Concentrating Subset Sums at k Points

Jerrold R. Griggs[1]
Department of Mathematics
University of South Carolina
Columbia, SC 29208 USA
email: griggs@math.sc.edu

Dedicated to the Memory of Professor Paul Erdős

Abstract

We consider the problem of maximimizing, over all choices of n nonzero elements $a_1, \ldots, a_n \in \mathbf{R}^m$, the number of the 2^n subset sums $\sum_{i \in I} a_i$, over all index sets I, belonging to some specified target set T. M. Miller, Roberts, and Simpson investigated the case $m = 1$ and $T = \{0, 1\}$ of this problem, and showed that the maximum in their case is $\binom{n+1}{\lfloor \frac{n+1}{2} \rfloor}$, but it remained open until now to prove the essential uniqueness of the extremal solutions a_i that achieve this maximum. More generally, we determine the maximum, as well as solutions achieving it, over n arbitrary elements a_i and target sets T of k arbitrary points in $\mathbf{R}^m$. We also obtain the same maximum number of sums when T is a union of k open balls of diameter $\min_i |a_i|$.

Running head: Concentrating Subset Sums

[1] Research supported in part by NSA/MSP Grants MDA904–92H3053 and 95H1024.

Section 1. Introduction.

In papers that previously appeared in this Bulletin, M. Miller, Roberts, and Simpson [13,14] determined how to maximize the usability of a particular statistical database. In their model, using the control mechanism called Audit Expert, they sought to maximize the number of Sum Queries that could be asked without compromising any individual entry of the database. Via some simple matrix theory, they reduced their problem to one about the distribution of sums of a collection of real numbers.

Specifically, they asked how to select n nonzero real numbers $a_1, \ldots, a_n$ so as to maximize the number of subset sums $\sum_{i \in I} a_i$ equal to 0 or 1, where I ranges over $2^{[n]}$, the collection of all 2^n subsets of $[n] := \{1, \ldots, n\}$. They obtained the best-possible bound for their problem, but were unable to prove the uniqueness of their extremal families. We solve this problem here, and extend the result to the maximum number of subset sums concentrated on a target set of k values. The solution, including the description of the maximum families, is then carried out in higher dimensions. We achieve these results by adapting methods of extremal set theory originally developed to tackle the closely-related problem of Littlewood and Offord [11]. For people not familiar with these methods, this paper may serve as an introduction. Paul Erdős published his seminal, frequently cited paper [5] on the Littlewood-Offord problem just over 50 years ago. In honor of this occasion, and to celebrate his remarkable contributions over the years, *we dedicate this paper to the memory of "Uncle Paul".*

After recalling the result of M. Miller *et al.* (Theorem 1 below), we note in Section 2 that their result is implied by Erdős' work on the Littlewood-Offord problem, and that it can be extended to solve the problem of the maximum number of subset sums equal to one of k fixed values. We see that this problem is closely related to the problem of finding the largest *k-family* of subsets in the Boolean lattice $B_n = (2^{[n]}, \subseteq)$.

Recall that an *antichain* of subsets is a collection of sets, no one containing any other, and a k-family of subsets is a collection with no $k+1$ of them on a chain. Equivalently, a k-family is the union of k antichains. The maximum for both problems, the number of subset sums and the size of a k-family, is the sum of the k middle binomial coefficients in n. The same maximum applies even if T a union of k half-open intervals on the real line, *e.g.*, when T is an open (or half-open) interval of length k.

The target set $T = \{0, 1\}$ of M. Miller *et al.* is one for which the maximum (with $k = 2$) is achieved, by taking all $|a_i| = 1$ with the appropriate number of a_i of each sign (roughly half each way). In a subsequent paper, K. Miller and Sarvate [12] proved that the solution above for $T = \{0, 1\}$ is unique, provided that the a_i are constrained to be integers. Branković and M. Miller [2] extended the uniqueness to the original case of real a_i's. (This paper is independent of [2], and was submitted before [2] appeared.) The proofs in the three papers [13,12,2] combine to yield an interesting and elegant application of the symmetric chain decomposition of the Boolean lattice.

