5

Inverse sum set theorems

In Chapter 2 we established the elementary theory of sum set estimates, showing how information on one sum $A + B$ can be used to control other sums such as $A - B$ or $nA - mA$. These estimates worked reasonably well even when the doubling constants of the sets involved were fairly large, since all the bounds were polynomial in this constant. On the other hand, we did not get detailed structural information on sets with small doubling constant; the best we could do is cover them by an approximate group (Proposition 2.26).

In this chapter we shall focus on the following question: given two additive sets A, B with $A + B$ very small, what is the strongest structural statement one can then conclude about A and B? One of the main results in this area is *Freiman's theorem* which (in the torsion-free case) asserts that an additive set A with small doubling constant $\sigma[A] = |2A|/|A|$ is contained in a progression of bounded rank which is not much larger than the original set. This theorem comes in a number of variants; we give several of them below. In doing so we shall also come across the useful concept of a *Freiman homomorphism*, which to a large extent frees the study of additive sets from the ambient group that they reside in, giving rise to a number of useful tricks, such as embedding the set inside a particularly nice group.

5.1 Minimal size of sum sets and the e-transform

Before we begin with inverse theorems, we first address an even more basic question: given the cardinalities $|A|, |B|$ of two additive sets A and B in an ambient group Z, what is the least possible cardinality $|A + B|$ of the sum set $A + B$? If we allow the group Z to be completely arbitrary, then the answer is given by (2.1) and Proposition 2.2, thus $|A + B| \geq \max(|A|, |B|)$, with equality if and only if one of the sets is contained inside a coset of a finite group G, and the other set is a finite union of cosets of G. However, for *specific* choices of Z, one can improve

this bound somewhat. For instance, if Z is the integers, then Z contains no finite subgroups other than the trivial one $\{0\}$, and so we expect to do better than (2.1) unless one of $|A|, |B|$ is equal to 1.

A very simple, but surprisingly powerful, tool for establishing the minimal size of sum sets is the *e-transform*, which we now define.

Definition 5.1 (*e*-transform) [73] Let A, B be additive sets in an ambient group Z, and let $e \in A - B$. We define the *e-transform* of the pair A, B to be the sets $A_{(e)} := A \cup (B + e)$ and $B_{(e)} := B \cap (A - e)$.

One can view this transform as removing the elements of $B \backslash (A - e)$ from B and transferring them to A (after translating them by e). The main point of the e-transform is that it shrinks (or keeps constant) the size $|A + B|$ of the sum set, while maintaining the total size $|A| + |B|$ of A and B. More precisely:

Lemma 5.2 *[73] Let A, B be additive sets in an ambient group Z, let $e \in A - B$, and let $A_{(e)}$, $B_{(e)}$ be the e-transform of A, B. Then $A_{(e)}$ and $B_{(e)}$ are also additive sets (i.e. finite and non-empty), and*

$$A_{(e)} + B_{(e)} \subseteq A + B. \tag{5.1}$$

Furthermore we have

$$|A_{(e)}| + |B_{(e)}| = |A| + |B|, \tag{5.2}$$

and more generally

$$\begin{aligned} |A_{(e)} \cap E| + |B_{(e)} \cap E| = {} & |A \cap E| + |B \cap E| \\ & + |(B \backslash (A - e)) \cap ((E - e) \backslash E)| \\ & - |(B \backslash (A - e)) \cap (E \backslash (E - e))| \end{aligned} \tag{5.3}$$

for any $E \subseteq Z$. Finally, we have

$$|A_{(e)}| \geq |A|; \quad |B_{(e)}| \leq |B| \tag{5.4}$$

with equality in either expression if and only if $B + e \subseteq A$.

We leave the easy proof of this lemma to Exercise 5.1.2. We now give some applications of this Lemma. First we obtain the minimal size of sum sets in the integers $\mathbf{Z}$ (cf. Lemma 3.18), taking advantage of the fact that the integers are ordered.

Lemma 5.3 *If A and B are additive sets in $\mathbf{Z}$, then we have $|A + B| \geq |A| + |B| - 1$.*

Proof Let $e := \max(A) - \min(B)$; then we see that $B_{(e)}$ is the singleton set $\{\min(B)\}$, and thus by (5.2) $|A_{(e)}| = |A| + |B| - 1$, so $|A_{(e)} + B_{(e)}| = |A| + |B| - 1$. The claim now follows from (5.1). □

Now we prove a similar result in a cyclic group $\mathbf{Z}_p$ of prime order. Here the key fact to exploit is that $\mathbf{Z}_p$ contains no non-trivial subgroups.

