} \frac{\nu_i}{\sum_{i \in I^+} \nu_i} & \text{if } i \in I^+, \\ -\frac{1}{2} \frac{\nu_i}{\sum_{i \in I^-} \nu_i} & \text{if } i \in I^-. \end{cases}$$

Note that $\nu_i' \gtrless 0$ for $i \in I^{\pm}$. Moreover, $\sum_{i=1}^{m} \nu_i' = 0$, so by assumption there exists a classifier $\xi \in \Xi$ for which $\sum_{i=1}^{m} \nu_i' \xi(\mathbf{x}_i) > 2\epsilon$. Clearly

$$\sum_{i=1}^{m} \nu_i \xi(x_i) = \sum_{i=1}^{m} \nu_i' \xi(x_i) + \sum_{i=1}^{m} (\nu_i - \nu_i') \xi(x_i)$$

$$\geq \sum_{i=1}^{m} \nu_i' \xi(x_i) - \left| \sum_{i=1}^{m} (\nu_i - \nu_i') \xi(x_i) \right|. \tag{18}$$

By the construction of ν' from ν we can now bound the second term of (18):

$$\left| \sum_{i=1}^{m} (\nu_i - \nu_i') \xi(x_i) \right| \leq \sum_{i=1}^{m} |\nu_i - \nu_i'|$$

$$= \sum_{i \in I^+} |\nu_i - \nu_i'| + \sum_{i \in I^-} |\nu_i - \nu_i'|$$

$$\stackrel{(a)}{=} \left| \sum_{i \in I^+} (\nu_i - \nu_i') \right| + \left| \sum_{i \in I^-} (\nu_i - \nu_i') \right|$$

$$= \left| \sum_{i \in I^+} \nu_i - \frac{1}{2} \right| + \left| \sum_{i \in I^-} \nu_i + \frac{1}{2} \right|$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2}. \tag{19}$$

where (a) used the observation that for all $i \in I^+$, $\nu_i - \nu_i'$ are of the same sign (and similarly for I^-). In the final line we used the following simple observation. If a and b

are two real numbers such that $|a| + |b| = 1$, $|a + b| \leq \epsilon$ and $a > 0$, then $|a - 1/2| \leq \epsilon/2$ & $|b + 1/2| \leq \epsilon/2$. The claim follows upon setting $a = \sum_{i \in I^+} \nu_i$ and $b = \sum_{i \in I^-} \nu_i$. Substituting (19) in (18) yields $\sum_{i=1}^{m} \nu_i \xi(\mathbf{x}_i) \geq 2\epsilon - \epsilon = \epsilon$, which establishes the claim. $\square$

B Proof of Lemma 4.2

We begin by recalling the definition of $F_n(d, \eta)$.

$$d^{-1} F_n(d, \eta) = -\binom{2n}{n} - 2 \sum_{\ell=1}^{n} (-1)^\ell (1 + \ell^2 (\eta/4nd)^2)^{1/2} \frac{(2n)!}{(n - \ell)!(n + \ell)!}. \quad (20)$$

Next, note that

$$\binom{2n}{n} + 2 \sum_{\ell=1}^{n} (-1)^\ell \frac{(2n)!}{(n - \ell)!(n + \ell)!} = 0. \quad (21)$$

This claim follows by noting that (21) is simply the binomial expansion of $(1 - 1)^{2n}$, where the sum has been re-arranged.

The first property in the Lemma follows from Lemma 3.3 and the definition of $F_n(d, \eta)$ in (30) as $-J_\Phi(\mathbf{x}, \mathbf{x}')$ where $\|\mathbf{x} - \mathbf{x}'\|_2 = d$.

For large η/d we have that $(1 + \ell^2(\eta/4nd)^2)^{1/2} \approx \ell\eta/4nd$. Thus, (20) is approximately:

$$d^{-1} F_n(d, \eta) \approx -\binom{2n}{n} - (\eta/2nd) \sum_{\ell=1}^{n} (-1)^\ell \ell \frac{(2n)!}{(n - \ell)!(n + \ell)!}. \quad (22)$$

This proves the second property for $b = -(1/2n) \sum_{\ell=1}^{n} (-1)^\ell \ell \frac{(2n)!}{(n-\ell)!(n+\ell)!}$ and $a = -\binom{2n}{n}$.