In Section 3, we solve the problem of determining all target sets of k real values and all choices $a_i \neq 0$ that achieve the maximum, integer or not, by applying the description of the maximum k-families in the Boolean lattice. We find that all a_i must be identical up to sign.

These results are extended to higher dimensions in Section 4, where we have $a_1, \ldots, a_n \in \mathbf{R}^m$. As with the original Littlewood-Offord problem, the k-family argument no longer works for $m \geq 2$, and a different strategy is needed, when the target set is a union of k balls. We apply Kleitman's method of inductive partition of $2^{[n]}$ into $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ collections, where no two sums in the same collection are "close". The sum of the k middle binomial coefficients in n is still the maximum, provided that the target set is a union of k open balls of diameter $\delta = \min_i |a_i|$.

For the case that the target set T consists of k points, we

obtain all extremal configurations–they are essentially just the one-dimensional ones–by reducing the problem once again to the k-family problem.

In a survey article currently in preparation, we plan to examine the surprising series of papers by M. Miller *et al.* on different models of database compromise under Audit Expert. For example, one important model [14] leads to the condition that the a_i's must also be distinct, which requires a dramatic change in the methods used. We take a different, simpler, approach to understanding these problems, and we discuss their interesting extensions to higher dimensions.

Section 2. The Application of Erdős' k-Family Theorem

Here is the original result of M. Miller *et al.* :

Theorem 1. [13] *If* $a_1, \ldots, a_n \in \mathbf{R}\backslash\{0\}$, *then*

$$\left| \left\{ I \subseteq [n] : \sum_{i \in I} a_i = 0 \quad \text{or} \quad 1 \right\} \right| \quad \leq \quad \binom{n+1}{\lfloor \frac{n+1}{2} \rfloor},$$

and this is best-possible.

Their bound is achieved by taking $|a_i| = 1$, where the number of $a_i = 1$ is $n/2$ or $(n/2)+1$, if n is even, and $(n+1)/2$, if n is odd.

The papers of M. Miller *et al.* both nicely apply the symmetric chain decomposition of the Boolean lattice B_n. To extend their results (especially, to obtain the extremal families) we instead use k-families of subsets. Indeed, Erdős' work on the Littlewood-Offord problem [5] immediately yields the following strengthening of Theorem 1, as we shall explain in this section:

Theorem 2. *Let* $a_1, \ldots, a_n \in \mathbf{R}\backslash\{0\}$. *Let* $\delta = \min_i |a_i|$. *Let* T *be a union of* k *half-open intervals* S_j *of width* δ. *Then the*

number of sums $\sum_{i\in I} a_i \in T$, $I \subseteq [n]$, *is at most the sum of the* k *middle binomial coefficients in* n.

Now in particular for $k = 2$, we can shrink the k intervals of width δ down to just two points, by taking $T = \{0, 1\}$, and we derive Theorem 1, since the two middle binomial coefficients in n can be combined:

$$\binom{n}{\lfloor \frac{n+1}{2} \rfloor} + \binom{n}{\lfloor \frac{n-1}{2} \rfloor} = \binom{n+1}{\lfloor \frac{n+1}{2} \rfloor}.$$

The Littlewood-Offord problem was posed by the number theorists back in the 1930's [11] in connection with their study of the roots of random algebraic equations. It asks for the maximum number of the 2^n sums $\sum_{i=1}^{n} \varepsilon_i a_i$, where each $\varepsilon_i = 1$ or -1, that lie inside any open ball $S \subseteq \mathbf{C}$ of unit radius, over all choices of S and the the numbers $a_i \in \mathbf{C}$, subject to the restriction that $|a_i| \geq 1$. An equivalent problem is to maximize the number of the 2^n sums $\sum_{i\in I} a_i \in S$, over $I \subseteq [n]$, where S is an open ball of unit diameter and $a_i \in \mathbf{C}$, $|a_i| \geq 1$.