Theorem 5.4 (Cauchy–Davenport inequality) *[47], [68] If p is a prime, and A, B are two additive sets in $\mathbf{Z}_p$, then*

$$|A + B| \geq \min(|A| + |B| - 1, p).$$

This result was first discovered by Cauchy [47] and then rediscovered 122 years later by Davenport [68]. We remark that the corresponding result for restricted summation $A \hat{+} B := \{a + b : a \in A, b \in B, a \neq b\}$ requires different methods to establish; see Section 9.2. We shall give alternative proofs of Theorem 5.4 in Section 9.2 and Section 9.8.

Proof We induce on the size of $|B|$; thus we suppose that the claim has already been proven for all smaller sets B (the case $|B| = 1$ is trivial). Suppose we can find an element $e \in A - B$ such that the e-transform $B_{(e)}$ of B is strictly smaller than B. Then we have $|A_{(e)}| + |B_{(e)}| \geq \min(|A_{(e)}| + |B_{(e)}| - 1, p)$ by the induction hypothesis, and the claim follows by (5.1) and (5.2). Thus we may assume that none of the e-transforms of B are strictly smaller than B. Using Lemma 5.2, this means that $B + e \subseteq A$ for all $e \in A - B$, so

$$A - B + B \subseteq A.$$

Using Proposition 2.2, we thus see that B is contained in a coset of a subgroup G of $\mathbf{Z}_p$, and A is a union of cosets of G. But since p is prime, the only subgroups G available are the trivial group $\{0\}$ and the full group $\mathbf{Z}_p$. In either case the Cauchy–Davenport inequality is easily verified. □

One can generalize Lemma 5.3 and Theorem 5.4. Recall from Definition 2.32 that the *symmetry group* $\mathrm{Sym}_1(A)$ of an additive set A in an ambient group Z was defined as $\mathrm{Sym}_1(A) := \{h \in Z : A + h = A\}$.

Theorem 5.5 (Kneser's theorem) *[211] For any additive sets A, B in an additive group Z, we have*

$$\begin{aligned} |A + B| &\geq |A + \mathrm{Sym}_1(A + B)| + |B + \mathrm{Sym}_1(A + B)| - |\mathrm{Sym}_1(A + B)| \\ &\geq |A| + |B| - |\mathrm{Sym}_1(A + B)|. \end{aligned}$$

Proof We use a triple induction. First we induce upward on $|A + B|$, thus assuming that the claim has been proven for all pairs A, B with a smaller value of $|A + B|$. Next, with $|A + B|$ fixed, we induce *downward* on $|A| + |B|$ (which is bounded above by $2|A + B|$), assuming the claim proven for larger values of $|A| + |B|$. Finally, with $|A + B|$ and $|A| + |B|$ fixed, we induce upward on $|B|$, assuming the claim proven for smaller values of $|B|$. This rather complex induction is forced

on us by the different reductions on A and B that we will use in the (surprisingly delicate) argument.

Let $G := \mathrm{Sym}_1(A+B)$. If G is not the trivial group $\{0\}$, then we can pass from Z to the quotient group Z/G, replacing A and B by $(A+G)/G$ and $(B+G)/G$ and reducing the size of $|A+B|$, and the claim then follows from the first induction hypothesis. Thus we may take $\mathrm{Sym}_1(A+B) = \{0\}$. Our task is then to show that $|A+B| \geq |A| + |B| - 1$.

Suppose that $B_{(e)} = B$ for all $e \in A - B$. Then we have $A - B + B \subseteq A$ as before, and so by Proposition 2.2, B is contained in a coset of a group H, and A is a union of cosets of H. Then $\mathrm{Sym}_1(A+B)$ contains H and hence $H = \{0\}$, which implies $|B| = 1$. The claim is then easily verified.

It remains to consider the case when $B_{(e)}$ is strictly smaller than B for at least one $e \in A - B$. Among all such e, we choose one which maximizes the value of $|B_{(e)}|$. By translating B (and $B_{(e)}$) by e if necessary we may normalize $e = 0$; thus $A_{(0)} = A \cup B$ and $B_{(0)} = A \cap B$. Note from (5.3) that $|A_{(0)} + B_{(0)}| \leq |A+B|$, that $|A_{(0)}| + |B_{(0)}| = |A| + |B|$, and $|B_{(0)}| < |B|$. Thus by the induction hypotheses we have