For small η/d one can use the following Taylor series (e.g., [GR94])

$$\sqrt{1 + x} = 1 + \sum_{k=1}^{\infty} (-1)^{k+1} a_k x^k,$$

where $a_1 = 1$ and

$$a_k = \frac{1 \cdot 1 \cdot 3 \cdots (2k - 3)}{2 \cdot 4 \cdot 6 \cdots (2k)} \quad (k > 1).$$

substituting this in 20 results in the following approximation:

$$d^{-1} F_n(d, \eta) = -\binom{2n}{n} - 2 \sum_{\ell=1}^{n} (-1)^\ell (1 + \sum_{k=1}^{\infty} (-1)^{k+1} a_k (\ell\eta/4nd)^{2k}) \frac{(2n)!}{(n - \ell)!(n + \ell)!}. \quad (23)$$

Changing the order of summation yields the following equation:

$$d^{-1}F_n(d,\eta) = -\binom{2n}{n} - 2\sum_{\ell=1}^{n}(-1)^{\ell}\frac{(2n)!}{(n-\ell)!(n+\ell)!}$$

$$+ 2\sum_{k=1}^{\infty}a_k\sum_{\ell=1}^{n}(-1)^{\ell}(\ell\eta/4nd)^{2k})\frac{(2n)!}{(n-\ell)!(n+\ell)!}. \tag{24}$$

The first term vanishes, as in (21). We proceed to study the elements in the second sum. The next combinatorial lemma characterize the behavior of the summed elements:

Lemma B.1 *If* $1 \le k < n$ *then* $\sum_{l=1}^{n}(-1)^{\ell}\ell^{2k}\frac{(2n)!}{(n+\ell)!(n-\ell)!} = 0$

Proof The proof is by induction over k. For $k = 1$ recall that $\sum_{\ell=0}^{2n}(-1)^{\ell}\ell^2\frac{(2n)!}{\ell!(2n-\ell)!} = 0$ [GR94] (0.154.3). We therefore have that:

$$0 = \sum_{\ell=0}^{2n}(-1)^{\ell}\ell^2\frac{(2n)!}{\ell!(2n-\ell)!}$$

$$= (-1)^n n^2\binom{2n}{n} + 2(-1)^n\sum_{\ell=1}^{n}((n-l)^2+(n+l)^2)(-1)^{\ell}\frac{(2n)!}{(n+\ell)!(n-\ell)!}$$

$$= (-1)^n n^2\binom{2n}{n} + 2n^2(-1)^n\sum_{\ell=1}^{n}(-1)^{\ell}\frac{(2n)!}{(n+\ell)!(n-\ell)!}$$

$$+ 2(-1)^n\sum_{\ell=1}^{n}(-1)^{\ell}\ell^2\frac{(2n)!}{(n+\ell)!(n-\ell)!}$$

The first two terms vanish, as in (21) so we have that $\sum_{\ell=1}^{n}(-1)^{\ell}\ell^2\frac{(2n)!}{(n+\ell)!(n-\ell)!} = 0$

We proceed to the induction step. Assume the result is true for $k' = 1, \ldots, k-1$, then from [GR94] (0.154.4) for $k < n$

$$\sum_{\ell=0}^{2n}(-1)^{\ell}\ell^k\frac{(2n)!}{\ell!(2n-\ell)!} = 0.$$

Rearranging and collecting terms yields:

$$0 = (-1)^n n^{2k} \binom{2n}{n} + 2(-1)^n \sum_{\ell=1}^{n} ((n-l)^{2k} + (n+l)^{2k})(-1)^{\ell} \frac{(2n)!}{(n+\ell)!(n-\ell)!}$$

$$= (-1)^n n^{2k} \binom{2n}{n} + 2n^{2k}(-1)^n \sum_{\ell=1}^{n} (-1)^{\ell} \frac{(2n)!}{(n+\ell)!(n-\ell)!}$$

$$+ 2(-1)^n \sum_{m=1}^{k-1} \binom{2k}{2m} n^{2k-2m} \sum_{\ell=1}^{n} (-1)^{\ell} \ell^{2m} \frac{(2n)!}{(n+\ell)!(n-\ell)!}$$

$$+ 2(-1)^n \sum_{\ell=1}^{n} (-1)^{\ell} \ell^{2k} \frac{(2n)!}{(n+\ell)!(n-\ell)!}$$

