Matching order for Erdős problem 790

Grant Huffman Jensen Kohlmeyer Will Howe Liam Kruer

12 November 2026

The four authors share equal contribution credit for the submission. Grant Huffman (Ghuffy1) is the repository owner and principal submitting contact. Historical theorem and formalization credits appear below and in the public contribution record. The contribution byline is separate from any prize allocation.

hexagon:2611.00778v1

Main theorem and scope

Theorem 1. For the distinct-summand-free extremal function on arbitrary finite integer sets, there are absolute constants c,C > 0 and N such that every integer n ≥ N satisfies c n/(log n)² ≤ l(n) ≤ C n/(log n)². The lower estimate is Samuel Korsky's theorem. Sections 2–5 prove the upper estimate.

Corollary 2. The ratio l(n)/$\sqrt{n}$ tends to infinity. For each $\epsilon > 0$, eventually l(n) ≥ n^(1−ϵ), and log(l(n))/log(n) tends to 1. In particular no fixed positive power saving is possible. These consequences already follow from Korsky's lower estimate.

The precise forbidden relation is $b_0 = b_1 + \cdots + b_r$ with r ≥ 2 and all participating integers pairwise distinct.

Let l(n) be the largest guaranteed size of a subset of any n distinct integers in which an element is never the sum of at least two pairwise distinct other elements. The argument gives absolute constants c,C>0 such that

$$c\,n/(\log n)^2 \le \ell(n) \le C\,n/(\log n)^2.$$

for all sufficiently large n. The lower bound is due to Samuel Korsky. The upper construction obtains the matching bound by propagating and refining progressions of exact mixed averages. The proof includes arbitrary signed integer hosts for the lower estimate and an exact lift to distinct summands for the upper witnesses.

1. The credited lower bound

For a positive real host A, repeatedly divide its sorted list into two consecutive parts of sizes differing by at most one. At each division pick a cut strictly between the parts and associate to each point in that node its distance to the cut. Each point receives at most ceil(log_2 n) distances. For two distinct points x,y, the first division separating them supplies two distances summing to |x-y|. The larger is in [|x-y|/2,|x-y|]. Also include the distance x at each point x.

For each associated distance delta of x, forbid every occupied dyadic bin [$2^{j,2}$(j+1)) other than x's own bin that meets [delta/(2n),2delta]. The entire list for each point has size at most D=O((log n)^2), with D>=1. Select occupied bins independently with probability 1/(2D) and retain each point whose own bin is selected and whose forbidden bins are absent. Its retention probability is at least 1/(4D) by the union bound.

In a forbidden positive relation order the head and summands as x>y>z>..., so z<=x-y<=nz and x>2z. If x,y occupy the same bin, one of their associated distances approximates x-y within factor two, and that point forbids the different bin of z. If the bins of x,y are adjacent, the bin of y is forbidden by the associated distance x because y>x/4>=x/(2n). If their bins differ by at least two, y<x/2 and hence z>x/(2n), so the distance x forbids the bin of z. Each case contradicts simultaneous retention. Thus some outcome gives at least n/(4D) safe points. For an arbitrary integer host, take the larger nonzero sign class, of size at least (n-1)/2, and reflect if necessary. Reflection preserves the equations. This is Korsky's argument, credited to his linked manuscript, Large Sum-Free Subsets of Sets of Integers, Theorem 1. The classical CKS lower bound already answers the square-root subsidiary question affirmatively, and Korsky's bound excludes every fixed power-saving upper.

