JOURNAL OF COMPUTER AND SYSTEM SCIENCES **52**, 374–389 (1996)
ARTICLE NO. 0028

Learning Binary Perceptrons Perfectly Efficiently*

SHAO C. FANG AND SANTOSH S. VENKATESH†

Department of Electrical Engineering, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Received May 17, 1994, revised October 10, 1995

The majority rule algorithm for learning binary weights for a perceptron is analysed under the uniform distribution on inputs. It is shown that even though the algorithm is demonstrably inconsistent on random samples for very small sample sizes, it nevertheless exhibits a curious and abrupt asymptotic transition to consistency at moderate sample sizes. Particular consequences are that the algorithm PAC-learns majority functions in linear time from small samples and that, while the general variant of binary integer programming embodied here is NP-complete, almost all instances of the problem are tractable given a sufficient number of inequalities to be satisfies.

1. INTRODUCTION

We consider the problem of PAC-learning percetrons with binary, ± 1 weights and zero threshold. For simplicity, write $\mathbb{B} = \{-1, 1\}$ and consider a binary perceptron (or McCulloch–Pitts neuron) characterised by a vector of binary weights $\mathbf{w} = (w_1, ..., w_n) \in \mathbb{B}^n$. The binary perceptron accepts literals $\mathbf{u} = (u_1, ..., u_n) \in \mathbb{B}^n$ as input and produces as output a Boolean value $f_{\mathbf{w}}(\mathbf{u}) \in \mathbb{B}$ given by the sign of a linear form of the inputs:

$$f_{\mathbf{w}}(\mathbf{u}) = \operatorname{sgn}\langle \mathbf{w}, \mathbf{u} \rangle = \operatorname{sgn}\left(\sum_{i=1}^{n} w_i u_i \right) = \begin{cases} -1, & \text{if } \sum_{i=1}^{n} w_i u_i < 0, \\ +1, & \text{if } \sum_{i=1}^{n} w_i u_i \geqslant 0. \end{cases}$$

We are concerned here with learning an arbitrary function in the class of functions $\{f_{\mathbf{w}} : \mathbf{w} \in \mathbb{B}^n\}$, i.e., learning an arbitrary majority function of a set of literals.

In an equivalent formulation, each weight vector $\mathbf{w} \in \mathbb{B}^n$ determines a positive half-space of vertices

$$\mathbb{B}^n_+(\mathbf{w}) = \{\mathbf{u} \in \mathbb{B}^n : \langle \mathbf{w}, \mathbf{u} \rangle \geqslant 0\}.$$

The concept class of interest is the set of indicators for the 2^n positive half-spaces $\{\mathbb{B}^n_+(\mathbf{w})\}$ corresponding to vertices $\mathbf{w} \in \mathbb{B}^n$. We hence identify the vectors $\mathbf{w}$—henceforth called perceptrons—with the corresponding majority functions $f_{\mathbf{w}}$. Our goal is to learn an arbitrary *target perceptron* (also called the *solution vector*) $\mathbf{w}^s \in \mathbb{B}^n$ from examples drawn at random from the vertices $\mathbb{B}^n$ of the cube.

Suppose $\mathbf{w}^s \in \mathbb{B}^n$ is some fixed (but unknown) target perceptron. We assume that the learner is provided with a random sample $U = \{\mathbf{u}^1, ..., \mathbf{u}^m\}$ of points in $\mathbb{B}^n$, together with the labels $l(\mathbf{u}^\alpha) = f_{\mathbf{w}^s}(\mathbf{u}^\alpha)$, $1 \leqslant \alpha \leqslant m$ induced by the unknown $\mathbf{w}^s$. *We will assume throughout that the examples are chosen independently from the uniform distribution on* $\mathbb{B}^n$. An easy consequence is that drawing labeled examples uniformly from $\mathbb{B}^n$ is equivalent to prescribing only *positive* examples $\hat{\mathbf{u}}$ from $\mathbb{B}^n_+ \triangleq \mathbb{B}^n_+(\mathbf{w}^s)$ obtained by reflecting every negative example $\mathbf{u}$ about the origin and relabelling it positive. The induced marginal distribution of positive examples $\hat{\mathbf{u}}$ is then uniform on $\mathbb{B}^n_+$.[1] Thus, in our considerations, we may replace the random m-set of positively and negatively labelled examples $U = \{\mathbf{u}^1, ..., \mathbf{u}^m\}$ by an equivalent m-set $\hat{U} = \{\hat{\mathbf{u}}^1, ..., \hat{\mathbf{u}}^m\}$ of positive examples drawn independently from the uniform distribution of points in the fixed positive half-space $\mathbb{B}^n_+$.

As a first step in the learning problem, we seek a solution to the problem of *loading* the set of examples, i.e., finding a perceptron $\mathbf{w} \in \mathbb{B}^n$ such that

$$\operatorname{sgn}\left(\sum_{i=1}^{n} w_i u_i^\alpha \right) = l(\mathbf{u}^\alpha), \qquad \alpha = 1, ..., m.$$

Equivalently, in terms of the reflected (positive) examples $\hat{U}$, the problem is to find a perceptron $\mathbf{w} \in \mathbb{B}^n$ solving the following set of linear inequalities:

$$\sum_{i=1}^{n} w_i \hat{u}_i^\alpha \geqslant 0, \qquad \alpha = 1, ..., m. \tag{1}$$

* This research was supported in part by the Air Force Office of Scientific Research under grants F49620-93-1-0120 and F49620-92-J-0344. Presented in part at the *Sixth ACM Conference on Computational Learning Theory*, *Santa Cruz*, *California*, *July*, 1993.

† E-mail address: fang@ee.upenn.edu; venkatesh@ee.upenn.edu.

[1] To be precise, this statement holds as stated only for *odd n*. For *even n* there is a probability $\mathcal{O}(n^{-1/2})$ of an example landing on the hyperplane orthogonal to $\mathbf{w}^s$. Such boundary effects will be negligible for large n.

0022-0000/96 $18.00

If the number m of examples is large enough, with high probability there is a *unique* solution to (1) given by the target perceptron $\mathbf{w}^s$ according to which the examples are drawn. Thus, if the sample complexity is large enough, loading the examples would be equivalent to learning the underlying majority function *exactly*; or, to put it in another way, with a sufficient number of examples *perfect* generalisation can be achieved, at least in principle. How many examples are required? Lyuu and Rivin [11] show that only about $1.45n$ examples are needed if time complexity is not an issue; in particular, if at least $1.45n$ examples are drawn from the uniform distribution on $\mathbb{B}^n$ and labelled by a target perceptron $\mathbf{w}^s \in \mathbb{B}^n$, the probability that there exists any other binary perceptron consistent with the examples is exponentially small. Thus, an exponential search through the vertices of the cube to find a perceptron $\mathbf{w}$ consistent on the examples will result in identification of the target perceptron with high confidence if m exceeds $1.45n$.

Can we find a polynomial time algorithm which guarantees learning with such small sample complexities? On the face of it, it appears unlikely. As we have seen, the problem of finding a binary perceptron $\mathbf{w}$ consistent on the examples requires a solution to the system of inequalities (1) with the components of the sought after n-dimensional vector $\mathbf{w} = (w_1, ..., w_n)$ constrained to $\mathbb{B}$. This is a binary integer programming problem, known to be NP-complete [12].[2] We might hence anticipate that there is an intractable worst-case for any algorithm. In this exposition we demonstrate, however, that when examples are drawn from the uniform distribution on the vertices of the cube, the very simple majority rule algorithm (Venkatesh [14]; also referred to as the clipped Hebb rule by Köhler *et al.* [10]) PAC-learns majority functions or perceptrons with binary weights while achieving the desideratum of low time and sample complexities.[3] This algorithm has also been the subject of recent nonrigourous investigations by Golea and Marchand [8]; Sections 5 and 6 contain brief discussions of their approach.

The paper is organised as follows. We first summarise our notation and for later ease collect the relevant technical facts that will be needed in Section 2. The majority rule algorithm is described in Section 3. Two lemmas central to the proof of the main theorems are proved in Section 4; the main results themselves are contained in Section 5 which deals with the issue of perfect generalisation and in Section 6 which deals with PAC-learning.

[2] Pitt and Valiant [12] proved the NP-completeness of learning Boolean threshold functions (i.e., perceptrons with 0/1 weights and integer thresholds) by reducing zero-one-integer programming as posed in Garey and Johnson [7, p. 245] to the learning problem. It is straightforward to reduce the 0/1-weight, integer-threshold problem to the ± 1-weight, zero-threshold problem considered here, thus proving the NP-completeness of the latter [2].

[3] Majority Rule is a linear time, off-line algorithm. For a randomised, on-line approach to the problem, see the Directed Drift algorithm of Venkatesh [14, 15].

2. PRELIMINARIES

As already indicated, we use $\mathbb{B}$ to denote the set $\{-1, 1\}$, with $\mathbb{B}^n = \{-1, 1\}^n$ the vertices of the n-cube. All logarithms are to base e. Also, ϕ denotes the standard Gaussian density, $\phi(x) = (1/\sqrt{2\pi})\, e^{-x^2/2}$, while Φ denotes the Gaussian distribution function, $\Phi(x) = \int_{-\infty}^{x} \phi(y)\, dy$, with Φ^{-1} its inverse. For any positive integer t, we adopt the "falling factorial" notation $(k)_t = k(k-1)\cdots(k-t+1)$. Also, for any real x, we denote by $\lfloor x \rfloor$ the greatest integer less than or equal to x, and by $\lceil x \rceil = -\lfloor -x \rfloor$ the smallest integer bigger than or equal to x. Throughout, $\mathbb{P}$ stands for probability measure on the underlying probability space and $\mathbb{E}$ denotes expectation. Finally, we use standard asymptotic order notation with the following caveats perhaps worth remarking: if $\{h_n\}$ and $\{g_n\}$ denote real sequences, by $h_n = \mathcal{O}(g_n)$ we mean that $|h_n|/|g_n|$ is bounded above; in particular, sign information is explicitly jettisoned in our use of the order notation. In addition, we will find it expedient to denote $h_n \ll g_n$ to mean $h_n = o(g_n)$ and $h_n \gg g_n$ to mean $h_n = \omega(g_n)$.

The following is a collection of known technical facts that will be needed in the exposition. Proofs are omitted.

Fact 1 (Taylor expansions of the natural logarithm). The series expansions

$$\log \frac{1}{1-t} = \sum_{j=1}^{\infty} \frac{t^j}{j}, \qquad \log \frac{1+t}{1-t} = 2 \sum_{j=1}^{\infty} \frac{t^{2j-1}}{2j-1}.$$

hold for every t in the range $|t| < 1$.

Fact 2 (Approximations of sums by integrals). If $f(x)$ is a monotonically increasing function of its argument, then

$$\int_{m-1}^{n} f(x)\, dx \leqslant \sum_{k=m}^{n} f(k) \leqslant \int_{m}^{n+1} f(x)\, dx.$$

Fact 3 (Stirling's formula). The inequalities

$$\sqrt{2\pi}\, n^{n+1/2} e^{-n} \cdot e^{(12n+1)^{-1}} < n! < \sqrt{2\pi}\, n^{n+1/2} e^{-n} \cdot e^{(12n)^{-1}}$$

hold for every positive integer n. In particular,

$$n! \sim \sqrt{2\pi}\, n^{n+1/2} e^{-n}$$

as $n \to \infty$.

Fact 4 (The central term of the binomial). As $n \to \infty$,

$$2^{-n} \binom{n}{\lfloor n/2 \rfloor} \sim \sqrt{2/\pi n}.$$

Fact 5 (Hoeffding's bound for the binomial tail). Let S_n denote the number of heads in n tosses of a fair coin. Then

$$\mathbb{P}\{S_n - n/2 \geqslant h\} \leqslant e^{-2h^2/n}$$

for every positive h.