Note that all the odd powers vanish, since the sum of binomial coefficients for $(n-l)^{2k} + (n+l)^{2k}$ consists of even powers only. The sum of the two first terms vanish as in (21), by the induction the third term vanishes too. The last term is the desired element and is therefore 0. □

As a result we know that the first $k-1$ elements in the sum of (24) vanish. The k-th element can be calculated by recalling [GR94] (0.154.4)

$$\sum_{\ell=0}^{2n} (-1)^{\ell} \ell^{2k} \frac{(2n)!}{\ell!(2n-\ell)!} = (2n)! .$$

Repeating the argumentation of Lemma B.1 results in

$$(2n)! = 2 \sum_{\ell=1}^{n} (-1)^{\ell} \ell^{2n} \frac{(2n)!}{(n+\ell)!(n-\ell)!} \tag{25}$$

By taking the first non vanishing element in (24) as the approximation we have that:

$$d^{-1} F_n(d, \eta) \approx a_k (\eta/4nd)^{2n} (2n)! \tag{26}$$

□

C Proof of Theorem 4.1

Let $I^+ (I^-)$ denote the indices of the positively (negatively) labeled points. Due to Lemma4.1 we assume without loss of generality that P is symmetric, namely $\sum_{i \in I^+} P_i = \sum_{i \in I^-} P_i$, implying that $\sum_i \nu_i = 0$, where $\nu_i = y_i P_i$. We start from the basic identity (8). Recall that the measure Φ is defined over the set $R = \{r_0, r_2, \dots, r_n\}$ and the convolution measure over $X \times R$ is given by $(\nu \star \Phi)(\mathbf{x}_i, r_k) = \nu(\mathbf{x}_i) \Phi(r_k)$. Let $X^+ = X_1^+ \cup \dots \cup X_{K_+}^+$, and similarly for X^-, be partitions of $X^{\pm}$. Denote the indices of points in $X_k^{\pm}$ by $I_k^{\pm}$.

Keeping in mind the negativity of $J_\Phi(\mathbf{x}_i, \mathbf{x}_j)$ (Lemma 3.3) and $I(\Phi)$ (Lemma 3.1), we have the expression,

$$-I(\nu \star \Phi) = |I(\Phi)| \sum \nu_i^2 - \sum_{j \in X^+} \sum_{i \in I^-} |J_\Phi(\mathbf{x}_i, \mathbf{x}_j)| |\nu_i| |\nu_j| \qquad (27)$$

$$+ \sum_{k=1}^{K_+} \sum_{\substack{i \neq j \\ i,j \in I_k^+}} |J_\Phi(\mathbf{x}_i, \mathbf{x}_j)| \nu_i \nu_j + \sum_{k=1}^{K_-} \sum_{\substack{i \neq j \\ i,j \in I_k^-}} |J_\Phi(\mathbf{x}_i, \mathbf{x}_j)| \nu_i \nu_j.$$

where we have used the fact that for points $\mathbf{x}_i$ and $\mathbf{x}_j$ belonging to the same class the weights ν_i and ν_j are of the same sign, while $\nu_i \nu_j < 0$ if $i \in I^+$ and $\nu_j \in I^-$. The remainder of the proof proceeds by first defining an appropriate measure Φ, and then bounding each of the terms in (27).

We begin with a useful lemma from [Ale94].

Lemma C.1 *Let ν be an atomic measure concentrated on a set of m points in $\mathbb{R}^d$. Let δ be the minimal distance between any two points. Then*

$$-I(\nu) \geq c_d \delta \sum_{i=1}^{\cdot} \nu_i^2,$$

where $c_2 = 0.02$.