2. The layered host and its uniform sum property

We use the following deterministic tent construction and uniform sum property. Let d be sufficiently large and put

$$
\begin{aligned}
&w = 6d, && q = 2^w, && L = dw + 1, && R = q^6, && T_0 = 2^{2d},\\
&p_0 = (100000L)^2, && Q = p_0!, && X = Qq^d,\\
&J = 2L\left\lceil 100\left(1 + \log(4X^2 2^{3d})\right)\right\rceil,\\
&k = J/(2L), && B_0 = 100p_0, && P = 2JB_0 + 1, && \lambda = \lceil \log_2 P \rceil.
\end{aligned}
$$

Here J,P are polynomial in d and lambda=O(log d). For 0<=h<d, partition the integer interval [0,T0) into cells [a,a+2M_h) of length 2M_h, where M_h=2^(d+h) and the endpoints are multiples of 2M_h. On that cell put

$$\theta_h(t) = \min\{2(t-a)+1,\ 2(a+2M_h-t)-1\}.$$

On the convex hull of the cell's integers this is a concave tent with slopes 2 and -2 and peak at a+M_h-1/2. Every higher theta_(h') with h'>h is affine there, since its half-integer slope breaks lie just before multiples of 2^(d+h') and none is internal to this cell's integer convex hull.

Set Gamma=2^(lambda+3) and U0=Q*floor(q^d/(8T0)). For t in [0,T0) and a binary word epsilon of length d form

$$z(t,\epsilon) = tU_0 + Q\sum_{h=0}^{d-1}(\Gamma\theta_h(t) + \epsilon_h)q^h.$$

The deterministic offsets are distinct multiples of Q in [0,X/4), with 2^(3d) choices. At fixed t the radix digits distinguish the word, and the total tent and digit contribution is at most X*(Gamma T0+1)/(q-1), much smaller than the step U0 of order X/T0. The bound on z follows also from t U0<=X/8.

There exist J auxiliary offsets b in [0,X/4), all coprime to Q, with distinct residues modulo Q and the following uniform phase property. For each deterministic z and each nontrivial character a/M with integer M in [X,2X) coprime to Q, fewer than k of the indices satisfy ||a(X+b+z)/M||<1/p0. Here the norm is distance to the nearest integer. To see this, sample the offsets independently from the rough residues in the interval. The interval is whole periods modulo Q and its allowed density is at least 1/(4 sqrt(p0)). Indeed the prime product for the density is bounded below by (1/2) times the product over all odd integers from 3 through p0, and log(1-1/j)>=-1/(j-1) bounds this product

using the harmonic sum. The order p=M/gcd(a,M)>p0. Before conditioning, at most 2p/p0+2 residue classes modulo p are bad, so the probability is at most (2/p0+2/p)(1+4p/X)<=36/p0. After conditioning it is below 1/(100L). For a fixed character and offset z the chance of k bad sample indices is at most (e/50)^k by choosing those indices. Union over at most 4 X^2 2^(3d) choices gives failure probability below 1/4 by the definition of J. The probability of a repeated residue modulo Q is at most binom(J,2)/phi(Q)<1/4 for large d, since phi(Q)>=Q/(4 sqrt(p0)) and J is polynomial in d. Fix a sample meeting both properties.

Put C={X+b+z(t,epsilon)}, with m=J T0 2^d points in [X,2X). Distinct residues separate the auxiliary offset fibers, and all points are coprime to Q. For any S subset C of size at least m/L, every integer N>=PX is a sum of at most N/X members of S, with repetitions allowed. Indeed fix a deterministic offset fiber with at least J/L=2k members and take one of its members as modulus u0. On that fiber independently vary each summand coefficient uniformly in {0,...,B0-1}. For a nontrivial character modulo u0 at least k factors have phase distance >=1/p0, so the corresponding normalized geometric sums have modulus at most 1/(2B0/p0)=1/200. Finite Fourier inversion gives positive probability for each residue, because the trivial character contributes 1 inside the 1/u0 factor and the total other contribution in absolute value is at most u0*200^(-k)<1/2. Choose coefficients attaining N modulo u0. Their nonnegative sum is below 2JB0 X<N, so append copies of u0. Every term is at least X.