Fact 6 (Large deviation central tendency in the random walk tail). Let $\{X_\alpha, 1 \leqslant \alpha \leqslant m\}$ be an independent, identically distributed sequence of Bernoulli random variables taking values $+1$ with probability p and -1 with probability $q = 1 - p$. (We allow p to depend on m but assume it is bounded away from both 0 and 1.) Consider the random walk $R_m = X_1 + \cdots + X_m$ and let

$$R_m^* = \frac{R_m - m(p-q)}{\sqrt{4mpq}}$$

denote the normalised walk. Then,

$$\mathbb{P}\{R_m^* < -h_m\} \sim \Phi(-h_m) \qquad (m \to \infty)$$

for any positive sequence $\{h_m\}$ satisfying $h_m = o(m^{1/6})$ as $m \to \infty$. If, in addition, $h_m \to \infty$, i.e., $1 \ll h_m \ll m^{1/6}$, then

$$\mathbb{P}\{R_m^* < -h_m\} \sim \frac{1}{h_m}\phi(h_m) \qquad (m \to \infty)$$

in view of the following estimate for the Gaussian tail.

Fact 7 (The Gaussian tail). As $x \to \infty$,

$$\Phi(-x) \sim x^{-1}\phi(x);$$

more precisely, the double inequality

$$[x^{-1} - x^{-3}]\,\phi(x) < \Phi(-x) < x^{-1}\phi(x)$$

holds for every $x > 0$.

Fact 8 (Markov's inequality). Let H be any nonnegative random variable with finite expectation. Then the inequality

$$\mathbb{P}\{H \geqslant h\} \leqslant \frac{\mathbb{E}H}{h}$$

holds for every positive h. In particular, if H assumes only nonnegative integer values, then

$$\mathbb{P}\{H \neq 0\} = \mathbb{P}\{H \geqslant 1\} \leqslant \mathbb{E}H.$$

Fact 9 (Bonferroni's inequalities). Let $A_1, ..., A_L$ be measurable subsets of a probability space. For $1 \leqslant k \leqslant L$, let $\mathfrak{s}_k$ be the sum of probabilities of all sets formed by intersecting k of the $A_1, ..., A_L$:

$$\mathfrak{s}_k = \sum_{1 \leqslant j_1 < j_2 < \cdots < j_k \leqslant L} \mathbb{P}\left(\bigcap_{i=1}^{k} A_{j_i}\right).$$

Then for every K, $1 \leqslant K \leqslant L$,

$$\mathbb{P}\left(\bigcup_{i=1}^{L} A_i\right) = \sum_{k=1}^{K} (-1)^{k-1}\,\mathfrak{s}_k + (-1)^K E_K$$

where $E_k \geqslant 0$.

Facts 1 and 2 are elementary. Proofs of the other results can be easily derived from similar results in the classic text of Feller [5].

3. MAJORITY RULE

Without loss of generality, assume the target perceptron is defined by

$$\mathbf{w}^s = (\underbrace{1, ..., 1}_{n}).$$

Let $U = \{\mathbf{u}^1, ..., \mathbf{u}^m\}$ be any m-set of vertices in $\mathbb{B}^n$, and, for each $\mathbf{u} \in U$, let

$$l(\mathbf{u}) = \begin{cases} -1 & \text{if } \langle \mathbf{w}^s, \mathbf{u} \rangle < 0, \\ +1 & \text{if } \langle \mathbf{w}^s, \mathbf{u} \rangle \geqslant 0, \end{cases}$$

denote the labels of the points induced by $\mathbf{w}^s$. (Recall that $\langle \mathbf{w}^s, \mathbf{u} \rangle \triangleq \sum_{i=1}^n w_i^s u_i = \sum_{i=1}^n u_i$.) The majority rule algorithm (cf. Venkatesh [14]) prescribes the weights w_i of a perceptron as a function of the labelled sample U as follows:

For $i = 1, ..., n$, let

$$U_i^+ = \{\mathbf{u} \in U : u_i = +l(\mathbf{u})\},$$
$$U_i^- = \{\mathbf{u} \in U : u_i = -l(\mathbf{u})\}.$$

Set

$$w_i = \begin{cases} +1 & \text{if } |U_i^+| \geqslant |U_i^-|, \\ -1 & \text{if } |U_i^+| < |U_i^-|. \end{cases}$$

The motivation behind the algorithm is easy to see if we consider the set of vertices $\hat{U} = \{l(\mathbf{u})\mathbf{u} : \mathbf{u} \in U\}$ obtained by reflecting negatively labelled examples about the origin. Clearly, finding a perceptron consistent[4] on the original set

[4] A learning algorithm is *consistent on a set of examples* if the hypothesis produced by the algorithm labels all the examples in the set correctly (i.e., for each example, the label generated by the hypothesis is consistent with that generated by the underlying target concept); the algorithm is *consistent* if it is consistent on any set of examples.

of examples U is completely equivalent to finding a perceptron consistent on the set of *positive* examples $\hat{U}$. Writing $\hat{\mathbf{u}} = l(\mathbf{u})\mathbf{u}$, we see that for the perceptron $\mathbf{w}$ generated by majority rule, $w_i = +1$ if the (reflected) points $\hat{\mathbf{u}}$ whose ith component is $+1$ are in the majority, and $w_i = -1$ otherwise. The motivation is readily seen from (1): each summand $w_i \hat{u}_i^\alpha$ is more likely to be positive so that the whole sum $\sum_i w_i \hat{u}_i^\alpha$ is also more likely to be positive.

Majority rule is a *local* algorithm; specifically, the ith component of $\mathbf{w}$ depends solely on the ith components of the input patterns in U and the desired labelling (output). The algorithm is also *homogeneous*; that is, the same procedure is employed to determine every component w_i of the perceptron $\mathbf{w}$. Locality and homogeneity are clearly desirable features contributing to a low complexity of specification. The algorithm requires $n(m-1)$ additions and n comparisons, so that it has time complexity linear in nm, the number of bits needed to specify the examples. Being an off-line algorithm, majority rule has a space requirement linear in nm.

Note that we can also write

$$\mathbf{w} = \operatorname{sgn}\left(\sum_{\alpha=1}^{m} l(\mathbf{u}^\alpha)\mathbf{u}^\alpha\right) = \operatorname{sgn}\left(\sum_{\alpha=1}^{m} \hat{\mathbf{u}}^\alpha\right),$$

where the sign operation on a vector is interpreted in a natural fashion as the vector obtained by taking the sign of each of the components. Thus, equivalently, majority rule can be interpreted as seeking a vertex close to the centre of mass of the (positive) sample $\hat{U}$. Hence, if $\hat{U}$ is "symmetric" about $\mathbf{w}^s$ then majority rule is guaranteed to return the target perceptron as is illustrated, for example, in the following result.

PROPOSITION 1. *If* $\hat{U} = \mathbb{B}^n_+$ *then* $\mathbf{w} = \mathbf{w}^s$.

Proof. Partition the positive half-space $\mathbb{B}^n_+$ into disjoint sets A_h, $0 \leqslant h \leqslant n/2$, where A_h consists of those vertices in $\mathbb{B}^n_+$ which have precisely h components taking value -1 and $n-h$ components taking value $+1$. Clearly, $|A_h| = \binom{n}{h}$. Consider majority rule applied to the ith component. The number of vertices in A_h whose ith component is "good," i.e., $+1$, is $\binom{n-1}{h}$, while the remaining vertices in A_h whose ith component is "bad," i.e., -1, number $\binom{n}{h} - \binom{n-1}{h} = \binom{n-1}{h-1}$ by an application of Pascal's triangle. Consequently,

$$\sum_{\hat{u} \in \mathbb{B}^n_+} \hat{u}_i = \sum_{h=0}^{\lfloor n/2 \rfloor} \sum_{\hat{u} \in A_h} \hat{u}_i = \sum_{h=0}^{\lfloor n/2 \rfloor} \left[\binom{n-1}{h} - \binom{n-1}{h-1}\right] = \binom{n-1}{\lfloor n/2 \rfloor}.$$

It follows that

$$w_i = \operatorname{sgn}\binom{n-1}{\lfloor n/2 \rfloor} = 1.$$

As this holds for any i, we have $\mathbf{w} = \mathbf{w}^s$. ∎

Of course, there is no guarantee that a given sample U will have such symmetries—much depends on the statistical model for example generation which we specify next.

We henceforth assume that examples $\mathbf{u}$ are independently generated from the uniform distribution on the vertices $\mathbb{B}^n$ of the cube and labelled positive or negative according to whether they lie in the positive or negative half-space of $\mathbf{w}^s$. Note that, if $l(\mathbf{u})$ denotes the label of a random example $\mathbf{u}$, then the marginal distribution of the (reflected) example $\hat{\mathbf{u}} = l(\mathbf{u})\mathbf{u}$ is uniform over the positive half-space $\mathbb{B}^n_+$ if n is odd, and "almost" uniform on $\mathbb{B}^n_+$ if n is even; the caveat "almost" for the case n even is thrown in to account for examples falling on the hyperplane orthogonal to $\mathbf{w}^s$, which event has asymptotically vanishing probability $2^{-n}\binom{n}{n/2} = \Theta(n^{-1/2})$.

The basic components of the analysis involve the use of a probabilistic sieve coupled with careful asymptotic calculations. The major complicating factor is that statistical dependencies, albeit somewhat weak, abound in the problem, and a considerable portion of the effort goes into quelling these dependencies with a firm hand.

EXAMPLE. *Dependencies across components.* Take $n = 5$ and $\mathbf{w}^s = (1\ 1\ 1\ 1\ 1)$. Denoting $+1$ simply by $+$ and -1 by $-$, the following is the complete set of points $\hat{\mathbf{u}}$ in $\mathbb{B}^n_+$:

$$\left\{\begin{array}{l}(+\,+\,+\,+\,+),\\ (+\,+\,+\,+\,-),\ (+\,+\,+\,-\,+),\ (+\,+\,-\,+\,+),\ (+\,-\,+\,+\,+),\ (-\,+\,+\,+\,+),\\ (+\,+\,+\,-\,-),\ (+\,+\,-\,+\,-),\ (+\,-\,+\,+\,-),\ (-\,+\,+\,+\,-),\ (+\,+\,-\,-\,+),\\ (+\,-\,+\,-\,+),\ (-\,+\,+\,-\,+),\ (+\,-\,-\,+\,+),\ (-\,+\,-\,+\,+),\ (-\,-\,+\,+\,+)\end{array}\right\}.$$

Since n is odd, generating examples $\mathbf{u}$ uniformly from $\mathbb{B}^n$ is completely equivalent to generating (reflected) examples $\hat{\mathbf{u}}$ uniformly from $\mathbb{B}^n_+$. Consequently, each positive example $\hat{\mathbf{u}}$ has probability $\frac{1}{16}$. Suppose we pick one point $\hat{\mathbf{u}}$ randomly from this set and apply majority rule to this singleton positive example. Then $\mathbf{w} = \hat{\mathbf{u}}$, and we have the following identities: for any choice $1 \leqslant i \leqslant n$,

$$\mathbb{P}\{w_i = w_i^s\} = \mathbb{P}\{\hat{u}_i = 1\} = \tfrac{11}{16};$$

if $i \neq j$, then

$$\mathbb{P}\{w_i = w_i^s \mid w_j = w_j^s\} = \mathbb{P}\{\hat{u}_i = 1 \mid \hat{u}_j = 1\} = \tfrac{7}{11} \neq \mathbb{P}\{\hat{u}_i = 1\};$$

for any three distinct indices i, j, and k,

$$\mathbb{P}\{w_i = w_i^s \mid w_j = w_j^s, w_k = w_k^s\} = \mathbb{P}\{\hat{u}_i = 1 \mid \hat{u}_j = \hat{u}_k = 1\} = \tfrac{4}{7};$$

for any four distinct indices i, j, k, and l,

$$\mathbb{P}\{w_i = w_i^s \mid w_j = w_j^s, w_k = w_k^s, w_l = w_l^s\} = \mathbb{P}\{\hat{u}_i = 1 \mid \hat{u}_j = \hat{u}_k = \hat{u}_l = 1\} = \tfrac{1}{2};$$

and, finally,

$$\mathbb{P}\{w_i = w_i^s \mid w_j = w_j^s, w_k = w_k^s, w_l = w_l^s, w_r = w_r^s\} = \mathbb{P}\{\hat{u}_i = 1 \mid \hat{u}_j = \hat{u}_k = \hat{u}_l = \hat{u}_r = 1\} = \tfrac{1}{2}$$

for any five distinct indices i, j, k, l, and r.