$$|A_{(0)} + B_{(0)}| \geq |A_{(0)} + H| + |B_{(0)} + H| - |H|, \tag{5.5}$$

where $H := \mathrm{Sym}_1(A_{(0)} + B_{(0)})$. Let $C := (A \cap B) + H$. By definition of H and $A_{(0)}$, $B_{(0)}$, we see that $A + C$ and $B + C$ are contained in $A_{(0)} + B_{(0)}$ and hence in $A + B$. So we can replace A and B by $A \cup C$ and $B \cup C$ without affecting $A + B$ or $\mathrm{Sym}(A+B)$. Thus we may assume that C is contained in both A and B, otherwise $|A + C| + |B + C|$ would exceed $|A| + |B|$ and the claim will follow from the second induction hypothesis. In particular we see that $A \cap B = C$ is the union of a non-zero number of cosets of H.

Suppose that $A_{(0)} + B_{(0)}$ is equal to $A + B$; then $H = \mathrm{Sym}(A+B) = \{0\}$, and the claim follows from (5.5) and (5.3). Thus we may assume that $A_{(0)} + B_{(0)}$ is strictly smaller than $A + B$.

Let A' denote those elements $a \in A$ such that $a + b \notin A_{(0)} + B_{(0)}$ for some $b \in B$. By the previous assumption, A' is non-empty; also observe that a (and hence $a + H$) is disjoint from $C = B_{(0)}$ for all $a \in A'$. Let b be such that $a + b \notin A_{(0)} + B_{(0)}$: then $a + b + H$ is disjoint from $A_{(0)} + B_{(0)}$ (by definition of H); since $b \in A_{(0)}$, we conclude that $a + H$ is disjoint from $A \cap B$. Also we have $((a+H) \cap A) + b$ disjoint from $A_{(0)} + B_{(0)}$ and contained in $A + B$; thus

$$|A + B| \geq |A_{(0)} + B_{(0)}| + |(a+H) \cap A|.$$

Since $A \cap B$ is disjoint from $a + H$, we have

$$\begin{aligned} |A_{(0)} + H| &\geq |A_{(0)}| + |(A_{(0)} + H \backslash A_{(0)}) \cap (a+H)| \\ &= |A_{(0)}| + |H| - |(a+H) \cap A| - |(a+H) \cap B| \end{aligned}$$

and hence by (5.5) and (5.3)

$$|A+B| \geq |A| + |B| - |(a+H) \cap B|.$$

Thus we will be done unless we have $|(a+H) \cap B| > 1$ for all $a \in A'$, which we now assume.

For each $a \in A'$, let $A_a := (a+H) \cap A$ and $B_a := (a+H) \cap B$. Suppose we can find $a, a' \in A'$ such that $A_a - B_a + B_{a'} \not\subseteq A_{a'}$. Then we can find $e \in A_a - B_a \subseteq H$ such that $B_{a'} + e \not\subseteq A_{a'}$. This shows that B is not contained in $A - e$, and thus $B_{(e)}$ is strictly smaller than B, and also contains both $B_{(0)} = C$ and the non-empty set $B_a \cap (A_a - e)$ (which lies in $a + H$ and is hence disjoint from C), and is thus strictly larger than $B_{(0)}$. This contradicts the maximality of $|B_{(0)}|$. Thus we must have $A_a - B_a + B_{a'} \subseteq A_{a'}$ for all $a, a' \in A'$. This implies in particular that $|A_a| = |A_{a'}|$ for all $a, a' \in A'$, which by Proposition 2.2 implies that the B_a are each contained in a coset of a fixed group K, and that the A_a are unions of cosets of K (in particular K is a subgroup of H). Since we are assuming that $|B_a| > 1$ for all $a \in A'$, we have $|K| > 1$. Since $A_a + B$ is the union of cosets of K for each a, and $A_{(0)} + B_{(0)}$ is a union of cosets of H, and hence K, we conclude that $A + B$ is the union of cosets of K. But this contradicts the hypothesis that $\mathrm{Sym}_1(A+B) = \{0\}$, and we are done. □

As one application of Kneser's theorem we give a complete classification of sets with very small doubling constant.

