A Proposed Solution to Erdős Problem 788

Shouqiao Wang

Columbia University Multiscalar Intelligence

Abstract

For $n \geq 1$, put

$$I_n = (n, 2n) \cap \mathbb{N}, \qquad J_n = (2n, 4n) \cap \mathbb{N},$$

let $f(n)$ be the largest integer t such that every $B \subseteq J_n$ admits a set $C \subseteq I_n$ with $c + c' \notin B$ for distinct $c, c' \in C$ and $|B| + |C| \geq t$. We prove that there is an absolute constant $c > 0$ such that, for all sufficiently large n,

$$c\sqrt{n \log n} \leq f(n) \leq n^{\frac{1}{2}+O\left(\left(\frac{\log\log n}{\log n}\right)^{1/3}\right)}.$$

In particular, $f(n) = n^{1/2+o(1)}$, which gives an affirmative answer to the principal question for every sufficiently large n.

The upper bound is obtained from a family of $p^{o(r)}$ surjective $\mathbb{F}_p$-linear maps $\mathbb{F}_p^{2r} \to \mathbb{F}_p^r$ which is a strong seeded extractor at min-entropy $r + o(r)$. Taking the union of their kernels gives a palette of $p^{r+o(r)}$ group sums with independence number $p^{r+o(r)}$. We give a self-contained mixed-alphabet Trevisan–RRV reconstruction argument establishing the extractor, and then lift the construction to ordinary integer sums while accounting for every base-p carry. An ordered block design uses the unused suffix of each hybrid to pay for its overlap excess, which gives the stated quantitative rate. The lower bound follows from the sparse-neighborhood coloring theorem of Alon, Krivelevich, and Sudakov. This proposed solution was found by GPT-5.

1 Introduction

Throughout, $\mathbb{N} = \{1, 2, \dots\}$, and all logarithms are natural unless a base is explicitly indicated. For $n \in \mathbb{N}$, put

$$I_n = (n, 2n) \cap \mathbb{N}, \qquad J_n = (2n, 4n) \cap \mathbb{N}.$$

For $B \subseteq J_n$, call $C \subseteq I_n$ *B-admissible* if $c + c' \notin B$ whenever $c, c' \in C$ are distinct, and define

$$f(n) = \max\{t \in \mathbb{Z} : \text{ every } B \subseteq J_n \text{ has a } B\text{-admissible } C \text{ with } |B| + |C| \geq t\}.$$

For a finite nonempty set W, write U_W for its uniform distribution. For probability distributions P, Q on the same finite set, write

$$d_{\mathrm{TV}}(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)|.$$

When Q has full support, also write

$$D(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

This is Problem 788 in the Erdős Problems collection [1]. The collection cites Erdős's 1973 survey [2], which credits Choi's 1971 paper [3] with raising the problem.

Choi proved $f(n) \ll n^{3/4}$ and conjectured the bound $f(n) \ll_\varepsilon n^{1/2+\varepsilon}$ [3]. Baltz, Schoen, and Srivastav improved the upper bound to $f(n) \ll (n \log n)^{2/3}$ [4]. The online problem record reports the subsequent bound $f(n) \le n^{3/5+o(1)}$, obtained by combining a random-Cayley-graph reduction with work of Alon and Pham [1, 5]. None of these upper bounds is used below; our construction directly gives the conjectured exponent for every sufficiently large interval length.

Our main theorem gives a little more quantitative information than the requested exponent.

Theorem 1.1 (Main theorem)**.** *There are absolute constants $c, C > 0$ such that, for all sufficiently large n,*

$$c\sqrt{n \log n} \le f(n) \le n^{\frac{1}{2}+C\left(\frac{\log\log n}{\log n}\right)^{1/3}}.$$

Consequently

$$f(n) = n^{1/2+o(1)}.$$

Remark 1.2 (Quantifiers in the upper bound). The theorem supplies a witness palette for every sufficiently large integer n, not merely along a subsequence. Equivalently, for each fixed $\varepsilon > 0$ and all sufficiently large n, one can choose a single $B \subseteq J_n$ such that every B-admissible $C \subseteq I_n$ obeys $|B| + |C| \le n^{1/2+\varepsilon}$.

The two sides of the proof use different information. The lower bound holds for every palette B. For the upper bound we construct one palette, depending on n, by combining coding theory, weak designs, and an exact carry closure.

The proof is organized as follows. Section 2 converts the problem into an exact minimization involving the independence number of a sum graph. Section 3 proves the universal lower bound. Section 4 constructs and verifies a mixed-alphabet linear seeded extractor, using an ordered weak design whose overlap excess is paid for by the remaining hybrid coordinates; Section 5 converts its kernels into a finite-field sum palette. Section 6 lifts that palette to ordinary integer sums without discarding any carry pattern. Section 7 chooses the parameters uniformly for every interval length, and Section 8 translates the construction back to the original intervals and completes the proof.

2 Exact graph formulation

For $s \in J_n$, let M_s be the graph on I_n whose edges are the unordered pairs $\{c, c'\}$ satisfying

$$c \ne c', \qquad c + c' = s.$$

Each M_s is a matching: after one endpoint is specified, the other is forced. A possible equality $c = s - c$ creates no loop, since the problem only constrains distinct elements. For $B \subseteq J_n$, put

$$G_B = \bigcup_{s \in B} M_s.$$

Proposition 2.1 (Min–max identity)**.** *For every $n \ge 1$,*

$$\boxed{f(n) = \min_{B \subseteq J_n} \left(|B| + \alpha(G_B)\right).}$$

Moreover, the minimum is unchanged if B is required to consist only of sums of two distinct elements of I_n.