The dependence across components complicates analysis of the algorithm. However, we will see that the dependence is *weak*. In fact, the components are independent *asymptotically*.

4. TWO LEMMAS

Consider a random pattern $\mathbf{u}$ chosen from the uniform distribution on $\mathbb{B}^n$, and let $l(\mathbf{u})$ be its label induced by $\mathbf{w}^s$. As before, let $\hat{\mathbf{u}} = l(\mathbf{u})\mathbf{u}$ denote the corresponding (reflected) point in the positive half-space $\mathbb{B}^n_+$ of $\mathbf{w}^s$. Let the random variable X denote the number of component matches between $\hat{\mathbf{u}}$ and $\mathbf{w}^s$; i.e.,

$$X = |\{i : \hat{u}_i = w_i^s, 1 \leqslant i \leqslant n\}|.$$

Equivalently, $X = n - D$, where D is the Hamming distance between $\mathbf{w}^s$ and $\hat{\mathbf{u}}$.

LEMMA 1. *The asymptotic estimate*

$$\mathbb{E}X = \left\lceil \frac{n}{2} \right\rceil + \frac{\sqrt{n}}{\sqrt{2\pi}} + \mathcal{O}(1)$$

holds as $n \to \infty$. In particular,

$$\frac{\mathbb{E}X}{n} = \frac{1}{2} + \frac{1}{\sqrt{2\pi n}} + \mathcal{O}\left(\frac{1}{n}\right)$$

as $n \to \infty$.

Proof. We have

$$\begin{aligned}\mathbb{E}X &= \sum_{k=\lceil n/2 \rceil}^{n} k\,\mathbb{P}\{X = k\} \\ &= \sum_{a=0}^{\lfloor n/2 \rfloor} \left(\left\lceil \frac{n}{2} \right\rceil + a \right) \mathbb{P}\left\{ X = \left\lceil \frac{n}{2} \right\rceil + a \right\} \\ &= \left\lceil \frac{n}{2} \right\rceil + 2^{-(n-1)} \sum_{a=1}^{\lfloor n/2 \rfloor} a \binom{n}{\lceil n/2 \rceil + a} \triangleq \left\lceil \frac{n}{2} \right\rceil + v.\end{aligned}$$

Fix $0 < \varepsilon < \frac{1}{8}$ and partition the sum for v into two terms, $v = \sum' + \sum''$, where the summation index for the first sum varies in the range $1 \leqslant a \leqslant n^{1/2+\varepsilon}$, while the summation index for the second sum varies over the residual range $n^{1/2+\varepsilon} < a \leqslant n/2$. We now proceed to estimate the two sums.

Let $S_n \sim \text{Binomial}(n, 1/2)$ denote the number of successes in n tosses of a fair coin. We have

$$\begin{aligned}\sum'' &\leqslant n \left[2^{-n} \sum_{n^{1/2+\varepsilon} < a \leqslant n/2} \binom{n}{\lceil n/2 \rceil + a} \right] \\ &= n\,\mathbb{P}\left\{ S_n > \left\lceil \frac{n}{2} \right\rceil + n^{1/2+\varepsilon} \right\} \leqslant n e^{-2n^{2\varepsilon}},\end{aligned}$$

the last inequality following by a direct application of Hoeffding's inequality to the tail of the binomial (Fact 5).

Consider the range of indices $1 \leqslant a \leqslant n^{1/2+\varepsilon}$ and recall that $0 < \varepsilon < \frac{1}{8}$ is held fixed. For these choices of a and ε, an application of Stirling's formula (Fact 3) yields

$$\begin{aligned}2^{-(n-1)} a \binom{n}{\lceil n/2 \rceil + a} &= 2^{-(n-1)} \frac{a n!}{(\lceil n/2 \rceil + a)!\,(\lfloor n/2 \rfloor - a)!} \\ &= \frac{2\sqrt{2}\,a}{\sqrt{\pi n}} e^{-2a^2/n + \mathcal{O}(a^4/n^3) + \mathcal{O}(1/n)} \\ &= \frac{2\sqrt{2}\,a}{\sqrt{\pi n}} e^{-2a^2/n} (1 + \mathcal{O}(n^{-1+4\varepsilon})),\end{aligned}$$

as $a^4/n^3 \leqslant n^{-1+4\varepsilon}$ which dominates the order n^{-1} term. It follows that

$$v = \left[\sum_{1 \leqslant a \leqslant n^{1/2+\varepsilon}} \frac{2\sqrt{2}\,a}{\sqrt{\pi n}} e^{-2a^2/n} \right] (1 + \mathcal{O}(n^{-1+4\varepsilon})) + \mathcal{O}(n e^{-2n^{2\varepsilon}}).$$

The sum enclosed in square parentheses does not have a simple closed form, but can be precisely estimated for large n by bracketing the sum with integrals (Fact 2) which are easy to evaluate. Note that, viewed as a function of a real parameter, the function

$$f(a) = \frac{2\sqrt{2}\,a}{\sqrt{\pi n}} e^{-2a^2/n}$$

has a unique maximum at $a = \sqrt{n}/2$. In particular, $f(a)$ increases monotonically in the range $1 \leqslant a < \sqrt{n}/2$, while $-f(a)$ increases monotonically in the range $-n^{1/2+\varepsilon} \leqslant a \leqslant -\sqrt{n}/2$. Two applications of Fact 2 in each of these two ranges leads to the estimate

$$[\,\cdot\,] = \frac{\sqrt{n}}{\sqrt{2\pi}} + \mathcal{O}(1).$$

Consequently,

$$v = \frac{\sqrt{n}}{\sqrt{2\pi}} + \mathcal{O}(1) \qquad (n \to \infty).$$

To complete the proof, note that replacing $\lceil n/2 \rceil$ by $n/2$ occasions an error of at most $\frac{1}{2}$ which is absorbed in the $\mathcal{O}(1)$ term. ∎

Note that the sign of each component of $\hat{\mathbf{u}}$ matches that of the corresponding component of $\mathbf{w}^s$ with probability $(1/n)\,\mathbb{E}X$. Now consider an independently drawn m-set of examples $U = \{\mathbf{u}^\alpha, 1 \leqslant \alpha \leqslant m\}$ and the corresponding m-set of (reflected) examples $\hat{U}$. Majority rule then generates the random vector $\mathbf{w} = \operatorname{sgn}(\sum_{\alpha=1}^m \hat{\mathbf{u}}^\alpha)$. Let the random variable H denote the Hamming distance between $\mathbf{w}$ and $\mathbf{w}^s$:

$$H = |\{i\colon w_i \neq w_i^s\}|.$$

The following is the central estimate we will need. We consider the situation, as $n \to \infty$, when $m = m_n$ is allowed to depend on n. For notational economy, we drop the subscript n and write simply m while keeping in mind that the number of examples is a function of n.

LEMMA 2. *If m increases with n such that $m = o(n^{3/2})$, then*

$$\mathbb{E}H \sim n\Phi(-\sqrt{2m/\pi n}) \qquad (n \to \infty).$$

If, in addition, $m = \omega(n)$, then

$$\mathbb{E}H \sim \frac{n^{3/2}}{2m^{1/2}}\, e^{-m/\pi n}$$

as $n \to \infty$.

Proof. Using Lemma 1, define

$$p \triangleq \mathbb{P}\{\hat{u}_i^\alpha = w_i^s\} = \frac{\mathbb{E}X}{n} = \frac{1}{2} + \frac{1}{\sqrt{2\pi n}} + \mathcal{O}\left(\frac{1}{n}\right),$$

$$q \triangleq \mathbb{P}\{\hat{u}_i^\alpha \neq w_i^s\} = 1 - p = \frac{1}{2} - \frac{1}{\sqrt{2\pi n}} + \mathcal{O}\left(\frac{1}{n}\right).$$

(Note that the order terms for p and q are equal and of opposite sign.) As the examples are drawn independently, for any i, the random sum $R_m = \sum_{\alpha=1}^m \hat{u}_i^\alpha$ represents a random walk with mean

$$\mathbb{E}R_m = m(p-q) = \frac{\sqrt{2}\,m}{\sqrt{\pi n}} + \mathcal{O}\left(\frac{m}{n}\right),$$

and variance

$$\operatorname{Var} R_m = 4mpq = m - \frac{2m}{\pi n} + \mathcal{O}\left(\frac{m}{n^{3/2}}\right) \qquad (n \to \infty).$$

With

$$R_m^* = \frac{R_m - m(p-q)}{\sqrt{4mpq}}$$

denoting the normalised random walk, we have

$$\mathbb{P}\{w_i \neq w_i^s\} = \mathbb{P}\{R_m < 0\} = \mathbb{P}\left\{R_m^* < -\frac{m(p-q)}{\sqrt{4mpq}}\right\}$$
$$= \mathbb{P}\left\{R_m^* < -\frac{\sqrt{2m}}{\sqrt{\pi n}} + \mathcal{O}\left(\frac{\sqrt{m}}{n}\right)\right\}.$$

For $m = o(n^{3/2})$, the requisite deviation from the mean is within the permissible range in the large deviation central limit theorem (Fact 6). We hence have

$$\mathbb{P}\{w_i \neq w_i^s\} \sim \Phi\left(-\frac{\sqrt{2m}}{\sqrt{\pi n}} + \mathcal{O}\left(\frac{\sqrt{m}}{n}\right)\right)$$
$$\sim \Phi\left(-\frac{\sqrt{2m}}{\sqrt{\pi n}}\right) \qquad (n \to \infty).$$

Consequently,

$$\mathbb{E}H = n\,\mathbb{P}\{w_i \neq w_i^s\} \sim n\Phi\left(-\frac{\sqrt{2m}}{\sqrt{\pi n}}\right) \qquad (n \to \infty).$$

The proof is completed by noting that the Gaussian tail estimate (Fact 7) applied to the above result is sharp if the range of interest is $n \ll m \ll n^{3/2}$. ∎

Remark. Note that the events $\{\{w_i \neq w_i^s\}, 1 \leqslant i \leqslant n\}$ are *exchangeable*, a fact which follows directly from the mode of generation of the examples.

5. PERFECT GENERALISATION

Let us begin with an ambitious question. Does majority rule generalise perfectly, i.e., produce the target perceptron as hypothesis, for a large enough sample size? (In the PAC setting, perfect generalisation corresponds to zero error and confidence one.) After all, the result of Lyuu and Rivin [11] shows that about $1.45n$ randomly chosen examples suffice to uniquely *identify* the target binary perceptron; thus, perfect generalisation is achievable with linear sample complexity for any consistent algorithm, such as, for instance, exhaustive search.[5] The catch here is that majority rule, in common with many low complexity heuristics, is *not* consistent. This observation has led Golea and Marchand in

[5] Literally and figuratively. This calls for an exponential time search.