Erdős [5] solved the restriction of this problem to the reals in 1945 via Sperner's Theorem. The similarity to our problem is apparent, even though now we have a lower bound on $|a_i|$. Given any $a_i \in \mathbf{R}$, $|a_i| \geq 1$, a nice observation is that if we replace, say, a_1, by its opposite, $-a_1$, then the full collection of the 2^n sums $\sum_{i\in I} a_i$, $I \subseteq [n]$, has the same relative location, but is translated by $-a_1$, since for any $I \subseteq \{2, 3, \ldots, n\}$, the pair

$$\sum_{i\in I} a_i, \qquad a_1 + \sum_{i\in I} a_i$$

can be replaced by the pair

$$-a_1 + \sum_{i\in I} a_i, \qquad \sum_{i\in I} a_i.$$

In particular, the maximum concentration of sums $\sum_{i\in I} a_i$ in any open unit interval S is the same. So it suffices to consider just the case that all $a_i > 0$, *i.e.*, $a_i \geq 1$.

The key insight is that in this case, for any open unit interval S, the collection of index sets

$$\left\{ I \subseteq [n] : \sum_{i \in I} a_i \in S \right\}$$

is an antichain, for if $I \subset J \subseteq [n]$, then

$$\sum_{i \in J} a_i - \sum_{i \in I} a_i = \sum_{i \in J - I} a_i \geq 1,$$

and not both $\sum_{i \in I} a_i$ and $\sum_{i \in J} a_i$ are in S. Sperner's Theorem [15], that $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ is the maximum size of any antichain $A \subseteq 2^{[n]}$, tells us that $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ is the maximum concentration of sums in S. In fact, this is easily achieved by taking all $a_i = 1$ and by centering S at $\lfloor n/2 \rfloor$.

More generally, Erdős observed that if S is an open interval of width $k \in \mathbf{Z}^+$, then no $k+1$ index sets I such that $\sum_{i \in I} a_i \in S$ form a chain. This is equivalent to saying that the collection $\{I \subseteq [n] : \sum_{i \in I} a_i \in S\}$ is a *k-family* (union of k antichains). Erdős proved

Theorem 3. [5] *The maximum size of a k-family in* $B_n = (2^{[n]}, \subseteq)$ *is the sum of the k middle binomial coefficients.*

Erdős' sharp bound is achieved by taking the family of subsets of k middle sizes. Specifically, for $n + k$ odd, take the consecutive sizes

$$\frac{n-k+1}{2}, \quad \cdots \quad , \frac{n+k-1}{2},$$

and for $n + k$ even, take the sizes

$$\frac{n-k}{2}, \quad \cdots \quad , \frac{n+k}{2} - 1,$$

or take the sizes

$$\frac{n-k}{2}+1, \quad \ldots \quad , \frac{n+k}{2}.$$

To derive Theorem 2 from Theorem 3, we argue similarly that we may assume all $a_i > 0$, and we observe that the collection $\{I \subseteq [n] : \sum_{i\in I} a_i \in T\}$ is a union of k antichains.

Section 3. Extremal Configurations for Theorem 1.

We can now use k-families to show that the only families achieving the maximum in Theorem 1 are those above. For general k, the possibilities for the target set T and the numbers a_i are very restricted.

Theorem 4. *Let $a_1, \ldots, a_n \in \mathbf{R}\backslash\{0\}$. Let $T = \{x_1, \ldots, x_k\} \subseteq \mathbf{R}$, $k \leq n+1$. The number of sums $\sum_{i\in I} a_i \in T$ is maximum, the sum of the k middle binomial coefficients in n, if and only if for some $\delta > 0$, $|a_i| = \delta$ for all i and, taking $\lambda = |\{i : a_i = \delta\}|$, T contains k middle values in the set*

$$\{(\lambda - n)\delta, (\lambda - n + 1)\delta, \ldots, \lambda\delta\}.$$

Proof. Let us first consider the case that all $a_i > 0$. Continuing the proof of Theorem 2, we find that

$$F := \left\{ I \subseteq [n] : \sum_{i\in I} a_i \in T \right\}$$

is a k-family of maximum size.

