DEPENDENT SETS OF CONSTANT WEIGHT BINARY VECTORS

NEIL J. CALKIN

ABSTRACT. We determine lower bounds for the number of random binary vectors, chosen uniformly from vectors of weight k, needed to obtain a dependent set.

1. INTRODUCTION

In this paper we determine lower bounds for the number of random binary vectors of weight k needed to obtain a dependent set of vectors with probability 1.

We denote by $S_{n,k}$ the set of binary vectors having k 1's. If we choose a random sequence $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m$ uniformly from $S_{n,k}$, how large must m be for these vectors to be dependent (over GF(2)) with probability 1?

In the case $k = 1$ this is exactly the birthday problem: given a set of n elements, how long must a sequence chosen (with replacement) be before an element occurs at least twice with probability close to 1. It is a standard combinatorics exercise to show that so long as $m/\sqrt{(n)} \to \infty$, a sequence of length m will almost surely contain a repetition as $n \to \infty$.

In the case $k = 2$, we can view the vectors of weight two as being edges in a graph on $\{1, 2, \ldots, n\}$: here a dependent set of vectors corresponds exactly to a set of edges which contain a cycle. There are two distinct modes of behaviour here: first, if the edges are chosen without replacement, and if the number of edges is cn then the probability that there is a cycle is strictly less than 1 as $n \to \infty$ if $c < 1/2$ and tends to 1 if $c \geq 1/2$[4]. If the edges are chosen with replacement, then if we choose cn edges, there is a positive probability that we get a repeated edge. Hence the probability increases up to $c = 1/2$, at which point we almost surely get a cycle.

In what follows, we will assume that k is a fixed integer greater than or equal to 3.

Denote by $p_{n,k}(m)$ the probability that $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m$ are linearly dependent. We will prove the following:

Theorem 1. *For each k there is a constant β_k so that if $\beta < \beta_k$ then*

$$\lim_{n\to\infty} p_{n,k}(\beta n) = 0.$$

Furthermore, $\beta_k \sim 1 - \frac{e^{-k}}{\log(2)}$ as $k \to \infty$.

We obtain this theorem as a corollary of the following: let r be the rank of the set $\{\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m\}$, and let $s = m - r$ (equivalently, the dimension of the kernel of the matrix having columns $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m$).

Theorem 2. *a) If $\beta < \beta_k$ and $m = m(n) < \beta n$ then $E(2^s) \to 1$ as $n \to \infty$. b) If $\beta > \beta_k$ and $m = m(n) > \beta n$ then $E(2^s) \to \infty$ as $n \to \infty$.*

Similar results have been obtained for different models by Balakin, Kolchin and Khokhlov [1, 5]: their methods are completely different.

Our approach is the following: we consider a Markov chain derived from a suitable random walk on the hypercube 2^n; using this we will determine an exact expression for $E(2^s)$. We then estimate $E(2^s)$ to determine β_k.

2. A random walk on the hypercube, and an associated Markov chain

We define a random walk on the hypercube 2^n as follows: let $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m, \ldots$ be vectors chosen uniformly at random from $S_{n,k}$. Define

$$\underline{x}_0 = \underline{0}, \quad \text{and} \quad \underline{x}_i = \underline{x}_{i-1} + \underline{u}_i$$

(so the steps in the walk correspond to flipping k random bits).

We associate with this random walk the following Markov chain: we define y_i to be the weight of $\underline{x}_1$. Then $y_0, y_1, \ldots, y_m$, is a Markov chain with states $\{0, 1, \ldots, n\}$. The transition matrix A for this chain, with $A = \{a_{pq}\}$, where a_{pq} is the probability of moving from state q to state p is given by

$$a_{pq} = \frac{\binom{q}{\frac{k-p+q}{2}}\binom{n-q}{\frac{k+p-q}{2}}}{\binom{n}{k}}$$

where the binomial coefficients are interpreted to be 0 if $k + p + q$ is odd.

