SIAM J. DISCRETE MATH.
Vol. 10, No. 3, pp. 482–498, August 1997

A THRESHOLD FUNCTION FOR HARMONIC UPDATE[*]

SHAO C. FANG[†] AND SANTOSH S. VENKATESH[†]

Abstract. Harmonic update is a randomized on-line algorithm which, given a random m-set of vertices $U(m) \subseteq \{-1,1\}^n$ in the n-dimensional cube, generates a random vertex $\mathbf{w} \in \{-1,1\}^n$ as a putative solution to the system of linear inequalities: $\sum_{i=1}^n w_i u_i \geq 0$ for each $\mathbf{u} \in U(m)$. Using tools from large deviation multivariate normal approximation and Poisson approximation, we show that $\sqrt{n}\big/\sqrt{\log n}$ is a threshold function for the property that the vertex $\mathbf{w}$ generated by harmonic update has positive inner product with each vertex in $U(m)$. More explicitly, let $P(n,m)$ denote the probability that $\sum_{i=1}^n w_i u_i \geq 0$ for each $\mathbf{u} \in U(m)$. Then, as $n \to \infty$, $P(n,m) \to 0$ or 1 according to whether $m = m_n$ varies with n such that $m \gg \sqrt{n}\big/\sqrt{\log n}$ or $m \ll \sqrt{n}\big/\sqrt{\log n}$, respectively. The analysis also exposes the fine structure of the threshold function.

Key words. polytopes, threshold function, randomized algorithm, harmonic update, binary integer programming, neural networks, large deviations, normal approximation, Poisson approximation

AMS subject classifications. 05C80, 60C05, 52B11, 60G85, 68R05

PII. S0895480195283701

1. Information and finite memory.

How much information can a single bit of memory updated on-line retain about a Bernoulli sequence? More specifically, the following problem was posed by J. Komlós. Write $\mathbb{B} \triangleq \{-1,1\}$ and let $\{\, u^{(t)}, t \geq 1 \,\}$ be a sequence of symmetric Bernoulli trials, where

$$u^{(t)} = \begin{cases} -1 & \text{with probability } 1/2, \\ +1 & \text{with probability } 1/2. \end{cases}$$

Suppose a single bit of memory $w \in \mathbb{B}$ is available to record this sequence; write $w^{(t)} \in \mathbb{B}$ for the state of the one-bit memory at epoch t (with $w^{(1)} \in \mathbb{B}$ being an arbitrary initial state of memory). We suppose that input bits $u^{(t)}$ arrive sequentially in time and memory updates $\bigl(w^{(t)}, u^{(t)}\bigr) \mapsto w^{(t+1)}$ proceed on-line governed by a sequence $\{\, f^{(t)} \colon \mathbb{B} \times \mathbb{B} \to \mathbb{B} \mid t \geq 1 \,\}$ of (possibly random) Boolean functions of two Boolean variables: $w^{(t+1)} = f^{(t)}\bigl(w^{(t)}, u^{(t)}\bigr)$. After m epochs, input bits $u^{(1)}, \ldots, u^{(m)}$ have been presented sequentially in time leading to the current state of memory $w^{(m+1)} \in \mathbb{B}$, which now constitutes the sole "record" (insofar as a single bit may be said to constitute a record) of the entire past, i.e., the sequence of bits $u^{(1)}, \ldots, u^{(m)}$. One measure of the efficacy of the update sequence $\{f^{(t)}\}$ in storing information up to this moment in the one-bit memory is the minimum covariance $\min_{1 \leq t \leq m} \mathbb{E}\bigl(w^{(m+1)} u^{(t)}\bigr)$: a minimum covariance of zero implies that there is at least one bit in the past about which the one-bit memory carries no information; a positive minimum covariance, on the other hand, indicates that the single bit of memory carries information about every one of the inputs in the past. In this context, Komlós posed the following

[*] Received by the editors March 24, 1995; accepted for publication (in revised form) June 25, 1996. This research was supported in part by Air Force Office of Scientific Research grants F49620-93-1-0120 and F49620-92-J-0344. A preliminary version of this paper was presented in part at the *Fourth ACM Conference on Computational Learning Theory*, Santa Cruz, CA, 1991.

http://www.siam.org/journals/sidma/10-3/28370.html

[†] Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104 (fang@ee.upenn.edu, venkatesh@ee.upenn.edu).

problem: what is $I_m \triangleq \max_{f^{(1)},\dots,f^{(m)}} \min_{1\le t\le m} \mathbb{E}\bigl(w^{(m+1)}u^{(t)}\bigr)$? The quantity I_m may be taken as an intrinsic measure of the amount of information that a single bit of memory updated on-line can retain about *each* past input.

In [12], Venkatesh and Franklin provide comprehensive answers to this and related questions. In particular, they show that

$$\tfrac{1}{m} \le I_m < \tfrac{2}{m},$$

whence $I_m = \Theta\bigl(m^{-1}\bigr)$. Their results also directly imply that any sequence of deterministic update rules $\bigl\{f^{(t)}\bigr\}$ yields exponentially small (in m) minimum covariances at best and hence that the optimal sequence of update rules $\bigl\{f^{(t)}_{\text{opt}}\bigr\}$ is necessarily randomized.

Consider an application of this notion of on-line information storage to the following classical problem in mathematical programming. Let $\mathbb{B}^n = \{-1,1\}^n$ denote the vertices of a cube in n dimensions and let $U(m) = \bigl\{\,\mathbf{u}^{(t)}, 1 \le t \le m\bigr\}$ be a random m-set of vertices in $\mathbb{B}^n$ obtained by independent sampling from the uniform distribution on $\mathbb{B}^n$. Write $\mathbf{u}^{(t)} = \bigl(u_1^{(t)},\dots,u_n^{(t)}\bigr)$. Does there exist a vertex $\mathbf{w} = (w_1,\dots,w_n) \in \mathbb{B}^n$ for which the inequalities

$$\begin{array}{c} w_1u_1^{(1)} + w_2u_2^{(1)} + \cdots + w_nu_n^{(1)} \ge 0 \\ w_1u_1^{(2)} + w_2u_2^{(2)} + \cdots + w_nu_n^{(2)} \ge 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ w_1u_1^{(m)} + w_2u_2^{(m)} + \cdots + w_nu_n^{(m)} \ge 0 \end{array} \tag{1.1}$$

are simultaneously satisfied? Let $\langle\cdot,\cdot\rangle$ denote the usual inner product in Euclidean n-space $\mathbb{R}^n$. Then each point $\mathbf{w}$ in $\mathbb{R}^n$ determines a *positive half-space* defined by $H^+(\mathbf{w}) = \{\,\mathbf{u} \in \mathbb{R}^n : \langle \mathbf{w}, \mathbf{u}\rangle \ge 0\,\}$. Geometrically speaking, our question is equivalent to asking whether there exists a vertex $\mathbf{w} \in \mathbb{B}^n$ such that the convex hull of $U(m)$ is contained in the positive half-space determined by $\mathbf{w}$.[1]

If $\mathbf{w}$ is allowed to range over $\mathbb{R}^n \setminus \{\mathbf{0}\}$, the problem is just an instance of linear programming and a solution to the system of inequalities (1.1), if one exists, can be found in polynomial time using interior point methods (cf. Karmarkar [6]). When, however, $\mathbf{w}$ is restricted to the vertices of the n-dimensional cube, the decision problem becomes NP-complete as an instance of binary integer programming (cf. Garey and Johnson [5], Pitt and Valiant [7]).

Consider an on-line programming scenario in which the random examples $\mathbf{u}^{(t)}$ comprising $U(m)$ arrive in sequence at epochs $t = 1,\dots,m$. In the information storage analogy, we are provided with n bits of memory and access to a sequence of memory update rules $\bigl\{\, f^{(t)}\colon \mathbb{B}^n \times \mathbb{B}^n \to \mathbb{B}^n \bigm| t \ge 1\,\bigr\}$. Starting from an arbitrary initial memory state $\mathbf{w}^{(1)} \in \mathbb{B}^n$, we then recursively generate memory states $\mathbf{w}^{(t+1)} = f^{(t)}\bigl(\mathbf{w}^{(t)}, \mathbf{u}^{(t)}\bigr)$ for $t \ge 1$. The on-line procedure is successful if, after presentation of the mth example $\mathbf{u}^{(m)}$, the vertex ("memory state") $\mathbf{w} \triangleq \mathbf{w}^{(m+1)}$ generated by the algorithm satisfies the system of linear inequalities (1.1).