Defining the measure Φ

Similarly to [Ale94] define the signed one-dimensional measure $\tilde{\Phi}_{n-1}$, concentrated on $\{0, 1, \dots, n\}$ as follows:

$$\tilde{\Phi}_{n-1}(k) = (-1)^k \binom{n}{k}$$

It is easily seen that $\tilde{\Phi}$ obeys the following following two conditions:

$$\|\tilde{\Phi}_{n-1}\| = 2^n,$$
$$\sum i^s \tilde{\Phi}_{n-1}(i) = 0 \qquad (0 \leq s \leq n-1).$$

The last property it easy to check by looking at the derivatives of the expression

$$(1-x)^n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} x^k,$$

and setting $x = 1$.

We then define a normalized measure Φ, supported on the set $\{0, \eta/4n, 2\eta/4n, \ldots, \eta/4\}$ by

$$\Phi(k\eta/4n) \triangleq 2^{-n}\tilde{\Phi}(k) = (-1)^k \binom{n}{k} 2^{-n} \qquad (k = 1, 2, \ldots, n) \tag{28}$$

Clearly $\|\Phi\|_1 = 1$ and the $n-1$ first moments of Φ vanish.

Lower bound on the first term in (27)

From Lemma C.1 we conclude that

$$|I(\Phi)| \geq 0.02(\eta/4n) \sum |\phi_k|^2$$

$$= (0.005\eta/n)2^{-2n} \sum_k \binom{n}{k}^2$$

$$\overset{(a)}{=} (0.005\eta/n)2^{-2n} \binom{2n}{n}$$

$$\geq 0.001\eta n^{-3/2}.$$

In (a) we used a standard identity for binomial coefficients, and the Stirling bound

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^{n+1/12n}$$

was used in the final step, in the form $\binom{2n}{n} \geq (1/2)(\pi n)^{-1/2}2^{2n}$.

Lower bound on the second term in (27)

Using a Taylor expansion for the square root function one may show (see Lemma 3.2) that

$$J_\Phi(\mathbf{x}, \mathbf{x}') = d \sum_{k=1} c_k I^{2k}(\phi)d^{-2k}$$

where $d = \|\mathbf{x} - \mathbf{x}'\|_2$ and $I^{2k}(\Phi)$ is defined in (10). Let $h = \eta/4$ be the diameter of the support of Φ. Keeping in mind that $|c_k| < 1$ and using the fact that $|I^k(\Phi)| = 0$ for $k = 0, 1, \ldots, n-1$, and $|I^{2k}(\Phi)| \leq \|\Phi\|_1^2 h^{2k}$ (see (11)) in general, we find that

$$|J_\Phi(\mathbf{x}, \mathbf{x}')| \leq d \sum_{k=n+1} h^{2k}d^{-2k}$$

$$= d[1 - h^2 d^{-2}]^{-1}\left(\frac{h}{d}\right)^{2(n+1)}$$

where we have used the fact that $\|\Phi\|_1 = 1$. Since $d \geq \eta$ (recall that $\mathbf{x}_i \in X^+$ and $\mathbf{x}_j \in X^-$) and $h = \eta/4$, it follows that $h/d \leq 1/4$ and thus $[1 - h^2 d^{-2}]^{-1} \leq 16/15$. We then find that

$$|J_\Phi(\mathbf{x}, \mathbf{x}')| \leq 1.07 \left(\frac{h}{d}\right)^{2n+1} h$$

$$\leq 1.07 \cdot 4^{-(2n+1)}(\eta/4)$$

Continuing we find that

$$-\sum_{i \in I^+} \sum_{j \in I^-} |J_\Phi(\mathbf{x}_i, \mathbf{x}_j)| \|\nu_i\| |\nu_j| \geq -0.06 \cdot 2^{-4n} \eta \sum_{j \in X^+} \sum_{i \in I^-} |\nu_i| |\nu_j|$$

$$\geq -0.06 \cdot 2^{-4n} \eta \sqrt{m_+ m_-} \left(\sum_{i \in I^+} \nu_i^2\right)^{1/2} \left(\sum_{j \in I^-} \nu_j^2\right)^{1/2},$$

where $m_\pm = |I^\pm|$. In the second step we used the Cauchy-Schwartz inequality to show that $\sum_i \sum_j |\nu_i| |\nu_j| \leq \sqrt{\sum_i \sum_j \nu_i^2 \nu_j^2 \sum_i \sum_j 1}$.

