DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

On the Distribution of Sums of Vectors in General Position

Jerrold R. Griggs and Günter Rote

Abstract. An analogue of the Littlewood-Offord problem posed by the first author is to find the maximum number of subset sums equal to the same vector over all ways of selecting n vectors in $\mathbb{R}^d$ in general position. This note reviews past progress and motivation for this problem, and presents a construction that gives a respectable new lower bound, $\Omega(2^n n^{1-3d/2})$, which compares for $d \geq 2$ to the previously known upper bound $O(2^n n^{-1-d/2})$.

One version of the famous Littlewood-Offord problem [**10**] asks how to select complex numbers $a_1, \ldots, a_n$, not necessarily distinct, with each $|a_i| \geq 1$, and a target open ball $T \subseteq \mathbb{C}$ of unit diameter to maximize the number of the 2^n subset sums $\sum_{i \in I} a_i$, where $I \subseteq [n]$, lying in T. Viewing $\mathbb{C}$ as $\mathbb{R}^2$, one can extend this problem to arbitrary dimension d, and ask the same thing, where now the a_i's are vectors in $\mathbb{R}^d$. By setting all a_i equal to the same vector, it is possible to have $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ subset sums lying inside T. Erdős [**3**] showed this was best-possible for the reals ($d = 1$); Katona [**7**] and Kleitman [**8**] independently proved the same for the original case of complex numbers ($d = 2$); Kleitman [**9**] later found an ingenious inductive proof that $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ is best-possible for general d.

Even if we restrict the target set to just a single point t, this bound is still achieved. But what if we must also spread out the vectors a_i in the sense of asking that any d of them be linearly independent? Had we only needed to hit a unit diameter ball target, the answer would have remained at $\binom{n}{\lfloor \frac{n}{2} \rfloor}$, but by shrinking the target to a single point, it will be tougher in general to get as many sums to hit the target. With a single point target, the restriction that each $|a_i| \geq 1$ no longer affects the answer, so it can be dropped. Therefore, we are now interested in the following:

General Position Subset Problem. *Given positive integers n, d, how can one select vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ and a target $t \in \mathbb{R}^d$ to achieve the maximum number $f_d(n)$ of the 2^n subset sums $\sum_{i \in I} a_i$, where $I \subseteq [n]$, equal to t, provided that every d of the vectors a_i are linearly independent?*

1991 *Mathematics Subject Classification.* Primary 52C10, 05D05; Secondary 11B75.
The first author was supported in part by NSF grant DMS–9701211.
The second author was supported in part by the Spezialforschungbereich "Optimierung und Kontrolle".

Griggs [**5**] arrived at exactly this problem in connection with a model of database security. In the database security studies of Mirka Miller *et al.* [**12, 11, 1**], *cf.* [**4**], there is a database of numerical records, $\{x_1, \dots, x_n\}$, *e.g.*, the salaries of the n members of a department. One may request the sums $\sum_{j \in J} x_j$ of certain subsets $J \subseteq \{1, \dots, n\}$, and an answer will be given by the control mechanism, provided that no "compromise" results. In the basic model, a compromise means that the requester is able to determine, by taking an appropriate linear combination of the answered queries (sums), some individual entry x_i. The problem is to maximize the number of queries that can be answered without compromising the database. Griggs proposed an extension of this problem to prevent compromise by anyone with prior knowledge of $d-1$ records: We say a compromise results whenever one can determine some linear combination of at most d records, $\sum_{j \in J} \alpha_j x_j$, where all $\alpha_j \neq 0$ and $0 < |J| \leq d$. It turns out that the maximum number of queries that can be answered without compromise is precisely $f_d(n)$.

In one dimension, $f_1(n)$ is equivalent to the real case of the Littlewood-Offord problem, and so

$$f_1(n) = \binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Determining $f_d(n)$ for fixed dimension $d \geq 2$ remains an intriguing and fundamental open problem with connections to many fields of mathematics. As shown in [**5**], an $O(2^n n^{-\lfloor d/2 \rfloor})$ upper bound on $f_d(n)$ can be deduced by a simple sphere-packing argument applied to the equivalent database security problem.