In [11], a sequence of randomized update rules $\bigl\{f^{(t)}\bigr\}$, dubbed *harmonic update*, is constructed starting from the following trivial observations: (i) for each t, the sum $\sum_{i=1}^n w_iu_i^{(t)}$ is more likely to be positive if the individual summands $w_iu_i^{(t)}$ are likely

[1] The above mathematical programming problem can also be formulated as a learning problem in a formal model of a neuron or perceptron (cf. Fang and Venkatesh [1]).

to be positive, i.e., if the random element $w_i \triangleq w_i^{(m+1)}$ is positively correlated with $u_i^{(t)}$, and (ii) the *smallest* of the inner products $\langle \mathbf{w}, \mathbf{u}^{(t)} \rangle$ must be nonnegative if (1.1) is to hold. Looking at the ith column ($1 \le i \le n$) in (1.1), these considerations suggest that the update rules be chosen so that $\min_{1\le t\le m} \mathbb{E}\bigl(w_i u_i^{(t)}\bigr)$ is maximized. The optimal update rules for Komlós's problem would do very nicely here if they could be determined explicitly; applying the optimal one-bit memory update rules n times, once for every component, will maximize the smallest expectation of the summands in each column, as desired. Harmonic update uses auxiliary randomization in an attempt to achieve this desideratum.

Harmonic update. Given examples $\mathbf{u}^{(t)} = \bigl(u_1^{(t)}, \dots, u_n^{(t)}\bigr) \in \mathbb{B}^n$ presented sequentially at epochs $t = 1, \dots, m$, the algorithm recursively generates a sequence of n-bit memory states $\mathbf{w}^{(t+1)} = \bigl(w_1^{(t+1)}, \dots, w_n^{(t+1)}\bigr) \in \mathbb{B}^n$, where $\mathbf{w}^{(t+1)}$ is a random function of $\mathbf{w}^{(t)}$ and $\mathbf{u}^{(t)}$ only. After m epochs, the algorithm returns the final n-bit memory state $\mathbf{w} \triangleq \mathbf{w}^{(m+1)}$ as a putative vertex solution to the system of inequalities (1.1).

H1. [Initialize.] Set $\mathbf{w}^{(1)} = \bigl(w_1^{(1)}, \dots, w_n^{(1)}\bigr)$ to be an arbitrary vertex in $\mathbb{B}^n$. Set $t \leftarrow 1$.

H2. [New example.] Obtain example $\mathbf{u}^{(t)}$.

H3. [Reinitialize component index.] Set $i \leftarrow 1$.

H4. [Update memory components.] If $w_i^{(t)} = u_i^{(t)}$, set $w_i^{(t+1)} = w_i^{(t)}$; else if $w_i^{(t)} = -u_i^{(t)}$, set

$$w_i^{(t+1)} = \begin{cases} -w_i^{(t)} & \text{with probability } 1/t, \\ +w_i^{(t)} & \text{with probability } 1 - 1/t. \end{cases}$$

H5. [Iterate.] Set $i \leftarrow i + 1$. If $i \le n$, go back to step H4; otherwise set $t \leftarrow t + 1$. If $t \le m$ go back to step H2; otherwise set $\mathbf{w} = \mathbf{w}^{(m+1)}$ and terminate the algorithm.

Remarks.

Computation. While the actual memory updates in step H4 are done in place, it is convenient to keep the notation $\mathbf{w}^{(t)}$ to identify the state of the memory at epoch t for purposes of later analysis.

Probability space. It is assumed implicitly that the auxiliary randomization in step H4 is independent across i and t; in particular, we can assume that a biased coin with the appropriate success probability is tossed independently each time step H4 is encountered.

The intuition behind the algorithm is as follows: at epoch t, the current state $w_i^{(t)}$ of the ith memory component presumably contains information about the ith components of the first $t-1$ examples. No problem arises in updating the state of the ith bit of memory if the ith component $u_i^{(t)}$ of the current example has the same sign as the current state $w_i^{(t)}$ of the ith memory component; setting $w_i^{(t+1)} = w_i^{(t)}$ adds $u_i^{(t)}$ to the knowledge base at no cost to the previously stored components. Complications arise, however, if $w_i^{(t)}$ and $u_i^{(t)}$ have opposite signs. In this case, retaining the sign of $w_i^{(t)}$ results in all information about $u_i^{(t)}$ being irrevocably lost; conversely, changing the sign of $w_i^{(t)}$ results in a loss of information about $u_i^{(1)}, \dots, u_i^{(t-1)}$. The solution in this case is to change the sign of the ith bit of memory probabilistically—and with increasing reluctance as time passes (when there is presumably considerable past history stored in the bit of memory). The exact measure of this reluctance to change the sign of the memory bit with increasing time is given probabilistically by the harmonic sequence[2] $1/t$. The effect of this randomized update rule is to ensure that each component of memory retains an equal amount of information about the corresponding component of every example.

As m increases, the probability that there exists any solution for the system of inequalities (1.1) decreases monotonically, and for large enough m the random vertex set $U(m)$ will fail to be linearly separable (in the sense that there is no solution for (1.1)) with high probability. In what follows we allow $m = m_n$ to depend implicitly on the dimensionality n. Our goal is to determine the "largest" rate of increase of m with n for which the system of inequalities (1.1) is satisfied with asymptotically high probability as $n \to \infty$. In the language of random graphs (cf. Spencer [10]), we wish to determine a threshold function for the property that (1.1) is satisfied.

Indeed, Füredi [4] showed that $2n$ is a threshold function for the property that there exists a *real* vector $\mathbf{w} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ satisfying (1.1). (Equivalently, $2n$ is a threshold function for the property that the convex hull of the random vertex set $U(m)$ contains the origin.) When restricted to vertex solutions for (1.1), we may expect the probability that there exists a solution $\mathbf{w} \in \mathbb{B}^n$ satisfying (1.1) to decay rather faster with m. In fact, a trite application of Boole's inequality readily establishes n as an upper bound for the rate of growth of m for *any* algorithm if we are to hold out hopes for a solution in $\mathbb{B}^n$ for (1.1). Informally, if m grows faster than n, then (1.1) admits no solution in $\mathbb{B}^n$ with probability $1 - \mathfrak{o}(1)$. Much sharper results can be shown for harmonic update, and the following theorem, which is our main result, exposes the fine structure of a threshold function for the algorithm.

Let $\mathbf{w} \in \mathbb{B}^n$ be the vertex generated by harmonic update and write $H^-(\mathbf{w}) = \{\, \mathbf{u} \in \mathbb{R}^n : \langle \mathbf{w}, \mathbf{u} \rangle < 0 \,\}$ for the negative half-space determined by $\mathbf{w}$. Let $Z = Z_{n,m} \triangleq \bigl|U(m) \cap H^-(\mathbf{w})\bigr|$ denote the number of $\mathbf{u}^{(t)}$ $(1 \le t \le m)$ that fall into the negative half-space determined by the vector $\mathbf{w}$ generated by harmonic update. Our main theorem shows that for a suitable rate of growth of $m = m_n$ with n, the random variable Z_{n,m_n} has a limiting Poisson distribution as $n \to \infty$. A sharp threshold for the event of interest $\{Z_{n,m} = 0\}$ that $\mathbf{w}$ has positive inner product with each vertex in the random m-set $U(m)$ follows immediately.

Main Theorem. *Let $\lambda > 0$ be any fixed positive number and suppose that $m = m_n$ grows with n such that*

$$m_n = \sqrt{\frac{n}{\log n}} \left\{ 1 + \frac{\log\log n + \log\bigl(\lambda\sqrt{2\pi}\,\bigr)}{\log n} + \mathcal{O}\biggl(\frac{\log\log n}{(\log n)^2}\biggr) \right\}. \tag{1.2}$$

Then Z_{n,m_n} tends in distribution to $\mathrm{Po}(\lambda)$, the Poisson distribution with parameter λ, as $n \to \infty$. In particular, for each fixed k,

$$\mathbb{P}\{Z_{n,m_n} = k\} \to \frac{\lambda^k}{k!}\, e^{-\lambda}$$

as $n \to \infty$.

Corollary. *Write $P(n,m) = \mathbb{P}\{Z_{n,m} = 0\}$ for the probability that the vertex $\mathbf{w} \in \mathbb{B}^n$ generated by harmonic update satisfies the system of inequalities* (1.1). *Then, with $m = m_n$ as in* (1.2), *$P(n,m) \to e^{-\lambda}$ as $n \to \infty$. In particular, for every fixed $\epsilon > 0$, the following assertions hold:*

[2] Hence we have the name harmonic update.

(a) *if m varies with n such that* $m \le (1-\epsilon)\sqrt{\frac{n}{\log n}}$ *then* $P(n,m) \to 1$ *as* $n \to \infty$,

(b) *if m varies with n such that* $m \ge (1+\epsilon)\sqrt{\frac{n}{\log n}}$ *then* $P(n,m) \to 0$ *as* $n \to \infty$.

In other words, $\sqrt{n}\big/\sqrt{\log n}$ is a threshold function for the attribute that the binary vector $\mathbf{w}$ generated by harmonic update satisfies the system of linear inequalities (1.1).

Remark. The Main Theorem provides a lower bound for the rate of growth of m with n for which a vertex solution exists for the system of inequalities (1.1). At this point, it is natural to wonder if the gap between the lower bound $\sqrt{n}\big/\sqrt{\log n}$ and the upper bound n can be reduced by increasing the computational complexity of the memory update rule. Indeed, substantial improvements in information storage can accrue if off-line procedures are permitted or the examples are recycled infinitely often in an on-line scenario [11, 12]. For instance, the majority rule algorithm introduced in [11] is an off-line procedure which, given the random m-set of examples $U(m)$, selects a vertex closest to the centroid of $U(m)$ as a putative solution to (1.1). In a companion exposition [2], we established a threshold function for majority rule at $\frac{n}{\pi \log n}$.

