Electronic Colloquium on Computational Complexity, Report No. 148 (2026)

An Output-Size-Optimal Algorithmic Balog–Szemerédi–Gowers Theorem

Zhao Song*

August 5, 2026

Abstract

We give a randomized algorithmic version of the Balog–Szemerédi–Gowers theorem for sets of integers. Let $A \subseteq [N]$ have size $n := |A| \geq 2$, let $1 \leq K \leq n$, and suppose that its additive energy satisfies $E(A) \geq n^3/K$. Reiher and Schoen proved existentially that, for every fixed $\epsilon \in (0, 1/2)$, there is a subset $A' \subseteq A$ satisfying $|A'| \geq (1-\epsilon)n/\sqrt{K}$ and $|A' - A'| \leq O_\epsilon(K^4)|A'|$; the size scale $n/\sqrt{K}$ is essentially optimal. We give an algorithmic counterpart: with probability at least $1 - n^{-10}$, our algorithm runs in time $nKN^{o(1)}$ and returns a subset $A' \subseteq A$ satisfying $|A'| \geq cn/\sqrt{K}$ and $|A' - A'| \leq CK^4|A'|$, where $c, C > 0$ are absolute constants. Thus the output attains the essentially optimal subset-size scale and the best-known K^4 dependence for the normalized difference set.

1 Introduction

The Balog–Szemerédi–Gowers principle [BS94, Gow98] converts high additive energy into a large subset with controlled additive growth. For a set $A \subseteq [N]$ of size n, its additive energy $E(A)$ is the number of ordered quadruples $(a_1, a_2, a_3, a_4) \in A^4$ satisfying $a_1 + a_2 = a_3 + a_4$; thus $2n^2 - n \leq E(A) \leq n^3$. Balog and Szemerédi [BS94] first proved that the hypothesis $E(A) \geq n^3/K$ forces a large subset with small additive growth. Gowers [Gow98, Proposition 12] replaced the regularity-based quantitative dependence by polynomial bounds. Tracking the constants in his proof under $E(A) \geq n^3/K$ gives a subset $A' \subseteq A$ with $|A'| \geq n/(2^{19}K^{12})$ and $|A' - A'| \leq 2^{38}nK^{24}$; equivalently, $|A' - A'| = O(K^{36}|A'|)$. Reiher and Schoen [RS24, Theorem 1.2] later proved that, for every fixed $\epsilon \in (0, 1/2)$, one can simultaneously achieve $|A'| \geq (1-\epsilon)n/\sqrt{K}$ and $|A' - A'| \leq O_\epsilon(K^4)|A'|$. Their $n/\sqrt{K}$ size scale is essentially optimal in order of magnitude. The coordinate-axis example of Reiher and Schoen [RS24] is stated over finite abelian groups and gives the sharp leading constant there. For an integer analogue, fix r and $M \gg r$, and let $A_0 \subseteq \mathbb{Z}^r$ be the union of the coordinate intervals $\{te_i : 1 \leq t \leq M\}$, where e_i is the ith standard basis vector. Then $|A_0| = rM$ and $E(A_0) = \Theta(rM^3)$, so $K = \Theta(r^2)$, while, for every fixed $\delta > 0$, every subset of more than $(1+\delta)M$ elements has $\Omega_\delta(M^2)$ distinct cross-axis differences. A sufficiently large-base encoding of the containing box into $\mathbb{Z}$ is a Freiman 2-isomorphism, so it preserves equality among differences and hence preserves both additive energy and difference-set cardinalities. Thus, for integer sets, $n/\sqrt{K}$ is the correct order of magnitude for the largest subset A' satisfying $|A' - A'| = O_K(|A'|)$; this transfer does not preserve the sharp leading constant 1. The exponent 4 is best known but is not proved optimal: the examples discussed in [RS24] only rule out replacing it by an exponent smaller than $\log(4)/\log(27/16) \approx 2.649$.

*magic.linuxkde@gmail.com. The author would like to thank Ce Jin for very helpful discussions.

ISSN 1433-8092

On the algorithmic side, Jin and Xu [JX24, Lemma 3.2] give a randomized algorithm running in $O(nK^{28}\operatorname{poly}\log N)$ time which returns a set $A' \subseteq A$ with $|A'| \geq \Omega(n/K^2)$ and $|A' + A'| \leq O(nK^{28}) = O(K^{30}|A'|)$. Independently, Abboud, Bringmann, and Fischer [ABF23, Theorem 3.3] give a randomized algorithm running in $\widetilde{O}(nK^{12})$ time and returning a set of size $\Omega(n/K^2)$ with $|A'+A'| \leq O(nK^{24}) = O(K^{26}|A'|)$. Fischer, Jin, and Xu [FJX25, Theorem 1.11] obtain a deterministic algorithm running in $(n^2/K^2 + K^{13}n)N^{o(1)}$ time and returning a set of size at least $n/(64K)$ with $|A' + A'| \leq K^{16}nN^{o(1)} = O(K^{17}N^{o(1)})|A'|$. Thus, prior algorithms either have a large polynomial dependence on K, return a set below the existential $n/\sqrt{K}$ scale, or incur a substantially weaker normalized additive bound.

It is natural to ask the following question.

Is it possible to design an $nKN^{o(1)}$-time algorithm that returns a set $A' \subseteq A$ with $|A'| \geq \Omega(n/\sqrt{K})$ and $|A' - A'| \leq O(K^4)|A'|$?

In this paper, we give a positive answer to this question.

1.1 Our results

We now state our main result in compact form, followed by a brief explanation of its proof and its comparison with the closest previous bounds. The complete statements and proofs appear later in the paper.