Deeper work of Halász provides a slight improvement, an $O(2^n n^{-1-d/2})$ upper bound for $d \geq 2$. More sophisticated analytical methods, especially Fourier analysis, are apparently needed to obtain this bound. An accessible proof would be valuable, in that it might adapt nicely to other variations of the Littlewood-Offord problem, such as when varying lower bounds on different vectors a_i are imposed.

Is the Halász bound correct for general d to within a constant factor? We still cannot say, not having a suitable construction achieving the bound for $d > 2$. In [**5**] two sets of vectors for $d = 2$ are presented, each achieving the Halász bound, to within a constant factor, for general n. Until the DIMATIA-DIMACS conference on "The Future of Discrete Mathematics" in Štiřín castle in May, 1997, no reasonable lower bound for dimensions $d > 2$ had been described. In this note, we provide such a construction.

We first review and extend the method which was applied in [**5**] to obtain a lower bound from a construction. Suppose we have a set of integer vectors $a_1, \dots, a_n \in \mathbb{R}^d$ such that any d of them are linearly independent, $n \geq d$. Consider any particular coordinate of the a_i's, say coordinate j, and denote this component of a_i by $u_i \in \mathbb{R}$. Then the distribution of the jth components of the 2^n subset sums $\sum_{i \in I} a_i$, when multiplied by 2^{-n}, is the same as the probability distribution for the random variable $X = \sum_{i=1}^n u_i X_i$, where the X_i are i.i.d. random variables, each equal to 0 or 1 with probability 1/2. One checks routinely that X has mean $\mu_j := \sum u_i/2$ and variance $\sigma_j^2 := \sum_i u_i^2/4$. Chebyshev's inequality implies that a proportion at most $1/K^2$ out of our 2^n subset sums have jth component differing from μ_j by more than $K\sigma_j$.

Applying this for all j with $K = \sqrt{2d}$, we learn that at least half of the 2^n subset sums vectors are within $K\sigma_j$ of μ_j for all j. That is, at least 2^{n-1} of the subset sums lie within a box, with sides parallel to the coordinate planes, with side lengths $2K\sigma_j$, $1 \le j \le d$. Since our vectors a_i have integer coordinates, the lattice points are the only possible subset sums. The number of lattice points in this box is roughly

$$\prod_{j=1}^{d} 2K\sigma_j = (8d)^{d/2} \prod_{j=1}^{d} \sigma_j.$$

Consequently, some lattice point occurs as a subset sum for at least

$$(1/2)2^n \left((8d)^{d/2} \prod_{j=1}^{d} \sigma_j \right)^{-1}$$

different subset sums.

Here is one construction that gives the desired independence, yet keeps the product of σ_j's under control: Let $a_i = (1, i, i^2, \ldots, i^{d-1})$, $1 \le i \le n$. Then any d of these vectors, say $a_{i_1}, \ldots, a_{i_d}$ with $i_1 < i_2 < \cdots < i_d$, are linearly independent, since they form a Vandermonde matrix, with determinant

$$\prod_{j<k} (i_k - i_j) \neq 0.$$

Thus, our n vectors are in general position. As for the bound they give, we have

$$\sigma_j = \left(\sum_i i^{2(j-1)}/4 \right)^{1/2} = \Theta(n^{\frac{2j-1}{2}}).$$

Thus, $\prod_{j=1}^{d} \sigma_j = \Theta(n^{d^2/2})$, giving us a lower bound on $f_d(n)$ of order $\Theta(2^n n^{-d^2/2})$.

For $d = 1$ and $d = 2$, this bound is best-possible, up to a constant factor. For $d > 2$ we can modify this construction so that it still works, while the vector coordinates stay much smaller ($< 2n$). Several months of continued reflection on the problem have not led to any further improvement, so we shall describe this progress now. Perhaps it is the upper bound that needs tightening, rather than the lower bound. Here is our new result.