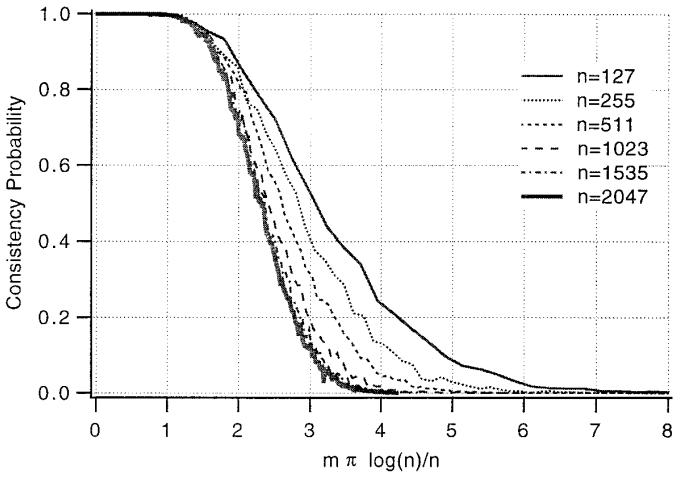

FIG. 1. The first threshold: consistency probability (or the probability of loading a random sample) versus normalised sample size. The algorithm fails to be consisting beyond the first threshold.

a recent paper [8] to conclude that majority rule cannot generalise perfectly.

In fact, the situation is rather worse than it may seem at first glance: when the sample size m exceeds $n/\pi \log n$, majority rule is (asymptotically) *guaranteed* to be inconsistent on the examples! Empirical evidence for this phenomenon is presented in Fig. 1, where an empirical estimate of the probability that majority rule is consistent on a random m-sample is plotted versus $m/(n/\pi \log n)$ for various values of n. The details of the computer simulations are as follows: the probability that majority rule is consistent on a random m-sample (for a fixed value of n) was estimated by the relative frequency of the number of times majority rule was consistent on a (pseudo-) random m-sample averaged over 1000 independent runs; the plots were generated for each n by varying m so that the normalised value $m/(n/\pi \log n)$ varied in the range 0 to 8.

Note the threshold phenomenon that develops around $m = n/\pi \log n$ when n becomes large. In particular, for any fixed small $\zeta > 0$ and large n, if $m = (1-\zeta)(n/\pi \log n)$, then the perceptron generated by majority rule classifies all the m examples correctly (i.e., the algorithm is consistent on the examples) for almost all choices of the examples; if, however, $m = (1+\zeta)(n/\pi \log n)$, then the perceptron generated by majority rule is guaranteed to classify at least one example incorrectly (i.e., the algorithm is inconsistent on the set of examples), again for almost all choices of the examples. The formal demonstration of this result is subtle and surprisingly messy and will appear elsewhere [4]. (For one half of the result see Venkatesh [14].) For our purposes here, note that neither of the two regimes in the figure appears particularly helpful for the purposes of learning the target perceptron: when $m = (1-\zeta)(n/\pi \log n)$, the algorithm is consistent on the examples with high probability, but the sample size is sublinear and not large enough to identify the target perceptron uniquely; and when $m = (1+\zeta)(n/\pi \log n)$, the algorithm is inconsistent on the set of examples itself, which does not augur well for generalisation, much less perfect generalisation.

In fact, as we will see shortly, this situation persists with a positive error measure between the target perceptron and the perceptron generated by majority rule as long as the sample size is linear in n. However, a most remarkable second threshold phenomenon develops around a sample complexity $m = \pi n \log n$ very slightly in excess of linear in n. Again, empirical evidence for this phenomenon is presented in Fig. 2, where an empirical estimate of the probability that majority rule is consistent on a random m-sample is plotted versus $m/(\pi n \log n)$ for various values of n. The details of the simulations are the same as those for Fig. 1, except that there is a different normalisation of the x-axis and the plots for each n were generated by varying m so that $m/(\pi n \log n)$ varied in the range 0 to 2. The behaviour of the algorithm in this range is the obverse of what was observed in Fig. 1. Note that the first threshold is still visible albeit scrunched up near the y-axis as a consequence of the different normalisation of the sample size.

The complete (asymptotic) picture is as indicated schematically in Fig. 3. The algorithm is consistent (asymptotically) when the sample size is less than the first threshold of $n/\pi \log n$, fails abruptly, i.e., is not consistent on the sample, when the sample size exceeds the first threshold, but recovers triumphantly from the dead when the sample size exceeds the second threshold of $\pi n \log n$. We dub this idiosyncratic behaviour *asymptotic consistency*. (While the finite n simulations of Figs. 1 and 2 indicate the emergence of two thresholds for large n, it is not immediately clear, perhaps, that the two thresholds are manifested *exactly* at

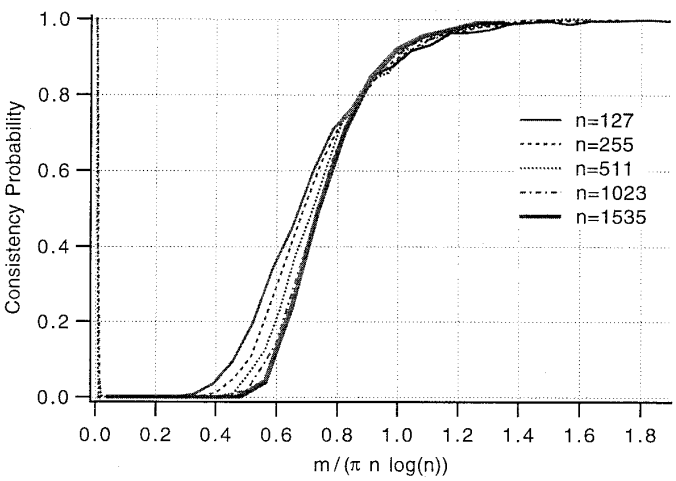

FIG. 2. The second threshold: consistency probability (or the probability of loading a random sample) versus normalised sample size. The algorithm sees the error of its ways abruptly at the second threshold.

$n/\pi \log n$ and $\pi n \log n$, respectively, as emphasised in Fig. 3; nor is it clear that the behaviour changes as abruptly at the thresholds as indicated. We will shortly place the second threshold phenomenon, which is of direct relevance to our problem, on a rigourous footing; a rigourous demonstration of the first threshold phenomenon appears elsewhere [4, 14].)

As already remarked, the first threshold, while interesting in its own right, does not have direct useful implications—except possibly pejoratively—to learning the target perceptron. However, the situation becomes much more interesting at the second threshold: the sample complexity $\pi n \log n$ at which consistency secures so remarkably exceeds the $1.45n$ information-theoretic estimate of Lyuu and Rivin (albeit only by a logarithmic amount) of the sample complexity at which the target perceptron is identified uniquely. Consequently, this implies that majority rule will in fact generate the target perceptron, i.e., generalise perfectly, when m exceeds $\pi n \log n$. (Of course, this also refutes the assertion of Golea and Marchand; see Section 6 for a brief discussion.) We formalise this assertion in the following theorem which is the main result of this section.

THEOREM 1. *The sequence $\pi n \log n$ is a threshold function for the perfect generalisation attribute that majority rule generates the target perceptron exactly, i.e., that $\mathbf{w} = \mathbf{w}^s$. More specifically, for every fixed $0 < \zeta < 1$, the following assertions hold*:

1. (Necessity). *If m grows with n such that $m \leqslant (1-\zeta)\,\pi n \log n$, then $\mathbb{P}(\mathbf{w} = \mathbf{w}^s\} \to 0$ as $n \to \infty$;*

2. (Sufficiency). *If m grows with n such that $m \geqslant (1+\zeta)\,\pi n \log n$, then $\mathbb{P}\{\mathbf{w} = \mathbf{w}^s\} \to 1$ as $n \to \infty$.*

We will devote the rest of this section to proving the theorem. To simplify notation, we take n odd for definiteness and agree, as before, to reflect all negative examples in U about the origin and relabel them as positive. The induced marginal distribution of the (reflected) examples $\hat{U}$ is uniform in $\mathbb{B}^n_+$ by symmetry (to each positive example there exists a unique negative example, its reflection, with both points having probability 2^{-n} in the original uniform distribution on $\mathbb{B}^n$). Majority rule applied to the sample $\hat{U}$ results in the (random) perceptron

$$\mathbf{w} = \operatorname{sgn}\left(\sum_{\alpha=1}^{m} \hat{\mathbf{u}}^\alpha\right).$$

As before, we denote by H the Hamming distance between $\mathbf{w}$ and the targer perceptron $\mathbf{w}^s$.

The proof of the theorem rests upon a Poissonisation argument which shows that the number of indices i for which $\{w_i \neq w_i^s\}$ is asymptotically Poisson. One half of the theorem (Sufficiency), however, is a ready consequence of Markov's inequality and we consider this case first. Suppose m increases with n such that $n \ll m \ll n^{3/2}$. Then

$$\mathbb{P}\{\mathbf{w} \neq \mathbf{w}^s\} = \mathbb{P}\{H > 0\} \leqslant \mathbb{E}H \sim \frac{n^{3/2}}{2m^{1/2}}\, e^{-m/\pi n} \qquad (n \to \infty),$$

where the second step is a consequence of Markov's inequality (Fact 8) and the last step follows from Lemma 2. Now fix $\delta > 0$ arbitrarily small. If m grows with n such that

$$m \geqslant \pi n \log n \left[1 - \frac{\frac{1}{2}\log\log n + \log 2\delta\sqrt{\pi}}{\log n} + \mathcal{O}\left(\frac{\log\log n}{\log^2 n}\right)\right] \tag{2}$$

then it is easy to see that the conditions of Lemma 2 are satisfied, viz., $n \ll m \ll n^{3/2}$, while simple substitution now shows that with such a choice of m,

$$\frac{n^{3/2}}{2m^{1/2}}\, e^{-m/\pi n} \leqslant \delta(1 + o(1)) \sim \delta \qquad (n \to \infty).$$

On the other hand, for any $\zeta > 0$, however small, the choice $m \geqslant (1+\zeta)\,\pi n \log n$ will ultimately dominate the right-hand side of (2) for any fixed choice of δ, however small, when n becomes large enough. Thus, for any $\zeta > 0$, as $n \to \infty$, $\mathbb{P}\{\mathbf{w} \neq \mathbf{w}^s\} \to 0$ if $m \geqslant (1+\zeta)\,\pi n \log n$. (Note that the argument above and the sample complexity estimate (2) actually provide a stronger sufficiency estimate than given in the theorem.)

Thus, we have proved one half of Theorem 1; a sample complexity slightly in excess of $\pi n \log n$ suffices to obtain perfect generalisation. To show that a sample complexity of $\pi n \log n$ is necessary, as well, requires somewhat more work. We will build up to the proof of the necessity through a sequence of auxiliary lemmas.

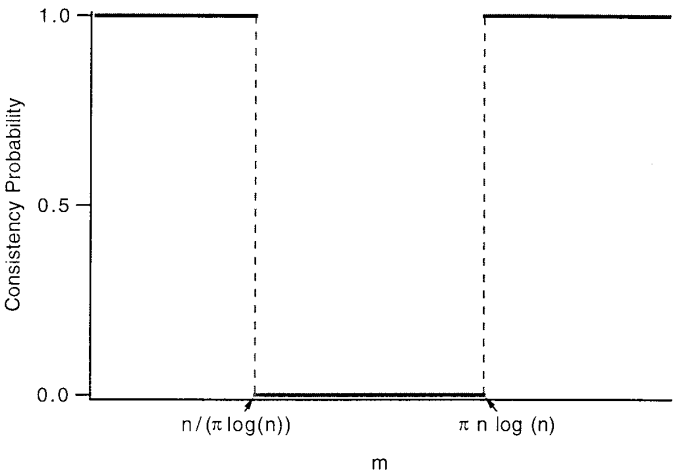

FIG. 3. A schematic emphasising the asymptotic consistency of majority rule and the two thresholds evidenced in the probability that the algorithm is consistent on a random m-sample.