Corollary 5.6 (Near-exact inverse sum set theorem) *Let A be an additive set in an ambient group Z. Then the following are equivalent:*

- $\sigma[A] < \frac{3}{2}$ *(i.e. $|A+A| < \frac{3}{2}|A|$);*
- $\delta[A] < \frac{3}{2}$ *(i.e. $|A-A| < \frac{3}{2}|A|$, or $d(A,A) < \log \frac{3}{2}$);*
- $|A+B| < \frac{3}{2}|A|$ *for some additive set B in Z with $|B| \geq |A|$;*
- $|nA - mA| < \frac{3}{2}|A|$ *for all non-negative integers n, m;*
- $A \subseteq x + G$ *for some $x \in Z$ and subgroup G of Z with $|G| < \frac{3}{2}|A|$.*

This should be compared with Proposition 2.7 and Exercise 2.6.5. The factor $\frac{3}{2}$ is sharp, as can be seen by the example $A = \{0, 1\}$ in the integers $\mathbf{Z}$, or more generally $A = \{0,1\} \times G$ in the group $\mathbf{Z} \times G$ for any finite group G.

Proof We shall only prove that the third claim implies the fifth; the other claims are similar or trivial and are left as an exercise. From Kneser's theorem we have

$$\frac{3}{2}|A| > |A+B| \geq |A| + |B| - |\mathrm{Sym}_1(A+B)| \geq 2|A| - |\mathrm{Sym}_1(A+B)|;$$

hence if we set $G := \mathrm{Sym}_1(A+B)$, then $|G| > |A|/2$. Since $|A+B| < \frac{3}{2}|A|$ and $A+B$ is a union of cosets of its symmetry group G, we thus see that $A+B$ is

equal to the union of at most two cosets in G, and $|G| < \frac{3}{2}|A|$. Suppose first that $A + B$ is the union of two cosets of G. Then $\frac{3}{2}|B| \geq \frac{3}{2}|A| > |A + B| = 2|G|$, which implies that neither A nor B can be contained in a single coset of G. But this contradicts Kneser's theorem again. Thus $A + B$ is a single coset of G, which implies that A is also contained in a coset of G. The claim follows. □

Now we return to the integers, and obtain a more advanced version of Lemma 5.3.

Theorem 5.7 (Mann's theorem) *[243] Let $N \geq 0$, let $0 < \alpha < 1$, and let A, B be additive sets in $\mathbf{Z}$ such that $0 \in A$, B and*

$$|A \cap [1, n]| + |B \cap [1, n]| \geq \alpha n \tag{5.6}$$

for all $0 \leq n \leq N$. Then

$$|(A + B) \cap [1, n]| \geq \alpha n \text{ for all } 0 \leq n \leq N.$$

Proof The claim is easily verified for $N = 0$, so let us assume inductively that $N \geq 1$ and the claim has already been proven for all smaller N. In particular from this induction hypothesis we already have

$$|(A + B) \cap [1, n]| \geq \alpha n \text{ for all } 0 \leq n < N$$

and so it suffices to prove that

$$|(A + B) \cap [1, N]| \geq \alpha N.$$

We now fix N and induce on $|B|$. If $|B| = 1$, then $B = \{0\}$ and the claim is easily verified, so suppose that $|B| > 1$ and the claim has already been proven for all smaller values of B. Without loss of generality we may take $A \subseteq [0, N]$ and $B \subseteq [0, N]$ as the additional elements of A and B are clearly harmless.

In light of Lemma 5.2 and the induction hypothesis, it will suffice to find an integer $e \in A \subseteq A - B$ such that $|B_{(e)}| < |B|$ and

$$|A_{(e)} \cap [1, n]| + |B_{(e)} \cap [1, n]| \geq \alpha n \text{ for all } 1 \leq n \leq N. \tag{5.7}$$

Note that the constraint $e \in A$ will ensure that both $A_{(e)}$ and $B_{(e)}$ contain 0.

Suppose first that B is not contained in A. Then we can simply choose $e = 0 \in A$, since $B_0 = A \cap B$ would then be strictly smaller than B, and from (5.3) and (5.6) we have

$$|A_0 \cap [1, n]| + |B_0 \cap [1, n]| = |A \cap [1, n]| + |B \cap [1, n]| \geq \alpha n$$

as desired.

Now we consider the harder case when B is contained in A. Here we take

$$e := \min\{a \in A : a + B \not\subseteq A\}.$$

Note the set on the right-hand side is non-empty since the largest element of A clearly belongs to this set. We have $e \in A$; by hypothesis, e is positive, and by construction we have

$$(A \cap [0, e)) + B \subseteq A. \tag{5.8}$$

Also by Lemma 5.2 $B_{(e)}$ is strictly smaller than B. Thus it remains to show (5.7). By (5.3) (and observing that $B \backslash (A - e)$ is disjoint from $[-e + 1, 0]$) we have

$$\begin{aligned} |A_{(e)} \cap [1, n]| + |B_{(e)} \cap [1, n]| &= |A \cap [1, n]| + |B \cap [1, n]| \\ &\quad - |(B \backslash (A - e)) \cap [n - e + 1, n]| \\ &\geq |A \cap [1, n]| + |B \cap [1, n - e]|. \end{aligned}$$