THEOREM. *For fixed $d \ge 2$, there exist constants $C, C' > 0$ such that the maximum number $f_d(n)$ of subset sums equal to the same value, for any set of n vectors in $\mathbb{R}^d$ in general position, satisfies*

$$C2^n n^{1-(3/2)d} < f_d(n) < C'2^n n^{-1-d/2}.$$

PROOF. As noted above, the upper bound follows from a result of Halász [**6**]. We present a construction for the lower bound. Choose a prime number p with $n \le p < 2n$, which is well-known to exist. Take the a_i as in the construction above, except reduce the coordinates modulo p, so that every coordinate belongs to $\{0, \ldots, p-1\}$. The new vectors are in general position: For any d of them, each entry of the corresponding determinant is the same, mod p, as before. Thus, the new determinant is congruent mod p to $\prod_{j<k} (i_k - i_j)$, which is not zero mod p since each of its factors is a positive integer $< n \le p$. It follows that the new determinant is not zero, since it is not divisible by p.

The first coordinates are all 1, so we have $\sigma_1^2 = n/4$. For $j > 1$, since all coordinates are at most $p-1$, we get that $\sigma_j = O((np^2)^{1/2}) = O(n^{3/2})$. Completing the analysis as before gives us the stated lower bound. $\square$

The point sets in general position with limited coordinates which were constructed above have also been proposed in computational geometry, where they were useful to deal with problems of geometric degeneracy by perturbation methods, see Emiris, Canny, and Seidel [**2**].

Acknowledgements

We are grateful to the centers DIMATIA in Prague and DIMACS in New Jersey for organizing the exciting workshop at Štiřín last May and for supporting its participants, including the authors, who had not met previously. The collaboration described in this note was a direct result of this contact.

We thank Torsten Thiele and Emo Welzl for discussions about the problem of finding many points with small coordinates in general position.

References

[1] L. Branković, M. Miller, and J. Širáň, *Towards a practical auditing method for the prevention of statistical database compromise,* Proc. 7th Australasian Database Conf., Austral. Comp. Sci. Commun.**18 (no. 2)**(1996), 177–184.

[2] I. Z. Emiris, J. F. Canny, and R. Seidel, *Efficient perturbations for handling geometric degeneracies,* Algorithmica **19**(1997), 219–242.

[3] P. Erdős, *On a lemma of Littlewood and Offord,* Bull. Amer. Math. Soc. (2nd ser.)**51**(1945), 898–902.

[4] J. R. Griggs, *Concentrating subset sums at k points,* Bull. Inst. Combin. Applns. **20**(1997), 65–74.

[5] J. R. Griggs, *Database security and the distribution of subset sums in* $\mathbb{R}^m$*,* Proc. Intern. Colloq. Combin. Graph Th. 1996, Balatonlelle, Hungary, to appear.

[6] G. Halász, *Estimates for the concentration function of combinatorial number theory and probability,* Periodica Math. Hungar. **8 (3–4)**(1977), 197–211.

[7] G. O. H. Katona, *On a conjecture of Erdős and a stronger form of Sperner's theorem,* Studia Sci. Math. Hungar. **1**(1966), 59–63.

[8] D. J. Kleitman, *On a lemma of Littlewood and Offord on the distribution of certain sums,* Math. Z. **90**(1965), 251–259.

[9] D. J. Kleitman, *On a lemma of Littlewood and Offord on the distributions of linear combinations of vectors,* Advances in Math. **5**(1970), 155–157.

[10] J. Littlewood and C. Offord, *On the number of real roots of a random algebraic equation III,* Mat. Sbornik **12**(1943), 277–285.

[11] M. Miller, I. Roberts, and J. Simpson, *Application of symmetric chains to an optimization problem in the security of statistical databases,* Bull. Inst. Combin. Applns. **2**(1991), 47–58.

[12] M. Miller and J. Seberry, *Relative compromise of statistical databases,* Austral. Computer J. **21 (2)**(1989), 56–61.

Department of Mathematics, University of South Carolina, Columbia, SC 29208 USA

E-mail address: `griggs@math.sc.edu`

Technische Universität Graz, Institut für Mathematik, Steyrergasse 30, A-8010 Graz, AUSTRIA

E-mail address: `rote@opt.math.tu-graz.ac.at`