Theorem 1.1 (Informal version of Theorem 5.1)**.** *Let $A \subseteq [N]$ be a set of $n \geq 2$ integers, let $1 \leq K \leq n$, and suppose that $E(A) \geq n^3/K$. There is a randomized algorithm which, with probability at least $1 - n^{-10}$, runs in time $nKN^{o(1)}$ and returns a set $A' \subseteq A$ such that*

$$|A'| \geq \Omega(n/\sqrt{K}) \qquad \text{and} \qquad |A' - A'| \leq O(K^4)|A'|.$$

We first use the deterministic popular-sums approximation of Lemma 3.3 with additive error $\Theta(n/K)$. Discarding differences below a constant multiple of this error loses only a small constant fraction of the energy. We then split the retained differences once at multiplicity $2n/\sqrt{K}$. If the high-multiplicity part carries a constant fraction of the energy, weighted sampling over all of its differences finds an energetic translate A_d of size $\Omega(n/\sqrt{K})$, and the cleaning argument gives the stronger normalized difference bound $O(K^3)$. Otherwise the low-multiplicity part carries a constant fraction of the energy. The light-tail selection argument in the proof of Theorem 5.1 combines its multiplicity levels into a thresholded tail whose relation has density $\Omega(1/\sqrt{K})$ and whose support is small enough for the four-difference path count. Lemma 4.1 then returns $\Omega(n/\sqrt{K})$ elements, and the path count gives $|A' - A'| \leq O(K^4)|A'|$. Sorting the retained multiplicities and carrying out the randomized cleaning steps fit within $nKN^{o(1)}$ time.

Thus Theorem 1.1 reaches, up to absolute constants, the essentially optimal existential size scale and the best-known K^4 dependence for the normalized difference set. Corollary 5.2 gives the corresponding bound $|A' + A'| \leq CK^8|A'|$. Since the theorem is randomized and one-set, it does not replace the asymmetric deterministic BSG theorem required in some applications.

Theorem 1.2 (Informal version of Corollary 5.2)**.** *Let $A \subseteq [N]$ be a set of $n \geq 2$ integers, let $1 \leq K \leq n$, and suppose that $E(A) \geq n^3/K$. There is a randomized algorithm which, with probability at least $1 - n^{-10}$, runs in time $nKN^{o(1)}$ and returns a set $A' \subseteq A$ such that*

$$|A'| \geq \Omega(n/\sqrt{K}) \qquad \text{and} \qquad |A' + A'| \leq O(K^8)|A'|.$$

Result	Refs	R/D	Running time	Gua. size	Sum/difference
Prop. 12	[Gow98]	–	–	n/K^{12}	$\lvert A'-A'\rvert \le \lvert A'\rvert K^{36}$
Thm. 1.2	[RS24]	–	–	$n/\sqrt{K}$	$\lvert A'-A'\rvert \le \lvert A'\rvert K^{4}$
Thm. 8	[Blo25]	–	–	$n/K^{1.24}$	$\lvert A'-A'\rvert \le \lvert A'\rvert K^{3.71}$
Lem. 3.2	[JX24]	Rand.	nK^{28}	n/K^2	$\lvert A'+A'\rvert \le \lvert A'\rvert K^{30}$
Thm. 3.3	[ABF23]	Rand.	nK^{12}	n/K^2	$\lvert A'+A'\rvert \le \lvert A'\rvert K^{26}$
Thm. 1.11	[FJX25]	Det.	$nK^{13}+n^2/K^2$	n/K	$\lvert A'+A'\rvert \le \lvert A'\rvert K^{17}N^{o(1)}$
Thm. 5.1	This work	Rand.	nK	$n/\sqrt{K}$	$\lvert A'-A'\rvert \le \lvert A'\rvert K^{4}$
Cor. 5.2	This work	Rand.	nK	$n/\sqrt{K}$	$\lvert A'+A'\rvert \le \lvert A'\rvert K^{8}$

Table 1: Comparison of existential and algorithmic BSG bounds, including Theorem 5.1. All rows take $A \subseteq [N] := \{1,\dots,N\}$, $n := |A| \ge 2$, $1 \le K \le n$, and $E(A) \ge n^3/K$, where $E(A)$ is the additive energy of A; $A' \subseteq A$ is the output. Here *Prop.*, *Thm.*, *Lem.*, *Cor.*, *Refs*, and *Gua.* abbreviate proposition, theorem, lemma, corollary, references, and guaranteed, respectively. The heading *R/D* records whether an algorithm is randomized or deterministic, abbreviated *Rand.* and *Det.*; dashes denote a nonalgorithmic result. The Gowers row records the quantitative bounds obtained at the end of the proof of Proposition 12, with absolute constant factors suppressed. The sets $A'+A'$ and $A'-A'$ are the sumset and difference set. The running-time column displays only the dependence on n and K: constant factors, $O(\cdot)$ or $\widetilde{O}(\cdot)$ notation, polylogarithmic factors, and $N^{o(1)}$ factors are suppressed there. The exponent $o(1)$ in the $N^{o(1)}$ factor retained in the last column tends to zero as $N \to \infty$; the second-to-last column suppresses $\Omega(\cdot)$, and every entry in the last column is written as $|A'|$ times a normalized factor and suppresses $O(\cdot)$; when a cited result is originally stated in terms of n, we use the guaranteed-size column to make this conversion. Corollary 5.2 follows from Lemma 3.2 and Theorem 5.1.

2 Technique Overview

Table 1 summarizes the comparison. This section explains the improvement of Theorem 5.1 and Corollary 5.2 over the algorithmic BSG results of Jin and Xu [JX24], Abboud, Bringmann, and Fischer [ABF23], and Fischer, Jin, and Xu [FJX25] along three directions: the running time, the guaranteed output size, and the additive bound. Our output-size and normalized additive guarantees improve the dependence on K in the prior algorithmic results. For running time, we compare the displayed dependence on n and K after suppressing logarithmic and $N^{o(1)}$ factors; without this convention, an unspecified $N^{o(1)}$ factor does not permit a uniform comparison when K is only polylogarithmic. The comparison with the deterministic result of [FJX25] is instead a tradeoff between running time and randomization.

