LINEAR INDEPENDENCE IN A RANDOM BINARY VECTOR MODEL

KIM BOWMAN, NEIL J. CALKIN, ZACH COCHRAN, TIMOTHY FLOWERS, KEVIN JAMES, AND SHANNON PURVIS

Abstract. We consider a natural model of random binary vectors with heavier heads than tails. In this model we determine a good upper bound for how many vectors we need to take to find a linearly dependent set of vectors.

Let $\mathbb{F}_2^k$ be the binary vector space of dimension k (we will regard the vectors as row vectors). We define the following probability space. Generate a vector $\underline{v}$ as follows: each coordinate of $\underline{v}$ is chosen independently, and

$$\Pr(v[j] = 1) = \frac{1}{p_j}$$

where p_j is the jth prime.

Now choose vectors $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_l$ independently from this distribution: the question we focus on is how big l should be so that with high probability the set $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_l$ is linearly dependent. This question arose in explorations of best stopping times for Pomerance's Quadratic Sieve [2] factorization algorithm. Although this turned out to be an inappropriate model for the sieve, the result we prove here seems of independent interest.

Our result is the following.

Theorem 1. *Let $\delta > 1/e$ be fixed. Let $l = k^\delta$. Then with high probability the set of vectors is linearly dependent.*

Proof: The simplest way to show that a set of vectors is linearly dependent is to show that there are more vectors than the dimension of the space in which the vectors live. We extend this trivial observation as follows: construct a $l \times k$ array A having rows $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_l$: then if A has more rows that it has *non-zero* columns then $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_l$ are linearly dependent.

Let the random variable X_j be 1 if the j^{th} column of A is nonzero and 0 otherwise. Then $X = \sum_j X_j$ is the number of non zero columns. So if $l > X$ then the rows of A are linearly dependent.

Date: May 13, 2005.

The authors wish to thank the National Science Foundation for their generous support through grant number DMS: 0504404.

The probability that the jth column of A is zero (that is, all its entries are zeros) is

$$\left(1 - \frac{1}{p_j}\right)^l \simeq \exp\left(\frac{-l}{p_j}\right).$$

Hence the expected number of non-zero columns of A is

$$E(X) = \sum_{j=1}^{k} \left(1 - \left(1 - \frac{1}{p_j}\right)^l\right) \simeq \sum_{j=1}^{k} \left(1 - \exp\left(-\frac{l}{p_j}\right)\right).$$

We analyze this sum by splitting it into three regions:

I: $1 \leq j \leq l/\log(l)^2$
II: $l/\log(l)^2 < j < l$
III: $l \leq j \leq k$.

We will show that regions I and II contribute a negligible amount to the sum, and estimate the contribution of region III using Mertens' theorem.

First, observe that region I has length $o(l)$: since the summands are certainly at most 1, and we wish to compare the sum to l, the contribution from region I will is negligible.

Next we consider region II. In this region, we approximate the sum by an integral: since j is large, we have $p_j \sim j \log j$, and so

$$\sum_{j=\frac{l}{(\log l)^2}}^{l} \left(1 - \exp\left(-\frac{l}{p_j}\right)\right) \simeq \int_{\frac{l}{(\log l)^2}}^{l} \left(1 - \exp\left(-\frac{l}{x \log x}\right)\right) \mathrm{d}x.$$

Furthermore, in this region, $\log x$ is essentially constant, with $\log x \simeq \log l = \delta \log k$: hence the contribution of region II is about

$$\int_{\frac{l}{(\log l)^2}}^{l} \left(1 - \exp\left(-\frac{l}{x \log l}\right)\right) \mathrm{d}x.$$

Substituting $u = \frac{l}{x \log l}$, this becomes

$$\begin{aligned} \frac{l}{\log l} \int_{\frac{1}{\log l}}^{\log l} \frac{1 - e^{-u}}{u^2} \, \mathrm{d}u &= \frac{l}{\log l} \int_{\frac{1}{\log l}}^{1} \frac{1 - e^{-u}}{u^2} \, \mathrm{d}u + \frac{l}{\log l} \int_{1}^{\log l} \frac{1 - e^{-u}}{u^2} \, \mathrm{d}u \\ &\leq \frac{l}{\log l} \int_{\frac{1}{\log l}}^{1} \frac{1}{u} \, \mathrm{d}u + \frac{l}{\log l} \int_{1}^{\log l} \frac{1}{u^2} \, \mathrm{d}u, \end{aligned}$$

since $1 - e^{-u} < u$ on $(0, 1)$. Hence the total contribution of region II is about

$$\frac{l \log \log l}{\log l} = o(l).$$

Finally, we consider region III. Here, $l/p_j = o(1)$, so

$$1 - \exp\left(-\frac{l}{p_j}\right) = \frac{l}{p_j} + O\left(\frac{l^2}{p_j^2}\right).$$

Now, by Mertens' theorem [1, Theorem 427],

$$\sum_{p<x} \frac{1}{p} = \log\log x + C + O\left(\frac{1}{\log x}\right).$$

Hence

$$\sum_{j=l}^{k} \frac{l}{p_j} = l(\log\log p_k - \log\log p_l) + O\left(\frac{l}{\log l}\right)$$

Since $l = k^\delta$,

$$\log\log p_k - \log\log p_l \simeq \log\log k - \log\log l = -\log\delta.$$

The error term coming from

$$\sum \frac{l^2}{p_j^2}$$

is easily seen to contribute $O(l/(\log l)^2)$, and so if $\delta > 1/e$, region III contributes less than l, and if $\delta < 1/e$ then region III contributes more than l.

Hence we have shown that since $\delta > 1/e$, $E(X) < l$, that is the expected number of non-zero columns is less than the number of rows. Computing the variance $V(X)$, it is easy to see that $V(X) \leq E(X)$. Hence a simple application of Tchebyschev's inequality shows that with high probability $X - E(X)$ is $o(l)$. Thus with high probability, the number of non-zero columns is less than the number of rows, and hence the rows of the array A are linearly dependent.

1. Acknowledgments

The authors gratefully acknowledge the support of the NSF for the 2004 Clemson REU in Number Theory and Combinatorics [DMS: 0504404], during which this research was performed.

References

[1] G. H. Hardy, E. M. Wright, "An Introduction to the Theory of Numbers". Fifth edition. Oxford University Press, New York, 1979.

[2] Carl Pomerance, *A tale of two sieves*. Notices Amer. Math. Soc. **43** (1996), no. 12, 1473–1485.

Department of Mathematical Sciences Clemson University, Clemson, SC 29634-0975

Department of Mathematical Sciences Clemson University, Clemson, SC 29634-0975

Department of Mathematics, University of Georgia, Athens, GA 30602

Department of Mathematical Sciences Clemson University, Clemson, SC 29634-0975

Department of Mathematical Sciences Clemson University, Clemson, SC 29634-0975

Department of Mathematical Sciences Clemson University, Clemson, SC 29634-0975