The rest of the paper is devoted to a proof of the Main Theorem. The main technical tools used in the proof are multivariate normal approximation and Poisson approximation, the former via a multivariate integral limit theorem for large deviations and the latter in the form of a probabilistic sieve. These technical results are collected in the next section for ease of later reference. The proof follows.

2. Preliminaries.

Notation. As already indicated, we use $\mathbb{B}$ to denote the set $\{-1,1\}$, with $\mathbb{B}^n = \{-1,1\}^n$ the vertices of the cube in n dimensions. The set of all integers is denoted by $\mathbb{Z}$, with $\mathbb{Z}^n$ denoting the corresponding set of lattice points in n dimensions. Also, we denote the real line by $\mathbb{R}$, with $\mathbb{R}^n$ denoting n-dimensional Euclidean space equipped with the usual inner product $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{i=1}^n x_i y_i$ and the induced norm $\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x}\rangle^{1/2}$. If $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ are points in $\mathbb{R}^n$, the vector inequality $\mathbf{x} \ge \mathbf{y}$ means that the scalar inequalities $x_i \ge y_i$ hold for each $i = 1, \dots, n$; likewise, the vector inequality $\mathbf{x} > \mathbf{y}$ means that the corresponding componentwise inequalities are strict: $x_i > y_i$ $(1 \le i \le n)$. We write $\mathbf{0}$ for the vector $(0, \dots, 0)$ with all components identically 0 and $\mathbf{1}$ for the vector $(1, \dots, 1)$ with all components identically 1. Superscripts r, s, and t and subscripts i and j are employed exclusively to index vectors and their components, respectively.

For purposes of definiteness in vector–matrix operations we assume that all vectors are *row* vectors; a prime ($'$) denotes vector and matrix transpose. We also reuse the notation $|V|$ to denote the determinant of a square matrix V as well as the cardinality of a set V. The usage will be clear from the context.

Throughout, $\mathbb{P}$ stands for probability measure on the underlying probability space, $\mathbb{E}$ denotes expectation, Var denotes variance, and Cov denotes covariance. For any integer $k \ge 1$, if $\mathbf{X} = (X_1, \dots, X_k)$ is a Gaussian (row) vector with zero mean, $\mathbb{E}\,\mathbf{X} = \mathbf{0}$, and nondegenerate covariance matrix $K = \operatorname{Cov}(\mathbf{X}) = \mathbb{E}\big(\mathbf{X}'\mathbf{X}\big)$, $|K| > 0$, we write

$$\phi_K(\mathbf{x}) \triangleq \frac{1}{(2\pi)^{k/2}|K|^{1/2}}\, e^{-\frac{1}{2}\mathbf{x}\,K^{-1}\mathbf{x}'}$$

for the multivariate Gaussian density, and likewise

$$\Phi_K(\mathbf{x}) \triangleq \int_{\mathbf{u}\le\mathbf{x}} \phi_K(\mathbf{u})\,d\mathbf{u}$$

for the multivariate Gaussian distribution function. For the univariate case we simply write $\phi(x)$ and $\Phi(x)$ for the standard $\mathcal{N}(0,1)$ Gaussian density and distribution, respectively.

All logarithms are to base e.

We use standard asymptotic order notation with the following caveats: 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 "big-oh" notation. In addition, we will find it expedient to occasionally use the more graphic $h_n \ll g_n$ and $h_n \gg g_n$ to mean $h_n = \mathfrak{o}(g_n)$ and $h_n = \omega(g_n)$, respectively.

In what follows we will be concerned with asymptotics as $n \to \infty$, and we will allow the number of elements m in the random vertex set $U(m)$ to depend on n. As a notational convention, however, we shall frequently write simply m instead of the more explicit m_n, while keeping in mind that m is to be thought of as a function of n.

Technical lemmas. The method of proof of the Main Theorem is to reduce the problem to the study of a random walk $\mathbf{S}_n$ where, for each n, $\mathbf{S}_n$ is the row sum of a triangular array of lattice variables. The principal result that will be needed is a sharp estimate for the probability of large deviations of the walk $\mathbf{S}_n$. The setup is as follows.

Let k be a fixed positive integer and consider a triangular array of k-dimensional lattice random vectors

$$\mathbf{X}_{ni} = \bigl(X_{ni}^{(1)}, \dots, X_{ni}^{(k)}\bigr) \qquad (i = 1, \dots, n;\ n = 1, 2, \dots),$$

where, for each n, the random vectors $\mathbf{X}_{n1}, \dots, \mathbf{X}_{nn}$ comprising the nth row of the array are independent, identically distributed lattice random vectors with probability one support in $\{0,1\}^k$ and with common distribution $p_n(\mathbf{x}) = \mathbb{P}\bigl\{\mathbf{X}_{ni} = \mathbf{x}\bigr\}$, where $p_n(\mathbf{0}) > 0$ and $p_n(\mathbf{e}_\nu) > 0$ for each of the canonical unit vectors $\mathbf{e}_\nu \in \{0,1\}^k$ $(1 \leq \nu \leq k)$.[3] Observe that the distribution of $\mathbf{X}_{ni}$ has minimal lattice $\mathbb{Z}^k$. Write $\boldsymbol{\mu}_n = \bigl(\mu_n^{(1)}, \dots, \mu_n^{(k)}\bigr) \triangleq \mathbb{E}\bigl(\mathbf{X}_{ni}\bigr)$ for the mean vector and $V_n \triangleq \operatorname{Cov}\bigl(\mathbf{X}_{ni}\bigr) = \mathbb{E}\bigl(\mathbf{X}'_{ni}\mathbf{X}_{ni}\bigr) - \boldsymbol{\mu}'_n\boldsymbol{\mu}_n$ for the covariance matrix.

Specializing to the case of interest, we assume that there exists a discrete probability distribution $p(\mathbf{x})$ with probability one support in the vertices of the cube $\mathbf{x} \in \{0,1\}^k$ such that $p_n(\mathbf{x}) \to p(\mathbf{x})$ as $n \to \infty$ for each $\mathbf{x} \in \{0,1\}^k$. We suppose that, for each n,

$$\operatorname{Cov}\bigl(X_{ni}^{(t)}, X_{ni}^{(s)}\bigr) = \mathbb{E}\bigl\{\bigl(X_{ni}^{(t)} - \mu_n^{(t)}\bigr)\bigl(X_{ni}^{(s)} - \mu_n^{(s)}\bigr)\bigr\} = \begin{cases} \sigma_n^2 & \text{if } t = s, \\ \psi_n & \text{if } t \neq s, \end{cases}$$

whence the covariance matrix of $\mathbf{X}_{ni}$ is of the form

$$V_n = \operatorname{Cov}\bigl(\mathbf{X}_{ni}\bigr) = \begin{bmatrix} \sigma_n^2 & \psi_n & \psi_n & \dots & \psi_n \\ \psi_n & \sigma_n^2 & \psi_n & \dots & \psi_n \\ \dots & \dots & \dots & \dots & \dots \\ \psi_n & \psi_n & \psi_n & \dots & \sigma_n^2 \end{bmatrix}.$$

We will suppose that, as $n \to \infty$, $\psi_n \to 0$ and $\sigma_n^2 \to \sigma^2$ for some positive constant σ^2. In particular, observe that V_n is nonsingular for sufficiently large n and $|V_n| \to \sigma^{2k}$ as $n \to \infty$.

[3] For $k > 1$ we can dispense with the condition $p_n(\mathbf{0}) > 0$.

For each n, form the lattice random vector $\mathbf{S}_n$ as the row sum

$$\mathbf{S}_n = \sum_{i=1}^{n} \mathbf{X}_{ni}.$$

We are interested in the tails of the random walk $\mathbf{S}_n$ in $\mathbb{Z}^k$. Write

$$\mathbf{S}_n^* = \frac{\mathbf{S}_n - n\boldsymbol{\mu}_n}{\sqrt{n}}$$

for the normalized row sum. The following result is a global version of an extension of a large deviation local limit theorem for sums of independent random vectors due to Richter [8, Theorem 2] to row sums of triangular arrays. The proof of the result follows easily along the lines of Richter's proof and is sketched in [2]. We omit the derivation here.

Lemma 2.1 (large deviation global limit theorem). *Suppose $\boldsymbol{\xi}_n = \bigl(\xi_n^{(1)}, \dots, \xi_n^{(k)}\bigr)$ is any sequence of positive real vectors whose components satisfy $1 \ll \xi_n^{(t)} \ll n^{1/6}$ $(1 \le t \le k)$ as $n \to \infty$. Then under the previous assumptions*

$$\mathbb{P}\{\mathbf{S}_n^* > \boldsymbol{\xi}_n\} \sim \Phi_{V_n}(-\boldsymbol{\xi}_n) \quad \textit{and likewise} \quad \mathbb{P}\{\mathbf{S}_n^* < -\boldsymbol{\xi}_n\} \sim \Phi_{V_n}(-\boldsymbol{\xi}_n)$$

as $n \to \infty$.