If $B \cap [n - e + 1, n]$ is empty then the claim (5.7) would now follow from (5.6), so we may assume $B \cap [n - e + 1, n]$ is non-empty. Then if we let b be the minimal element in $B \cap [n - e + 1, n]$, then $b \in B \subseteq A$, and also since $e \in A \subseteq [0, N]$ we see that $n - b \leq e - 1 < N$. We can now continue the previous calculation using two applications of (5.8) and the induction hypothesis as

$$\begin{aligned} &|A_{(e)} \cap [1, n]| + |B_{(e)} \cap [1, n]| \\ &\quad \geq |A \cap [1, n]| + |B \cap [1, n - e]| \\ &\quad = |A \cap [1, b - 1]| + 1 + |A \cap [b + 1, n]| + |B \cap [1, b - 1]| \\ &\quad \geq |A \cap [1, b - 1]| + |B \cap [1, b - 1]| + 1 + |((A \cap [0, e)) + B) \cap [b + 1, n]| \\ &\quad \geq \alpha(b - 1) + 1 + |((A \cap [0, e)) + b) \cap [b + 1, n]| \\ &\quad \geq \alpha b + |A \cap [1, n - b]| \\ &\quad \geq \alpha b + |((A \cap [0, e)) + B) \cap [1, n - b]| \\ &\quad \geq \alpha b + |(A + B) \cap [1, n - b]| \\ &\quad \geq \alpha b + \alpha(n - b) \\ &\quad = \alpha n \end{aligned}$$

as desired. □

For further discussion of Mann's theorem and several variants, see [168].

The e-transform method also allows one to characterize when the above inequalities are sharp. We begin with an inverse theorem for Lemma 5.3.

Proposition 5.8 *Let A and B be additive sets in $\mathbf{Z}$ such that $|A|, |B| \geq 2$. Then $|A + B| = |A| + |B| - 1$ if and only if A, B are arithmetic progressions of the same step.*

Proof The "if" part is clear, so we prove the "only if" part. Let $e := \max(A) - \min(B)$. From the proof of Lemma 5.3 we see that we must have

$$A + B = A_{(e)} + B_{(e)} = (A \cup (B + e)) + \min(B) = (A + \min(B)) \cup (B + \max(A)).$$

Now let $\min(B) + v$ be the second smallest element of B, after $\min(B)$; then $v > 0$ and for any $a \in A \backslash \{\max(A)\}$ we have

$$\begin{aligned} a + \min(B) + v \subseteq A + B &= (A + \min(B)) \cup (B + \max(A)) \\ &= (A + \min(B)) \cup (B \backslash \{\min(B)\} + \max(A)). \end{aligned}$$

Note that since $a < \max(A)$ and $\min(B) + v$ is the minimal value of $B \backslash \{\min(B)\}$, than $a + \min(B) + v$ cannot lie in $(B \backslash \{\min(B)\} + \max(A))$. We conclude that

$$a + v \in A \text{ for all } a \notin \max(A).$$

From this it is easy to see that A is an arithmetic progression of step v. In particular $\max(A) - v$ is the second largest value of A after $\max(A)$, and by adapting the previous argument we see that B is also an arithmetic progression of step v, and we are done. □

Now we give an inverse theorem for the Cauchy–Davenport inequality.

Theorem 5.9 (Vosper's theorem) *[375] Let p be a prime, and let A, B be additive sets in $\mathbf{Z}_p$ such that $|A|, |B| \geq 2$ and $|A + B| \leq p - 2$. Then $|A + B| = |A| + |B| - 1$ if and only if A and B are arithmetic progressions with the same step.*

A similar theorem has recently been proven [174] in the case when $|A + B| = |A| + |B|$. A version of Vosper's theorem exists for arbitrary groups Z but is more complicated to state; see [201], [231]. See also Exercise 5.1.11.

Proof The "if" part is easy, so we prove the "only if" part. We first prove this claim when A is an arithmetic progression $\{a, a + v, \ldots, a + nv\}$ for some $n \geq 1$. Then by Cauchy–Davenport

$$\begin{aligned} |B| + n &= |A| + |B| - 1 \\ &= |A + B| \\ &= |\{a, a + v, \ldots, a + (n-1)v\} + \{0, v\} + B| \\ &\geq |B + \{0, v\}| + n - 1, \end{aligned}$$

and hence (by Cauchy–Davenport again) we have $|B + \{0, v\}| = |B| + 1$. Thus B and $B + v$ only differ by at most one element, which implies that B is a progression of length v (see Exercise 3.2.7). By symmetry we have the same claim when the roles of A and B are reversed.