Theorem 3. *The eigenvalues λ_i and corresponding eigenvectors $\underline{e}_i$ for A, $i = 0, 1, \ldots, n$, are given by*

$$\lambda_i = \sum_{t=0}^{k} (-1)^t \frac{\binom{i}{t}\binom{n-i}{k-t}}{\binom{n}{k}} \tag{1}$$

and the jth component of $\underline{e}_i$ is given by

$$\underline{e}_i[j] = \sum_{t=0}^{j} (-1)^t \binom{i}{t}\binom{n-i}{j-t}.$$

Proof: We first show that $\underline{e}_i$ is an eigenvector for A with eigenvalue λ_i: indeed the jth coefficient of $A\underline{e}_i$ is

$$\sum_{l=0}^{n} \frac{\binom{l}{\frac{k-j+l}{2}}\binom{n-l}{\frac{k+j-l}{2}}}{\binom{n}{k}} \sum_{t=0}^{j} (-1)^t \binom{i}{k}\binom{n-i}{l-t}$$

and the jth coefficient of $\lambda_i \underline{e}_i$ is

$$\sum_{s=0}^{k} (-1)^s \frac{\binom{i}{s}\binom{n-i}{k-s}}{\binom{n}{k}} \sum_{t=0}^{j} (-1)^t \binom{i}{t}\binom{n-i}{j-t}$$

Observe now that

$$\sum_{t=0}^{j} (-1)^t \binom{i}{t}\binom{n-i}{j-t} = \sum_{t=0}^{j} (-1)^{i+t} 2^t \binom{i}{t}\binom{n-t}{j},$$

since each is the coefficient of x^j in

$$(1-x)^i(1+x)^{n-i} = \left(1 - \frac{2}{1+x}\right)^i (1+x)^n.$$

Hence it is sufficient to show that

$$\sum_{l=0}^{n} \binom{l}{\frac{k-j+l}{2}}\binom{n-l}{\frac{k+j-l}{2}} \sum_{t=0}^{j} (-1)^{t+i} 2^t \binom{i}{t}\binom{n-t}{l} = \sum_{s=0}^{k} (-1)^s \binom{i}{s}\binom{n-i}{k-s} \sum_{t=0}^{k} (-1)^t \binom{i}{t}\binom{n-i}{j-t}.$$

We show this by multiplying both sides by $x^j y^k$ and summing over j and k. Writing $j = l - 2r + k$, the left hand side becomes

$$\sum_{l,k,r,t} \binom{l}{r}\binom{n-m}{k-r}\binom{i}{t}\binom{n-t}{l} (-1)^{i+t} 2^t x^{l+k-2r} y^k$$

$$= \sum_{l,r,t} \binom{l}{r}\binom{i}{t}\binom{n-t}{l} (-1)^{i+t} 2^t x^{l-r} (1+xy)^{n-m} y^r$$

$$= \sum_{l,t} \binom{i}{t}\binom{n-t}{m} (-1)^{i+t} 2^t (1+xy)^{n-m} (x+y)^m$$

$$= \sum_{t} (-1)^{i+t} \binom{i}{t} 2^t (1+xy)^t (1+x)^{n-t} (1+y)^{n-t}$$

$$= (1+x)^n (1+y)^n \left(\frac{2(1+xy)}{(1+x)(1+y)} - 1\right)^i$$

$$= (1-x)^i (1+x)^{n-i} (1-y)^i (1+y)^{n-i}.$$

Similarly the right hand side becomes

$$\sum_{j,k,s,t} \binom{i}{s}\binom{n-i}{k-s}\binom{i}{t}\binom{n-i}{j-t}(-1)^{s+t}x^j y^k$$

$$= \sum_{s,t} \binom{i}{s}\binom{i}{t}(-1)^{s+t}x^t y^s (1+x)^{n-i-t}(1+y)^{n-i-s}$$

$$= (1-x)^i(1+x)^{n-i}(1-y)^i(1+y)^{n-i}$$

as required. Hence $\underline{e}_i$ is an eigenvector with eigenvalue λ_i for each i.