Remark. Observe that asymptotic normality persists for deviations of $\mathbf{S}_n$ from the mean as large as $\mathfrak{o}\bigl(n^{2/3}\bigr)$, which admits of deviations much larger than the $\mathcal{O}\bigl(\sqrt{n}\,\bigr)$ deviations, which are the province of the standard central limit theorem.

It will be convenient in the analysis to find an elementary estimate of the multivariate Gaussian tail in Lemma 2.1. For the univariate case, the classical estimate of the tail $\Phi(-x)$ of the Gaussian can be expressed in terms of Mill's ratio in the form

$$\frac{\Phi(-x)}{\phi(x)} \sim x^{-1} \qquad (x \to \infty).$$

(See Feller [3, Lemma VII.1.2], for example.) The following specialization of a result of Ruben yields an analogous result for the multivariate case. We refer the reader to Ruben's paper [9] for the proof.

Lemma 2.2 (multivariate Mill's ratio). *Let $\{A(\rho_n)\}$ be a sequence of $k \times k$ covariance matrices, where $A(\rho_n)$ has unity as its diagonal elements and ρ_n as its off-diagonal elements, and suppose $\rho_n \to 0$ as $n \to \infty$. Let $\{x_n\}$ be any positive sequence satisfying $x_n \to \infty$ as $n \to \infty$. Then, writing $\mathbf{1} \in \mathbb{R}^k$ for the vector with all components identically 1, we have the asymptotic estimate*

$$\Phi_{A(\rho_n)}(-x_n\mathbf{1}) \sim x_n^{-k}\phi_{A(\rho_n)}(x_n\mathbf{1})$$

as $n \to \infty$.

Observe that the classical estimate for the tail of the univariate Gaussian follows directly with $k = 1$ and $K = [1]$.

The final technical result that will be needed is a probabilistic sieve. Suppose $\bigl\{B_n^{(t)}\bigr\}$ is a triangular array of events in a probability space with m_n events $(1 \le t \le m_n)$ in the nth row, and let $\bigl\{\mathfrak{z}_n^{(t)}\bigr\}$ denote the corresponding triangular array of indicator random variables for these events. Let $Z_{n,m_n} = \sum_{t=1}^{m_n} \mathfrak{z}_n^{(t)}$ denote the number of events that occur simultaneously in the nth row. We will be interested in the limiting distribution of the row sums Z_{n,m_n}.

Additionally, for each $k = 1, \dots, m_n$ and each $n = 1, 2, \dots$, define

$$S^{(k)}_{n,m_n} = \sum \mathbb{P}\bigl\{B^{(t_1)}_n \cap \cdots \cap B^{(t_k)}_n\bigr\} = \sum \mathbb{E}\bigl(\mathfrak{z}^{(t_1)}_n \cdots \mathfrak{z}^{(t_k)}_n\bigr),$$

where the sum is over all subsets $\{t_1, \dots, t_k\}$ of cardinality k drawn from $\{1, \dots, m_n\}$. Observe that $S^{(1)}_{n,m_n} = \mathbb{E}\, Z_{n,m_n}$, while, in general, $S^{(k)}_{n,m_n} = \mathbb{E}\bigl(Z^{[k]}_{n,m_n}/k!\bigr)$, where $Z^{[k]}_{n,m_n}$ denotes the number of ordered k-sets of events (no repetition) in the nth row for which all k events occur simultaneously.

LEMMA 2.3 (Poisson tendency). *Suppose there is a constant λ such that, for every fixed k, $S^{(k)}_{n,m_n} \to \lambda^k/k!$ as $n \to \infty$. Then Z_{n,m_n} converges in distribution to* $\mathrm{Po}(\lambda)$*, the Poisson distribution with parameter λ. In particular, for every fixed k,*

$$\mathbb{P}\{Z_{n,m_n} = k\} \to \frac{\lambda^k}{k!}\, e^{-\lambda}$$

as $n \to \infty$.

The proof follows directly from inclusion and exclusion by use of Bonferroni's inequalities to bound on both sides the probability $\mathbb{P}\{Z_{n,m_n} = k\}$ that exactly k of the events in the nth row occur simultaneously (cf. Feller [3, Theorem IV.3.1], for instance). We omit the standard proof.

3. Proof of Main Theorem.

We are interested in the probability

$$P(n,m) = \mathbb{P}\biggl\{\bigcap_{t=1}^{m} \bigl\{\langle \mathbf{w}, \mathbf{u}^{(t)}\rangle \geq 0\bigr\}\biggr\} = 1 - \mathbb{P}\biggl\{\bigcup_{t=1}^{m} \bigl\{\langle \mathbf{w}, \mathbf{u}^{(t)}\rangle < 0\bigr\}\biggr\}$$

of the event that $\mathbf{w} = \mathbf{w}^{(m+1)}$ has positive inner product with each vertex in $U(m)$. The following gives a thumbnail sketch of the principal ideas involved in the estimation of $P(n,m)$. We begin by showing via elementary arguments that the random summands $Y^{(t)}_i \triangleq w_i u^{(t)}_i$ $(1 \leq t \leq m)$ are exchangeable and then evaluate the first two mixed moments. Invoking Lemma 2.1, we then proceed to show that the events $\bigl\{\langle \mathbf{w}, \mathbf{u}^{(t)}\rangle < 0\bigr\}$ are governed asymptotically by a normal law even though the tail probabilities of interest correspond to deviations from the mean rather larger than the $\mathcal{O}\bigl(\sqrt{n}\,\bigr)$ deviations that fall under the usual province of the central limit theorem. Direct calculations of the relevant probabilities are still difficult, however, because of insidious statistical dependencies, albeit somewhat weak, evinced in the events $\bigl\{\langle \mathbf{w}, \mathbf{u}^{(t)}\rangle < 0\bigr\}$. The next stage in the proof involves quelling these dependencies with a firm hand using Lemma 2.2 to conclude that the events of interest are "asymptotically independent." The method of inclusion and exclusion embodied in Lemma 2.3 then allows us to conclude that, in the range of interest, the distribution of errors $\bigl\{\langle \mathbf{w}, \mathbf{u}^{(t)}\rangle < 0\bigr\}$ approaches a Poisson distribution asymptotically. The final stage of the calculation is a relatively straightforward bootstrap which rapidly produces an estimate of the critical sample size m by successive approximation.

A. Exchangeable random variables.

Since the m-set of examples $U(m) = \bigl\{\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(m)}\bigr\}$ is generated by independent sampling from the uniform distribution on the vertices $\mathbb{B}^n$, it follows that the example components $\bigl\{u^{(t)}_i, 1 \leq t \leq m, 1 \leq i \leq n\bigr\}$ are independent, identically distributed random variables taking values -1 and $+1$ only, each with probability $1/2$. Now, for each i, the ith memory component

$w_i = w_i^{(m+1)}$ is a (random) function solely of $u_i^{(1)}, \dots, u_i^{(m)}$. It is clear by symmetry that for every sample path which results in $w_i = +1$ there exists a sample path of equal probability (its reflection) which results in $w_i = -1$, and vice versa. Consequently, w_i is a symmetric Bernoulli random variable taking values -1 and $+1$ only, each with probability 1/2. Furthermore, as i runs through 1 to n, the sets $\{u_i^{(1)}, \dots, u_i^{(m)}\}$ partition the set of mn example components $\{u_i^{(t)}\}$ into n disjoint, identically distributed subsets. It follows that the memory components $\{\, w_i, 1 \le i \le n \,\}$ are independent, identically distributed symmetric binary random variables.

We begin with a preliminary result. Fix the index i and recall that $w_i^{(r)} \in \mathbb{B}$ represents the state of the ith bit of memory following the presentation of the $(r-1)$th example $\mathbf{u}^{(r-1)}$. Write

$$p_{r;t_1,\dots,t_k} \triangleq \mathbb{P}\{w_i^{(r)} = u_i^{(t_1)} = \cdots = u_i^{(t_k)}\}$$

for every choice of epochs $t_1, \dots, t_k$, r.

Assertion 1. *Let k and r be any fixed integers in the range $1 \le k < r \le m+1$. Then*

$$p_{r;t_1,\dots,t_k} = 2^{-k}\bigl(1 + \tfrac{k}{r-1}\bigr)$$

for every selection of k distinct epochs $t_1, \dots, t_k$ satisfying $1 \le t_j \le r-1$ $(1 \le j \le k)$.

Proof. To keep the notation unencumbered, suppress the subscript i and simply write $w^{(r)}$ and $u^{(t)}$ instead of the explicit $w_i^{(r)}$ and $u_i^{(t)}$. We may also assume without loss of generality that the indices are so ordered that $1 \le t_1 < t_2 < \cdots < t_k \le r-1$ as $p_{r;t_1,\dots,t_k}$ is invariant with respect to the permutation of the indices $t_1, \dots, t_k$. The proof of the assertion is by induction over r, k, and $t_1, \dots, t_k$.