Set F=union_(j=0)^(L-1) 2^j C, a disjoint union with |F|=mL. Temporarily allow repeated summands of arity 2 through R. For a subset B of F avoiding all such relations let U_j={u in C:2^j u in B}. A heavy level has |U_j|>=m/L. The light levels together contribute less than m. If i is heavy and j occupied with j-i in [lambda,6w-1], any u in U_j gives target 2^(j-i)u in [PX,RX), a forbidden sum from level i. Group heavy levels into clusters by joining successive gaps smaller than lambda. Distinct clusters are separated by at least 6w. A cluster has diameter smaller than lambda, since the first index reaching distance lambda from its start would have distance less than 2lambda<6w-1. A base point cannot occur at two levels in one cluster, since the higher dilate is a repeated sum of copies of the lower with arity <=R.

For a cluster starting at j0 pick one heavy representative i from each earlier cluster and associate

$$h = d - \lfloor (j_0 - i)/w \rfloor + 2.$$

These positions are distinct, in [2,d-4]. At any j in this cluster,

$$(j - i) + w(h - d) \in [2w, 4w).$$

Let H be their set, with r=|H|<=d/6. At a fixed j, b, and t, two points whose largest differing digit is h in H would have positive difference D between Q q^h/2 and 2Q q^h. Then 2^(j-i)D/X is between q^2/2 and 2q^4, so a repeated sum from earlier heavy i gives a forbidden relation with the two high points. Thus fixing b,t and epsilon outside H leaves at most one selected binary profile per level. With more high bits fixed, at most 2^s occurrences of any fixed t are possible across the whole cluster if s bits remain free, since no base point is repeated across levels.

3. Mixed attainable means

Work in one cluster with fixed b, epsilon outside H and all bits of H higher than a chosen h. Work in one h-cell of the t domain, with halves of M=M_h integers each. A mixed mean at time j is the exact weighted average of t values of a multiset of selected points in this subproblem at levels k<j, with each such term's weight 2^(k-j), total weight one, and at least one term from each t half. The same mean remains attainable at j+1 by doubling all multiplicities. If a mixed mean x at j and a

selected point with t value y at level j are both available, their ordinary midpoint is a mixed mean at j+1.

A selected point at level j cannot have its t equal to a mixed mean at j in its subproblem. Indeed the near terms cancel the X+b and t U0 parts. Higher tent functions are affine in this cell and higher binary bits are fixed, so all higher digits cancel. At h the tent deficit is at least $2 \cdot 2^{-\lambda}$. To check this, extend the affine branch on the head's side across the cell. Each opposite-side term has weight at least 2^(-lambda), distance at least 1/2 from the peak, and its tent value lies below that extension by four times that distance. Thus the positive contribution after multiplying by Gamma is at least 16. The binary difference has magnitude at most one. The absolute residual coefficient of any digit is at most Gamma T0+1, so the tail below h is less than Q q^h for large d. In head-level units the positive residual therefore lies between Q q^h and Q q^(h+1). For i associated to h, multiplication by 2^(j-i) gives an integer N because every near level is above i, with $q^{2<=N/X<=q}5$. The earlier heavy level supplies the residual in at most q^5 terms. The near terms number at most 2^lambda by the weight sum and smallest possible weight. The combined arity is at most R and is at least two. This is a forbidden repeated relation.

We prove a bound for the later-starting side of this cell. If both sides are occupied, choose the side whose first level is smaller as the earlier side, breaking a tie arbitrarily, and fix one of its points c at its first level. Write S_j for the set of t values selected on the later side at level j. The point c's level is no larger than any j with nonempty S_j. At most one point occurs per t at each level, and at most 2^s occurrences per t across the cluster if s bits are free. We will show, for an absolute C2, that

$$\sum_j |S_j| \le M\big(2^s \eta_s + C_2 (s+1)^2\big), \qquad \eta_s = 2^{-s/4-4}. \tag{1}$$

The proof uses intervals of mixed means in an arithmetic progression, an exact additive charge while a residue class persists, and a strict decrease of the modulus whenever a selected point leaves that class.