Moreover, we see that the $\underline{e}_i$'s are linearly independent (as vectors over Q): indeed: we have:

Lemma 1. *Let U be the matrix whose columns are $\underline{e}_0, \underline{e}_1, \ldots, \underline{e}_n$. Then $U^2 = 2^n I$, and if Λ is the diagonal matrix of eigenvalues, then $A = 1/2^n U\Lambda U$.*

Proof: The ijth entry of U^2 is

$$\sum_{l=0}^{n} \underline{e}_l[i]\underline{e}_j[l] = \sum_{l,s,t}(-1)^s\binom{l}{s}\binom{n-l}{i-s}(-1)^t\binom{j}{t}\binom{n-j}{l-t}.$$

Multiplying by x^i and summing over i we obtain

$$\sum_{i,l,s,t}(-1)^{s+t}\binom{l}{s}\binom{n-l}{i-s}\binom{j}{t}\binom{n-j}{l-t}x^i$$

$$= \sum_{l,s,t}(-1)^{s+t}\binom{l}{s}\binom{j}{t}\binom{n-j}{l-t}(1+x)^{n-l}x^s$$

$$= \sum_{l,t}(-1)^t\binom{j}{t}\binom{n-j}{l-t}(1+x)^{n-l}(1-x)^l$$

$$= \sum_{t}(-1)^t\binom{j}{t}(1+x)^j 2^{n-j}\left(\frac{1-x}{1+x}\right)^t$$

$$= 2^n x^j$$

from which we see that $U^2 = 2^n I$. Hence the eigenvectors are linearly independent as claimed.

Observation: the eigenvectors do not depend upon k: hence the matrices A and A' corresponding to distinct values of k commute. This corresponds roughly to the idea that when walking around the hypercube it doesn't matter if you take a step of size l then a step of size k, or a step of size k then a step of size l.

We can now compute the probability that $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_t$ sum to $\underline{0}$: indeed, this is exactly the 00th coefficient in A^t, which is equal to

$$\sum_{i=0}^{n} \frac{1}{2^n} \lambda_i^t \binom{n}{i}$$

(since $A = 1/2^n U \Lambda U$).

Hence if $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_m$ are vectors with k 1's chosen independently at random, then the expected number of subsequences $\underline{u}_{a_1}, \underline{u}_{a_2}, \ldots, \underline{u}_{a_t}$ which sum to $\underline{0}$ is exactly

$$E(2^s) = \sum_{t=0}^{m} \binom{m}{t} \sum_{i=0}^{n} \frac{1}{2^n} \lambda_i^t \binom{n}{i} = \sum_{i=0}^{n} \frac{1}{2^n} \binom{n}{i} (1+\lambda_i)^m .$$

3. Asymptotics of λ_i

In order to estimate the size of $E(2^s)$, we require asymptotics for the value of λ_i.

Lemma 2. *a)* $|\lambda_i| < 1$ *for all* $0 \le i \le n$.
b) If $i > \frac{n}{2}$ *then* $\lambda_i = (-1)^k \lambda_{n-i}$.
c) Let $0 < c < \frac{1}{2}$. *If* $i = cn$ *then*

$$\lambda_i = \left(1 - \frac{2i}{n}\right)^k - \frac{4\binom{k}{2}}{n} \left(1 - \frac{2i}{n}\right)^{k-2} \frac{i}{n} \left(1 - \frac{i}{n}\right) + O\left(\frac{k^4}{c^2 n^2}\right)$$