One can show, with standard methods in extremal set theory, that a k-family G of maximum size must consist of all subsets of the k middle sizes given explicitly in the displays at the end of the last section. This is an instance of what has been called the *strict k-Sperner property* [4, p.49]; Engel [3] gave a

general result, which includes our case of the Boolean lattice B_n. A simpler direct proof is to note that by the so-called LYM inequality in B_n (see such surveys as [7] or [1]), G can contain no sets in levels strictly smaller than the k middle levels in B_n. In fact, G must consist of the k middle levels when $n+k$ is odd, and when $n+k$ is even, G contains the $k-1$ middle levels plus some elements of the two levels, one on each side of the middle, that surround these. An argument similar to Sperner's original proof [15] of the case $k=1$ implies that G contains all of one and none of the other of these two surrounding levels.

For the sums $\sum_{i\in I} a_i$, $I \in F$, to generate just the k values in the target set T, it must be that all a_i are identical, say all $a_i = \delta > 0$, and T contains the k middle multiples of δ among $0, \delta, \ldots, n\delta$.

In the general case where some $a_i < 0$, we replace such a_i by $-a_i$ and translate the target set T by $-a_i$, to get an equivalent problem. Working backwards from the case that all $a_i > 0$, we obtain the stated extremal families. ∎

Section 4. Extensions to Higher Dimensions.

We now consider our problem in the space $\mathbf{R}^m$ of general dimension m, where we take $|a_i|$ to denote the Euclidean norm. The original Littlewood-Offord problem concerned sums of numbers $a_i \in \mathbf{C}$, which is equivalent to looking at it in $\mathbf{R}^2$. Not until 20 years after Erdős solved the one-dimensional case, did Katona and Kleitman (independently) [8,9] prove that in $\mathbf{R}^2$, the maximum number of sums $\sum_{i\in I} a_i$ inside any open ball S of unit diameter is still just $\binom{n}{\lfloor \frac{n}{2} \rfloor}$. Extending this result to $\mathbf{R}^m$ for general m, required a radically different approach: Kleitman [10] observed that for given a_i, the 2^n index sets I can be partitioned into collections with the property that no two sets have sums which are at distance less than one. The same idea, scaled by $\delta = \min_i |a_i|$, allows us to extend our Theorem 2 to higher dimensions:

Theorem 5. *Let $a_1, \ldots, a_n \in \mathbf{R}^m \backslash \{0\}$. Let $\delta = \min_i |a_i|$. Let T be a union of k open balls S_j of diameter δ. Then the number of sums $\sum_{i \in I} a_i \in T$, $I \subseteq [n]$, is at most the sum of the k middle binomial coefficients in n.*

Before presenting the proof, we describe how to obtain the bound in Theorem 5, for the special case that the target sets T consist of k points, by the following simple reduction to $\mathbf{R}^1$: Given any problem instance, with $a_1, \ldots, a_n \in \mathbf{R}^m \backslash \{0\}$ and $T = \{x_1, \ldots, x_n\} \subseteq \mathbf{R}^m$, select any $b \in \mathbf{R}^m$ such that b is not on any of the k hyperplanes

$$\{b \in \mathbf{R}^m : b \cdot a_i = 0\}.$$

Then $b \cdot a_1, \ldots, b \cdot a_n \in \mathbf{R}^m \backslash \{0\}$ and an index set $I \subseteq [n]$ satisfies $\sum_{i \in I} a_i = x_j$ only if it satisfies $\sum_{i \in I} b \cdot a_i = b \cdot x_j$ (*i.e.*, we project the problem onto the line from the origin through b). This means the sum $\sum_{i \in I} b \cdot a_i$ hits the target set $T_b := \{b \cdot x_1, \ldots, b \cdot x_k\}$ consisting of k (not necessarily distinct) values in $\mathbf{R}$. By Theorem 2, the number of sums meeting T is at most the sum of the k middle binomial coefficients in n.