`Induction base:` With $k = 1$, $t_1 = 1$, and $r = 2$, we have

$$p_{2;1} = \mathbb{P}\{w^{(2)} = u^{(1)}\} = 1 = \tfrac{1}{2}(1+1).$$

`Induction hypothesis:` For some $2 \le r \le m$, suppose that

$$p_{r;t_1,\dots,t_k} = 2^{-k}\bigl(1 + \tfrac{k}{r-1}\bigr) \qquad (1 \le k \le r-1;\ 1 \le t_1 < \cdots < t_k \le r-1).$$

Now consider $p_{r+1;t_1,\dots,t_k}$. We break the induction into two cases.

Case 1: $1 \le k \le r-1$, $1 \le t_1 < \cdots < t_k \le r-1$. Conditioning on $w^{(r)}$, we obtain

$$\begin{aligned}\mathbb{P}\{&w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\}\\ &= \mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = 1\}\,\mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\}\\ &\quad + \mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = -1\}\,\mathbb{P}\{w^{(r)} = -1, u^{(t_1)} = \cdots = u^{(t_k)} = 1\}\end{aligned}$$

as $w^{(r+1)}$ is conditionally independent of $u^{(t_1)}, \dots, u^{(t_k)}$ given $w^{(r)}$ for $1 \le t_1 < \cdots < t_k \le r-1$. The conditional probabilities are readily evaluated: condition on $u^{(r)}$ and exploit the independence of $u^{(r)}$ and $w^{(r)}$ to obtain

$$\begin{aligned}\mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = 1\} &= \tfrac{1}{2}\,\mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = 1, u^{(r)} = 1\}\\ &\quad + \tfrac{1}{2}\,\mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = 1, u^{(r)} = -1\}\\ &= \tfrac{1}{2} + \tfrac{1}{2}\bigl(1 - \tfrac{1}{r}\bigr) = 1 - \tfrac{1}{2r}.\end{aligned} \tag{3.1}$$

The reflection principle shows that the random variables $\{ w^{(r)}, 2 \le r \le m+1 \}$ have symmetric marginal distributions

$$\mathbb{P}\{w^{(r)} = -1\} = \mathbb{P}\{w^{(r)} = +1\} = \tfrac{1}{2} \qquad (2 \le r \le m+1).$$

A simple application of Bayes's rule hence yields

$$\begin{aligned}\mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = -1\} &= \mathbb{P}\{w^{(r)} = -1 \mid w^{(r+1)} = 1\} = 1 - \mathbb{P}\{w^{(r)} = 1 \mid w^{(r+1)} = 1\} \\ &= 1 - \mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = 1\} = \tfrac{1}{2r},\end{aligned}$$

the last step following from (3.1). It follows that

$$\begin{aligned}\mathbb{P}\{w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\} &= \big(1 - \tfrac{1}{2r}\big)\, \mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\} \\ &\quad + \tfrac{1}{2r}\, \mathbb{P}\{w^{(r)} = -1, u^{(t_1)} = \cdots = u^{(t_k)} = 1\}.\end{aligned}$$

Now observe that

$$\mathbb{P}\{w^{(r)} = -1, u^{(t_1)} = \cdots = u^{(t_k)} = 1\} = 2^{-k} - \mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\},$$

whence

$$\mathbb{P}\{w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\} = \tfrac{2^{-k}}{2r} + \big(1 - \tfrac{1}{r}\big)\, \mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\}.$$

Likewise,

$$\mathbb{P}\{w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_k)} = -1\} = \tfrac{2^{-k}}{2r} + \big(1 - \tfrac{1}{r}\big)\, \mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_k)} = -1\}.$$

It follows that for $1 \le k \le r-1$ and $1 \le t_1 < \cdots < t_k \le r-1$

$$p_{r+1;t_1,\dots,t_k} = \tfrac{2^{-k}}{r} + \big(1 - \tfrac{1}{r}\big) p_{r;t_1,\dots,t_k} = 2^{-k}\big(1 + \tfrac{k}{r}\big)$$

by the induction hypothesis.

Case 2: $1 \le k \le r$, $1 \le t_1 < \cdots < t_k = r$. Conditioning on $w^{(r)}$ again, we obtain

$$\begin{aligned}\mathbb{P}\{w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\} &= \mathbb{P}\{w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_{k-1})} = u^{(r)} = 1\} \\ = \tfrac{1}{2}\, \mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = -1, u^{(r)} = 1\}&\, \mathbb{P}\{w^{(r)} = -1, u^{(t_1)} = \cdots = u^{(t_{k-1})} = 1\} \\ + \tfrac{1}{2}\, \mathbb{P}\{w^{(r+1)} = 1 \mid w^{(r)} = u^{(r)} = 1\}&\, \mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_{k-1})} = 1\}\end{aligned}$$

as $w^{(r+1)}$ is conditionally independent of $u^{(t_1)}, \dots, u^{(t_{k-1})}$ given $w^{(r)}$ and $u^{(r)}$, and $u^{(r)}$ is independent of $w^{(r)}$ and $u^{(t_1)}, \dots, u^{(t_{k-1})}$.[4] The conditional probabilities above are completely determined by the auxiliary randomization in the algorithm. Hence

$$\begin{aligned}\mathbb{P}\{w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\} &= \tfrac{1}{2}\, \mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_{k-1})} = 1\} \\ &\quad + \tfrac{1}{2r}\, \mathbb{P}\{w^{(r)} = -1, u^{(t_1)} = \cdots = u^{(t_{k-1})} = 1\}.\end{aligned}$$

Now observe that

$$\mathbb{P}\{w^{(r)} = -1, u^{(t_1)} = \cdots = u^{(t_{k-1})} = 1\} = 2^{-(k-1)} - \mathbb{P}\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_{k-1})} = 1\},$$

[4] As usual, we identify the joint event $\bigcap_{j=1}^{k-1} \{u^{(t_j)} = 1\}$ with the certain event if $k = 1$.

whence

$$\mathbb{P}\bigl\{w^{(r+1)} = u^{(t_1)} = \cdots = u^{(t_k)} = 1\bigr\} = \tfrac{2^{-k}}{r} + \bigl(\tfrac{1}{2} - \tfrac{1}{2r}\bigr)\,\mathbb{P}\bigl\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_{k-1})} = 1\bigr\}.$$

A completely analogous argument gives

$$\begin{aligned}\mathbb{P}\bigl\{w^{(r+1)} = u^{(t_1)} = \ldots = u^{(t_{k-1})} = u^{(r)} = -1\bigr\}&\\ = \tfrac{2^{-k}}{r} + \bigl(\tfrac{1}{2} - \tfrac{1}{2r}\bigr)\,\mathbb{P}\bigl\{w^{(r)} = u^{(t_1)} = \cdots = u^{(t_{k-1})} = -1\bigr\}&.\end{aligned}$$

It follows that

$$p_{r+1;t_1,\ldots,t_{k-1},r} = \tfrac{2^{-k+1}}{r} + \tfrac{1}{2}\bigl(1 - \tfrac{1}{r}\bigr)p_{r;t_1,\ldots,t_{k-1}}.$$

Applying the induction hypothesis, it follows that

$$p_{r+1;t_1,\ldots,t_k} = 2^{-k}\bigl(1 + \tfrac{k}{r}\bigr) \qquad (1 \le k \le r;\ 1 \le t_1 < \cdots < t_{k-1} < t_k = r),$$

as was to be shown. The two cases taken together completes the induction. □

Now define the random variables

$$Y_i^{(t)} \triangleq w_i u_i^{(t)} \qquad (1 \le t \le m;\ 1 \le i \le n).$$

Observe that, for each t, $\bigl\langle \mathbf{w}, \mathbf{u}^{(t)} \bigr\rangle = \sum_{i=1}^n Y_i^{(t)}$ is a sum of independent, identically distributed ±1 random variables, i.e., a random walk on the line. The walk is asymmetric with a positive drift, as we shall see shortly.

In what follows it will be slightly more convenient to consider the related $(0,1)$ random variables

$$X_i^{(t)} = \tfrac{1}{2}\bigl(1 + Y_i^{(t)}\bigr) \qquad (1 \le t \le m;\ 1 \le i \le n),$$

whence $\bigl\langle \mathbf{w}, \mathbf{u}^{(t)} \bigr\rangle = 2\sum_{i=1}^n X_i^{(t)} - n$. Consider the random set $\mathcal{X}_i \triangleq \bigl\{X_i^{(1)}, \ldots, X_i^{(m)}\bigr\}$. Since w_i is a function only of $u_i^{(1)}, \ldots, u_i^{(m)}$, it follows that $\mathcal{X}_i$ is also determined only by $u_i^{(1)}, \ldots, u_i^{(m)}$. Consequently, the random sets $\mathcal{X}_1, \ldots, \mathcal{X}_n$ are statistically independent and have identical (joint) distributions. Assertion 1 now allows us to explicitly characterize the joint distribution of the random set $\mathcal{X}_i$.