Now we use a duality trick to claim the following variant: if the *sum set* $A + B$ is a proper arithmetic progression, and $|A + B| = |A| + |B| - 1$, then so is A and B, and all three progressions have the same step. To see this, set $C := -(\mathbf{Z}_p \backslash (A + B))$. Then C is also an arithmetic progression with the same step as $A + B$ and with cardinality $|C| = p - |A + B| = p + 1 - |A| - |B| \geq 2$. Observe also that $C + B \subseteq -(\mathbf{Z}_p \backslash A)$, because if any element $-a$ of $-A$ was contained in $C + B$, then

C would intersect $-a - B \subset -(A+B)$, a contradiction. Thus $|C+B| \leq p - |A| = |C| + |B| - 1$, and hence by Cauchy–Davenport $|C+B| = |C| + |B| - 1$. Since C was an arithmetic progression of length at least 2, we see from the previous discussion that B is also, and has the same step as C. Similarly for A.

To summarize, we have now proven Vosper's theorem in the cases when at least one of A, B, or $A+B$ is an arithmetic progression. Now we handle the general case. We induce on the size of B. If $|B| = 2$ then B is an arithmetic progression already, and the claim has already been proved. Now suppose that $|B| > 2$ and the claim has already been proven for smaller B. Suppose first that we can find an $e \in A - B$ such that the e-transform $B_{(e)}$ of B has size $1 < |B_{(e)}| < |B|$. Since $|A+B| = |A| + |B| - 1$; by hypothesis, we see from (5.1), (5.2) and the Cauchy–Davenport inequality that we must have $A_{(e)} + B_{(e)} = A + B$ and

$$|A_{(e)} + B_{(e)}| = |A_{(e)}| + |B_{(e)}| - 1.$$

Using the induction hypothesis, we thus see that $A_{(e)}$ and $B_{(e)}$ are arithmetic progressions with the same step v, and hence $A + B = A_{(e)} + B_{(e)}$ is also an arithmetic progression, and the claim follows by the preceding discussion.

The only remaining case is if we have $|B_{(e)}| = 1$ or $|B_{(e)}| = |B|$ for all $e \in A - B$. But if $E \subseteq A - B$ denotes all the $e \in A - B$ such that $|B_{(e)}| = |B|$, then by Lemma 5.2 we have $B + E \subseteq A$, and hence $|E| \leq |A| - |B| + 1$ by Cauchy–Davenport. Since $|A - B| \geq |A| + |B| - 1$ by Cauchy–Davenport, we thus see that $|B_{(e)}| = 1$ for at least $2|B| - 2$ values of e. Since $B_{(e)}$ is a singleton subset of B, we thus see from the pigeonhole principle that there exists $e, e' \in A - B$ and $b \in B$ such that $B_{(e)} = B_{(e')} = \{b\}$. Since $|A+B| = |A| + |B| - 1$ by hypothesis, we see from (5.1), (5.2) that

$$A + B = A_{(e)} + b = A_{(e')} + b$$

and hence

$$A \cup (B + e) = A \cup (B + e').$$

Since A intersects $B + e$ only in $b + e$, and A intersects $B + e'$ only in $b + e'$, we thus see that $B + e$ and $B + e'$ differ by at most one element. But this forces B to be a progression (of step $e' - e$), and the claim follows. □

We now develop an inverse theorem for sets A, B of integers with fairly small sum set. We need a preliminary lemma.

Lemma 5.10 *Let A be an additive set in $\mathbf{Z}$ such that $0 \in A$, let $N \geq 1$ be an integer, and let $\phi_N : \mathbf{Z} \to \mathbf{Z}_N$ be the canonical quotient map. For each $x \in \phi_N(A)$, let $\mu_x := |\{a \in A : \phi_N(a) = x\}|$ denote the multiplicity of ϕ_N at x, and denote $m := \min_{x \in \phi_N(A) \backslash \{0\}} \mu_x$. Then*

$$|2A| \geq |A| + |\phi_N(A)|(\mu_0 - 2m) + |2\phi_N(A)|(2m - 1)$$

Proof We split (using Lemma 5.3 and the observation $\sum_{x\in\phi_N(A)} \mu_x = |A|$)