Running-time comparison. Our algorithm runs in time $nKN^{o(1)}$, with linear dependence on n and K up to the $N^{o(1)}$ factor; the comparison bounds are $O(nK^{28}\operatorname{poly}\log N)$ obtained by tracking the proof of [JX24, Lemma 3.2], $\widetilde{O}(nK^{12})$ for the algorithm of [ABF23], and $(n^2/K^2+K^{13}n)N^{o(1)}$ for the algorithm of [FJX25]. Strictly speaking, Lemma 3.2 of Jin and Xu [JX24] assumes $E(A) > n^3/K$. Under our convention $E(A) \ge n^3/K$, apply it with $K' := \min\{2K, n\}$. If $K < n$, then $K' > K$ and hence $E(A) > n^3/K'$; if $K = n$, then $E(A) \ge 2n^2 - n > n^2 = n^3/K'$. Since $K' \le 2K$, this changes every displayed bound only by an absolute constant. The improvement has three

sources.

First, the only step of our algorithm that examines the global additive structure of A is the deterministic popular-sums approximation of Lemma 3.3, invoked once at accuracy $\epsilon := \Theta(1/K)$; its cost $\epsilon^{-1} n N^{o(1)} = nKN^{o(1)}$ is the main bottleneck. Thresholding the approximation leaves $O(nK)$ relevant differences, so sorting them and scanning their prefix sums costs only $nKN^{o(1)}$ time.

Second, in the low-multiplicity branch, the light-tail selection argument in the proof of Theorem 5.1 constructs a relation whose density ρ is known within a constant factor and satisfies $\rho \geq c/\sqrt{K}$. The codegree classifier, pivot search, and vertex cleaning of Lemma 4.1 together cost $\widetilde{O}(n\rho^{-2} + \rho^{-4}) = \widetilde{O}(nK + K^2) = \widetilde{O}(nK)$, where the last step uses $K \leq n$. The high-multiplicity branch instead constructs an energetic translate A_d directly and does not invoke sampled relation cleaning. In the applications underlying [JX24, Lemma 3.2] and [ABF23, Theorem 3.3], the Chan–Lewenstein graph lemma [CL15, Lemmas 7.1 and 7.2] is applied to a relation of density at least c/K^2, using the lower-bound parameter $\alpha := c/K^2 = \Theta(K^{-2})$. The implementation stated in [JX24, Lemma 3.4] runs in $\widetilde{O}(n\alpha^{-5}) = \widetilde{O}(nK^{10})$, whereas the formulation used in [ABF23, Theorem B.1] has running time $\widetilde{O}(n\alpha^{-6}) = \widetilde{O}(nK^{12})$; both analyses rely on the Chan–Lewenstein length-3-path guarantee. Thus the quoted prior cleaning bounds contribute K-exponents 10 and 12, respectively, while our total sampled-cleaning bound has linear dependence on K.

Third, we never materialize or certify the output's additive structure: Theorem 5.1 guarantees its bounds with high probability without ever computing $A' - A'$, whose size may exceed the running time. By contrast, the proof of [JX24, Lemma 3.2] verifies $|A' + A'| \leq nK^{28}$ by computing the sumset with sparse convolution, so its running time cannot drop below the cost of that certification.

Against the deterministic theorem of [FJX25], the comparison must be between the stated upper bounds. After suppressing logarithmic and $N^{o(1)}$ factors, their ratio is $n/K^3 + K^{12}$, minimized near $K = n^{1/15}$ at $n^{4/5}$. Thus, in the polynomial-universe regime $N = n^{O(1)}$, our stated bound is smaller by a factor of $n^{4/5-o(1)}$ uniformly for $1 \leq K \leq n$. This does not establish a lower bound on the actual running time of the algorithm in [FJX25], and the comparison trades determinism for randomization.

Guaranteed output size. Our output has size $|A'| \geq cn/\sqrt{K}$, matching up to an absolute constant the $n/\sqrt{K}$ scale that Reiher–Schoen [RS24] show is essentially optimal for guaranteeing a subset A' with $|A' - A'| = O_K(|A'|)$. In the applications underlying [JX24, Lemma 3.2] and [ABF23, Theorem 3.3], the popular-sums graph has density at least c/K^2, and the Chan–Lewenstein cleaning step [CL15, Lemmas 7.1 and 7.2] is invoked with the lower-bound parameter $\alpha := c/K^2 = \Theta(K^{-2})$, guaranteeing $\Omega(\alpha n) = \Omega(n/K^2)$ vertices. Our low-multiplicity branch instead combines all levels above a suitable threshold and produces a working relation of density $\rho \geq c/\sqrt{K}$, so Lemma 4.1 returns $\Omega(\rho n) = \Omega(n/\sqrt{K})$ vertices. The high-multiplicity branch works directly with an energetic translate A_d of the same size. Thus our guarantee improves the K-dependence of the bounds in [JX24, ABF23] by a factor of $K^{3/2}$. The theorem of [FJX25] guarantees $n/(64K)$, whereas our $n/\sqrt{K}$ scale improves its K-dependence by a factor of $\sqrt{K}$.