Lower bound on the third and fourth terms in (27)

We now move on to deal with the final terms in (27). For this purpose we need a lower bound on $|J_\Phi(\mathbf{x}_i, \mathbf{x}_j)|$. From (9) we have the explicit expression

$$J_\Phi(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=0}^{n} \sum_{l=0}^{n} \|\mathbf{q}_{ik} - \mathbf{q}_{jl}\|_2 \phi_k \phi_l,$$

where $\phi_k = \Phi(r_k)$ and $\mathbf{q}_{ik} = (\mathbf{x}_i^T, r_k)^T \in \mathbb{R}^{d+1}$.

Let $\mathbf{x}$ and $\mathbf{x}'$ be points in $\mathbb{R}^d$ such that $\|\mathbf{x} - \mathbf{x}'\|_2 = d$. Let $\xi = \eta/4nd$, then the following relations hold for the measure Φ defined in (28).

$$d^{-1} J_\Phi(\mathbf{x}, \mathbf{x}') = d^{-1} \sum_{k=10}^{n} \sum_{l=0}^{n} (-1)^{k+\ell}(d^2 + (\eta/4n)^2(k-\ell)^2)^{1/2} \binom{n}{k}\binom{n}{\ell}$$

$$= \sum_{k=0}^{n} \sum_{l=0}^{n} (-1)^{k+\ell}(1 + \xi^2(k-\ell)^2)^{1/2} \binom{n}{k}\binom{n}{\ell}$$

$$\overset{(a)}{=} \sum_{k=0}^{n} \binom{n}{k}^2 - 2(1+\xi^2)^{1/2} \sum_{k=1}^{n} \binom{n}{k}\binom{n}{k-1} + 2(1+4\xi^2)^{1/2} \sum_{k=2}^{n} \binom{n}{k}\binom{n}{k-2} + \cdots$$

$$\overset{(b)}{=} \sum_{k=0}^{n} \binom{n}{k}^2 + 2\sum_{\ell=1}^{n}(-1)^\ell(1+\ell^2\xi^2)^{1/2} \sum_{k=1}^{n} \binom{n}{k}\binom{n}{k-\ell}$$

$$= \binom{2n}{n} + 2\sum_{\ell=1}^{n}(-1)^\ell(1+\ell^2\xi^2)^{1/2} \frac{(2n)!}{(n-\ell)!(n+\ell)!} \tag{29}$$

In step (a) we separated the sum into terms with $k = \ell$, $|k - \ell| = 1$, $|k - \ell| = 2$, etc. In step (b) a standard formula for the sum of binomial coefficients was used, and a re-arrangement of the sum was performed.

For the specific measure Φ introduced in (28), define

$$F_n(d, \eta) = -J_\Phi(\mathbf{x}, \mathbf{x}') > 0 \qquad (\|\mathbf{x} - \mathbf{x}'\|_2 = d). \qquad (30)$$

From Lemma 3.3 we conclude that $F_n(d, \eta)$ is a positive and strictly decreasing function of d. Thus, the third term in (17) is bounded from below by

$$\sum_{k=1}^{K_+} F_n(d_k^+, \eta) \sum \sum |\nu_i||\nu_j|$$

where

$$d_k^+ = \max_{i,j \in I_k^+} |\mathbf{x}_i - \mathbf{x}_j|.$$

A similar bound holds for the fourth term in (17).

Concluding the proof

Combining all the results derived, we have the following lower bound,

$$-I(\nu) \geq \eta \left[0.001 n^{-3/2} \sum_i \nu_i^2 - 0.06 \cdot 2^{-4n} \sqrt{m_+ m_-} \|\nu\|_+ \|\nu\|_- \right]$$
$$+ \sum_{k=1}^{K_+} F_n(d_k^+, \eta) \sum \sum |\nu_i||\nu_j| + \sum_{k=1}^{K_-} F_n(d_k^-, \eta) \sum \sum |\nu_i||\nu_j|, \qquad (31)$$

where $\|\nu\|_\pm = \left(\sum_{i \in I^\pm} \nu_i^2 \right)^{1/2}$. The parameter n has been free up to now. In order to guarantee that the second term is smaller than the first, we need to select n large enough so that the second term is smaller than (say) half the first term, namely

$$0.0005 n^{-3/2} \sum \nu_i^2 > 0.06 \cdot 2^{-4n} \sqrt{m_+ m_-} \|\nu\|_+ \|\nu\|_-.$$

Simple mathematical manipulation shows that this is equivalent to the requirement

$$n \geq \frac{3}{8} \log_2 \frac{n}{a}$$
$$a = \left[\frac{1}{120} \frac{\|\nu\|^2}{\sqrt{m_+ m_-} \|\nu\|_+ \|\nu\|_-} \right]^{2/3}$$

We use the following simple lemma.