$$\begin{aligned}
|2A| &= \sum_{x\in\phi_N(2A)} \big|2A\cap\phi_N^{-1}(\{x\})\big| \\
&\geq \sum_{x\in\phi_N(2A)} \sup_{y,z\in\phi_N(A):y+z=x} \big|\big(A\cap\phi_N^{-1}(\{y\})\big)+\big(A\cap\phi_N^{-1}(\{z\})\big)\big| \\
&\geq \sum_{x\in\phi_N(2A)} \sup_{y,z\in\phi_N(A):y+z=x} \big(\big|A\cap\phi_N^{-1}(\{y\})\big|+\big|A\cap\phi_N^{-1}(\{z\})\big|-1\big) \\
&= \left(\sum_{x\in\phi_N(2A)} \sup_{y,z\in\phi_N(A):y+z=x} \mu_y+\mu_z\right) - |\phi_N(2A)| \\
&\geq \left(\sum_{x\in\phi_N(A)} \mu_0+\mu_x\right)+\left(\sum_{x\in\phi_N(2A)\backslash\phi_N(A)} m+m\right) - |\phi_N(2A)| \\
&= \mu_0\phi_N(A)+|A|+(|\phi_N(2A)|-|\phi_N(A)|)2m-|\phi_N(2A)|
\end{aligned}$$

as desired (noting that $2\phi_N(A)=\phi_N(2A)$). □

Now we give the inverse theorem.

Theorem 5.11 **($3k-3$ theorem)** *[116] Let A be an additive set in $\mathbf{Z}$ such that $|2A|<3|A|-3$. Then there exists a proper arithmetic progression $P = a+[0,|2A|-|A|]\cdot v$ of length $|2A|-|A|+1$ that contains A.*

Proof We use an argument from [233]. By translating A we may assume that $\min(A)=0$. We may also assume that the set A has no common divisor $d>1$, since otherwise we could replace A by $\frac{1}{d}\cdot A$. We will assume that $|A|\geq 3$ as the cases $|A|=1,2$ can be verified directly.

Write $N := \max(A)$, thus $A\subseteq[0,N]$ and $0,N\in A$. It will suffice to show that $N\leq|2A|-|A|$. Suppose for contradiction that $N>|2A|-|A|$. We now apply Lemma 5.10. Observe in this case that $\mu_0=2$ and $m=1$, and hence

$$|2A|\geq|A|+|2\phi_N(A)|. \tag{5.9}$$

Since we are assuming $N>|2A|-|A|$, we conclude that

$$|2\phi_N(A)|<N. \tag{5.10}$$

By Exercise 2.1.6 and the hypothesis $|2A|<3|A|-3$ we have

$$|2\phi_N(A)|<2|A|-3=2|\phi_N(A)|-1.$$

If N were prime then we could apply the Cauchy–Davenport inequality to conclude the desired contradiction. But in general we must rely instead on Kneser's theorem. Let $H := \mathrm{Sym}_1(2\phi_N(A))$, then by Kneser's theorem we have

$$|2\phi_N(A)|\geq 2|\phi_N(A)+H|-|H|$$

and hence if we set $k := |\phi_N(A) + H| - |\phi_N(A)|$, then

$$0 \leq k \leq \frac{|H| - 2}{2}. \tag{5.11}$$

In particular $|H| \geq 2$. Also from (5.10) we have $|H| < N$. Since H is a subgroup of $\mathbf{Z}_N$, we see that $H = (h \cdot Z)/(N \cdot Z)$ for some $2 \leq h < N$ which is a factor of N.

Note that $\phi_N(A)$ contains zero, but cannot be contained entirely inside H as this would mean that A has a common divisor of h, contradicting our hypothesis. So we know that $\phi_N(A)$ contains at least two cosets of H, or equivalently that $|\phi_h(A)| \geq 2$.

Now we apply Lemma 5.10 again, but with N replaced by h. From (5.11) we see that if $x + H \subseteq \mathbf{Z}_N$ is any non-trivial coset of H, then $H \cup (x + H)$ intersects $\phi_N(A)$ in at least $2|H| - k$ points; since $\phi_N(0) = \phi_N(N) = 0 \in H \cup (x + H)$, this implies that $\phi_N^{-1}(H \cup (x + H)) = \phi_h^{-1}(\{0, x \bmod h\})$ intersects A in at least $2|H| - k + 1$ points. In other words we have