Normalized additive bound. Set $e := \eta n/K$ for a sufficiently small absolute constant $\eta > 0$, and retain the differences satisfying $g(d) \geq 4e$. The discarded coordinates contribute only a small constant fraction of the energy. We split the retained set once at $2n/\sqrt{K}$. If the low-multiplicity part carries a constant fraction of the energy, The light-tail selection argument in the proof of Theorem 5.1 finds a threshold $t = \Omega(n/K)$ and a tail D with $m := |D|$ whose relation has density

ρ satisfying

$$\rho \geq \frac{c}{\sqrt{K}}, \qquad \rho \geq \frac{cn^2}{K^2t^2}, \qquad m \leq \frac{C\rho n^2}{t}.$$

Lemma 4.1 guarantees that every $a_1 - a_2 \in A' - A'$ has many representations

$$a_1 - a_2 = (a_1 - u) - (b - u) + (b - v) - (a_2 - v)$$

by four elements of D. The resulting path count gives

$$\frac{|A' - A'|}{|A'|} \leq \frac{Cm^4}{\rho^6 n^4} \leq \frac{Cn^4}{\rho^2 t^4} \leq CK^4,$$

where the first step follows from the four-difference representation count, the second step from $m \leq C\rho n^2/t$, and the third step from $\rho \geq cn^2/(K^2t^2)$. If the high-multiplicity part carries a constant fraction of the energy, weighted shift sampling and the direct representation count give the stronger bound $|A' - A'|/|A'| = O(K^3)$. Therefore $O(K^4)$ is the uniform guarantee. Lemma 3.2 then squares this normalized difference-set bound, giving $|A' + A'|/|A'| = O(K^8)$.

The larger powers of K in the earlier randomized bounds arise at two explicit stages. In the proof of [JX24, Lemma 3.2], the popular-sums graph has density at least c/K^2, and the Chan–Lewenstein length-3-path lemma [CL15, Lemmas 7.1 and 7.2], invoked with the lower-bound parameter $\alpha := c/K^2 = \Theta(K^{-2})$, supplies $\Omega(\alpha^5 n^2)$ paths per output pair. Since the popular-sum set has size $O(nK)$, representation counting first gives $|A' + B'| = O(nK^{13})$; the Ruzsa triangle inequality and $|B'| = \Omega(n/K^2)$ then give $|A' + A'| = O(nK^{28})$. In the proof of [ABF23, Theorem 3.3], the same parameter in [ABF23, Theorem B.1] gives $|A' + B'| = O(K^{10}n) = O(K^{12}|B'|)$, and the Plünnecke–Ruzsa conversion squares the normalized K^{12} loss to obtain $O(K^{24}n)$. The deterministic result of [FJX25] states $|A' + A'| \leq K^{16}nN^{o(1)}$ and $|A'| \geq n/(64K)$, hence its corresponding normalized loss is $O(K^{17})N^{o(1)}$.

3 Preliminaries

We write $[N] := \{1, \ldots, N\}$. For finite sets $A, B \subseteq \mathbb{Z}$, the sumset and the difference set are

$$A + B := \{a + b : a \in A,\, b \in B\}, \qquad A - B := \{a - b : a \in A,\, b \in B\},$$

and the difference representation function is

$$r_{A-B}(d) := |\{(a, b) \in A \times B : a - b = d\}|, \qquad d \in \mathbb{Z}.$$

Definition 3.1 (Additive energy)**.** *The additive energy of a finite set A is*

$$E(A) := |\{(a_1, a_2, a_3, a_4) \in A^4 : a_1 + a_2 = a_3 + a_4\}| = \sum_d r_{A-A}(d)^2,$$

where the second expression follows by grouping quadruples according to the common difference $d := a_1 - a_4 = a_3 - a_2$.

Every set of n integers satisfies $2n^2 - n \leq E(A) \leq n^3$.

Throughout, we assume $n \geq 2$. Unless stated otherwise, c and C, possibly subscripted, denote positive absolute constants, and c_ϵ, C_ϵ, and O_ϵ indicate constants that may depend only on the fixed parameter ϵ. We write $\operatorname{poly}\log N := (\log N)^{O(1)}$; $\widetilde{O}(\cdot)$ and $\widetilde{O}_\epsilon(\cdot)$ suppress factors polylogarithmic in n; logarithms are base 2; and $\mathbb{E}$ denotes expectation.

3.1 Tools from previous work

Lemma 3.2 (Ruzsa sum–difference inequality; see, e.g., [TV06]). *Every finite nonempty set X in an abelian group satisfies $\frac{|X+X|}{|X|} \le (\frac{|X-X|}{|X|})^2$.*

The indicator vector of A is $\mathbf{1}_A$, the discrete convolution of finitely supported vectors f and g is $(f \star g)(x) := \sum_y f(y)g(x-y)$, and $\|f\|_0$ denotes the number of nonzero entries of f. In particular, $r_{A-A} = \mathbf{1}_A \star \mathbf{1}_{-A}$.

We use the following deterministic coordinatewise approximation to a convolution.

Lemma 3.3 (Fischer–Jin–Xu [FJX25], Theorem 1.9). *Let $A, B \subseteq [N]$ be multisets with $|A|, |B| \le N$, and let $\epsilon > 0$. There is a deterministic algorithm running in time*

$$(\epsilon^{-1}|A| + |B|)N^{o(1)}$$

that computes a vector f satisfying

$$\|f\|_0 \le O(\epsilon^{-1}|A| \log^2 N), \qquad \|f - \mathbf{1}_A \star \mathbf{1}_B\|_\infty \le \epsilon|B|.$$

Here $\mathbf{1}_A \star \mathbf{1}_B$ denotes discrete convolution and $\|f\|_0$ denotes the sparsity of f.

4 Supporting relation-cleaning lemma

The following lemma is the relation-cleaning ingredient used in the proof of Theorem 5.1.