Proof. A set $C \subseteq I_n$ is B-admissible exactly when it is independent in G_B. Thus, for fixed B, the largest possible value of $|B| + |C|$ is $|B| + \alpha(G_B)$. An integer t has the universal property in the definition of $f(n)$ exactly when

$$t \le |B| + \alpha(G_B) \qquad \text{for every } B \subseteq J_n.$$

The family of choices of B is finite, so taking the largest such t gives the displayed minimum.

If s is not the sum of two distinct elements of I_n, then M_s is empty. Deleting such an s from B leaves G_B unchanged and decreases $|B|$. Hence no minimizer needs an unattainable sum. $\square$

It will be useful to normalize the vertices. Put

$$N = n - 1, \qquad [N]_0 = \{0, 1, \dots, N-1\}.$$

Write $c = n + 1 + x$. Then

$$c + c' = 2n + 2 + x + y.$$

For $N \geq 2$, the sums of two distinct members of $[N]_0$ are exactly

$$\{1, 2, \dots, 2N-3\}.$$

For $A \subseteq \{1, \dots, 2N-3\}$, let H_A be the graph on $[N]_0$ in which distinct x, y are adjacent exactly when $x + y \in A$. Translation identifies H_A with G_B for

$$B = 2n + 2 + A \subseteq \{2n+3, \dots, 4n-3\} \subseteq J_n.$$

We will construct and estimate H_A and translate back only at the end. If $b = |A|$, then

$$\Delta(H_A) \leq b, \tag{2.1}$$

because each selected sum supplies at most one neighbor of any fixed vertex.

3 Universal lower bounds

We first record an elementary bound which makes the square-root scale visible.

Lemma 3.1 (Edge sums control chromatic number)**.** *Let a simple graph have distinct real labels on its vertices, and suppose that its edges use b distinct sums of endpoint labels. Then*

$$\chi(G) \leq \left\lfloor \frac{b+3}{2} \right\rfloor.$$

Proof. Put $k = \chi(G)$ and pass to a vertex-minimal induced subgraph K with chromatic number k. Then $\delta(K) \geq k - 1$. Let u and v be the least and greatest labels in K. The at least $k-1$ sums on edges incident with u are distinct and at most $u + v$. The at least $k-1$ sums on edges incident with v are distinct and at least $u + v$. The two collections intersect in at most the one value $u + v$. Hence $b \geq 2k - 3$, which is equivalent to the assertion. $\square$

Taking the largest color class in H_A, where $b = |A|$, gives

$$\alpha(H_A) \geq \left\lceil \frac{N}{\lfloor (b+3)/2 \rfloor} \right\rceil.$$

Optimizing this elementary expression yields, for $N \geq 4$,

$$b + \alpha(H_A) \geq \lceil \sqrt{8N} \rceil - 3.$$

The case $b = 0$ gives the value N and cannot minimize when $N \geq 4$ (the choice $A = \{N-1\}$ has $b = 1$ and gives $1 + \lceil N/2 \rceil \leq N$). For $b \geq 1$ the relevant values are $b = 2q - 3$, and

$$\min_{q \geq 2} \left(2q - 3 + \left\lceil \frac{N}{q} \right\rceil \right) = \lceil \sqrt{8N} \rceil - 3.$$

For completeness, put $t=\lceil\sqrt{8N}\rceil$. The inequality $2q+N/q\ge\sqrt{8N}$ gives the lower direction. Conversely,

$$\max_{q\in\mathbb{Z}} q(t-2q)=\left\lfloor\frac{t^2}{8}\right\rfloor\ge N.$$

For $N\ge 4$ a maximizing positive integer q can be chosen with $q\ge 2$; then $N/q\le t-2q$ and hence $2q+\lceil N/q\rceil\le t$. This proves the identity. This already proves $f(n)\ge 2\sqrt{2n}-O(1)$. We next gain a logarithmic factor.

Lemma 3.2 (Triangle sum triples)**.** *If H_A uses b selected sums and has T triangles, then*

$$T\le\binom{b}{3}.$$

Proof. A triangle with vertices x,y,z has the three pairwise distinct edge sums

$$x+y,\qquad x+z,\qquad y+z.$$

Their unordered set determines the triangle: if the three incident sums are denoted by s_{xy},s_{xz},s_{yz}, then

$$x=\frac{s_{xy}+s_{xz}-s_{yz}}{2},\quad y=\frac{s_{xy}+s_{yz}-s_{xz}}{2},\quad z=\frac{s_{xz}+s_{yz}-s_{xy}}{2}.$$

Equivalently, from an unordered triple $\{a,b,c\}$ one recovers the unordered vertex set

$$\left\{\frac{a+b-c}{2},\frac{a+c-b}{2},\frac{b+c-a}{2}\right\}.$$

Thus triangles inject into the three-element subsets of A. □

We use the following conservative form of a theorem of Alon, Krivelevich, and Sudakov [6, Theorem 1.1].

Theorem 3.3 (Sparse-neighborhood coloring theorem, safe form)**.** *There is an absolute $K>0$ with the following property. If a graph has maximum degree at most d and every vertex neighborhood spans at most d^2/F edges, where $2\le F\le d^2$, then*

$$\chi(G)\le K\frac{d}{\log F}.$$

The closed endpoint $F=d^2$ follows from the usual strict-range formulation by applying it with $F'=d^2/2$ and changing the absolute constant; bounded d is absorbed by greedy coloring. The primary statement calls d the maximum degree. If that convention is read as requiring equality rather than an upper bound, one may adjoin a disjoint star $K_{1,d}$: this makes the maximum degree exactly d, adds no edges to any neighborhood, and does not affect the coloring bound for the original component.

Proposition 3.4 (Universal lower bound)**.** *There is an absolute $c>0$ such that, for every sufficiently large N and every $A\subseteq\{1,\dots,2N-3\}$,*

$$|A|+\alpha(H_A)\ge c\sqrt{N\log N}.$$

Consequently $f(n)\gg\sqrt{n\log n}$.