For every choice of epochs $t_1, \ldots, t_{\mathfrak{h}}$, $t_{\mathfrak{h}+1}, \ldots, t_{\mathfrak{h}+\mathfrak{k}}$, write

$$q_{m;t_1,\ldots,t_{\mathfrak{h}};t_{\mathfrak{h}+1},\ldots,t_{\mathfrak{h}+\mathfrak{k}}} \triangleq \mathbb{P}\bigl\{X_i^{(t_1)} = 1, \ldots, X_i^{(t_{\mathfrak{h}})} = 1, X_i^{(t_{\mathfrak{h}+1})} = 0, \ldots, X_i^{(t_{\mathfrak{h}+\mathfrak{k}})} = 0\bigr\}.$$

It will also be convenient to define

$$f_m(\mathfrak{h}, \mathfrak{k}) \triangleq 2^{-(\mathfrak{h}+\mathfrak{k})}\bigl(1 + \tfrac{\mathfrak{h}-\mathfrak{k}}{m}\bigr).$$

Observe that $p_{m+1;t_1,\ldots,t_k} = f_m(k, 0)$ by Assertion 1. As an immediate consequence, we have the following.

Assertion 2. *The* $(0,1)$ *random variables* $X_i^{(1)}, \ldots, X_i^{(m)}$ *are exchangeable. In particular,*

$$q_{m;t_1,\ldots,t_{\mathfrak{h}};t_{\mathfrak{h}+1},\ldots,t_{\mathfrak{h}+\mathfrak{k}}} = f_m(\mathfrak{h}, \mathfrak{k}) \tag{3.2}$$

for every pair of nonnegative integers $\mathfrak{h}$ *and* $\mathfrak{k}$ *with* $\mathfrak{h} + \mathfrak{k} \le m$ *and* $\mathfrak{h} + \mathfrak{k}$ *distinct indices* $t_1, \ldots, t_{\mathfrak{h}+\mathfrak{k}}$ *in* $\{1, \ldots, m\}$.

Remark. In accordance with usual convention, we identify the joint event

$$\bigcap_{j=1}^{\mathfrak{h}} \bigl\{X_i^{(t_j)} = 1\bigr\} \cap \bigcap_{j=\mathfrak{h}+1}^{\mathfrak{h}+\mathfrak{k}} \bigl\{X_i^{(t_j)} = 0\bigr\}$$

with the certain event if $\mathfrak{h} = \mathfrak{k} = 0$. The assertion then holds when $\mathfrak{h} = \mathfrak{k} = 0$ as well when the desired probability is identically $f_m(0, 0) = 1$.

Proof. The result follows quickly by induction on $\mathfrak{k}$.

`Induction base:` When $\mathfrak{k} = 0$, it follows immediately that for every $0 \le \mathfrak{h} \le m$,

$$q_{m;t_1,\dots,t_\mathfrak{h}} = \mathbb{P}\bigl\{X_i^{(t_1)} = 1, \dots, X_i^{(t_\mathfrak{h})} = 1\bigr\} = p_{m+1;t_1,\dots,t_\mathfrak{h}} = 2^{-\mathfrak{h}}\bigl(1 + \tfrac{\mathfrak{h}}{m}\bigr) = f_m(\mathfrak{h}, 0)$$

depends only on $\mathfrak{h}$ and m and is independent of the choice of (distinct) indices $t_1, \dots, t_\mathfrak{h}$. The proof is completed by induction over $\mathfrak{k}$.

`Induction hypothesis:` Suppose that for some choice of $\mathfrak{k} \ge 0$, (3.2) holds for every $\mathfrak{h} \ge 0$ with $\mathfrak{h} + \mathfrak{k} \le m$ and every choice of $\mathfrak{h} + \mathfrak{k}$ distinct indices $t_1, \dots, t_\mathfrak{h}$, $t_{\mathfrak{h}+1}, \dots, t_{\mathfrak{h}+\mathfrak{k}}$ in $\{1, \dots, m\}$. It follows that, for every $\mathfrak{h} \ge 0$ with $\mathfrak{h} + \mathfrak{k} + 1 \le m$ and every distinct collection of $\mathfrak{h} + \mathfrak{k} + 1$ indices $t_1, \dots, t_\mathfrak{h}$, $t_{\mathfrak{h}+1}$, $t_{\mathfrak{h}+2}, \dots, t_{\mathfrak{h}+\mathfrak{k}+1}$ in $\{1, \dots, m\}$,

$$\begin{aligned} q_{m;t_1,\dots,t_\mathfrak{h};t_{\mathfrak{h}+1},t_{\mathfrak{h}+2},\dots,t_{\mathfrak{h}+\mathfrak{k}+1}} &= q_{m;t_1,\dots,t_\mathfrak{h};t_{\mathfrak{h}+2},\dots,t_{\mathfrak{h}+\mathfrak{k}+1}} - q_{m;t_1,\dots,t_\mathfrak{h},t_{\mathfrak{h}+1};t_{\mathfrak{h}+2},\dots,t_{\mathfrak{h}+\mathfrak{k}+1}} \\ &= f_m(\mathfrak{h}, \mathfrak{k}) - f_m(\mathfrak{h}+1, \mathfrak{k}) \\ &= f_m(\mathfrak{h}, \mathfrak{k}+1), \end{aligned}$$

the penultimate step following from the induction hypothesis and the last step following by the definition of f_m. This completes the induction. □

The moments of the random variables $X_i^{(1)}, \dots, X_i^{(m)}$ are now readily determined. In particular,

$$\begin{aligned} \mu &\triangleq \mathbb{E}\bigl(X_i^{(t)}\bigr) = f_m(1,0) = \tfrac{1}{2} + \tfrac{1}{2m}, \\ \sigma^2 &\triangleq \operatorname{Var}\bigl(X_i^{(t)}\bigr) = f_m(1,0) - f_m(1,0)^2 = \tfrac{1}{4} - \tfrac{1}{4m^2}, \\ \psi &\triangleq \operatorname{Cov}\bigl(X_i^{(s)}, X_i^{(t)}\bigr) = f_m(2,0) - f_m(1,0)^2 = -\tfrac{1}{4m^2} \qquad (s \ne t), \\ \rho &\triangleq \tfrac{\psi}{\sigma^2} = -\tfrac{1}{m^2-1} = \mathcal{O}\bigl(m^{-2}\bigr). \end{aligned}$$

B. Normal tendency. Now consider the random walks

$$\bigl\langle \mathbf{w}, \mathbf{u}^{(t)} \bigr\rangle = \sum_{i=1}^{n} w_i u_i^{(t)} = 2\sum_{i=1}^{n} X_i^{(t)} - n \qquad (1 \le t \le m).$$

Recall that the n random sets $\mathcal{X}_i = \bigl\{X_i^{(1)}, \dots, X_i^{(m)}\bigr\}$ $(1 \le i \le n)$ are mutually independent with identical joint distributions. It now follows as a consequence of Assertion 2 that the random variables $\bigl\langle \mathbf{w}, \mathbf{u}^{(1)} \bigr\rangle, \dots, \bigl\langle \mathbf{w}, \mathbf{u}^{(m)} \bigr\rangle$ are exchangeable.

Let k be any fixed positive integer and consider any distinct set of k indices $t_1, \dots, t_k$ in $\{1, \dots, m\}$. Write

$$\mathfrak{P}_{n,m}(k) \triangleq \mathbb{P}\bigl\{\bigl\langle \mathbf{w}, \mathbf{u}^{(t_1)} \bigr\rangle < 0, \dots, \bigl\langle \mathbf{w}, \mathbf{u}^{(t_k)} \bigr\rangle < 0\bigr\}.$$

Note that $\mathfrak{P}_{n,m}(k)$ is just the probability that any given k inequalities in (1.1) are violated, where the random $\mathbf{w}$ is determined by harmonic update. Since the random walks $\langle \mathbf{w}, \mathbf{u}^{(t)} \rangle$ $(1 \le t \le m)$ are exchangeable random variables, $\mathfrak{P}_{n,m}(k)$ does not depend on the specific choice of indices t_j and we may suppose without any loss of generality that $t_j = j$ $(1 \le j \le k)$. Thus,

$$\mathfrak{P}_{n,m}(k) = \mathbb{P}\bigl\{\langle \mathbf{w}, \mathbf{u}^{(1)} \rangle < 0, \dots, \langle \mathbf{w}, \mathbf{u}^{(k)} \rangle < 0\bigr\} = \mathbb{P}\biggl\{\sum_{i=1}^{n} X_i^{(1)} < \tfrac{n}{2}, \dots, \sum_{i=1}^{n} X_i^{(k)} < \tfrac{n}{2}\biggr\}.$$

Let us now explicitly allow $m = m_n$ to vary with n and acknowledge this dependence on n by writing,

$$X_i^{(t)} = X_{ni}^{(t)} = \tfrac{1}{2}\bigl(1 + w_i u_i^{(t)}\bigr) \qquad (1 \le t \le m_n;\ 1 \le i \le n),$$

where $u_i^{(t)} = u_{ni}^{(t)}$ $(1 \le t \le m_n; 1 \le i \le n)$ are the components of the random examples and $w_i = w_{ni}$ $(1 \le i \le n)$ are the components of the memory state determined by harmonic update. Now consider the triangular array of k-dimensional lattice random vectors

$$\mathbf{X}_{ni} = \bigl(X_{ni}^{(1)}, \dots, X_{ni}^{(k)}\bigr) \qquad (i = 1, \dots, n;\ n = 1, 2, \dots).$$