Lemma 4.1 (Sampled relation cleaning). *Fix $\epsilon > 0$, and let A be a finite set of size $n := |A| \ge 2$. Suppose membership in a relation $R \subseteq A^2$ can be tested in $O(\log n)$ time, and suppose a number $\rho > 0$ is given such that $\rho n^2 \le |R| \le 2\rho n^2$. There is a randomized algorithm which, with probability at least $1 - n^{-20}$, returns a set $A' \subseteq A$ satisfying $|A'| \ge c_\epsilon \rho n$ and the following path property: for every ordered pair $(a_1, a_2) \in (A')^2$, there are at least $c_\epsilon \rho^4 n^2 |A'|$ triples $(u, b, v) \in A^3$ for which $(a_1, u), (b, u), (b, v), (a_2, v) \in R$. Its running time is $\widetilde{O}_\epsilon(n\rho^{-2} + \rho^{-4})$.*

Proof. For $a, a' \in A$, put $c_R(a, a') := |\{z \in A : (a, z), (a', z) \in R\}|$. Choose constants $\eta, \theta > 0$ depending only on ϵ, with θ sufficiently small compared with η, and set $T := \theta\rho^2 n$. Draw once a pool Z of $q := \lceil C_\epsilon \rho^{-2} \log n \rceil$ independent uniform elements of A. For every unordered pair $\{a, a'\}$ with $a, a' \in A$, allowing $a = a'$, define

$$\widehat{c}_R(a, a') := \frac{n}{q} \sum_{z \in Z} \mathbf{1}_R(a, z)\mathbf{1}_R(a', z),$$

and classify the pair as bad when $\widehat{c}_R(a, a') < 3T/2$, and as good otherwise. The same pool and the same unordered-pair convention are used in every query, so the resulting classifier is fixed and symmetric. By the multiplicative Chernoff bound, after increasing C_ϵ, with probability at least $1 - n^{-30}$ the following implications hold simultaneously for all pairs in A^2:

$$c_R(a, a') \le T \Longrightarrow (a, a') \text{ is classified bad}, \qquad c_R(a, a') \ge 2T \Longrightarrow (a, a') \text{ is classified good}.$$

We condition on this event for the rest of the proof. Taking the union bound over all n^2 pairs also makes the guarantee valid for the adaptive queries made below.

For $x \in A$, let $N(x) := \{a \in A : (a, x) \in R\}$, set $d_x := |N(x)|$, and let b_x be the number of ordered pairs in $N(x)^2$ that are classified bad. If $B \subseteq A^2$ denotes the symmetric ordered relation of classified-bad pairs, then

$$\sum_{x \in A} |N(x)^2 \cap B| = \sum_{(a,a') \in B} c_R(a, a') \leq 2Tn^2,$$

where the first step follows from double-counting ordered triples (x, a, a') with $a, a' \in N(x)$ and $(a, a') \in B$, and the second step from $c_R(a, a') < 2T$ on B and $|B| \leq n^2$. Also,

$$\sum_{x \in A} d_x^2 \geq \frac{1}{n}(\sum_{x \in A} d_x)^2 \geq \rho^2 n^3,$$

where the first step follows from Cauchy–Schwarz and the second step from $\sum_{x \in A} d_x = |R| \geq \rho n^2$. Consequently, for a sufficiently small choice of θ, the average of the potential $\Phi(x) := d_x^2 - \eta^{-1} b_x$ is at least $(1 - \eta)\rho^2 n^2$. Since $\Phi(x) \leq n^2$, comparing the average with this upper bound shows that at least $\frac{\eta\rho^2}{1-(1-2\eta)\rho^2} n \geq \eta\rho^2 n$ choices of x satisfy $\Phi(x) \geq (1 - 2\eta)\rho^2 n^2$. For every such x, we have $d_x \geq (1 - 2\eta)^{1/2}\rho n$ and $b_x \leq \eta d_x^2$.

Sample $O_\epsilon(\rho^{-2} \log n)$ pivots independently and uniformly from A, with replacement. For each pivot, compute $N(x)$ exactly by scanning A. If $d_x < (1 - 3\eta)\rho n$, reject the pivot immediately. Otherwise, estimate b_x/d_x^2 to additive error at most $\eta/4$ using $O_\epsilon(\log n)$ independent ordered-pair samples from $N(x)^2$ and the fixed classifier. Accept a pivot only if its estimated bad-pair proportion is at most $3\eta/2$. Chernoff bounds show that, with probability at least $1 - n^{-30}$, a sampled high-potential pivot is accepted and every accepted pivot satisfies $b_x \leq 2\eta d_x^2$.

If no pivot is accepted, output $A' := \varnothing$ and halt. Otherwise, fix an accepted pivot x and write $N := N(x)$ and $d := |N|$. For every $a \in N$, estimate the proportion of classified-bad partners of a in N to additive error at most $\sqrt{\eta}/2$, using $O_\epsilon(\log n)$ independent uniform partner samples. Retain all vertices whose estimated proportion is at most $3\sqrt{\eta}$. With probability at least $1 - n^{-30}$, every retained vertex has at most $4\sqrt{\eta}d$ classified-bad partners. Moreover, at most $2\sqrt{\eta}d$ vertices have more than $\sqrt{\eta}d$ bad partners, because there are at most $2\eta d^2$ bad ordered pairs. Thus, after decreasing η as a function of ϵ, the retained set A' satisfies $|A'| \geq (1 - 2\sqrt{\eta})d \geq c_\epsilon \rho n$.

Fix $a_1, a_2 \in A'$. All but at most $8\sqrt{\eta}d$ choices of $b \in A'$ make both (a_1, b) and (b, a_2) classified good. For each remaining b, both codegrees exceed T. Hence the number of triples (u, b, v) with $(a_1, u), (b, u), (b, v), (a_2, v) \in R$ is at least $(|A'| - 8\sqrt{\eta}d)T^2$. Since $|A'| \geq (1 - 2\sqrt{\eta})d$, choosing η sufficiently small gives $|A'| - 8\sqrt{\eta}d \geq c_\epsilon |A'|$, and hence the last display is at least $c_\epsilon \rho^4 n^2 |A'|$. This proves the path property.