Proof. Write $b=|A|$. For a vertex v, let t_v be the number of edges spanned by its neighborhood, equivalently the number of triangles containing v. Lemma 3.2 gives

$$\sum_v t_v=3T\le 3\binom{b}{3}\le\frac{b^3}{2}.$$

Therefore the set

$$U = \left\{ v : t_v \le \frac{b^3}{N} \right\}$$

has at least $N/2$ vertices. The induced graph $H = H_A[U]$ has maximum degree at most b, and every neighborhood in H spans at most b^3/N edges.

Suppose $2 \le b \le N/2$, and put

$$F_0 = \min \left\{ b^2, \frac{N}{b} \right\} \ge 2.$$

In either case $b^3/N \le b^2/F_0$. Applying Theorem 3.3 to H and taking a largest color class gives

$$\alpha(H_A) \ge c_0 \frac{N}{b} \log\left(\min \left\{ b^2, \frac{N}{b} \right\} \right) \tag{3.1}$$

for an absolute $c_0 > 0$.

Now split into three ranges. If $b \ge \sqrt{N \log N}$, the palette itself proves the result. If $b < N^{1/4}$, the maximum-degree bound (2.1), followed by greedy coloring, gives

$$\alpha(H_A) \ge \frac{N}{b+1} \gg \sqrt{N \log N}.$$

In the remaining range

$$N^{1/4} \le b < \sqrt{N \log N},$$

we have

$$\min \left\{ b^2, \frac{N}{b} \right\} \ge \sqrt{\frac{N}{\log N}},$$

and hence its logarithm is at least $(1/3) \log N$ for large N. Equation (3.1) now yields

$$b + \alpha(H_A) \ge b + c_1 \frac{N \log N}{b} \ge 2\sqrt{c_1 N \log N}.$$

The cases $b = 0, 1$ are covered by the maximum-degree estimate. The claim for f follows from Proposition 2.1. □

4 A linear seeded extractor over a growing odd field

The upper bound rests on the following finite-field statement. All logarithms used in min-entropy in this section are to base p.

Theorem 4.1 (Linear extractor family)**.** *Let p run through odd primes and let $r \to \infty$ in such a way that*

$$\left(\log(pr)\right)^2 \log(2 + \log(pr)) = o(r \log p).$$

Then there are integers $d, s = o(r)$ and an indexed family

$$\{F_y : \mathbb{F}_p^{2r} \to \mathbb{F}_p^r : y \in \mathcal{Y}\}, \qquad |\mathcal{Y}| \le p^d,$$

such that every F_y is surjective and $\mathbb{F}_p$-linear, and every random variable X on $\mathbb{F}_p^{2r}$ with

$$H_{\infty,p}(X) := -\log_p \max_x \Pr[X = x] \ge r + s$$

satisfies

$$\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} d_{\mathrm{TV}}(F_y(X), U_{\mathbb{F}_p^r}) < \frac{1}{3}.$$

More quantitatively, there is an auxiliary integer $\ell = O(\log(pr))$ for which one may take

$$d, s = O\left(\frac{\ell\sqrt{r} + \ell^2 \log(\ell+1)}{\log p} + \log_p(pr) + 1\right).$$

The family is indexed: repeated maps are allowed.

The construction is a mixed-alphabet adaptation of the linear form of Trevisan's extractor [8] and the weak-design reconstruction analysis of Raz, Reingold, and Vadhan [7]. The code and every fixed-seed output map are $\mathbb{F}_p$-linear, but code coordinates and seed coordinates are indexed by binary strings. This decoupling is useful: an overlap of t seed bits has 2^t possible inputs, independently of the growing output alphabet. We prove the required code, design, and reconstruction statements with this distinction explicit.

The proof proceeds in four stages. We first construct a short linear code, and then a binary weak design controlling its overlap moments. A hybrid argument and reconstruction count combine these objects into a strong linear extractor. Finally we choose the parameters and discard rank-deficient seeds; Section 5 turns the retained maps into a sum palette by taking the union of their kernels.

4.1 A short linear list-decodable code

We identify a word of length 2^ℓ with a function $\{0,1\}^\ell \to \mathbb{F}_p$. Agreement always means the fraction of coordinates on which two words have the same symbol.

Lemma 4.2 (Short binary-coordinate linear code)**.** *Let p be a prime, $m \geq 1$, and $0 < \eta < 1/2$. There exist a positive integer ℓ and an injective $\mathbb{F}_p$-linear map*

$$\mathcal{E} : \mathbb{F}_p^m \longrightarrow \mathbb{F}_p^{\{0,1\}^\ell}$$

such that, for every word $Q : \{0,1\}^\ell \to \mathbb{F}_p$,

$$\#\left\{x \in \mathbb{F}_p^m : \Pr_{z \sim U_{\{0,1\}^\ell}}[\mathcal{E}(x)(z) = Q(z)] > \frac{1}{p} + \eta\right\} \leq J := \frac{2}{\eta^2} + 1.$$

One can arrange

$$2^\ell \leq \frac{C_0 m \log p}{\eta^4} \tag{4.1}$$

for an absolute constant C_0.