Proof: Parts a) and b) are immediate from the definition of λ_i. To prove part c), since k is fixed, we have

$$\binom{n}{k} = \frac{n^k}{k!} \left(1 - \frac{\binom{k}{2}}{n} + O\left(\frac{k^4}{n^2}\right)\right)$$

$$\binom{i}{k} = \frac{i^t}{t!} \left(1 - \frac{\binom{t}{2}}{i} + O\left(\frac{t^4}{i^2}\right)\right)$$

$$\binom{n}{k} = \frac{(n-i)^{k-t}}{(k-t)!} \left(1 - \frac{\binom{k-t}{2}}{n-i} + O\left(\frac{(k-t)^4}{(n-i)^2}\right)\right).$$

Hence

$$\frac{\binom{i}{t}\binom{n-i}{k-t}}{\binom{n}{k}} = \left(\frac{i}{n}\right)^t \left(1 - \frac{i}{n}\right)^{k-t} \binom{k}{t} \left(1 + \frac{\binom{k}{2}}{n} - \frac{\binom{t}{2}}{i} - \frac{\binom{k-t}{2}}{n-i} + O\left(\frac{k^4}{c^2 n^2}\right)\right)$$

and

$$\lambda_i = \sum_{t=0}^{k} (-1)^k \left(\frac{i}{n}\right)^t \left(1 - \frac{i}{n}\right)^{k-t} \binom{k}{t} \left(1 + \frac{\binom{k}{2}}{n} - \frac{\binom{t}{2}}{i} - \frac{\binom{k-t}{2}}{n-i} + O\left(\frac{k^4}{c^2 n^2}\right)\right)$$

$$= \left(1-\frac{2i}{n}\right)^k + \frac{\binom{k}{2}}{n}\left(1-\frac{2i}{n}\right)^k - \frac{\binom{k}{2}}{i}\left(\frac{i}{n}\right)^2\left(1-\frac{2i}{n}\right)^{k-2}$$
$$-\frac{\binom{k}{2}}{n-i}\left(\frac{n-i}{n}\right)^2\left(1-\frac{2i}{n}\right)^{k-2} + O\left(\frac{k^4}{c^2n^2}\right)$$
$$= \left(1-\frac{2i}{n}\right)^k + \frac{\binom{k}{2}}{n}\left(1-\frac{2i}{n}\right)^{k-2}\left(\left(1-\frac{2i}{n}\right)^2 - \frac{i}{n} - \frac{n-i}{n}\right) + O\left(\frac{k^4}{c^2n^2}\right)$$
$$= \left(1-\frac{2i}{n}\right)^k - \frac{4\binom{k}{2}}{n}\left(1-\frac{2i}{n}\right)^{k-2}\left(-\frac{4i}{n}+4\frac{i^2}{n}\right) + O\left(\frac{k^4}{c^2n^2}\right)$$
$$= \left(1-\frac{2i}{n}\right)^k - \frac{4\binom{k}{2}}{n}\left(1-\frac{2i}{n}\right)^{k-2}\left(\frac{i}{n}\right)\left(1-\frac{i}{n}\right) + O\left(\frac{k^4}{c^2n^2}\right)$$

as claimed.

Observe that since we are assuming that $k \geq 3$ throughout, when i is close to $\frac{n}{2}$, say $\frac{n}{2} - i = \frac{n^\theta}{2}$, we have

$$\lambda_i = \left(\frac{1}{n^{1-\theta}}\right)^k - \frac{4\binom{k}{2}}{n}\left(\frac{1}{n^{1-\theta}}\right)^{k-2} + O\left(\frac{k^4}{n^2}\right)$$

Then, provided that $\theta < 1 - \frac{1}{k}$, we see that if $\frac{n}{2} - i = \frac{n^\theta}{2}$, then $\lambda_i n \to 0$ as $n \to \infty$. In the estimation of $E(2^s)$ we will use this to show that the middle part of the sum is asymptotic to 1.