It remains to justify the running time. There are $\widetilde{O}_\epsilon(\rho^{-2})$ pivot candidates. Their exact neighborhoods cost $\widetilde{O}_\epsilon(n\rho^{-2})$ time. Estimating the bad-pair fraction for all candidates costs $\widetilde{O}_\epsilon(\rho^{-4})$, because one classifier query uses $\widetilde{O}_\epsilon(\rho^{-2})$ relation tests. The final cleaning uses $\widetilde{O}_\epsilon(n\rho^{-2})$ additional tests. A union bound over the preceding failure events gives total success probability at least $1 - n^{-20}$. □

5 A randomized algorithmic BSG theorem

Theorem 5.1 (Randomized algorithmic BSG)**.** *There are absolute constants $c, C > 0$ with the following property. Let $A \subseteq [N]$ be a set of $n \geq 2$ integers, let $1 \leq K \leq n$, and assume that $E(A) = \sum_d r_{A-A}(d)^2 \geq n^3/K$. There is a randomized algorithm (Algorithm 1) which, with probability at*

least $1 - n^{-10}$*, runs in time* $nKN^{o(1)}$ *and returns a set* $A' \subseteq A$ *satisfying*

$$|A'| \geq \frac{cn}{\sqrt{K}} \quad \textit{and} \quad |A' - A'| \leq CK^4|A'|.$$

Algorithm 1 A randomized BSG algorithm

Require: $A \subseteq [N]$, $n := |A| \geq 2$, $1 \leq K \leq n$, and $E(A) \geq n^3/K$
Ensure: A set $A' \subseteq A$

procedure RANDOMIZEDBSG(A, K) ▷ Theorem 5.1
  Fix sufficiently large absolute constants $C_7, C_{17}, C_{\text{est}} > 0$, set $C_8 := C_7 + 2$, $\gamma_0 := 1/4$, $a_0 := 3\gamma_0/16$, and $c_0 := 3\gamma_0^2/64$, and choose sufficiently small absolute $\eta, \epsilon_0 > 0$
  $\epsilon := \eta/K$, $e := \epsilon n$, $M_1 := \lceil C_{\text{est}}\eta^{-2} \log n \rceil$, and $M_2 := \lceil C_{\text{est}}\eta^{-1} \log n \rceil$
  Translate $-A$ into $[N]$, apply Lemma 3.3 at accuracy ϵ, and translate back to obtain f with $\|f - r_{A-A}\|_\infty \leq e$
  $g_0(d) := \max\{f(d), 0\}$ and $g(d) := (g_0(d) + g_0(-d))/2$, treating missing entries as zero
  Store A and the explicit support of g as sorted arrays; use binary search for membership and value queries, and take $g(d) := 0$ outside its explicit support
  $S := \{d : g(d) \geq 4e\}$, $H := \{d \in S : g(d) > 2n/\sqrt{K}\}$, and $Q := S \setminus H$
  $F_H := \sum_{d \in H} g(d)^2$ and $F_Q := \sum_{d \in Q} g(d)^2$
  if $F_Q \geq F_H$ **then** ▷ Low-multiplicity case
    Sort Q in nonincreasing order of $g(d)$, group equal values, and scan prefix sums only at the ends of tie blocks to find $t \geq a_0 n/K$ such that, for $M(t) := \sum_{d \in Q : g(d) \geq t} g(d)$, $M(t)/n^2 \geq c_0/\sqrt{K}$ and $M(t)/n^2 \geq c_0 n^2/(K^2 t^2)$
    $D := \{d \in Q : g(d) \geq t\}$ and $M := \sum_{d \in D} g(d)$
    Store D as a sorted array, use binary search for membership, and represent $R_D := \{(a, b) \in A^2 : a - b \in D\}$ implicitly
    $\widehat{\delta} := M/n^2$ and $\rho := 3\widehat{\delta}/4$
    return the output of Lemma 4.1 on (A, R_D, ρ) with parameter ϵ_0
  end if
  Build a prefix-sum sampler on H that returns d with probability $g(d)^2/F_H$, and set $B(x, y) := \mathbf{1}[g(x - y) \leq 2e]$
  for $s := 1, \ldots, \lceil C_{17} \log n \rceil$ **do**
    Draw $d \in H$ independently from the prefix-sum sampler, with replacement, and form an array listing $A_d := \{a \in A : a - d \in A\}$
    Let $\widehat{q}_d$ be the fraction with $B(x, y) = 1$ among M_1 fresh independent uniform ordered pairs from A_d^2
    if $\widehat{q}_d \leq (C_7 + 1)\eta$ **then**
      for each $a \in A_d$ **do**
        Let $\widehat{q}_a$ be the fraction with $B(a, y) = 1$ among M_2 fresh independent uniform elements of A_d
      end for
      return $\{a \in A_d : \widehat{q}_a \leq (2C_8 + 1)\sqrt{\eta}\}$
    end if
  end for
  return $\emptyset$
end procedure

Proof. We analyze Algorithm 1. Fix sufficiently small absolute constants $\eta, \epsilon_0 > 0$. Apply the deterministic popular-sums approximation of Lemma 3.3 at accuracy $\epsilon := \eta/K$, and set $e := \epsilon n = \eta n/K$. After truncating and symmetrizing as in the algorithm, the explicitly supported vector g satisfies $g(d) = g(-d)$ and $\|g - r_{A-A}\|_\infty \le e$. The approximation takes $nKN^{o(1)}$ time.

Let $S := \{d : g(d) \ge 4e\}$. On its complement, $r_{A-A}(d) < 5e$, and hence

$$\sum_{d \notin S} r_{A-A}(d)^2 \le 5e \sum_d r_{A-A}(d) = 5en^2 = \frac{5\eta n^3}{K}.$$

The first step follows from $r_{A-A}(d) < 5e$, the second from $\sum_d r_{A-A}(d) = n^2$, and the third from $e = \eta n/K$. For $d \in S$, the approximation guarantee gives $3g(d)/4 \le r_{A-A}(d) \le 5g(d)/4$. Therefore, after decreasing η,

$$\sum_{d \in S} g(d)^2 \ge \frac{16}{25} \sum_{d \in S} r_{A-A}(d)^2 \ge \frac{n^3}{2K}.$$

Moreover, $g(d) \le 4r_{A-A}(d)/3$ on S, so $\sum_{d \in S} g(d) \le 4n^2/3$ and $|S| \le nK/(3\eta)$. Thus the retained support can be sorted and scanned in $nKN^{o(1)}$ time.