Proof. Set

$$\tau = \frac{p\eta^2}{2(p-1)}, \qquad \Delta = 1 - \frac{1}{p} - \tau.$$

We first find a linear code of relative distance at least Δ. Write

$$H_p(a) = a\log_p(p-1) - a\log_p a - (1-a)\log_p(1-a)$$

for the p-ary entropy function, with the usual continuous extension at the endpoints. Then

$$1 - H_p(\Delta) = \frac{D(P \| U_{\mathbb{F}_p})}{\log p},$$

where

$$P = \left(\frac{1}{p} + \tau, \frac{1}{p} - \frac{\tau}{p-1}, \ldots, \frac{1}{p} - \frac{\tau}{p-1}\right).$$

The total variation distance between P and $U_{\mathbb{F}_p}$ is τ. Pinsker's inequality, with natural logarithms [9], gives

$$1 - H_p(\Delta) \geq \frac{2\tau^2}{\log p}.$$

Let $R = 2^\ell$ be the least power of 2 strictly exceeding

$$\frac{(m+1)\log p}{2\tau^2}.$$

The displayed quantity exceeds 1, so $\ell \geq 1$. Choose an $R \times m$ matrix T uniformly over $\mathbb{F}_p$. For every fixed nonzero x, the word Tx is uniform on $\mathbb{F}_p^R$. The standard p-ary Hamming-ball estimate [9] is exact in the form

$$\sum_{0 \leq j \leq aR} \binom{R}{j} (p-1)^j \leq p^{RH_p(a)} \qquad \left(0 < a \leq 1 - \frac{1}{p}\right).$$

Indeed, with $z = a/((p-1)(1-a)) \leq 1$, the left side is at most $z^{-aR}(1+(p-1)z)^R = p^{RH_p(a)}$. Taking $a = \Delta$ and using the preceding Pinsker bound show that

$$\Pr\left[\mathrm{wt}(Tx) < \Delta R\right] \leq p^{-R(1-H_p(\Delta))}.$$

A union bound over fewer than p^m nonzero x is strictly below one. We may therefore fix T for which every nonzero codeword has relative weight at least Δ. This also makes T injective. Index the R coordinates by $\{0,1\}^\ell$ and call the resulting encoder $\mathcal{E}$. The choice of R, the formula for τ, and rounding to the next power of 2 give, explicitly,

$$R \leq 2\frac{(m+1)\log p}{2\tau^2} \leq \frac{8m\log p}{\eta^4},$$

which proves (4.1).

It remains to prove the list bound. Represent the p symbols by unit vectors v_a forming a regular simplex:

$$\langle v_a, v_b \rangle = \begin{cases} 1, & a = b, \\ -1/(p-1), & a \neq b. \end{cases}$$

Represent a word by the normalized concatenation of its simplex vectors. Two words with agreement fraction a then have inner product

$$\frac{pa-1}{p-1}.$$

Suppose L codewords agree with Q on more than $1/p + \eta$ of the coordinates. Let their unit vectors be $u_1, \ldots, u_L$, and let w be the vector of Q. By this identity,

$$\langle u_i, w \rangle > \gamma := \frac{p\eta}{p-1}.$$

The minimum distance gives, for $i \neq j$,

$$\langle u_i, u_j \rangle \leq \beta := \frac{p\tau}{p-1} = \frac{\gamma^2}{2}.$$

Cauchy–Schwarz now implies

$$L^2\gamma^2 < \left\langle \sum_i u_i, w \right\rangle^2 \leq \left\| \sum_i u_i \right\|^2 \leq L + L(L-1)\beta.$$

Thus

$$L \leq \frac{1-\beta}{\gamma^2 - \beta} \leq \frac{2}{\gamma^2} \leq \frac{2}{\eta^2},$$

as required. $\square$

4.2 An ordered binary weak design

Lemma 4.3 (Suffix-slack design)**.** *For all integers $\ell, r \geq 1$, there are an integer D and ℓ-subsets $S_1, \dots, S_r \subseteq [D]$ such that, for every i,*

$$e_i := \sum_{j<i} \left(2^{|S_i \cap S_j|} - 1\right) \leq r - i, \tag{4.2}$$

and

$$D = O\left(\ell\sqrt{r} + \ell^2 \log(\ell + 1)\right).$$

The implied constant is absolute, and the sets and their ordering may be chosen deterministically from (r, ℓ).

Proof. Partition the r rows into consecutive blocks. If R_b rows remain before block b, put

$$m_b = \left\lceil \frac{R_b}{2} \right\rceil, \qquad R_{b+1} = \left\lfloor \frac{R_b}{2} \right\rfloor.$$

Thus $R_{b+1} \geq m_b - 1$. Give different blocks disjoint seed-coordinate universes.

For a block of size $m = m_b$, let q be the least prime at least

$$Q = \max\{2, \ell, \lceil \sqrt{m} \rceil\}.$$

Bertrand's postulate gives $q < 2Q$ [10]. Choose an ℓ-set $X \subseteq \mathbb{F}_q$ and use the universe $X \times \mathbb{F}_q$. There are $q^2 \geq m$ affine functions $x \mapsto ax + b$; select any m distinct ones and associate to (a, b) its graph

$$S_{a,b} = \{\, (x, ax + b) : x \in X \,\}.$$

Two distinct affine graphs meet in at most one point.

Let a row have zero-based local index t in its block and one-based global index i. Rows in earlier blocks have disjoint universes and contribute zero to the excess in (4.2). Each of the t earlier block-mates contributes at most $2^1 - 1 = 1$. Moreover, the exact number of global rows after this row is

$$r - i = (m_b - t - 1) + R_{b+1}.$$

Consequently

$$e_i \leq t \leq m_b - 1 \leq R_{b+1} \leq r - i.$$

This includes the final global row, for which the last block has size one and the excess is zero.

The universe of block b has size

$$\ell q = O\big(\ell \max\{\ell, \sqrt{m_b}\}\big).$$

The block sizes decrease geometrically, so $\sum_b \sqrt{m_b} = O(\sqrt{r})$. Once $m_b < \ell^2$, only $O(\log(\ell + 1))$ nonempty blocks remain, each using $O(\ell^2)$ coordinates. The disjoint union of the block universes therefore has the claimed size; identify it with $[D]$. Taking the least eligible prime and the first affine functions in lexicographic order makes the construction deterministic. □

The subtracted 1 in (4.2) isolates the unavoidable baseline: even disjoint sets contribute $2^0 = 1$ to the reconstruction count. Equivalently,

$$\sum_{j<i} 2^{|S_i \cap S_j|} \leq r - 1.$$

The remaining $r - i$ output coordinates will pay for the overlap excess at hybrid row i.