Let $\hat{\mathbf{u}}$ be a random positive example generated, as usual, from the uniform distribution on the positive half-space $\mathbb{B}_+^n$ of the target perceptron $\mathbf{w}^s$. Let t be any fixed positive integer, and let $i_1, ..., i_t$ denote distinct indices in $\{1, ..., n\}$. Define

$$g(t) \triangleq \mathbb{P}\{\hat{u}_{i_1} = w^s_{i_1}, ..., \hat{u}_{i_t} = w^s_{i_t}\}.$$

Note that by exchangeability, $g(t)$ depends solely on t (and, of course, n) and not on the specific choice of indices $i_1, ..., i_t$.

LEMMA 3. *For every fixed t,*

$$g(t) = \frac{1}{2^t} + \frac{t}{2^{t-1}} \frac{1}{\sqrt{2\pi n}} + \frac{\gamma_n(t)}{n},$$

where $|\gamma_n(t)| \leqslant \Gamma(t)$ for an absolute positive constant $\Gamma(t)$ which depends solely on t and not on n.

Proof. The proof is by induction on t. The base case $t = 1$ is covered by Lemma 1. Now, as an induction hypothesis, assume

$$g(t-1) = \frac{1}{2^{t-1}} + \frac{t-1}{2^{t-2}} \frac{1}{\sqrt{2\pi n}} + \frac{\gamma_n(t-1)}{n},$$

where $|\gamma_n(t-1)| \leqslant \Gamma(t-1)$ for an absolute positive constant $\Gamma(t-1)$ independent of n. As before, let the random variable X denote the number of component matches between $\hat{\mathbf{u}}$ and $\mathbf{w}^s$,

$$X = |\{i : \hat{u}_i = w^s_i, 1 \leqslant i \leqslant n\}|.$$

Define the conditional joint probability

$$g(t|k) \triangleq \mathbb{P}\{u_{i_1} = w^s_{i_1}, ..., \hat{u}_{i_t} = w^s_{i_t} \mid X = k\}.$$

We then have

$$g(t) = \sum_{k \geqslant n/2} g(t|k)\, \mathbb{P}\{X = k\} = \sum_{k \geqslant n/2} g(t|k) \binom{n}{k} 2^{-(n-1)}.$$

We now provide a recursive estimate for $g(t|k)$. In the sequel $\gamma'_n(t)$ and $\gamma_n(t)$ both denote quantities bounded in absolute value by a positive quantity $\Gamma(t)$ depending solely on t. We have

$$g(t|k) = \frac{\binom{n-t}{k-t}}{\binom{n}{t}} = \frac{(k)_t}{(n)_t} = \frac{k \cdot (k)_{t-1}}{(n)_t} - \frac{(t-1)}{(n-t+1)} \frac{(k)_{t-1}}{(n)_{t-1}}$$

$$= \frac{k \cdot (k)_{t-1}}{(n)_t} - \frac{(t-1)}{(n-t+1)}\, g(t-1|k).$$

For $k = \Theta(n)$, the first term on the right-hand side is given by $k^t/n^t + \mathcal{O}(1/n)$, while the second term is $\mathcal{O}(1/n)$ by induction hypothesis. Consequently,

$$\begin{aligned} g(t) &= \frac{1}{n^t} \sum_{k \geqslant n/2} k^t \binom{n}{k} 2^{-(n-1)} + \frac{\gamma'_n(t)}{n} \\ &= \frac{1}{n^t} \sum_{a \geqslant 0} \left(\left\lceil \frac{n}{2} \right\rceil + a \right)^t \binom{n}{\lceil n/2 \rceil + a} 2^{-(n-1)} + \frac{\gamma'_n(t)}{n} \\ &= \frac{1}{n^t} \left[\frac{n^t}{2^t} + \frac{t n^{t-1}}{2^{t-1}} \sum_{a \geqslant 0} a \binom{n}{\lceil n/2 \rceil + a} 2^{-(n-1)} + \mathcal{O}(n^{t-1}) \right] \\ &\quad + \frac{\gamma'_n(t)}{n} \\ &= \frac{1}{2^t} + \frac{t}{2^{t-1}} \frac{1}{\sqrt{2\pi n}} + \frac{\gamma_n(t)}{n}, \end{aligned}$$

where the order term $\mathcal{O}(n^{t-1})$ in the penultimate step arises from the simple observation that the variance of the binomial is linear in n, which gives a crude, but satisfactory, bound on the remaining terms in the sum, and the last equality follows from the proof of Lemma 1. This completes the induction. ∎

Now fix positive integers t and r and let $i_1, ..., i_t$, and $j_1, ..., j_r$ denote any distinct set of $t + r$ indices. Define

$$f(t, r) = \mathbb{P}\left(\bigcap_{h=1}^{t} \{\hat{u}_{i_h} = w^s_{i_h}\} \cap \bigcap_{k=1}^{r} \{\hat{u}_{j_k} \neq w^s_{j_k}\} \right),$$

the probability of the event that $\hat{u}_i = w^s_i$ in t specific components and $\hat{u}_i \neq w^s_i$ in r specific components. (Again, by exchangeability, for each n, $f(t, r)$ depends only on t and r and not on the specific selection of indices i_h and j_k.) Note that $f(t, 0) = g(t)$.

LEMMA 4. *For every fixed choice of positive integers t and r,*

$$f(t, r) = \frac{1}{2^{t+r}} + \frac{t-r}{2^{t+r-1}} \frac{1}{\sqrt{2\pi n}} + \frac{\lambda_n(t, r)}{n},$$

where $|\lambda_n(t, r)| \leqslant \Lambda(t, r)$ for an absolute positive constant $\Lambda(t, r)$ which depends solely on t and r and not on n. In particular, $|f(t, r) - f(1, 0)^t f(0, 1)^r| = \mathcal{O}(n^{-1})$.

Proof. The proof is by double induction. Fix t and r with $(t, r) \neq (0, 0)$.

Base. We have already shown

$$f(t, 0) = \frac{1}{2^t} + \frac{t}{2^{t-1}} \frac{1}{\sqrt{2\pi n}} + \frac{\lambda_n(t, 0)}{n} \qquad (t \geqslant 1),$$

$$f(0, 1) = \frac{1}{2} - \frac{1}{\sqrt{2\pi n}} + \frac{\lambda_n(0, 1)}{n},$$

where the first equation is just the content of Lemma 3, while the second equation follows from Lemma 1.

Induction hypothesis. Assume the assertion holds for $f(t, r)$ and $f(t+1, r)$. Then

$$\begin{aligned} f(t, r+1) &= f(t, r) - f(t+1, r) \\ &= \left(\frac{1}{2^{t+r}} + \frac{t-r}{2^{t+r-1}} \frac{1}{\sqrt{2\pi n}}\right) \\ &\quad - \left(\frac{1}{2^{t+r+1}} + \frac{t+1-r}{2^{t+r}} \frac{1}{\sqrt{2\pi n}}\right) \\ &\quad + \frac{\lambda_n(t, r) - \lambda_n(t+1, r)}{n} \\ &= \frac{1}{2^{t+r+1}} + \frac{t-r-1}{2^{t+r}} \frac{1}{\sqrt{2\pi n}} + \frac{\lambda_n(t, r+1)}{n}, \end{aligned}$$

which concludes the induction. It is now simple to verify that $|f(t, r) - f(1, 0)^t f(0, 1)^r| = \mathcal{O}(n^{-1})$ by direct substitution. ∎

Remark. In particular, any fixed number of events from the set $\{u_i \neq w_i^s, 1 \leqslant i \leqslant n\}$ are *asymptotically independent* as n approaches infinity.

Note that when $m = 1$, i.e., when there is only one example $\hat{\mathbf{u}}^1$ in the training set, the perceptron generated by majority rule is exactly $\hat{\mathbf{u}}^1$. Thus, equivalently, any fixed number of events from $\{w_i \neq w_i^s, 1 \leqslant i \leqslant n\}$ are asymptotically independent when $m = 1$. By virtue of independent generation of the examples $\hat{\mathbf{u}}^\alpha$, $\alpha = 1, ..., m$, we anticipate, as indeed works out to be the case, that this asymptotic independence is preserved for $m > 1$. This in turn leads to the main lemma which states that the number of component errors in the perceptron generated by majority rule has a Poisson distribution in the limit. First some notation.

For any fixed positive integer τ and selection of indices $i_1, ..., i_\tau$, define

$$\rho(\tau) \triangleq \mathbb{P}\{w_{i_1} \neq w_{i_1}^s, ..., w_{i_\tau} \neq w_{i_\tau}^s\}.$$

Again, by exchangeability, $\rho(\tau)$ does not depend on the specific choices of indices $i_1, ..., i_\tau$, but only on τ. (There is also, of course, the usual dependence on m and n which we suppress for notational economy.) Also, for simplicity, set

$$\rho \triangleq \frac{\sqrt{n}}{2\sqrt{m}} e^{-m/\pi n}.$$

LEMMA 5. *Suppose m grows with n such that $n \ll m \ll n^{3/2}$. Then, for any fixed positive integer τ,*

$$\rho(\tau) \sim \rho^\tau$$

as $n \to \infty$.

Proof. The proof is again by induction. The base case is handled by Lemma 2 which gives us the asymptotically tight estimate $\rho(1) \sim \rho$ when m increases appropriately with n. Now suppose that $\rho(\tau - 1) \sim \rho^{\tau-1}$ as induction hypothesis. Without loss of generality, suppose $i_j = j$ for $j = 1, ..., \tau$. Let

$$\mathbf{v} = (v_j^\alpha \in \mathbb{B}, 2 \leqslant j \leqslant \tau, 1 \leqslant \alpha \leqslant m)$$

denote a fixed vector of $m(\tau - 1)$ signs, and define the events

$$\begin{aligned} \mathscr{F}_{\tau-1} &= \{w_2 \neq w_2^s, ..., w_\tau \neq w_\tau^s\}, \\ \mathscr{G}_{\tau-1}(\mathbf{v}) &= \{u_j^\alpha = v_j^\alpha, 2 \leqslant j \leqslant \tau, 1 \leqslant \alpha \leqslant m\}. \end{aligned}$$

Now consider

$$\begin{aligned} &\mathbb{P}\{w_1 \neq w_1^s \mid w_2 \neq w_2^s, ..., w_\tau \neq w_\tau^s\} \\ &\quad \overset{(a)}{=} \mathbb{P}\left\{\sum_{\alpha=1}^m \hat{u}_1^\alpha < 0 \mid \mathscr{F}_{\tau-1}\right\} \\ &\quad \overset{(b)}{=} \mathbb{P}\left\{\sum_\alpha \hat{u}_1^\alpha < 0, \mathscr{G}_{\tau-1}(\mathbf{v}) \mid \mathscr{F}_{\tau-1}\right\} \\ &\quad \overset{(c)}{=} \sum_v \mathbb{P}\left\{\sum_\alpha \hat{u}_1^\alpha < 0 \mid \mathscr{G}_{\tau-1}(\mathbf{v})\right\} \mathbb{P}\{\mathscr{G}_{\tau-1}(\mathbf{v}) \mid \mathscr{F}_{\tau-1}\}, \end{aligned} \quad (3)$$

where (a) follows from the definition of majority rule and the choice of $\mathbf{w}^s = (1, ..., 1)$, the sum over $\mathbf{v}$ exhausts all possible assignments of values ± 1 to $\{\hat{u}_j^\alpha, 2 \leqslant j \leqslant \tau, 1 \leqslant \alpha \leqslant m\}$ in (b), and (c) follows because the σ-algebra generated by the random variables $\{\hat{u}_j^\alpha, 2 \leqslant j \leqslant \tau, 1 \leqslant \alpha \leqslant m\}$ is clearly a refinement of the σ-algebra generated by the random variables $\{w_2, ..., w_\tau\}$. Now fix $\mathbf{v}$ and consider the term

$$\mathbb{P}\left\{\sum_\alpha \hat{u}_1^\alpha < 0 \mid \mathscr{G}_{\tau-1}(\mathbf{v})\right\}. \quad (4)$$