Split S into $H := \{d \in S : g(d) > 2n/\sqrt{K}\}$ and $Q := S \setminus H$. At least one of $F_H := \sum_{d \in H} g(d)^2$ and $F_Q := \sum_{d \in Q} g(d)^2$ is at least $n^3/(4K)$. The algorithm chooses the corresponding case using these computable quantities.

Low-multiplicity case. Suppose that $F_Q \ge n^3/(4K)$. We first prove the light-tail selection used by the algorithm. Put $\gamma := 1/4$, $a := 3\gamma/16$, $c_0 := 3\gamma^2/64$, $t_0 := an/K$, $t_* := n/K^{3/4}$, and $T := 2n/\sqrt{K}$. For $s \ge 0$, define $M(s) := \sum_{d \in Q : g(d) \ge s} g(d)$. The layer-cake identity gives

$$F_Q = \int_0^T M(s)\mathrm{d}s.$$

We claim that some $t \in [t_0, T]$ satisfies

$$\frac{M(t)}{n^2} \ge \frac{c_0}{\sqrt{K}} \quad \text{and} \quad \frac{M(t)}{n^2} \ge \frac{c_0 n^2}{K^2 t^2}.$$

Otherwise, for every $s \in [t_0, T]$, $M(s) < c_0 n^2 \max\{1/\sqrt{K}, n^2/(K^2 s^2)\}$. The two terms agree at $s = t_*$. Since $M(s) \le 4n^2/3$, integration gives

$$F_Q < \frac{4n^2 t_0}{3} + \frac{c_0 n^4}{K^2} \int_{t_0}^{t_*} \frac{\mathrm{d}s}{s^2} + \frac{c_0 n^2}{\sqrt{K}} \int_{t_*}^{T} \mathrm{d}s \le (\frac{\gamma}{4} + \frac{c_0}{a} + 2c_0)\frac{n^3}{K} < \frac{\gamma n^3}{K},$$

contradicting $F_Q \ge \gamma n^3/K$. The first step follows by splitting the layer-cake integral at t_0 and t_* and applying the mass and tail bounds. The second step evaluates the integrals using the definitions of t_0, t_* and T, and the third step follows from $a = 3\gamma/16$ and $c_0 = 3\gamma^2/64$. It suffices to test thresholds equal to observed values of g: rounding a valid threshold upward to the smallest included value leaves $M(t)$ unchanged and only makes the second required inequality easier. Hence sorting Q, grouping equal values, and scanning prefix sums at the ends of tie blocks finds such a threshold. This selector is a convenient algorithmic form of the low-multiplicity selection in [RS24, Lemma 2.2].

Fix the threshold found by the algorithm, and set $D := \{d \in Q : g(d) \ge t\}$, $m := |D|$, $M := M(t)$, and $\rho := 3M/(4n^2)$. Because $t \ge an/K$ and $e = \eta n/K$, choosing $\eta \le a/4$ gives

$e \le t/4$. Consequently, the density of $R_D := \{(x,y) \in A^2 : x - y \in D\}$ lies between ρ and 2ρ. The selector also gives

$$\rho \ge \frac{c_1}{\sqrt{K}}, \rho \ge \frac{c_1 n^2}{K^2 t^2}, m \le \frac{M}{t} = \frac{4\rho n^2}{3t}$$

for an absolute constant $c_1 > 0$. Apply Lemma 4.1 to R_D with parameter ϵ_0. With probability at least $1 - n^{-20}$ it returns $A' \subseteq A$ satisfying $|A'| \ge c_2 \rho n \ge c_3 n/\sqrt{K}$.

The path guarantee in that lemma supplies at least $c_4 \rho^4 n^2 |A'|$ triples (u, b, v) for each ordered pair in $(A')^2$. Each triple expresses

$$a_1 - a_2 = (a_1 - u) - (b - u) + (b - v) - (a_2 - v)$$

using four elements of D. For fixed a_1, a_2 and four values in D, the triple is uniquely determined. Choosing one ordered pair for each value in $A' - A'$ therefore yields

$$\frac{|A' - A'|}{|A'|} \le \frac{C_1 m^4}{\rho^6 n^4} \le \frac{C_2 n^4}{\rho^2 t^4} \le C_3 K^4.$$

The first step follows from the preceding path count and $|A'| \ge c_2 \rho n$, the second from $m \le 4\rho n^2/(3t)$, and the third from $\rho \ge c_1 n^2/(K^2 t^2)$. The running time of the cleaner is $\widetilde{O}(n\rho^{-2} + \rho^{-4}) = \widetilde{O}(nK + K^2) = \widetilde{O}(nK)$ because $K \le n$.

High-multiplicity case. Suppose that $F_H \ge n^3/(4K)$. For $d \in H$, define $A_d := \{a \in A : a - d \in A\}$. Since $g(d) > 2n/\sqrt{K}$ and $e \le \eta n/\sqrt{K}$, we have $|A_d| = r_{A-A}(d) > 3n/(2\sqrt{K})$ after decreasing η. Call an ordered pair $(x, y) \in A^2$ bad when $g(x - y) \le 2e$. Because g is even, the bad relation is symmetric, and every bad difference s satisfies $r_{A-A}(s) \le 3e$. If B_d denotes the number of bad ordered pairs in A_d^2, double counting gives

$$\sum_{d \in H} B_d \le \sum_{s : g(s) \le 2e} r_{A-A}(s)^2 \le 3en^2 = \frac{3\eta n^3}{K}.$$

The first step follows because each bad ordered pair of difference s lies in at most $r_{A-A}(s)$ sets A_d, the second from $r_{A-A}(s) \le 3e$ on bad differences and $\sum_s r_{A-A}(s) = n^2$, and the third from $e = \eta n/K$.