4.3 The reconstruction argument

Fix

$$\varepsilon = \frac{1}{20}, \qquad h = \frac{\varepsilon}{r(p-1)}, \qquad \eta = \frac{h}{2}. \tag{4.3}$$

Apply Lemma 4.2 with input dimension $2r$, and let $\mathcal{E}$ be the resulting encoder, with coordinate set $\{0,1\}^\ell$ and list size $J = 2/\eta^2 + 1$. Apply Lemma 4.3 with these values of ℓ and r. Put

$$d = \left\lceil D \log_p 2 \right\rceil,$$

so that the binary seed set has size $2^D \le p^d$. For $x \in \mathbb{F}_p^{2r}$ and $y \in \{0,1\}^D$, define

$$\mathrm{Ext}(x,y)_i = \mathcal{E}(x)(y|_{S_i}), \qquad 1 \le i \le r. \tag{4.4}$$

The coordinates of each S_i are read in increasing order.

Proposition 4.4 (Extractor inequality)**.** *Suppose the integer s satisfies*

$$s \ge d + \left\lceil \log_p \frac{4J}{h} \right\rceil + 2, \qquad r + s \le 2r. \tag{4.5}$$

If X is any random variable on $\mathbb{F}_p^{2r}$ with $H_{\infty,p}(X) \ge r + s$, and Y is uniform on $\{0,1\}^D$ independently of X, then

$$d_{\mathrm{TV}}\big((Y, \mathrm{Ext}(X,Y)), (Y, U_{\mathbb{F}_p^r})\big) \le \varepsilon. \tag{4.6}$$

Here the copy of $U_{\mathbb{F}_p^r}$ on the right is independent of Y. For each fixed seed y, the map $x \mapsto \mathrm{Ext}(x,y)$ is $\mathbb{F}_p$-linear with codomain $\mathbb{F}_p^r$.

Proof. Write $Z = \mathrm{Ext}(X,Y) = (Z_1, \dots, Z_r)$ and suppose that (4.6) fails. Define the hybrids

$$\mathsf{H}_i = (Y, Z_1, \dots, Z_i, U_{i+1}, \dots, U_r), \qquad 0 \le i \le r,$$

where the displayed U_j are independent uniform symbols, also independent of (X,Y,Z). The triangle inequality gives an i such that

$$\mathbb{E}_{(Y,Z_{<i})} d_{\mathrm{TV}}\big(\mathcal{L}(Z_i \mid Y, Z_{<i}), U_{\mathbb{F}_p}\big) > \frac{\varepsilon}{r}.$$

For a distribution μ on $\mathbb{F}_p$,

$$\max_a \mu(a) - \frac{1}{p} \ge \frac{d_{\mathrm{TV}}(\mu, U_{\mathbb{F}_p})}{p-1},$$

because the total positive deviation is the total variation distance and has at most $p-1$ positive summands. At every context choose a most likely symbol. The preceding inequality produces a deterministic predictor

$$\mathcal{P} : \{0,1\}^D \times \mathbb{F}_p^{i-1} \to \mathbb{F}_p$$

such that

$$\Pr[\mathcal{P}(Y, Z_{<i}) = Z_i] > \frac{1}{p} + h.$$

The index i and the predictor $\mathcal{P}$ may depend on the fixed law of X, but from now on they are fixed and do not depend on the point x.

For fixed x, let q_x be the success probability in the preceding predictor bound over Y, with $Z = \mathrm{Ext}(x,Y)$, and put

$$\mathcal{B} = \left\{ x : q_x > \frac{1}{p} + \frac{h}{2} \right\}.$$

Since $q_x \leq 1$, averaging that predictor bound gives

$$\Pr[X \in \mathcal{B}] > \frac{h}{2}. \tag{4.7}$$

Indeed, if $\theta = \Pr[X \in \mathcal{B}]$, then

$$\frac{1}{p} + h < \mathbb{E}q_X \leq \frac{1}{p} + \frac{h}{2} + \theta\left(1 - \frac{1}{p} - \frac{h}{2}\right),$$

which implies $\theta > h/2$.

We now count $\mathcal{B}$. Fix $x \in \mathcal{B}$. Regard a binary assignment on S_i as a code coordinate in $\{0,1\}^\ell$, using the increasing order on S_i. By averaging over the seed coordinates outside S_i, we obtain a fixing $a \in \{0,1\}^{[D]\setminus S_i}$ for which the predictor agrees with the word $\mathcal{E}(x)$ on more than a $1/p + h/2 = 1/p + \eta$ fraction of assignments $z \in \{0,1\}^{S_i}$.

After this outside fixing, the earlier output $\mathcal{E}(x)(y|_{S_j})$ is a function of only the $|S_i \cap S_j|$ overlap coordinates of z. We may enumerate every such function: a function $\{0,1\}^t \to \mathbb{F}_p$ has p^{2^t} possibilities. Thus the number of descriptions consisting of the outside fixing and all earlier-output functions is at most

$$\begin{aligned} 2^{D-\ell} \prod_{j<i} p^{2^{|S_i \cap S_j|}} &\leq p^{d + \sum_{j<i} 2^{|S_i \cap S_j|}} \\ &= p^{d+(i-1)+e_i} \leq p^{d+r-1}. \end{aligned}$$