Consider any α in the set $\{1, ..., m\}$ and suppose that in the vector $(v_2^\alpha, ..., v_\tau^\alpha)$ there are t components taking value $+1$ and $r = \tau - 1 - t$ components taking value -1. As a consequence of the independent generation of the (positive) examples $\hat{\mathbf{u}}^\alpha$, we have

$$\begin{aligned} &\mathbb{P}\{\hat{u}_1^\alpha = 1 \mid \mathscr{G}_{\tau-1}(\mathbf{v})\} \\ &\quad = \mathbb{P}\{\hat{u}_1^\alpha = 1 \mid \hat{u}_j^\alpha = v_j^\alpha, 2 \leqslant j \leqslant \tau\} = \frac{f(t+1, r)}{f(t, r)} \\ &\quad = \frac{f(1, 0)^{t+1} f(0, 1)^r + c_1/n}{f(1, 0)^t f(0, 1)^r + c_2/n} = f(1, 0) + \frac{c_3}{n}, \end{aligned}$$

where, by Lemma 4, the terms c_1, c_2, and c_3 are all bounded in absolute value by a positive constant which depends solely on t and r and, hence, is completely determined by the

binary vector $(v_2^\alpha, ..., v_\tau^\alpha)$. Thus, by another application of Lemma 4, we have

$$\begin{aligned}\mathbb{P}\{\hat{u}_1^\alpha = +1 \mid \mathscr{G}_{\tau-1}(\mathbf{v})\} &= \frac{1}{2} + \frac{1}{\sqrt{2\pi n}} + \frac{c^\alpha}{n},\\ \mathbb{P}\{\hat{u}_1^\alpha = -1 \mid \mathscr{G}_{\tau-1}(\mathbf{v})\} &= \frac{1}{2} - \frac{1}{\sqrt{2\pi n}} - \frac{c^\alpha}{n},\end{aligned} \tag{5}$$

where c^α is a term bounded above in absolute value by a positive constant $C(v_2^\alpha, ..., v_\tau^\alpha)$ determined solely by the vector of signs $(v_2^\alpha, ..., v_\tau^\alpha)$. As the vector of signs $(v_2^\alpha, ..., v_\tau^\alpha)$ varies, the variation in the upper bound for $|c^\alpha|$ is hence determined by a fixed function $C\colon \mathbb{B}^{\tau-1} \to [0, \infty)$. As the domain is finite, this function takes values only in a finite range and consequently has a finite maximum, say C^*. Thus, $|c^\alpha| \leqslant C^*$ *uniformly for all* α. Now, let $(v_1^\alpha \in \mathbb{B}, 1 \leqslant \alpha \leqslant m)$ be any fixed assignment of signs. Again, by virtue of independent generation of the examples, we have

$$\begin{aligned}&\mathbb{P}\{\hat{u}_1^\alpha = v_1^\alpha, 1 \leqslant \alpha \leqslant m \mid \mathscr{G}_{\tau-1}(\mathbf{v})\}\\ &\quad= \prod_{\alpha=1}^{m} \mathbb{P}\{\hat{u}_1^\alpha = v_1^\alpha \mid \hat{u}_j^\alpha = v_j^\alpha, 2 \leqslant j \leqslant \tau\}\\ &\quad= \prod_{\alpha=1}^{m} \left(\frac{1}{2} + \frac{v_1^\alpha}{\sqrt{2\pi n}} + v_1^\alpha \frac{c^\alpha}{n}\right),\end{aligned}$$

where $\max_\alpha |c^\alpha| \leqslant C^* < \infty$.

Returning to a consideration of (4), we see that the term to be evaluated is simply the left tail probability of a conditional random walk over m steps whose individual summands are conditionally independent ± 1 random variables with individual (in general, nonidentical) marginal distributions governed by (5). The differences in the distributions of the summands are evinced only in the order terms, however, and the situation is ripe for use of the large deviation central limit theorem (Fact 6). We finesse the nonidentical distributions of the summands of the conditional random walk by exploiting the fact that the order terms c^α/n are uniformly bounded; thus: define two sequences $\{\xi_\pm^\alpha, 1 \leqslant \alpha \leqslant m\}$ of i.i.d., ± 1 random variables satisfying

$$\begin{aligned}\mathbb{P}\{\xi_\pm^\alpha = +1\} &= \frac{1}{2} + \frac{1}{\sqrt{2\pi n}} \pm \frac{C^*}{n},\\ \{\xi_\pm^\alpha = -1\} &= \frac{1}{2} + \frac{1}{\sqrt{2\pi n}} \mp \frac{C^*}{n},\end{aligned}$$

and let $R_\pm^{(m)} = \sum_{\alpha=1}^m \xi_\pm^\alpha$ be the two corresponding random walks over m steps. It is easy to see now that the distribution of the left tail of the conditional random walk (4) lies above the distribution of the corresponding tail of the random walk $R_+^{(m)}$ and below that of the random walk $R_-^{(m)}$, i.e.,

$$\mathbb{P}\{R_+^{(m)} < 0\} \leqslant \mathbb{P}\left\{\sum_\alpha \hat{u}_1^\alpha < 0 \mid \mathscr{G}_{\tau-1}(\mathbf{v})\right\} \leqslant \mathbb{P}\{R_-^{(m)} < 0\}.$$

Now to impose constraints on the rate of growth of m: suppose m grows with n such that $n \ll m \ll n^{3/2}$. Running through the proof of Lemma 2 again, we now obtain identical asymptotic estimates for both the upper and the lower bounds,

$$\mathbb{P}\{R_\pm^{(m)} < 0\} \sim \Phi\left(-\frac{\sqrt{2m}}{\sqrt{\pi n}}\right) \sim \frac{\sqrt{n}}{2\sqrt{m}}\, e^{-m/\pi n},$$

whence it follows that

$$\mathbb{P}\left\{\sum_\alpha \hat{u}_1^\alpha < 0 \mid \mathscr{G}_{\tau-1}(\mathbf{v})\right\} \sim \frac{\sqrt{n}}{2\sqrt{m}}\, e^{-m/\pi n} = \rho.$$

As this estimate holds uniformly for all choices of $\mathbf{v}$, we can now substitute back in (3) to obtain

$$\mathbb{P}\{w_1 \neq w_1^s \mid w_2 \neq w_2^s, ..., w_\tau \neq w_\tau^s\} \sim \rho.$$

Now recall that by the induction hypothesis,

$$\mathbb{P}\{w_2 \neq w_2^s, ..., w_\tau \neq w_\tau^s\} \sim \rho^{\tau-1},$$

whence the asymptotic relation

$$\mathbb{P}\{w_1 \neq w_1^s, w_2 \neq w_2^s, ..., w_\tau \neq w_\tau^s\} \sim \rho^\tau$$

obtains as $n \to \infty$ with $n \ll m \ll n^{3/2}$. The induction is complete. ∎

All the pieces are now in place. To complete the proof of the theorem, start with the observation

$$\mathbb{P}\{\mathbf{w} \neq \mathbf{w}^s\} = \mathbb{P}\left(\bigcup_{i=1}^n \{w_i \neq w_i^s\}\right).$$

Now let T be a fixed *even* positive integer which we will subsequently choose suitably large. Using Bonferroni's inequalities (Fact 9), together with the exchangeability of the events $\{w_i \neq w_i^s\}$, we now have

$$\begin{aligned}&\sum_{\tau=1}^{T} (-1)^{\tau-1} \binom{n}{\tau} \rho(\tau)\\ &\qquad\leqslant \mathbb{P}\left(\bigcup_{i=1}^n \{w_i \neq w_i^s\}\right) \leqslant \sum_{\tau=1}^{T-1} (-1)^{\tau-1} \binom{n}{\tau} \rho(\tau).\end{aligned} \tag{6}$$

Finally, let us fix the rate of growth of m with n. Let $\mathfrak{d}$ denote any fixed positive quantity satisfying $0 < \mathfrak{d} < 1$, and set

$$m = \pi n \log n \left[1 - \frac{\frac{1}{2}\log\log n + \log\log(1/\mathfrak{d}) + \log 2\sqrt{\pi}}{\log n} + \mathcal{O}\left(\frac{\log\log n}{\log^2 n}\right)\right]. \tag{7}$$

Clearly, m satisfies the conditions $n \ll m \ll n^{3/2}$. It is now simple to verify that

$$n\rho = \frac{n^{3/2}}{2m^{1/2}} e^{-m/\pi n} \sim \log\frac{1}{\mathfrak{d}}$$

as $n \to \infty$. Now consider the bounds in (6). As T is fixed (but arbitrary), as $n \to \infty$, we have the asymptotic estimates $\binom{n}{\tau} \sim n^\tau/\tau!$ for the binomial coefficients in the summands. Further, with m as prescribed above and with T fixed, Lemma 5 is applicable to the terms $\rho(\tau)$ in the summands. Consequently, for any fixed T,

$$\sum_{\tau=1}^{T} (-1)^{\tau-1} \binom{n}{\tau} \rho(\tau) \sim \sum_{\tau=1}^{T} (-1)^{\tau-1} \frac{(n\rho)^\tau}{\tau!} = 1 - \sum_{\tau=0}^{T} \frac{(-n\rho)^\tau}{\tau!} \sim 1 - \sum_{\tau=0}^{T} \frac{\log^\tau \mathfrak{d}}{\tau!}$$

as $n \to \infty$. We notice that the sequence of partial sums $\{\sum_{\tau=0}^{T} (\log^\tau \mathfrak{d}/\tau!), T \geqslant 0\}$ converges uniformly (in any bounded range) with respect to $\mathfrak{d}$ to $e^{\log \mathfrak{d}} = \mathfrak{d}$. Hence, by choosing T large enough and then allowing $n \to \infty$, both the lower and the upper bounds in (6) can be brought as close to $1 - \mathfrak{d}$ as desired. We thus have

$$\mathbb{P}\{\mathbf{w} \neq \mathbf{w}^s\} \to 1 - \mathfrak{d} \qquad (n \to \infty)$$

for a choice of m given by (7). For any choice of $\mathfrak{d} > 0$, however small, and any choice of $1 > \zeta > 0$, a sample complexity of $m = (1 - \zeta)\,\pi n \log n$ will be eventually dominated by the right-hand side of (7) so that it is easy to see by monotonicity that $\mathbb{P}\{\mathbf{w} \neq \mathbf{w}^s\}$ will approach one as $n \to \infty$. This proves *necessity* of the sample complexity in the theorem. Conversely, for any choice of $\mathfrak{d} < 1$, however close to 1, and any choice of $1 > \zeta > 0$, a sample complexity of $m = (1 + \zeta)\,\pi n \log n$ will eventually dominate the right-hand side of (7) so that, by analogous reasoning, $\mathbb{P}\{\mathbf{w} \neq \mathbf{w}^s\}$ will approach zero as $n \to \infty$. This also gives us *sufficiency* of the sample complexity in the theorem and completes the proof. Note that in (7) is embodied a somewhat sharper estimate of the second threshold than that given in the theorem.