For each n, the random vectors $\mathbf{X}_{n1}, \dots, \mathbf{X}_{nn}$ comprising the nth row of the array are independent, identically distributed lattice random vectors with probability one support in $\{0,1\}^k$. Let $p_n(\mathbf{x}) = \mathbb{P}\bigl\{\mathbf{X}_{ni} = \mathbf{x}\bigr\}$ $(\mathbf{x} \in \{0,1\}^k)$ denote the distribution of $\mathbf{X}_{ni}$. Write $|\mathbf{x}|$ for the number of components of $\mathbf{x} \in \{0,1\}^k$ that take value one (this is just the L^1 vector norm in this case). Observe then, as a consequence of Assertion 2, that

$$p_n(\mathbf{x}) = f_{m_n}\bigl(|\mathbf{x}|, k - |\mathbf{x}|\bigr) = \frac{1}{2^k}\left(1 + \frac{2|\mathbf{x}| - k}{m_n}\right) \qquad \bigl(\mathbf{x} \in \{0,1\}^k\bigr).$$

Suppose $m_n \to \infty$ as $n \to \infty$. Then, for sufficiently large n, $p_n(\mathbf{x}) > 0$ for every $\mathbf{x} \in \{0,1\}^k$; in particular, $\mathbf{X}_{ni}$ takes values $\mathbf{0}$, $\mathbf{e}_1, \dots, \mathbf{e}_k$ (where $\mathbf{e}_\nu$ $(1 \le \nu \le k)$ denotes the canonical unit vectors in $\{0,1\}^k$) with positive probability. Furthermore, $p_n(\mathbf{x}) \to 2^{-k}$ (uniformly) for all $\mathbf{x} \in \{0,1\}^k$, whence the distribution of $\mathbf{X}_{ni}$ becomes uniform over $\{0,1\}^k$ in the limit.

Let us also explicitly write

$$\begin{gathered}\mu = \mu_n = \tfrac{1}{2} + \tfrac{1}{2m_n},\\ \sigma^2 = \sigma_n^2 = \tfrac{1}{4} - \tfrac{1}{4m_n^2},\\ \rho = \rho_n = -\tfrac{1}{m_n^2 - 1} = \mathcal{O}\bigl(m_n^{-2}\bigr)\end{gathered}$$

for the mean, variance, and correlation coefficient, respectively, of the $(0,1)$ random variables $X_{ni}^{(t)}$ $(1 \le t \le m_n)$. We then have

$$\begin{gathered}\mathbb{E}\bigl(\mathbf{X}_{ni}\bigr) = \boldsymbol{\mu}_n = \mu_n \mathbf{1},\\ \operatorname{Cov}\bigl(\mathbf{X}_{ni}\bigr) = V_n = \sigma_n^2 A(\rho_n),\end{gathered}$$

where, as before, $A(\rho_n)$ denotes a $k \times k$ covariance matrix which has unity as its diagonal elements and ρ_n as its off-diagonal elements. Observe that $\mu_n \to \frac{1}{2}$, $\sigma_n^2 \to \frac{1}{4}$, and $\rho_n \to 0$ if $m_n \gg 1$ as $n \to \infty$.

Everything is now set for an application of Lemma 2.1.

ASSERTION 3. *If $m = m_n$ increases with n in such a way that $n^{1/3} \ll m_n \ll n^{1/2}$ then, for every fixed positive integer k,*

$$\mathfrak{P}_{n,m_n}(k) \sim \Phi_{A(\rho_n)}\Bigl(-\tfrac{\sqrt{n}}{m_n}\,\mathbf{1}\Bigr)$$

as $n \to \infty$.

Proof. Form the row sums and the corresponding normalized row sums

$$\mathbf{S}_n = \sum_{i=1}^{n} \mathbf{X}_{ni}, \qquad \mathbf{S}_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bigl(\mathbf{X}_{ni} - \boldsymbol{\mu}_n\bigr),$$

and define the sequence

$$\xi_n \triangleq \bigl(\mu_n - \tfrac{1}{2}\bigr)\sqrt{n} = \frac{\sqrt{n}}{2m_n}.$$

We then obtain

$$\mathfrak{P}_{n,m_n}(k) = \mathbb{P}\bigl\{\mathbf{S}_n < \tfrac{n}{2}\,\mathbf{1}\bigr\} = \mathbb{P}\bigl\{\mathbf{S}_n^* < -\xi_n\mathbf{1}\bigr\}.$$

Observe that in the range $n^{1/3} \ll m_n \ll n^{1/2}$ we have $1 \ll \xi_n \ll n^{1/6}$ as $n \to \infty$. Applying Lemma 2.1, we hence obtain

$$\mathfrak{P}_{n,m_n}(k) \sim \Phi_{V_n}\Bigl(\tfrac{-\sqrt{n}}{2m_n}\,\mathbf{1}\Bigr) = \Phi_{A(\rho_n)}\Bigl(\tfrac{-\sqrt{n}}{2\sigma_n m_n}\,\mathbf{1}\Bigr) \qquad (n \to \infty).$$

Now observe that

$$\tfrac{\sqrt{n}}{2\sigma_n m_n} = \tfrac{\sqrt{n}}{m_n}\bigl(1 - m_n^{-2}\bigr)^{-1/2} = \tfrac{\sqrt{n}}{m_n}\bigl[1 + \mathcal{O}\bigl(m_n^{-2}\bigr)\bigr] = \tfrac{\sqrt{n}}{m_n} + \mathcal{O}\bigl(\tfrac{\sqrt{n}}{m_n^3}\bigr) = \tfrac{\sqrt{n}}{m_n} + \mathfrak{o}(1)$$

when $m_n \gg n^{1/3}$. Furthermore, $\rho_n \to 0$ as $n \to \infty$, whence the covariance matrix $A(\rho_n)$ and its inverse converge componentwise to the identity matrix. Consequently, $\mathfrak{P}_{n,m_n}(k) \sim \Phi_{A(\rho_n)}\bigl(\tfrac{-\sqrt{n}}{m_n}\,\mathbf{1}\bigr)$ as $n \to \infty$. □

We can parlay the asymptotic normality of finite sets of the random variables $\bigl\langle \mathbf{w}, \mathbf{u}^{(t)} \bigr\rangle$ into a statement about the distribution of errors. This follows next.

C. Poisson tendency. We begin by exploiting the fact that the covariances between the random variables $X_i^{(t)}$ $(1 \le t \le m_n)$ vanish asymptotically. Coupling this with the exchangeability of the events $\bigl\{\bigl\langle \mathbf{w}, \mathbf{u}^{(t)} \bigr\rangle < 0\bigr\}$, we will indeed be able to show that, for a suitable rate of increase of $m = m_n$ with n, the distribution of errors $\bigl\langle \mathbf{w}, \mathbf{u}^{(t)} \bigr\rangle < 0$ is asymptotically Poisson. This will suffice to complete the proof of the Main Theorem.

It will be convenient to first define the double sequence

$$\mathfrak{Q}_{n,m} \triangleq \frac{m}{\sqrt{2\pi n}} \exp\Bigl\{-\frac{n}{2m^2}\Bigr\}$$

before launching into the result.

ASSERTION 4. *Suppose $m = m_n$ increases with n such that $n^{1/3} \ll m_n \ll n^{1/2}$. Then, for every fixed positive integer k, $\mathfrak{P}_{n,m_n}(k) \sim \mathfrak{P}_{n,m_n}(1)^k \sim \mathfrak{Q}_{n,m_n}^k$ as $n \to \infty$.*

Proof. Write $x_n \triangleq m_n^{-1}\sqrt{n}$. Note that $x_n \to \infty$ as $n \to \infty$ for the given rate of growth of m_n with n. Applying Assertion 3 to the case $k = 1$ yields $\mathfrak{P}_{n,m_n}(1) \sim \Phi(-x_n)$, while Lemma 2.2 shows that the right-hand side is asymptotic to $\mathfrak{Q}_{n,m_n}$ as

$n \to \infty$. Now suppose k is any fixed positive integer. Recall that $\rho_n = \mathcal{O}\bigl(m_n^{-2}\bigr) = \mathfrak{o}(1)$ as $m_n \gg 1$. We may hence apply Lemma 2.2 again to obtain

$$\begin{aligned}\mathfrak{P}_{n,m_n}(k) \sim \Phi_{A(\rho_n)}(-x_n\mathbf{1}) &\sim x_n^{-k}\phi_{A(\rho_n)}(x_n\mathbf{1})\\ &= (2\pi)^{-k/2}x_n^{-k}|A(\rho_n)|^{-1/2}e^{-\frac{1}{2}x_n^2\mathbf{1}A(\rho_n)^{-1}\mathbf{1}'}.\end{aligned}$$