For a description $\mathcal{D} = (a, (g_j)_{j<i})$ and an assignment $z \in \{0,1\}^{S_i}$, let $y(a,z)$ be the unique seed whose restrictions to S_i and $[D] \setminus S_i$ are respectively z and a, and define

$$Q_{\mathcal{D}}(z) = \mathcal{P}(y(a,z), (g_j(z|_{S_i \cap S_j}))_{j<i}).$$

Via the increasing-order identification $\{0,1\}^{S_i} \cong \{0,1\}^\ell$, this is one received word $Q_{\mathcal{D}} : \{0,1\}^\ell \to \mathbb{F}_p$. For each $x \in \mathcal{B}$, the description obtained above gives a word $Q_{\mathcal{D}}$ which agrees with $\mathcal{E}(x)$ on more than a $1/p + \eta$ fraction of coordinates. Thus x belongs to the agreement list of at least one counted description. Lemma 4.2 bounds every such list by J, so

$$|\mathcal{B}| \leq Jp^{d+r-1}.$$

The min-entropy assumption now gives

$$\begin{aligned} \Pr[X \in \mathcal{B}] &\leq p^{-r-s}|\mathcal{B}| \\ &\leq Jp^{d-s-1} < \frac{h}{4}, \end{aligned}$$

where (4.5) gives explicitly

$$Jp^{d-s-1} \leq Jp^{-\lceil \log_p(4J/h) \rceil - 3} \leq \frac{h}{4p^3} < \frac{h}{4}.$$

This contradicts (4.7), proving (4.6).

Finally, for fixed y, every coordinate in (4.4) is one coordinate of the $\mathbb{F}_p$-linear encoding $\mathcal{E}$. The tuple of the r selected coordinates is therefore $\mathbb{F}_p$-linear. Surjectivity is not asserted yet; it is enforced below. $\square$

4.4 Parameters and surjectivity

From Lemma 4.2, (4.3), and the input dimension $2r$,

$$2^\ell = O\big(r \log p\,(pr)^4\big).$$

Therefore

$$\ell = O(\log(pr)). \tag{4.8}$$

Lemma 4.3 gives

$$D = O\left(\ell\sqrt{r} + \ell^2 \log(\ell+1)\right)$$

and therefore

$$d = O\left(\frac{\ell\sqrt{r} + \ell^2\log(\ell+1)}{\log p} + 1\right).$$

Also

$$\log_p(4J/h) = O(1 + \log_p(pr)),$$

so (4.5) holds with

$$s = O\left(\frac{\ell\sqrt{r} + \ell^2\log(\ell+1)}{\log p} + \log_p(pr) + 1\right).$$

The growth hypothesis in Theorem 4.1, together with (4.8), makes both d and s equal to $o(r)$. Indeed, if $u = \log(pr)$ and $\lambda = \log p$, then

$$\left(\frac{u}{\lambda\sqrt{r}}\right)^2 = \frac{u^2}{r\lambda^2} = o(1), \qquad \frac{u^2\log(2+u)}{r\lambda} = o(1),$$

while $\log_p(pr)/r = o(1)$. These three estimates control respectively the $\ell\sqrt{r}$, $\ell^2\log(\ell+1)$, and remaining terms above. For the first estimate, use the second one together with $\lambda \geq \log 3$ and $\log(2+u) > 1$. This verifies all parameter claims in Theorem 4.1, except surjectivity.

The full set of 2^D binary seeds is an indexed family of linear maps. Apply Proposition 4.4 to the source $U_{\mathbb{F}_p^{2r}}$. For any source X, the joint distance in (4.6) is exactly the fixed-seed average

$$d_{\mathrm{TV}}\big((Y, \mathrm{Ext}(X,Y)), (Y, U_{\mathbb{F}_p^r})\big) = \frac{1}{2^D} \sum_{y\in\{0,1\}^D} d_{\mathrm{TV}}(\mathrm{Ext}(X,y), U_{\mathbb{F}_p^r}).$$

If a seed map has rank $t < r$, its output is uniform on a p^t-element subspace, and therefore

$$d_{\mathrm{TV}}(\mathrm{Ext}(U_{\mathbb{F}_p^{2r}}, y), U_{\mathbb{F}_p^r}) = 1 - p^{t-r} \geq 1 - \frac{1}{p}.$$

If b is the fraction of rank-deficient seeds, (4.6) gives

$$b \leq \frac{\varepsilon}{1 - 1/p} \leq \frac{3}{40}.$$

Retain the good seeds, with their original multiplicities, and put the uniform distribution on this retained index set $\mathcal{Y}$. For each $y \in \mathcal{Y}$, set

$$F_y(x) = \mathrm{Ext}(x, y).$$

Then $|\mathcal{Y}| \leq p^d$, every F_y is surjective, and for every source covered by Proposition 4.4,

$$\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{Y}} d_{\mathrm{TV}}(F_y(X), U_{\mathbb{F}_p^r}) \leq \frac{\varepsilon}{37/40} = \frac{2}{37} < \frac{1}{3}.$$

This completes the proof of Theorem 4.1.

5 From the extractor to a finite-field sum palette

Put

$$G = \mathbb{F}_p^{2r}, \qquad V = \mathbb{F}_p^r,$$

and let the maps $F_y : G \to V$ be those of Theorem 4.1. Define

$$S = \bigcup_{y \in \mathcal{Y}} \ker F_y \subseteq G.$$

Let Γ_S be the simple graph on G in which distinct x, x' are adjacent exactly when $x + x' \in S$.