4. Asymptotics of $E(2^s)$

Define

$$f(\alpha,\beta) = -\log 2 - \alpha\log(\alpha) - (1-\alpha)\log(1-\alpha) + \beta\log(1+(1-2\alpha)^k)$$

and let (α_k, β_k) be the root of

$$\begin{aligned} f(\alpha,\beta) &= 0 \\ \frac{\partial f(\alpha,\beta)}{\partial\alpha} &= 0 \end{aligned}$$

We shall show:

Lemma 3. *If $\beta < \beta_k$ and $m < \beta n$ then $\sum_i 2^{-n}\binom{n}{i}(1+\lambda_i)^m \to 1$ as $n \to \infty$, and if $\beta > \beta_k$ and $m > \beta n$ then $\sum_i 2^{-n}\binom{n}{i}(1+\lambda_i)^m \to \infty$ as $n \to \infty$.*

Proof: we proceed as follows: since our goal is to show that the behaviour of $E(2^s)$ changes when m goes from below $\beta_k n$ to above $\beta_k n$, and since our value β_k is less than 1, we may assume that $\frac{m}{n} < 1 - \delta$ for some $\delta > 0$. We shall show:
a) the extreme tails of the sum for $E(2^s)$ are small
b) the middle range of the sum contributes 1 to the sum
c) and d) the rest of the sum is small if $\frac{m}{n} < \beta < \beta_k$ and large if $\frac{m}{n} > \beta > \beta_k$.
a) there is an $\epsilon > 0$ so that

$$\sum_{i=0}^{\epsilon n} 2^{-n} \binom{n}{i} (1+\lambda_i)^m \to 0 \text{ as } n \to \infty$$

Indeed,

$$\begin{aligned}\sum_{i=0}^{\epsilon n} 2^{-n} \binom{n}{i} (1+\lambda_i)^m &< \sum_{i=0}^{\epsilon n} 2^{m-n} \binom{n}{i} \\ &< n\epsilon 2^{m-n} \binom{n}{\epsilon n}\end{aligned}$$

and provided ϵ is sufficiently small, this tends to 0 (indeed, if $-\delta \log 2 - \epsilon \log \epsilon + \epsilon < 0$ then the sum tends to 0).

Similarly,

$$\sum_{i=(1-\epsilon)n}^{n} 2^{-n} \binom{n}{i} (1+\lambda_i)^m \to 0 \text{ as } n \to \infty.$$

Hence, if $E(2^s) \to \infty$ for some $m < (1-\delta)n$, we must have the major contribution from

$$\sum_{i=\epsilon n}^{(1-\epsilon)n} 2^{-n} \binom{n}{i} (1+\lambda_i)^m .$$

b) We now show that the middle range of the sum contributes 1 to $E(2^s)$. Indeed, in the range $\frac{n}{2} - n^{4/7} < i < \frac{n}{2} + n^{4/7}$

$$(1+\lambda_i)^m = \left(1 + O\left(\frac{k^4}{n^2}\right)\right)^m = 1 + O\left(\frac{k^4}{n}\right)$$

we have

$$\sum_{i=\frac{n}{2}-n^{4/7}}^{\frac{n}{2}+n^{4/7}} 2^{-n} \binom{n}{i} (1+\lambda_i)^m \sim \sum_{1=\frac{n}{2}-n^{4/7}}^{\frac{n}{2}+n^{4/7}} 2^{-n} \binom{n}{i} \to 1 \text{ as } n \to \infty.$$

c) We now show that we can widen the interval about the middle:

$$\sum_{i=\frac{n}{2}(1-\epsilon)}^{\frac{n}{2}(1+\epsilon)} 2^{-n} \binom{n}{i} (1+\lambda_i)^m \to 1.$$

Since $\lambda_{n-i} = (-1)^k \lambda_i$, it suffices to show that

$$\sum_{i=\frac{n}{2}(1-\epsilon)}^{\frac{n}{2}} 2^{-n} \binom{n}{i} (1+\lambda_i)^m \to 1.$$

In this range,

$$\lambda_i < \epsilon^k - \frac{\binom{k}{2}}{n} \epsilon^{k-2} + O\left(\frac{k^4}{n^2}\right).$$