Pushing this method farther, we can obtain all extremal configurations in $\mathbf{R}^m$ for target sets of k points. However, an extension of our earlier k-family method is nicer. We return to this below in Theorem 6. First, we prove Theorem 5, which extends this bound for a target set of k points to a target set of k (small) balls.

Proof of Theorem 5. We prove by induction on p that for $1 \leq p \leq n$, the 2^p index sets $I \subseteq \{1, \ldots, p\}$ can be partitioned into $\binom{p}{\lfloor \frac{p}{2} \rfloor}$ collections we call here *anticlusters*, which are collections $A \subseteq 2^{[n]}$ such that for distinct $I, J \in A$, $|\sum_{i \in I} a_i - \sum_{j \in J} a_j| \geq \delta$. Further, we require that exactly $\binom{p}{q} - \binom{p}{q-1}$ of these anticlusters have size $p + 1 - 2q$, $0 \leq q \leq \lfloor p/2 \rfloor$. This distribution of sizes is in fact identical to that for the sizes of the chains in the symmetric chain decomposition of the Boolean lattice B_p: Both distributions result from the same induction.

We start the induction at $p = 1$ by simply taking the single anticluster $\{\emptyset, \{1\}\}$. For general p, $1 \leq p < n$, assume that we have a partition of $2^{[p]}$ into anticlusters $A_1, A_2, \ldots$ of the correct sizes.

For each anticluster A_l, let I_l be a set in A_l that maximizes, over all $I \in A_l$, the inner product $\left(\sum_{i \in I} a_i\right) \cdot a_{p+1}$. Geometrically, $\sum_{i \in I_l} a_i$ is a last point over all such sums met by a hyperplane sweeping in direction a_{p+1}. Then the collection

$$A'_l := A_l \cup \left\{ I_l \cup \{p+1\} \right\}$$

is also an anticluster, as $\sum_{i \in I_l} a_i + a_{p+1}$ is at distance at least $|a_{p+1}|$ from the sums $\sum_{i \in I} a_i, I \in A_l$. The collection

$$A''_l := \left\{ I \cup \{p+1\} : I \in A_l, I \neq I_l \right\}$$

is also an anticluster, for it consists of translates by a_{p+1} of sums in A_l. We have a partition of $2^{[p+1]}$ into anticlusters $A'_1, A''_1, A'_2, A''_2, \ldots$, and one can check that their sizes

$$|A'_l| = |A_l| + 1, \qquad |A''_l| = |A_l| - 1$$

satisfy the size condition. We discard $A''_l = \emptyset$ if $|A_l| = 1$. This completes the induction.

For $p = n$, we get a partition into $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ anticlusters A_l. Now for any anticluster A and any open ball S in $\mathbf{R}^m$ of diameter δ, at most one sum $\sum_{i \in I} a_i$, $I \in A$, lies inside S. Thus, at most $\min\{k, |A|\}$ of the sums $\sum_{i \in I} a_i$, $I \in A$, belong to $T = \cup_{j=1}^{k} S_j$. So we find that

$$\begin{aligned} \left| \left\{ I \subseteq [n] : \sum_{i \in I} a_i \in T \right\} \right| &= \sum_l \left| \left\{ I \in A_l : \sum_{i \in I} a_i \in T \right\} \right| \\ &\leq \sum_l \min(k, |A_l|). \end{aligned}$$

This sum is at most the sum of the k middle binomial coefficients in n, which is most easily seen by noticing that the last sum above is the same as the sum, for a symmetric chain decomposition $\{C_l\}$ of B_n, which is $\sum_l \min(k, |C_l|)$. This sum, in turn, is the sum over l of the number of subsets in C_l of the k middle sizes. ∎

We can slightly extend Theorem 5 by permitting the balls S_j to include at most one point from each pair of antipodal points on its boundary. The proof will be the same as above, but now Theorem 2 is the one-dimensional version.