Proposition 5.1 (Group palette)**.** *With $d, s = o(r)$ as above,*

$$|S| \le p^{r+d}, \qquad \alpha(\Gamma_S) < p^{r+s}.$$

Proof. Every retained map is surjective, so every kernel has size p^r. The definition of S therefore gives

$$|S| \le |\mathcal{Y}| p^r \le p^{r+d}.$$

Suppose that $A \subseteq G$ is independent. Fix $y \in \mathcal{Y}$. If $z \neq 0$ and both z and $-z$ occur in $F_y(A)$, choose $a, a' \in A$ with $F_y(a) = z$ and $F_y(a') = -z$. Since p is odd, $z \neq -z$, so $a \neq a'$; but then $a + a' \in \ker F_y \subseteq S$, a contradiction. The zero fibre contains at most one member of A, because two distinct members of that fibre would also have their sum in $\ker F_y$. The nonzero elements of V form $(p^r - 1)/2$ antipodal pairs. Hence

$$|F_y(A)| \le \frac{p^r + 1}{2}.$$

Any distribution P supported on $T \subseteq V$ satisfies

$$d_{\mathrm{TV}}(P, U_V) \ge 1 - \frac{|T|}{|V|},$$

by testing the event T. If $|A| \ge p^{r+s}$ and X is uniform on A, then $H_{\infty,p}(X) \ge r + s$, while the preceding support bound gives, for every retained seed,

$$d_{\mathrm{TV}}(F_y(X), U_V) \ge \frac{p^r - 1}{2p^r} \ge \frac{1}{3}.$$

This contradicts Theorem 4.1. Therefore $|A| < p^{r+s}$. $\square$

The zero-fibre argument is where the absence of loops must be handled carefully: it forbids two distinct zero-fibre points, but it does not forbid a single point a merely because $2a \in S$.

6 Carry closure and ordinary integer sums

Set

$$M = p^{2r}$$

and identify $[M]_0$ with $\mathbb{F}_p^{2r}$ by the base-p digit bijection

$$\psi(x) = (x_0, \ldots, x_{2r-1}), \qquad x = \sum_{i=0}^{2r-1} x_i p^i.$$

For $u \in \mathbb{F}_p^{2r}$, define its ordinary carry fibre by

$$\mathcal{C}(u) = \{x + x' : 0 \le x, x' < M,\ \psi(x) + \psi(x') = u\}, \tag{6.1}$$

where addition inside $\psi(x) + \psi(x')$ is coordinatewise modulo p.

Lemma 6.1 (Carry count)**.** *For every* $u \in \mathbb{F}_p^{2r}$*,*

$$|\mathcal{C}(u)| \le 2^{2r}.$$

Proof. In the ordinary addition of x and x', let $c_i \in \{0,1\}$ be the carry entering digit i, with $c_0 = 0$, and let c_{2r} be the final carry. Once u and the vector

$$(c_1, \dots, c_{2r}) \in \{0,1\}^{2r}$$

are fixed, the ith ordinary output digit is forced to be $(u_i + c_i) \bmod p$, and the final digit is c_{2r}. Thus the resulting integer sum is uniquely determined. Some carry vectors may be inconsistent with any pair of summands, which only reduces the count. □

Define the integer palette

$$B_M = \bigcup_{u \in S} \mathcal{C}(u) \subseteq \{0, \dots, 2M-2\}.$$

Lemma 6.1 and Proposition 5.1 give

$$|B_M| \le 2^{2r} p^{r+d}.$$

Let K_{B_M} be the simple graph on $[M]_0$ in which distinct integers x, x' are adjacent when $x + x' \in B_M$. If x, x' form an edge of Γ_S under the digit bijection, then $\psi(x) + \psi(x') \in S$, so the carry-fibre definition (6.1) and the definition of B_M place their ordinary sum in B_M. Therefore K_{B_M} contains the group graph and

$$\alpha(K_{B_M}) \le p^{r+s}.$$

No carry has been discarded, and only pairs of distinct vertices are used in this graph comparison.

7 Parameters for every interval length

We now choose the field and dimension separately for each N. This also gives the quantitative error term in Theorem 1.1. Let N be sufficiently large, put

$$L = \log N, \qquad \lambda_0 = \left(\frac{L}{\log L}\right)^{1/3}, \qquad P = \lceil \exp(\lambda_0) \rceil,$$

and choose a prime $p \in [P, 2P]$, which exists by Bertrand's postulate [10]. Set

$$r = \left\lceil \frac{\log N}{2 \log p} \right\rceil, \qquad M = p^{2r}. \tag{7.1}$$

Then p is odd and tends to infinity, and

$$\lambda := \log p = \lambda_0 + O(1), \qquad r = \Theta(L/\lambda).$$

Since $\log r = O(\log L) = o(\lambda)$, one has $\log(pr) = O(\lambda)$, and (4.8) gives $\ell = O(\lambda)$. Moreover,

$$\big(\log(pr)\big)^2 \log(2 + \log(pr)) = O(\lambda^2 \log \lambda) = o(L),$$

whereas $r \log p = \Theta(L)$. Thus the hypothesis of Theorem 4.1 holds with room to spare. The design construction and the quantitative bounds in that theorem give

$$\begin{aligned} D &= O\left(\lambda\sqrt{L/\lambda} + \lambda^2 \log \lambda\right) = O\left(L^{2/3} (\log L)^{1/3}\right), \\ d, s &= O\left(L^{1/3} (\log L)^{2/3}\right), \\ \frac{d}{r}, \frac{s}{r} &= O\left(\left(\frac{\log L}{L}\right)^{1/3}\right). \end{aligned}$$

Also, by the ceiling in (7.1),

$$N \le M < Np^2.$$

Build the M-vertex palette B_M above and restrict to the first N vertices. Delete the normalized sums which cannot come from two distinct members of $[N]_0$ by putting

$$B_N = B_M \cap \{1, \ldots, 2N-3\}. \tag{7.2}$$

Let Γ_N be the subgraph of Γ_S induced by $[N]_0$. Every edge of Γ_N has ordinary sum in $B_M \cap \{1, \ldots, 2N-3\} = B_N$, so

$$\Gamma_N \subseteq H_{B_N}.$$

Therefore

$$\alpha(H_{B_N}) \le \alpha(\Gamma_N) \le \alpha(\Gamma_S) < p^{r+s}.$$

Together with the palette-size bound, this gives

$$|B_N| \le 2^{2r} p^{r+d}, \qquad \alpha(H_{B_N}) \le p^{r+s}.$$

We estimate these expressions relative to N. Put $\lambda = \log p$. By (7.1), $r = L/(2\lambda) + O(1)$. At these parameters the bounds for d and s give $d\lambda, s\lambda = O(D + \log(pr)) = O(D + \lambda)$, so

$$\begin{aligned}
\frac{\log(2^{2r} p^{r+d})}{L} &= \frac{2r \log 2}{L} + \frac{(r+d)\lambda}{L} \le \frac{1}{2} + O\left(\frac{1}{\lambda} + \frac{D}{L} + \frac{\lambda}{L}\right), \\
\frac{\log(p^{r+s})}{L} &\le \frac{1}{2} + O\left(\frac{D}{L} + \frac{\lambda}{L}\right).
\end{aligned}$$