6. ε-GENERALISATION

In the previous section, we saw that with a log-linear sample complexity in n majority rule generates the target perceptron *exactly*, i.e., the algorithm generalises perfectly in the sense that the error is zero and the confidence one, asymptotically. Are there savings to be made concomitant with a more generous measure of error tolerance and confidence?

Let ε and δ be fixed positive quantities denoting permissible error and confidence parameters, respectively. Write $\mathbf{w} \triangle \mathbf{w}^s$ to denote the symmetric difference between the positive half-spaces of $\mathbf{w}$ and $\mathbf{w}^s$, i.e., $\mathbf{w} \triangle \mathbf{w}^s$ is the set of vertices $\mathbf{u}$ for which either $[\langle \mathbf{w}, \mathbf{u}\rangle < 0] \wedge [\langle \mathbf{w}^s, \mathbf{u}\rangle \geqslant 0]$ or $[\langle \mathbf{w}, \mathbf{u}\rangle \geqslant 0] \wedge [\langle \mathbf{w}^s, \mathbf{u}\rangle < 0]$. Let the random variable $\mathscr{E} = \mathbb{P}(\mathbf{w} \triangle \mathbf{w}^s)$ denote the (error) probability that a random test example $\mathbf{u}$ selected independently from the uniform distribution on $\mathbb{B}^n$ is misclassified by the (random) perceptron $\mathbf{w}$ generated by majority rule. We are interested in quantifying conditions on the sample complexity m for which $\mathbb{P}\{\mathscr{E} > \varepsilon\} < \delta$. The setting is that of PAC learning under the uniform distribution.

Let $\varepsilon(h)$ denote the error probability conditioned on the event that the Hamming distance between $\mathbf{w}^s$ and the random $\mathbf{w}$ is h, i.e.,

$$\varepsilon(h) = \mathbb{P}(\mathbf{w} \triangle \mathbf{w}^s \mid H = h).$$

While ostensibly still a random variable, a consideration of the symmetry of the situation shows that $\varepsilon(h)$ is a fixed, nonrandom quantity and is exactly equal to the probability that any *fixed* vector in $\mathbb{B}^n$ at Hamming distance h from $\mathbf{w}^s$ misclassifies a randomly selected example $\mathbf{u}$. It is now not difficult to write down an expression for the latter quantity. A direct combinatorial argument shows that

$$\varepsilon(h) = \frac{1}{2^{n-1}} \sum_k \sum_l \binom{h}{(h+k)/2} \binom{n-h}{(n-h+l)/2}, \tag{8}$$

where the outer sum ranges over $1 \leqslant k \leqslant h$, the inner sum ranges over $-k \leqslant l \leqslant k-1$, and we use the convention that $\binom{a}{b} = 0$ if b is not integer. In particular, a simple evaluation shows

$$\varepsilon(1) = \frac{1}{2^{n-1}} \binom{n-1}{\lfloor (n-1)/2 \rfloor} \sim \sqrt{2/\pi n} \qquad (n \to \infty),$$

where the asymptotic estimate obtains from the estimate for the central term of the binomial (Fact 4).

Despite the forbidding appearance of (8), sharp asymptotic estimates can be obtained for $\varepsilon(h)$ when h is relatively large by application of the large-deviation central limit theorem (Fact 6) applied successively to the inner and outer sums in (8).

LEMMA 6. *If h increases with n such that $h = \Theta(n)$, then*

$$\varepsilon(h) \sim \frac{2}{\pi} \sin^{-1} \sqrt{h/n} \tag{9}$$

as $n \to \infty$.

For a demonstration of the result see Lyuu and Rivin [11] who utilise the above estimate in a characterisation of the space of possible solutions when a random sample is generated according to a target binary perceptron. (The result may be extended to arbitrary h as we will see subsequently.)

THEOREM 2. *Let ε and δ be fixed positive quantities. If the sample complexity m grows with n such that*

$$m = \frac{\pi n}{2} \left[\Phi^{-1} \left(1 - \delta \sin^2 \frac{\pi\varepsilon}{2} \right) \right]^2$$

then, as $n \to \infty$, with confidence at least $1 - \delta$, the probability that the perceptron $\mathbf{w}$ generated by majority rule misclassifies a randomly selected example (whose true classification is determined by the target perceptron $\mathbf{w}^s$) is less than ε. In particular, if $n \to \infty$ and $\varepsilon \to 0$, a sample complexity of

$$m = 2\pi n \log \left(\frac{2}{\pi\varepsilon \sqrt{\delta}} \right) \left[1 + \mathcal{O} \left(\frac{\log \log \varepsilon^{-1}}{\log \varepsilon^{-1}} \right) \right]$$

suffices to ensure that the error probability $\mathscr{E}$ is less than ε with confidence at least $1 - \delta$.

Remarks. Thus, for small error probabilities ε, a sample complexity of the order of $n \log(\varepsilon \sqrt{\delta})^{-1}$ is sufficient for majority rule to generate, with high confidence, a perceptron with small error probability. This estimate compares very favourably with the distribution-free sample complexities of the order of $(n/\varepsilon) \log(n/\varepsilon)$ demanded by the Vapnik–Červonenkis theory.

Golea and Marchand [8] have also recently reported sample complexities of order $n \log(\varepsilon^{-1})$ for the majority rule algorithm based on a nonrigourous "average case" analysis of the expectation of the error probability $\mathscr{E}$. Their analysis relies upon the following two simplifying assumptions which have the virtue of avoiding technical difficulties, albeit at the expense of mathematical rigour: (a) assume that the (discrete) joint distribution of $\langle \mathbf{w}, \mathbf{u} \rangle$ and $\langle \mathbf{w}^s, \mathbf{u} \rangle$ can be replaced by a bivariate normal; (b) assume that a satisfactory estimate for $\mathbb{E}\mathscr{E} = \mathbb{E}\varepsilon(H)$, the expectation of the error probability, can be obtained by evaluating $\varepsilon(\mathbb{E}H)$ instead. Both assumptions are difficult to justify mathematically. The first assumption ignores the range of validity of the central limit theorem and the messy technical details that result from truncations to keep ranges within those permissible by the central limit theorem (cf. Lyuu and Rivin [11] for the mess that results). The second assumption, somewhat more drastically, "linearises" the problem and appears much harder to justify formally; note that by Lemma 6, the second assumption is equivalent to replacing $\mathbb{E} \sin^{-1} \sqrt{H/n}$ by the "estimate" $\sin^{-1} \sqrt{\mathbb{E}H/n}$. This nonrigourous analysis, in common with the approaches borrowed from statistical physics, appears to produce an estimate in the right ballpark, although the sweeping brush of the approximations eliminates the confidence parameter δ from consideration. The approximations fail more seriously, however, in the regime of zero error or perfect generalisation; in their paper, Golea and Marchand are led to the erroneous conclusion that majority rule, being inconsistent, does not exhibit a "phase transition" to perfect generalisation. As we have seen in Section 5, however, majority rule does in fact exhibit a curious and abrupt transition to perfect generalisation.

Proof. We utilise Lemma 6 for h in the range $\Theta(n)$. Suppose the admissible error probability ε is fixed. By inverting (9) we can then obtain the corresponding range of Hamming distances $h = h(\varepsilon)$ as a function of ε; thus,

$$h(\varepsilon) = n \sin^2 \frac{\pi\varepsilon}{2} + o(n).$$

A simple application of Markov's inequality (Fact 8), together with Lemma 2, then yields

$$\mathbb{P}\{\mathscr{E} \geqslant \varepsilon\} = \mathbb{P}\{H \geqslant h(\varepsilon)\} \leqslant \frac{\mathbb{E}H}{h(\varepsilon)} \sim \frac{\Phi(-\sqrt{2m/\pi n})}{\sin^2(\pi\varepsilon/2)}$$

as $n \to \infty$. Suppose $0 < \delta < 1$ denotes a fixed confidence parameter. It follows that, for large enough n, $\mathbb{P}\{\mathscr{E} < \varepsilon) \gtrsim 1 - \delta$ if m grows linearly with n such that

$$m = \frac{\pi n}{2} \left[\Phi^{-1} \left(1 - \delta \sin^2 \frac{\pi\varepsilon}{2} \right) \right]^2.$$

A slightly clearer picture may be arrived at in the limit of small error tolerances ε. Note first that the Taylor expansion for the sine function yields

$$1 - \delta \sin^2 \frac{\pi\varepsilon}{2} = 1 - \frac{\pi^2 \varepsilon^2 \delta}{4} + \mathcal{O}(\varepsilon^4) \qquad (\varepsilon \to 0).$$

Using the asymptotic estimate for the Gaussian tail (Fact 7) it is now not difficult to verify by bootstrapping that

$$\Phi^{-1} \left(1 - \delta \sin^2 \frac{\pi\varepsilon}{2} \right) = \sqrt{4 \log(2/\pi\varepsilon \sqrt{\delta})} + \mathcal{O} \left(\frac{\log \log \varepsilon^{-1}}{\sqrt{\log \varepsilon^{-1}}} \right) \to \sqrt{4 \log(2/\pi\varepsilon \sqrt{\delta})} \qquad (\varepsilon \to 0).$$

Consequently, as $n \to \infty$ and $\varepsilon \to 0$, a choice of sample complexity

$$m = 2\pi n \log\left(\frac{2}{\pi\varepsilon\sqrt{\delta}}\right)\left[1 + \mathcal{O}\left(\frac{\log\log \varepsilon^{-1}}{\log \varepsilon^{-1}}\right)\right]$$

guarantees error probability less than ε on a randomly selected test example with confidence in excess of $1-\delta$. ∎

It is perhaps worth noting that the range of applicability of the theorem can be expanded to the case when the error tolerance ε is allowed to depend on n. In particular, the *theorem continues to hold for all choices of $\varepsilon = \varepsilon_n$ which are such that $n^{-1/2} \ll \varepsilon < 1$*. The latter estimate of the sample complexity in the theorem becomes particularly potent in the range $n^{-1/2} \ll \varepsilon \ll 1$ when $\varepsilon \to 0$ as $n \to \infty$.

To prove this assertion, it will suffice to extend the range of applicability of Lemma 6 from $h = \Theta(n)$ to $1 \ll h \leqslant n$. (Note that in this range of h, $(2/\pi)\sin^{-1}\sqrt{h/n} = \omega(n^{-1/2})$ as needed.) The key to the analysis is an exact recurrence relation for the error probabilities $\varepsilon(h)$ due to Baum and Lyuu [1].[6]

LEMMA 7. *The following recurrence holds for the conditional error probabilities $\varepsilon(h)$:*

Base. $\varepsilon(0) = 0$.

Recurrence. For even h,

$$\varepsilon(h) - \varepsilon(h-1) = 0,$$
$$\varepsilon(h+1) - \varepsilon(h) = \frac{1}{2^{n-1}}\binom{h}{h/2}\binom{n-h-1}{\lfloor (n-h-1)/2 \rfloor}.$$

To apply the recurrence, fix h even and note that we can write $\varepsilon(h+1)$ as a telescoping series:

$$\begin{aligned}\varepsilon(h+1) &= \varepsilon(1) + [\varepsilon(2) - \varepsilon(1)] + [\varepsilon(3) - \varepsilon(2)] \\ &\quad + \cdots + [\varepsilon(h+1) - \varepsilon(h)] \\ &= \varepsilon(1) + [\varepsilon(3) - \varepsilon(2)] + [\varepsilon(5) - \varepsilon(4)] \\ &\quad + \cdots + [\varepsilon(h+1) - \varepsilon(h)] \\ &= \varepsilon(1) + \sum_{i=1}^{h/2} [\varepsilon(2i+1) - \varepsilon(2i)],\end{aligned}$$

where the second equality follows from the above recurrence for $\varepsilon(h)$. Now, note that an application of Stirling's formula (Fact 3) in the range $1 \ll i \ll n$ yields the asymptotic estimates

$$\varepsilon(2i+1) - \varepsilon(2i) \sim \frac{2}{\pi}\frac{1}{\sqrt{2i(n-2i)}} \qquad (n \to \infty).$$

[6] Baum and Lyuu actually treat the n even case. The n odd case is completely analogous.