Hence

$$(1+\lambda_i)^m < e^{n\epsilon^k} e^{\binom{k}{2}\epsilon^{k-2}}$$

and since $k \geq 3$, the $n\epsilon^k$ term in the exponent is dominated by the $-n\epsilon^2$ term from the binomial coefficient, provided that ϵ is sufficiently small.
d) We now consider the remainder of the sum (or rather, the part in $(0, \frac{n}{2})$: if k is even, the remaining part follows by symmetry, and if k is odd, then $(1+\lambda_i)^m < 1$ for $i > n/2$, and the remaining part tends to 0).

Define

$$f(\alpha, \beta) = -\log 2 - \alpha \log \alpha - (1-\alpha)\log(1-\alpha) + \beta \log(1 + (1-2\alpha)^k).$$

Then if $f(\frac{i}{n}, \frac{m}{n}) < \gamma < 0$ the corresponding term of the sum is expontentially small, and if $f(\frac{i}{n}, \frac{m}{n}) > \gamma > 0$ the corresponding term of the sum is exponentially large. Thus, if $f(\alpha, \frac{m}{n}) < \gamma < 0$ for all α in $(\epsilon, 1-\epsilon)$, we have

$$\sum_{i=\epsilon n}^{\frac{n}{2}(1-\epsilon)} 2^{-n} \binom{n}{i} (1+\lambda_i)^m < n e^{\gamma n + o(n)} \to 0,$$

and if $f(\alpha, \frac{m}{n}) > \gamma > 0$ for some α in $(\epsilon, 1-\epsilon)$, then

$$\sum_{i=\epsilon n}^{\frac{n}{2}(1-\epsilon)} 2^{-n} \binom{n}{i} (1+\lambda_i)^m > \binom{n}{\alpha n} (1+\lambda_{\alpha n})^m 2^{-n} > e^{\gamma n + o(n)} \to \infty.$$

Now let β_k be so that if $\beta < \beta_k$ then $f(\alpha, \beta) < 0$ for all α in $(\epsilon, 1-\epsilon)$, and if $\beta > \beta_k$ then there is an alpha in $(\epsilon, 1-\epsilon)$ so that $f(\alpha, \beta) > 0$. Thus we wish to find α_k, β_k so that

$$f(\alpha_k, \beta_k) = 0 \text{ and } \frac{\partial}{\partial \alpha} f(\alpha, \beta) = 0.$$

As k goes to ∞, the value of β_k is asymptotic to

$$1 - \frac{e^{-k}}{\log 2} - \frac{1}{2\log 2}(k^2 - 2k + \frac{2k}{\log 2} - 1)e^{-2k} + O(k^4)e^{-3k}.$$

This completes the proof of the lemma. Now, since $E(2^s) = \sum_i 2^{-n} \binom{n}{i} (1+\lambda_i)^m$ this completes the proof of theorem 2, and Theorem 1 follows by the simple observation that since s is integer valued, the probability that $2^s > 1$ is less than $E(2^s) - 1$.

References

1. G. V. Balakin, V. F. Kolchin and V. I. Khokhlov. Hypercycles in a random hypergraph Diskretnaya Matematika 3, No. 3, 1-2–108 (1991) (in Russian).
2. P. Diaconis. Group representations in probability and statistics Vol 11, Inst. Math. Statist., Hayward, Calif. 1988.
3. P. Diaconis and R. L. Graham. Asypmtotic analysis of a random walk on the hypercube with many dimensions. *Random Structures and Algorithms* 1,51–72 (1990)
4. P. Erdös and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.* 5,17–61 (1960).
5. V. F. Kolchin and V. I. Khokhlov. On the number of cycles in a non-equiprobably random graph. *Diskretnaya Matematika* 2, No.3,137–145 (1990) (in Russian).

School Of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332
E-mail address: calkin@math.gatech.edu