Here the carry factor contributes $2r \log 2/L = (\log 2)/\lambda + O(L^{-1})$, and the ceiling in (7.1) contributes $O(\lambda/L)$. By the choices above,

$$\frac{1}{\lambda}, \frac{D}{L}, \frac{\lambda}{L} = O\left(\left(\frac{\log L}{L}\right)^{1/3}\right).$$

Therefore, for an absolute constant C_1 and all sufficiently large N,

$$\begin{aligned}
|B_N| &\le N^{\frac{1}{2} + C_1\left(\frac{\log\log N}{\log N}\right)^{1/3}}, \\
\alpha(H_{B_N}) &\le N^{\frac{1}{2} + C_1\left(\frac{\log\log N}{\log N}\right)^{1/3}}.
\end{aligned} \tag{7.3}$$

8 Completion of the proof

Take $N = n - 1$ and the normalized palette B_N from (7.2). Select the original sums

$$B = \{\, 2n + 2 + a : a \in B_N \,\}.$$

Because $1 \le a \le 2N - 3 = 2n - 5$, this is a subset of

$$\{2n+3, \ldots, 4n-3\} \subseteq J_n.$$

The translation $x \mapsto n + 1 + x$ identifies the graph generated by B_N on $[N]_0$ with G_B on I_n. Hence (7.3) gives

$$|B| + \alpha(G_B) \le n^{\frac{1}{2} + O\left(\left(\frac{\log\log n}{\log n}\right)^{1/3}\right)}.$$

The factor 2 from adding the two bounds contributes only

$$\frac{\log 2}{\log n} = o\left(\left(\frac{\log\log n}{\log n}\right)^{1/3}\right)$$

to the exponent. Replacing the scale $N = n-1$ by n changes the exponent by a smaller amount. These errors are absorbed in the displayed term. By Proposition 2.1, this is the desired upper bound for $f(n)$. Proposition 3.4 supplies the lower bound, and Theorem 1.1 follows.

For clarity about the quantifiers, the construction above chooses one set B for each n. The preceding estimate for $|B| + \alpha(G_B)$ therefore holds with $\alpha(G_B)$ replaced by $|C|$ for *every* B-admissible C. Since the entire exponent correction displayed above tends to zero, it is smaller than ε for sufficiently large n. This proves the stated assertion for every fixed $\varepsilon > 0$ and every sufficiently large n, not merely along a subsequence.

9 Concluding remarks

The proof separates three interfaces that are easy to conflate. The extractor has p-ary linear outputs but binary code coordinates and a binary seed; the weak-design overlap moment is therefore 2^t, not p^t. Ordering the design sets exposes a further resource: at hybrid row i, the remaining $r-i$ output coordinates pay for the excess above the unavoidable baseline overlap. The union of the resulting kernels controls finite-field sum graphs, and the carry closure in Section 6 then transfers that control to ordinary integer addition. Keeping these interfaces explicit yields the quantitative exponent correction $O((\log\log n/\log n)^{1/3})$.

The argument determines the exponent of $f(n)$ but not its finer order of growth. The universal lower bound is $\Omega(\sqrt{n\log n})$, whereas the constructed upper bound is

$$\sqrt{n}\,\exp\bigl(O((\log n)^{2/3}(\log\log n)^{1/3})\bigr).$$

Closing this remaining subpolynomial gap would require a sharper palette, a stronger universal lower bound, or both.

References

[1] T. F. Bloom, Erdős Problem #788, https://www.erdosproblems.com/788, accessed July 18, 2026.

[2] P. Erdős, Problems and results on combinatorial number theory, in *A Survey of Combinatorial Theory* (J. N. Srivastava et al., eds.; Proc. Internat. Sympos., Colorado State Univ., Fort Collins, 1971), North-Holland, 1973, pp. 117–138, doi:10.1016/B978-0-7204-2262-7.50017-X.

[3] S. L. G. Choi, On a combinatorial problem in number theory, *Proc. London Math. Soc.* (3) **23** (1971), 629–642, doi:10.1112/plms/s3-23.4.629.

[4] A. Baltz, T. Schoen, and A. Srivastav, Probabilistic construction of small strongly sum-free sets via large Sidon sets, *Colloq. Math.* **86**, no. 2 (2000), 171–176, doi:10.4064/cm-86-2-171-176.

[5] N. Alon and H. T. Pham, Random Cayley graphs and random sumsets, arXiv:2509.02561, 2025, https://arxiv.org/abs/2509.02561.

[6] N. Alon, M. Krivelevich, and B. Sudakov, Coloring graphs with sparse neighborhoods, *J. Combin. Theory Ser. B* **77** (1999), 73–82, doi:10.1006/jctb.1999.1910.

[7] R. Raz, O. Reingold, and S. Vadhan, Extracting all the randomness and reducing the error in Trevisan's extractors, *J. Comput. System Sci.* **65** (2002), 97–128, doi:10.1006/jcss.2002.1824.

[8] L. Trevisan, Extractors and pseudorandom generators, *J. ACM* **48** (2001), 860–879, doi:10.1145/502090.502099.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2006.

[10] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 6th ed., revised by D. R. Heath-Brown and J. H. Silverman, Oxford University Press, Oxford, 2008.