Seeding an interval of mixed means

Use eta=eta_s. If the cumulative union of S_j never reaches size eta M, the bound follows from the multiplicity limit. Otherwise let a be the first level where its cumulative union A reaches eta M. Before a, there are fewer than 2^s eta M occurrences. Let u=min A, v=max A, W=v-u, and g=gcd{t-u:t in A}. Then W>=eta M/2 and 1<=g<=2/eta, uniformly for large d. Each t in A can be used as one term at effective level a by repeating a selected point at its actual level with the appropriate power of two. The opposite point c can be used in the same way.

Here is the needed integer sumset fact. Put D=A-u subset [0,W], which contains 0,W. Modulo W its residues P generate the subgroup of multiples of g and have size |A|-1. Within that subgroup every residue is a sum of at most K0=ceil(2W/(|A|-1))<=5/eta elements of P. By finite Kneser's theorem, if K0 P is proper and has stabilizer V, its size is at least |V|+K0*(|P+V|-|V|). The quotient by V is aperiodic at the final sum and hence at every earlier partial sum, so this is the iterated two-set Kneser inequality. Since P generates the subgroup and contains zero, it meets at least two V-cosets. Therefore |P+V|-|V|>=|P|/2, whether |V| is at most or exceeds |P|/2. The resulting lower bound exceeds W, a contradiction. This uses Kneser's addition theorem, in the form stated and proved by Matt DeVos.

Consequently any multiple N of g in [K0 W,(b-K0)W] is the sum of exactly b terms of D, with repetitions. First use at most K0 terms for its residue modulo W, of sum at most K0 W. Then append copies of W to reach N and zeros to make the number of terms b. The upper bound on N ensures the count fits.

Take ($b_0 = 2^{s_0}$) the least power of two at least 100/etaˆ2. At time a+s_0 every integer t in the central half [u+W/4,v-W/4] with t congruent u modulo g is a mixed mean. To realize it, use g effective copies of c and b_0-g terms from A. The desired difference from (b_0-g)u is

$$N = b_0(t-u) - g(c-u).$$

It is divisible by g. Since |c-u|<=2M, W>=eta M/2, g<=2/eta and K0<=5/eta, the central-half inequalities put N in [K0 W,(b_0-g-K0)W]. The preceding fact applies. Both sides are represented. Thus we have a full progression of mixed integer means in the later half, of span at least eta M/8, in the coset u+g Z. Call this its initial core. The delay s_0 is O(s+1), costing at most s_0 M further occurrences. If the time is beyond the cluster, this already proves (1).

Charging selected points while the coset persists

At any seeded time j keep a set T_j of mixed means inside the later half, a union of consecutive lattice intervals in the current lattice u+g Z. Start with the core progression. All selected S_j avoid T_j by the tent argument. At a stage with S_j inside the current coset, define T_(j+1) to consist of T_j carried forward and every midpoint (x+t)/2 in that coset with x in T_j and t in S_j. These are attainable mixed means at j+1. Maximal consecutive lattice components of T_(j+1) contain either an old component or an image of an old component of at least half its lattice size rounded down, since taking x of one parity in lattice coordinates and averaging with fixed t gives a consecutive lattice interval.

In each gap of T_j in the later half of the lattice, choose an adjacent component, using either boundary if two exist. Its two boundary points nearest the gap have opposite parities in lattice coordinates. For t in that gap at distance at least two lattice steps from the boundary, averaging with the parity-matching boundary point or its neighbor yields a lattice point strictly inside the same gap. It is newly attained. The map is at most two-to-one in each gap, since only two seed values are used. At most one t per gap is excepted. Hence

$$|T_{j+1}| - |T_j| \geq (|S_j| - \text{number of gaps of } T_j)/2. \tag{2}$$

The size lower bound on components used here holds uniformly even after all modulus changes below. The core span will always be at least etaˆ3 M/32, and the modulus at most 2/eta. At the start of a stable period the core has at least etaˆ4 M/64 lattice points. Through at most lambda steps, every component therefore has at least etaˆ4 M/(128*2ˆlambda) points for large d, by the halving bound. This is far larger than two since s<=d/6, M>=2ˆd and lambda=O(log d). The number of gaps is at most 1+128*2$^{\text{lambda*eta}}$(-4). Even its product with lambda is o(M) uniformly, because eta$^{(-4)=2}$(s+16)<=2ˆ(d/6+16). Summing (2) in any stable period thus charges at most 3M occurrences, including the final propagation to the next time if the cluster ends.