Consequently, let v_n be any slowly growing function of n satisfying $1 \ll v_n \ll h$. (For definiteness, we may take $v_n = \log h$.) Then, in the range $1 \ll h \ll n$, we obtain

$$\sum_{v_n < i \leqslant h/2} [\varepsilon(2i+1) - \varepsilon(2i)] = \left[\frac{2}{\pi}\sum_{v_n < i \leqslant h/2}\frac{1}{\sqrt{2i(n-2i)}}\right](1 + o(1)).$$

It is now simple to estimate the term within square parentheses by bounding the sum above and below by integrals (Fact 2). (Note that the function $f(x) = [x(n-x)]^{-1/2}$ decreases monotonically in the range $0 < x \leqslant n/2$, so that Fact 2 may be applied to the function $-f(x)$.) Noting the following simple evaluation of the indefinite integral

$$\int \frac{dx}{\sqrt{2x(n-2x)}} = \sin^{-1}\sqrt{2x/n},$$

it is easy to ascertain that in the range $1 \ll h \ll n$, both integral bounds evaluate to $(2/\pi)\sqrt{h/n} + \mathcal{O}(\sqrt{v_n/n})$, where we have used the fact that, for this range of h, the Taylor series expansion for $\sin^{-1}$ gives

$$\frac{2}{\pi}\sin^{-1}\sqrt{h/n} \sim \frac{2}{\pi}\sqrt{h/n} \qquad (n \to \infty).$$

It follows that

$$\left[\frac{2}{\pi}\sum_{v_n < i \leqslant h/2}\frac{1}{\sqrt{2i(n-2i)}}\right] = \frac{2}{\pi}\sqrt{h/n} + \mathcal{O}(\sqrt{v_n/n})$$

also. A similar argument using the Stirling bounds (Fact 3) yields the estimate

$$\varepsilon(1) + \sum_{1 \leqslant i \leqslant v_n} [\varepsilon(2i+1) - \varepsilon(2i)] = \Theta(\sqrt{v_n/n}).$$

Putting everything together, for even values of h satisfying $1 \ll h \ll n$, we obtain

$$\varepsilon(h+1) = \frac{2}{\pi}\sqrt{h/n} + \mathcal{O}(\sqrt{v_n/n}) \sim \frac{2}{\pi}\sqrt{h/n} \qquad (n \to \infty).$$

We can consequently extend the region of validity of Lemma 6.

LEMMA 6′. *If h increases with n such that $1 \ll h \leqslant n$, then*

$$\varepsilon(h) \sim \frac{2}{\pi}\sin^{-1}\sqrt{h/n} \tag{10}$$

as $n \to \infty$.

Inverting (10) gives

$$h(\varepsilon) \sim n \sin^2 \frac{\pi\varepsilon}{2} \qquad (n \to \infty),$$

so that the rest of the analysis flows exactly as before. We consequently have the following extended version of Theorem 2.

THEOREM 2′. *Let δ be a fixed positive quantity and suppose ε satisfies $n^{-1/2} \ll \varepsilon < 1$. If the sample complexity m grows with n such that*

$$m = \frac{\pi n}{2}\left[\Phi^{-1}\left(1 - \delta \sin^2 \frac{\pi\varepsilon}{2}\right)\right]^2$$

then, as $n \to \infty$, with confidence at least $1-\delta$, the probability that the perceptron $\mathbf{w}$ generated by majority rule misclassifies a randomly selected example (whose true classification is determined by the target perceptron $\mathbf{w}^s$) is less than ε. In particular, if $n^{-1/2} \ll \varepsilon \ll 1$, then, as $n \to \infty$, a sample complexity of

$$m = 2\pi n \log\left(\frac{2}{\pi\varepsilon\sqrt{\delta}}\right)\left[1 + \mathcal{O}\left(\frac{\log\log \varepsilon^{-1}}{\log \varepsilon^{-1}}\right)\right]$$

suffices to ensure that the error probability $\mathscr{E}$ is less than ε with confidence at least $1-\delta$.

While the sample complexity estimate becomes larger when tighter and tighter error tolerances are demanded, the increase is very modest. Thus, even if $\varepsilon \to 0$ as $n \to \infty$, the sample complexity needed to learn the underlying binary perceptron (with vanishingly small error probability as $n \to \infty$) is only slightly more than linear in n.

We can further seek to extend these results to the case $\varepsilon = \Theta(n^{-1/2})$ (or equivalently, $h = \Theta(1)$). The Stirling bounds start becoming a little less precise in this range, however, and the integral approximation technique only shows that $\varepsilon(h) = \Theta(n^{-1/2})$. More precise results can be obtained in this range, however, using the techniques of Section 5.

7. SUMMARY

Recently, several researchers have studied the phenomenon of perfect generalisation in perceptrons with binary weights in $\{-1, 1\}$. The main result that has been shown is that when the number of randomly and uniformly chosen examples is $m = \alpha n$ and labeled according to a target perceptron, there exists a critical α_c such that when $m > \alpha_c n$ (or $\alpha > \alpha_c$), perfect generalisation is attained as n approaches infinity; namely, the target perceptron is the only one consistent with the sample set. Borrowing from persuasive, but nonrigourous, methods from statistical physics. Sompolinsky, Tishby, and Seung [13] and Györgyi [9] both found α_c to be around 1.24. A rigourous demonstration was first provided by Baum and Lyuu [1] who showed formally that $\alpha_c < 2.0821$; this estimate was subsequently improved by Lyuu and Rivin [11] (cf. also earlier nonrigourous work of Gardner and Derrida [6]) who proved the following: for $\alpha > 1.44797$, the expected number of nontarget perceptrons consistent with the examples is $2^{-\Theta(\sqrt{n})}$. Consequently, by Markov's inequality, when $m > 1.44797n$, the probability that any consistent algorithm (such as exhaustive search) will generalise perfectly to a target binary perceptron is at least $1 - 2^{-\Theta(\sqrt{n})}$.

These efforts have produced a (partial) resolution of the following information-theoretic question: how many random examples are sufficient to uniquely identify an underlying target perceptron? The results imply that any consistent learning algorithm, such as exponential search through the vertices of the cube, will generate the target perceptron as output with high confidence provided $m > \alpha_c n$. The NP-completeness of the problem, however, suggests that it may not be possible to polynomially bound the time complexity of any guaranteed consistent algorithm.

Majority rule is an example of an *inconsistent* learning algorithm which nonetheless exhibits superior generalisation performance. The algorithm has very modest computational requirements, linear time and space complexity, and operates off-line. Our main result is the demonstration that, for large n, the algorithm exhibits a curious transition to zero-error, perfect generalisation abruptly when the sample complexity exceeds $\pi n \log n$ random examples chosen from the uniform distribution on the vertices of the cube. This sample complexity estimate exceeds the information-theoretic minimum by only a logarithmic factor, this added cost being offset, perhaps, by the extreme simplicity of the algorithm. An implication of this result is that almost all instances of the variant of binary integer programming considered here with $\Omega(n \log n)$ inequalities to be satisfied are tractable.

We have also shown, in a more traditional PAC learning setting, that if an error probability $\varepsilon > 0$ is permissible, then the sample complexity needed to PAC-learn an underlying target binary perceptron becomes linear in n; more specifically, if the sample complexity m grows with n such that

$$m = \frac{\pi n}{2}\left[\Phi^{-1}\left(1 - \delta \sin^2 \frac{\pi\varepsilon}{2}\right)\right]^2$$

then, as $n \to \infty$, with confidence at least $1-\delta$, the probability that the perceptron $\mathbf{w}$ generated by majority rule misclassifies a randomly selected example (whose true classification is determined by the target perceptron $\mathbf{w}^s$) is

less than ε. If, in particular, $n \to \infty$ and $\varepsilon \to 0$, a sample complexity of

$$m = 2\pi n \log\left(\frac{2}{\pi\varepsilon\sqrt{\delta}}\right)\left[1 + \mathcal{O}\left(\frac{\log\log\varepsilon^{-1}}{\log\varepsilon^{-1}}\right)\right]$$

suffices to ensure that the error probability is less than ε with confidence at least $1 - \delta$. These sample complexity estimates obtained for examples drawn at random from the uniform distribution on the vertices of the cube are substantially lower than the distribution-free sample complexities of the order of $(n/\varepsilon)\log(n/\varepsilon)$ demanded by the Vapnik–Červonenkis theory.

ACKNOWLEDGMENTS

We thank the referees of a preliminary version of this paper [3] for helpful suggestions and for pointing out Golea and Marchand's paper [8] to us.

REFERENCES

1. E. B. Baum and Y.-D. Lyuu, The transition to perfect generalisation in perceptrons, *Neural Comput*. **3** (1991), 386–401.
2. S. C. Fang, "The Complexity of Learning Binary Perceptons," Ph.D. dissertation, Dept. of Electrical Eng., University of Pennsylvania, 1995.
3. S. C. Fang and S. S. Venkatesh, On the average tractability of binary integer programming and the curious transition to perfect generalisation in learning majority functions, *in* "Proceedings, 6th ACM Conf. on Computational Learning Theory," pp. 310–316, ACM Press, New York, 1993.
4. S. C. Fang and S. S. Venkatesh, The capacity of Majority Rule, Random Structures and Algorithms, to appear.
5. W. Feller, "An Introduction to Probability Theory and Its Applications," Vol. I, 3rd ed., Wiley, New York, 1968.
6. E. Gardner and B. Derrida, Three unfinished works on the optimal storage capacity of networks, *J. Phys. A*: *Math. Gen*. **22** (1989), 1983–1994.
7. M. Garey and D. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness," Freeman, San Francisco, 1979.
8. M. Golea and M. Marchand, On learning perceptrons with binary weights, *Neural Comput*. **5** (1993), 767–782.
9. G. Györgi, First order transition to perfect generalisation in a neural network with binary synapses, *Phys. Rev. A* **41** (1990), 7097–7100.
10. W. Köhler, S. Diederich, W. Kinzel, and M. Opper, Learning algorithm for a neural network with binary synapses, *Z. Phys. B* **78** (1990), 333–342.
11. Y.-D. Lyuu and I. Rivin, Tight bounds on transition to perfect generalisation in perceptrons, *Neural Comput*. **4** (1992), 854–862.
12. L. Pitt and L. G. Valiant, Computational limitations on learning from examples, *J. Assoc. Comput. Mach*. **35**, No. 4 (1988), 965–984.
13. H. Sompolinsky, N. Tishby, and H. Seung, Learning from examples in large neural networks, *Phys. Rev. Lett*. **65** (1990), 1683–1686.
14. S. S. Venkatesh, On learning binary weights for majority functions, *in* "Proceedings, Fourth Workshop on Computational Learning Theory" (L. G. Valiant and M. K. Warmuth, Eds.), Morgan Kaufmann, San Mateo, CA, 1991.
15. S. S. Venkatesh, Directed drift: A new linear threshold algorithm for learning binary weights on-line, *J. Comput. Systems Sci.* **46**, No. 2 (1993), 198–217.

Printed in Belgium