Lemma C.2 *(Vidyasagar, Lemma 4.4) Suppose $\alpha, \beta > 0$, $\alpha\beta > 2$, $n \geq 1$. Then*

$$n \geq 2\alpha \log \alpha\beta \quad \Rightarrow \quad n > \alpha \log \beta n.$$

We thus conclude that a sufficient condition on n is that

$$n \geq \frac{1}{4} \log_2 \frac{5^6 m_+ m_- \|\nu\|_+ \|\nu\|_-}{\|\nu\|^2}$$

Since $\|\nu\|_+ \|\nu\|_- \leq \|\nu\|^2$, and $m_\pm = \alpha_\pm m$, it suffices that

$$n \geq \frac{1}{2} \log_2 (125\sqrt{\alpha_+ \alpha_-} m)$$

Thus, the first two terms in (31) can be lower bounded by $0.004 \log_2(125\sqrt{\alpha_+ \alpha_-} m) \|\nu\|^2$. Substituting (31) we obtain the desired result. $\qquad\qquad\square$

References

[AB99]   M. Anthony and P.L. Bartlett. *Neural Network Learning; Theoretical Foundations.* Cambridge University Press, 1999.

[Ale90]   R. Alexander. Geometric methods in the study of irregularities of distribution. *Combinatorica*, 10(2):115–136, 1990.

[Ale91]   R. Alexander. Principles of a new method in the study of irregularities of distribution. *Invent. Math.*, 103:279–296, 1991.

[Ale94]   R. Alexander. The effect of dimension on certain geometric problems of irregularities of distribution. *Pac. J. Math.*, 165(1):1–15, 1994.

[BBD99]   P. Bartlett and S. Ben-David. On the hardness of learning with neural networks. In *Proceedings of the fourth European Conference on Computational Learning Theory*, 99.

[Bre96]   L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.

[FS96]   Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proceeding of the thirteenth international conference on machine learning*, pages 148–156, 1996.

[GR94]   I.S. Gradshteyn and I.M. Ryzhik. *Tables of Integrals, Series and Products.* Academic Press, New York, 5th edition, 1994.

[JFT00]   T. Hastie J. Friedman and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, To appear, 2000.

[JJ94]   M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[KM96]   M. Kearns and Y. Mansour. On the boosting ability og top-down decision tree learning algorithms. In *Proc. 28th ACM Symposium on the Theory of Computing,*, pages 459–468. ACM Press, 1996.

[LMF00]   J. Baxter L. Mason, P. Bartlett and M. Frean. Functional gradient techniques for combining hypotheses. In B. Schölkopf A. Smola, P. Bartlett and D. Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 2000.

[Mat99]   J. Matoušek. *Geometric Discrepancy: An Illustrated Guide*. Springer Verlag, 1999.

[MM00]   S. Mannor and R. Meir. Weak learners and improved rates of convergence in boosting. Preprint, 2000.

[Rog94]   A.D. Rogers. A functional from geometry with applications to discrepancy estimates and the radon transform. *Trans. Math. Soc.*, 341(1):275–313, 1994.

[San76]   L.A. Santaló. *Integral Geometry and Geometric Probability*. Addison-Wesley, 1976.

[Sch37]   I.J. Schoenberg. On certain metric spaces arising from euclidean spaces by change of metric and their embedding into hilbert space. *Ann. Math.*, 38:787–793, 1937.

[Sch90]   R. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[SFBL98] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[SS99]   R.E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[Vap82]   V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Verlag, New York, 1982.

[Vap95]   V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[VC71]   V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

The *Center for Communication and Information Technologies* (CCIT)

is managed by the Department of Electrical Engineering.

This Technical Report is listed also as

EE PUB #1257, August 2000.