Refining the lattice

If S_j contains a point b outside the current coset u+g Z, use it to replace g by g'=gcd(g,b-u)<g. Keep the persistent core interval in the old coset, with span E and endpoints p,q1. Choose ($a_1 = 2^{s_1}$) the least power of two at least 16gM/E. For any target t congruent u modulo g' in the central half of [p,q1], choose 0<=v1<g/g' such that

$$a_1 t \equiv (a_1 - v_1)u + v_1 b \pmod{g}.$$

This is possible because b-u generates the subgroup of multiples of g' modulo g. The real number (a_1 t-v1 b)/(a_1-v1) lies in [p,q1], since its deviation from t is at most 2g M/a_1<=E/8.

The progression of core means therefore realizes the desired integer sum with a_1-v1 terms by full interval addition and the displayed congruence. Use v1 copies of b. These are a_1 blocks of effective weight 2^j, with at least one old mixed core block. They give t as a mixed mean at time j+s_1. The new core has step g' and span at least E/4 for large d, after rounding central endpoints. A modulus can decrease at most log_2 g_initial times, and g_initial<=2/eta. Thus the initial span eta M/8 decreases by a total factor at most g_initial$^{2<=4/\mathrm{eta}}$2, proving the claimed bound eta^3 M/32. Each delay s_1 is O(s+1) by the modulus and core-span bounds. During a delay charge at most s_1 M occurrences starting with the triggering stage, or all remaining stages if the new time falls outside the cluster. Resume with the new core and reset the propagated set to it. The charge 3M per stable period and the total delay O((s+1)^2)M prove (1). Every propagated or refined mean is represented by actual earlier points of this subproblem with the required exact weight sum. At any head time in the cluster its repetition count is automatically bounded by 2^lambda.

4. Summing the cluster masses

Keep b and epsilon outside H fixed. At a recursion node the t domain is a dyadic interval, some highest bits in H have been fixed, and s bits remain free. We bound its mass by (f_s 2^s) times its t interval length, with absolute nonnegative f_s. For s=0 choose f_0=1 because a t determines one base point and that point cannot repeat across levels in the cluster.

For s>0 partition the interval into cells of the largest remaining h. In each cell, if both sides occur use the first-level ordering of Section 3. If only one occurs call it earlier. Bound the later side by (1). On the earlier side fix epsilon_h to either value and recurse with s-1 free bits, summing the two bounds. All smaller cells partition its half, so the recurrence is

$$f_s \le f_{s-1}/2 + \eta_s/2 + C_2(s+1)^2 2^{-s-1}.$$

Choose f_s by equality on the right. Its sum over s>=0 is finite, since the last two terms are summable and the first has coefficient 1/2. For example the partial sums are bounded by twice the sum of the starting and forcing terms. Summing over all b and the fixed exterior words gives mass at most m f_r for the cluster with r earlier clusters. Each r occurs at most once. Light mass is less than m. We obtain an absolute bound

$$\alpha_{\mathrm{rep}}(F; \text{arities } 2, \dots, R) \le C_3 m.$$

5. Distinct summands and exact host size

Use Kelley–Meka, Strong Bounds for 3-Progressions, Theorem 1.2. A set of density at least η' in an integer interval of size T has at least exp(-C4*(1+log(1/eta'))^12) T^2 ordered midpoint triples. Set eta'=1/L and choose integer T>(R+1) exp(C4*(1+log L)^12) with log T=O(d). Choose H0>T max F and form

$$A_d = \{a(H_0 + u) : a \in F,\ 1 \le u \le T\}.$$

There are N_d=T m L distinct positive integers because b(H0+1)>a(H0+T) whenever b>a. For any B subset A_d let U_a be its selected digits in fiber a. Call v in U_a good if it has at least floor(R/2) unordered nontrivial centered pairs in U_a. These pairs are pairwise disjoint and avoid the center. The bad set in U_a has size less than eta'T. Otherwise the progression theorem on that bad set gives more than (R+1)T ordered triples, while each bad center supplies at most R+1 even in U_a. Therefore some v is good in at least |B|/T-eta'|F| coarse labels.