Induction on k readily yields explicit expressions for the determinant and inverse of the covariance matrix $A(\rho_n)$,

$$\begin{aligned}|A(\rho_n)| &= \bigl(1+(k-1)\rho_n\bigr)(1-\rho_n)^{k-1},\\ A(\rho_n)^{-1} &= a_nA(b_n),\end{aligned}$$

where a_n and b_n are specified by

$$a_n = \frac{1+(k-2)\rho_n}{(1-\rho_n)[1+(k-1)\rho_n]} \quad \text{and} \quad b_n = \frac{-\rho_n}{1+(k-2)\rho_n}.$$

We hence have

$$\begin{aligned}\mathfrak{P}_{n,m_n}(k) &\sim (2\pi)^{-k/2}(1-\rho_n)^{-(k-1)/2}\bigl(1+(k-1)\rho_n\bigr)^{-1/2}x_n^{-k}e^{-kx_n^2/2\{1+(k-1)\rho_n\}}\\ &= (2\pi)^{-k/2}(1-\rho_n)^{-(k-1)/2}\bigl(1+(k-1)\rho_n\bigr)^{-1/2}x_n^{-k}e^{-\frac{1}{2}kx_n^2+\mathcal{O}(x_n^2\rho_n)}\\ &= (2\pi)^{-k/2}x_n^{-k}e^{-\frac{1}{2}kx_n^2}\bigl\{1+\mathcal{O}(\rho_n)+\mathcal{O}(x_n^2\rho_n)\bigr\}\\ &= \mathfrak{Q}_{n,m_n}^k\bigl\{1+\mathcal{O}\bigl(m_n^{-2}\bigr)+\mathcal{O}\bigl(nm_n^{-4}\bigr)\bigr\}.\end{aligned}$$

All order terms on the right-hand side approach zero asymptotically for the given rate of growth of m_n, whence $\mathfrak{P}_{n,m_n}(k) \sim \mathfrak{Q}_{n,m_n}^k$ as $n \to \infty$. □

In slightly imprecise language—the events $\bigl\{\bigl\langle \mathbf{w}, \mathbf{u}^{(t)}\bigr\rangle < 0\bigr\}$ are asymptotically independent.

All the pieces are now in place. We complete the proof of the Main Theorem by invoking Lemma 2.3.

For each n, suppose $m = m_n$ random vertices $\mathbf{u}^{(t)} = \mathbf{u}_n^{(t)}$ $(1 \le t \le m_n)$ are generated by independent sampling from the uniform distribution on $\mathbb{B}^n$ and let $\mathbf{w} = \mathbf{w}_n \in \mathbb{B}^n$ be the corresponding vertex generated by harmonic update. Consider the triangular array of "error" events

$$B_n^{(t)} = \bigl\{\bigl\langle \mathbf{w}_n, \mathbf{u}_n^{(t)}\bigr\rangle < 0\bigr\} \qquad (t = 1, \dots, m_n;\ n = 1, 2, \dots),$$

where the nth row has $m = m_n$ elements, and let $\bigl\{\mathfrak{z}_n^{(t)}\bigr\}$ be the corresponding triangular array of indicator random variables for these events. Thus,

$$\mathfrak{z}_n^{(t)} = \begin{cases} 0 & \text{if } \bigl\langle \mathbf{w}_n, \mathbf{u}_n^{(t)}\bigr\rangle \ge 0,\\ 1 & \text{if } \bigl\langle \mathbf{w}_n, \mathbf{u}_n^{(t)}\bigr\rangle < 0.\end{cases}$$

Form the row sums $Z_{n,m_n} = \sum_{t=1}^{m_n}\mathfrak{z}_n^{(t)}$. For each n, Z_{n,m_n} is just the number of examples $\mathbf{u}^{(t)} = \mathbf{u}_n^{(t)}$ $(1 \le t \le m_n)$ which fall into the negative half-space determined by the vector $\mathbf{w} = \mathbf{w}_n$ generated by harmonic update. Alternatively, for given n, Z_{n,m_n} is the number of examples $\mathbf{u}^{(t)} = \mathbf{u}_n^{(t)}$ for which the corresponding inequality in (1.1) is violated. For each $k = 1, \dots, m_n$ and each $n = 1, 2, \dots$, define

$$S_{n,m_n}^{(k)} = \sum \mathbb{E}\bigl(\mathfrak{z}_n^{(t_1)} \cdots \mathfrak{z}_n^{(t_k)}\bigr) = \sum \mathbb{P}\bigl\{B_n^{(t_1)} \cap \cdots \cap B_n^{(t_k)}\bigr\},$$

where the sum is over all subsets $\{t_1, \dots, t_k\}$ of cardinality k drawn from $\{1, \dots, m_n\}$. For each n, the events $B_n^{(t)}$ $(1 \le t \le m_n)$ are exchangeable, so that each of the summands above is just $\mathfrak{P}_{n,m_n}(k)$. We hence obtain

$$S_{n,m_n}^{(k)} = \binom{m_n}{k} \mathfrak{P}_{n,m_n}(k)$$

for every choice of k and n.

Let us now fix the rate of growth of $m = m_n$ with n. Let λ denote any fixed positive quantity and suppose

$$m_n = \sqrt{\frac{n}{\log n}} \left\{ 1 + \frac{\log\log n + \log\bigl(\lambda\sqrt{2\pi}\,\bigr)}{\log n} + \mathcal{O}\biggl(\frac{\log\log n}{(\log n)^2}\biggr) \right\}. \tag{3.3}$$

Clearly, m_n satisfies the conditions $n^{1/3} \ll m_n \ll n^{1/2}$ as $n \to \infty$. Invoking Assertion 4, it is now simple to verify that

$$S_{n,m_n}^{(1)} = m_n \mathfrak{P}_{n,m_n}(1) \sim m_n \mathfrak{Q}_{n,m_n} = \frac{m_n^2}{\sqrt{2\pi n}} \exp\biggl\{ -\frac{n}{2m_n^2} \biggr\} \to \lambda$$

as $n \to \infty$. Now, fix any value of k and allow m_n to grow as in (3.3). Observe that $\binom{m_n}{k} \sim \frac{m_n^k}{k!}$ as $n \to \infty$. Invoking Assertion 4 again, we hence obtain the asymptotic estimate

$$S_{n,m_n}^{(k)} \sim \binom{m_n}{k} \mathfrak{Q}_{n,m_n}^k \sim \frac{(m_n \mathfrak{Q}_{n,m_n})^k}{k!} \to \frac{\lambda^k}{k!} \qquad (n \to \infty).$$

We can now directly apply Lemma 2.3 to conclude that Z_{n,m_n} converges in distribution to the Poisson distribution with parameter λ. This completes the proof of the Main Theorem.

Finally, for the rate of growth given in (3.3), $P(n, m_n) = \mathbb{P}\{Z_{n,m_n} = 0\} \to e^{-\lambda}$ as $n \to \infty$. Now, for any choice of $\lambda > 0$, however small, and any choice of $1 > \epsilon > 0$, a sample size of $m \le (1-\epsilon)\sqrt{n}\big/\sqrt{\log n}$ will be eventually dominated by the right-hand side of (3.3), so that $P(n, m)$ will approach one as $n \to \infty$ by monotonicity. Conversely, for any choice of $\lambda < \infty$, however large, and any choice of $1 > \epsilon > 0$, a sample size of $m \ge (1+\epsilon)\sqrt{n}\big/\sqrt{\log n}$ will eventually dominate the right-hand side of (3.3), so that, by analogous reasoning, $P(n, m)$ will approach zero as $n \to \infty$. This establishes the corollary to the Main Theorem.

REFERENCES

[1] S. C. Fang and S. S. Venkatesh, *Learning binary perceptrons perfectly efficiently*, J. Comput. System Sci., 52 (1996), pp. 374–389.

[2] S. C. Fang and S. S. Venkatesh, *The capacity of Majority Rule*, Random Structures Algorithms, to appear.

[3] W. Feller, *An Introduction to Probability Theory and Its Applications*, Volume I, 3rd ed., John Wiley, New York, 1968.

[4] Z. Füredi, *Random polytopes in the d-dimensional cube*, Discrete Comput. Geom., 1 (1986), pp. 315–319.

[5] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman, San Francisco, CA, 1979.

[6] N. Karmarkar, *A new polynomial-time algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[7] L. Pitt and L. G. Valiant, *Computational limitations on learning from examples*, J. Assoc. Comput. Mach., 35 (1988), pp. 965–984.

[8] W. Richter, *Multidimensional local limit theorems for large deviations*, Theory Probab. Appl., 3 (1958), pp. 100–106.

[9] H. Ruben, *An asymptotic expansion for a class of multivariate normal integrals*, J. Austral. Math. Soc., 2 (1962), pp. 253–264.

[10] J. Spencer, *Ten Lectures on the Probabilistic Method*, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

[11] S. S. Venkatesh, *On learning binary weights for majority functions*, in Proc. of the Fourth Annual Workshop on Computational Learning Theory, San Mateo, CA, Morgan Kaufmann, 1991.

[12] S. S. Venkatesh and J. Franklin, *How much information can one bit of memory retain about a Bernoulli sequence?*, IEEE Trans. Inform. Theory, 37 (1991), pp. 1595–1604.