We conclude by extending Theorem 4, the description of extremal families when T is a finite set, to higher dimensions. We employ the old trick, of multiplying some a_i's by minus one, in a novel way, in order to be able to use the existing theory on k-families to solve this problem.

Theorem 6. *Let $a_1, \ldots, a_n \in \mathbf{R}^m \backslash \{0\}$. Let $T = \{x_1, \ldots, x_k\} \subseteq \mathbf{R}^m$, $k \leq n+1$. The number of sums $\sum_{i \in I} a_i \in T$ is maximum, the sum of the k middle binomial coefficients in n, if and only if each $a_i = a_1$ or $-a_1$, and letting $\lambda = |\{i : a_i = a_1\}|$, the set T contains k middle points in the sequence*

$$\{(\lambda - n)a_1, (\lambda - n + 1)a_1, \ldots, \lambda a_1\}.$$

Proof. We use the notation $a_i = (a_{i1}, \ldots, a_{im})$ to represent the coordinates of a_i. As in one dimension, we replace a_i by $-a_i$ and translate the target set T by $-a_i$, this time for all i such that the first nonzero component a_{ij} is negative.

So we now assume each a_i has positive first nonzero component. This property is shared by any sum of elements a_i, so that for any $I \subset J \subseteq [n]$, the sum $\sum_{i \in J-I} a_i \neq 0$. It follows that for all j,

$$D_j := \left\{ I \subseteq [n] : \sum_{i \in I} a_i = x_j \right\}$$

is an antichain in B_n. Then

$$F := \bigcup_j D_j = \left\{ I \subseteq [n] : \sum_{i \in I} a_i \in T \right\}$$

is a k-family. As with Theorem 4, it follows that F consists of all subsets of the k middle sizes (there are at most two choices for F). One can check that each antichain D_j consists of all subsets of a certain size, and this in turn forces all a_i to be identical. The set T must then consist of k middle values in the sequence $0, a_1, 2a_1, \ldots, na_1$.

The treatment for the case of general a_i now follows the same argument as in the earlier proof of the one-dimensional version, Theorem 4. ∎

References

1. I. Anderson, *Combinatorics of Finite Sets*, Clarendon Press (1987).

2. L. Branković and M. Miller, An application of combinatorics to the security of statistical databases, *Austral. Math. Soc. Gazette* **22** (1995), 173–177.

3. K. Engel, Strong properties in partially ordered sets II, *Disc. Math.* **48** (1984), 187–196.

4. K. Engel and H.-D. Gronau, *Sperner Theory in Partially Ordered Sets*, Teubner (1985).

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

6. P. Frankl and Z. Füredi, The Littlewood-Offord problem in higher dimensions, *Annals Math.* **128** (1988), 259–270.

7. C. Greene and D. Kleitman, Proof techniques in the theory of finite sets, in *Studies in Combinatorics* (G.-C. Rota, ed.), Math. Assn. America (1978) 22–79.

8. 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.

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

10. D. J. Kleitman, On a lemma of Littlewood and Offord on the distributions of linear combinations of vectors, *Advances in Math.* **5** (1970), 1–3.

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

12. K. Miller and D. Sarvate, Application of symmetric chains to a statistical database compromise prevention problem, *Bull. ICA* **13** (1995), 57–64.

13. M. Miller, I. Roberts, and J. Simpson, Application of symmetric chains to an optimization problem in the security of statistical databases, *Bull. ICA* **2** (1991), 47–58.

14. M. Miller, I. Roberts, and J. Simpson, Prevention of relative compromise in statistical databases using audit expert, *Bull. ICA* **10** (1994), 51–62.

15. E. Sperner, Ein Satz über Untermengen einer endlichen Menge, *Math. Z.* **27** (1928), 544–548.