Sample $d \in H$ with probability proportional to $g(d)^2$. Since $g(d)^2$ and $r_{A-A}(d)^2$ differ by only an absolute factor on H, the preceding estimate and $F_H \ge n^3/(4K)$ imply

$$\mathbb{E}_d[\frac{B_d}{|A_d|^2}] \le C_4 \eta.$$

Consequently, a constant fraction of the sampling mass consists of shifts with $B_d \le C_7 \eta |A_d|^2$. Draw $C_{17} \log n$ shifts, build each A_d by scanning A, and estimate its bad-pair proportion to additive error at most η using $O_\eta(\log n)$ sampled pairs. Accept the first sampled shift whose estimate is at most $(C_7 + 1)\eta$. With probability at least $1 - n^{-12}$, a shift is accepted and every accepted shift satisfies $B_d \le C_8 \eta |A_d|^2$, where $C_8 := C_7 + 2$.

For every $a \in A_d$, estimate its proportion of bad partners in A_d to additive error at most $\sqrt{\eta}$, using $O_\eta(\log n)$ independent partner samples. Retain the vertices whose estimate is at most $(2C_8 + 1)\sqrt{\eta}$. With probability at least $1 - n^{-14}$, all estimates are simultaneously accurate, and every retained vertex has at most $(2C_8 + 2)\sqrt{\eta}|A_d|$ bad partners. Moreover, Markov's inequality and $B_d \le C_8 \eta |A_d|^2$ show that all but at most $\sqrt{\eta}|A_d|/2$ vertices are retained. Thus, for sufficiently small η, the retained set satisfies $|A'| \ge c_5 |A_d| \ge c_6 n/\sqrt{K}$.

Fix $a_1, a_2 \in A'$. At least $c_7|A'|$ choices of $b \in A'$ make both (a_1, b) and (b, a_2) nonbad. For every such b, $r_{A-A}(a_1 - b) > e$ and $r_{A-A}(b - a_2) > e$. Choosing representations of these two differences produces at least $c_7|A'|e^2$ quadruples in A^4 whose two differences sum to $a_1 - a_2$. For a fixed representative a_1, a quadruple determines b, while the quadruple also determines the resulting difference. Therefore choosing one representative ordered pair for every value in $A' - A'$ gives disjoint families of quadruples, so

$$|A' - A'|c_7|A'|e^2 \le n^4.$$

It follows that

$$\frac{|A' - A'|}{|A'|} \le \frac{C_5 n^4}{|A'|^2 e^2} \le C_6 K^3 \le C_6 K^4,$$

where the first step follows from the packing bound, the second from $|A'| \ge c_6 n/\sqrt{K}$ and $e = \eta n/K$, and the third from $K \ge 1$.

The prefix-sum sampler is built in $O(|H|)$ time and produces each draw in $O(\log |H|)$ time. Each A_d is built by scanning the sorted array A and using binary search to test whether $a - d \in A$. All sampled tests and vertex cleaning take $n(\log n)^{O(1)}$ time. Building the sorted arrays adds only logarithmic factors, so these costs are absorbed by $nKN^{o(1)}$. Combining the two cases proves the bounds and running time. Adding the relation-cleaning, shift-selection, and vertex-estimation failure probabilities gives total failure probability at most n^{-10} in either branch. □

Corollary 5.2 (Randomized algorithmic BSG with a small sumset)**.** *There are absolute constants $c, C > 0$ with the following property. Let $A \subseteq [N]$ be a set of $n \ge 2$ integers, let $1 \le K \le n$, and assume that $E(A) = \sum_d r_{A-A}(d)^2 \ge n^3/K$. There is a randomized algorithm which, with probability at least $1 - n^{-10}$, runs in time $nKN^{o(1)}$ and returns a set $A' \subseteq A$ satisfying*

$$|A'| \ge \frac{cn}{\sqrt{K}} \quad \text{and} \quad |A' + A'| \le CK^8|A'|.$$

Proof. On the success event in Theorem 5.1, which has probability at least $1 - n^{-10}$, Algorithm 1 returns a nonempty set A' satisfying $|A' - A'|/|A'| \le C_0 K^4$ for an absolute constant $C_0 > 0$. Applying Lemma 3.2 with $X = A'$ gives $|A' + A'|/|A'| \le C_0^2 K^8$. Absorbing C_0^2 into C proves the additive bound. The output-size and running-time guarantees hold on the same success event. □

Acknowledgment

Gemini Pro 3.1, GPT 5.6 and Claude code 5.5 are used for language editing and grammar checking.

References

[ABF23] Amir Abboud, Karl Bringmann, and Nick Fischer. Stronger 3-SUM lower bounds for approximate distance oracles via additive combinatorics. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)*, pages 391–404. ACM, 2023.

[Blo25] Thomas F Bloom. Control and its applications in additive combinatorics. *arXiv preprint arXiv:2501.09470*, 2025.

[BS94] Antal Balog and Endre Szemerédi. A statistical theorem of set addition. *Combinatorica*, 14(3):263–268, 1994.

[CL15] Timothy M. Chan and Moshe Lewenstein. Clustered integer 3SUM via additive combinatorics. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 31–40, 2015.

[FJX25] Nick Fischer, Ce Jin, and Yinzhan Xu. New applications of 3SUM-counting in fine-grained complexity and pattern matching. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4547–4595. SIAM, 2025.

[Gow98] W. T. Gowers. A new proof of szemerédi's theorem for arithmetic progressions of length four. *Geometric and Functional Analysis*, 8(3):529–551, 1998.

[JX24] Ce Jin and Yinzhan Xu. Removing additive structure in 3SUM-based reductions. *SIAM Journal on Computing*, pages STOC23–1–STOC23–56, 2024.

[RS24] Christian Reiher and Tomasz Schoen. Note on the theorem of balog, szemerédi, and gowers. *Combinatorica*, 44(3):691–698, 2024.

[TV06] Terence Tao and Van H. Vu. *Additive Combinatorics*, volume 105 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006. See the official errata for Corollary 6.20.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il