If those labels have a repeated relation a_0=sum_a m_a a of arity from 2 through R, positivity ensures the head label differs from every summand label. For each summand label choose m_a distinct digits of average v by taking disjoint centered pairs and including v when the multiplicity is odd. The actual lifted points are distinct and sum to a_0(H0+v), a distinct selected head. Thus any distinct-summand-safe B has size at most T(C3 m+eta'|F|)=O(Tm). This is the original relation, with the intermediate repeated terms fully removed.

Here log m=O(d), log R=O(d), log T=O(d) and L=6d^2+1. In particular N_d<=exp(C5 d) for an absolute C5 and large d. For arbitrary sufficiently large n choose d=floor(log n/(2C5)), so N_d<=sqrt n. Take floor(n/N_d) disjoint integer dilates of A_d and add fewer than N_d extra positive points to reach n. Each dilate retains its bound and a safe set in the union restricts to safe sets in every dilate. Its size is at most O(n/L)+N_d=O(n/(log n)^2). Positive integer witnesses suffice for the minimum over arbitrary integer hosts, and Section 1 covers all signs and zero for the lower estimate. The two subsidiary answers follow in their original scopes.

Formal proof correspondence

The machine-checked main conclusion is Erdos790.matching_order in Erdos790/Upper/Full.lean. The upper estimate is Erdos790.matching_upper_bound. The lower estimate is Erdos790.korsky, imported from the attributed lower-bound development. Erdos790JSP/OriginalStatement.lean states the finite-set condition independently and Erdos790JSP/StatementBridge.lean proves its equivalence to the submitted definitions and transfers the conclusions.

Section 5 uses Kelley–Meka's quantitative three-progression theorem as an informal input. The formal proof uses the pinned OpenAI theorem OAI.Erdos3.manuscriptQuantitativeDensityTheorem. Its AP-free extremal estimate has the form $r_k(N) \le C\, N \exp(-c\,(\log\log N)^{1+\eta})$ for positive constants. Erdos790/Upper/ProgressionInput.lean proves the finite progression and centered-pair bridge needed here, with lift length 2^(100d). Both inputs supply the required density decay. The formal dependency is stated precisely here to distinguish the two sufficient routes.

The manuscript and Lean development were prepared with Codex assistance. The recorded verification report documents the Windows build and the 18 direct target and axiom checks, which passed. It also records that the additional fresh whole-closure replay timed out.

References and previous results

- Samuel Korsky, *Large Sum-Free Subsets of Sets of Integers*, Theorem 1, 13 September 2026. Korsky attributes the principal bisection idea to GPT Astra and the completed verification to his own work. Public claim and manuscript link
- S. L. G. Choi, J. Komlós and E. Szemerédi, *On sum-free subsequences*, Transactions of the American Mathematical Society 212 (1975). The classical upper bound is O(n/log n). Paper
- Matt DeVos, *A short proof of Kneser's addition theorem for abelian groups*, Theorem 1. Preprint
- Zander Kelley and Raghu Meka, *Strong Bounds for 3-Progressions*, Theorem 1.2. Preprint
- OpenAI mathematics development, OAI/Combinatorics/Progressions, revision 4eaee87548250b2682affda7d05d947c05554bae. Source
- The original problem is Erdős 790 and the corresponding prize catalog item is JSP-000649. Original statement

The new mathematical claim is the upper bound and the resulting matching-order synthesis. The public priority search found no earlier matching upper estimate. Global first priority and award attribution await external review and public evidence.