$$\mu_0 + m \geq 2|H| - k + 1.$$

A similar argument gives

$$m \geq |H| - k.$$

But since H was the symmetry group of $2\phi_N(A)$, we see that $2\phi_h(A)$ has trivial symmetry group; furthermore from (5.10) we see that $|2\phi_h(A)| < h$. Thus by Kneser's theorem we have $|2\phi_h(A)| \geq 2|\phi_h(A)| - 1$. Inserting all these facts into Lemma 5.10, we obtain

$$\begin{aligned}
|2A| &\geq |A| + |\phi_h(A)|(\mu_0 - 2m) + (2|\phi_h(A)| - 1)(2m - 1) \\
&\geq |A| + |\phi_h(A)|(2|H| - k - 3m + 1) + (2|\phi_h(A)| - 1)(2m - 1) \\
&= |A| + |\phi_h(A)|(2|H| - k - 1) + (|\phi_h(A)| - 2)m + 1 \\
&\geq |A| + |\phi_h(A)|(2|H| - k - 1) + (|\phi_h(A)| - 2)(|H| - k) + 1 \\
&= |A| + 3|\phi_h(A)||H| - 2k|\phi_h(A)| - |\phi_h(A)| - 2|H| + 2k + 1 \\
&\geq |A| + 3|\phi_h(A)||H| - (|H| - 2)|\phi_h(A)| - |\phi_h(A)| - 2|H| + 2k + 1 \\
&= |A| + 2|\phi_h(A)||H| + |\phi_h(A)| - 2|H| + 2k + 1 \\
&= |A| + 2(|A| + k) + |\phi_h(A)| - 2|H| + 2k + 1 \\
&= 3|A| + |\phi_h(A)| - 2|H| + 4k - 1 \\
&\geq 3|A| + 2 - 2|H| + 4\frac{|H| - 2}{2} - 1 \\
&\geq 3|A| - 3
\end{aligned}$$

which contradicts the hypothesis $2|A| < 3|A| - 3$. □

Note that we have used a result on torsion groups to imply a result in the torsion-free case; this phenomenon will also come up in later proofs of Freiman's theorem. The original proof of Freiman was somewhat different; see [116], [257]. A treatment of the case $|2A| = 3|A| - 3$ appears in [113], [28]. For some partial progress in the case $|2A| = 3|A| + o(|A|)$, see [193]. There has also been much work on generalizing the $3k - 3$ theorem to pairs of sets [111], [336], [333], [233]. For instance one has the following result.

Theorem 5.12 *[233] Let A, B be additive sets in $\mathbf{Z}$ such that $|A + B| < |A| + |B| + \min(|A|, |B|) - 3$. Then A is contained in an arithmetic progression of length at most $|A + B| - |B| + 1$ and B is contained in an arithmetic progression of length at most $|A + B| - |A| + 1$, where both progressions have the same difference.*

For some further refinements to this theorem, see [233].

Exercises

5.1.1 Prove the remaining claims in Corollary 5.6.

5.1.2 Prove Lemma 5.1.

5.1.3 Show that Kneser's theorem implies Lemma 5.3 and the Cauchy–Davenport inequality.

5.1.4 [211] Let A, B be additive sets in an ambient group. Show that if $|A + B| < |A| + |B|$ then

$$|A+B| = |A+\mathrm{Sym}_1(A+B)|+|B+\mathrm{Sym}_1(A+B)| - |\mathrm{Sym}_1(A+B)|.$$

5.1.5 [244] Let A, B be additive sets in an ambient group Z such that $|A + B| < |A| + |B| - 1$. Show that $|(A + \mathrm{Sym}_1(A + B))\backslash A| < |\mathrm{Sym}_1(A + B)| - 1$; thus A is rather close to being a union of cosets of $\mathrm{Sym}_1(A + B)$.

5.1.6 [243] If A is a (possibly infinite) set of integers, define the *Schnirelmann density* $\sigma(A)$ of A to be the quantity

$$\sigma(A) := \inf_{N>0} \mathbf{E}_{x\in[1,N]}(x \in A) = \inf_{N>0} \frac{|A \cap [1, N]|}{|[1, N]|}.$$

(Note that this is distinct from the lower density $\underline{\sigma}(A)$ defined in Definition 1.21, due to the use of the inf rather than the lim inf.) Show that if A and B are any sets of integers with $0 \in A, B$, then $\sigma(A + B) \geq \min(\sigma(A) + \sigma(B), 1)$. (Hint: use Theorem 5.7.) Conclude that if $0 \in A$ and $\sigma(A) \geq 1/k$ for some integer $k > 0$, then $kA \subset \mathbf{Z}^+$. Thus every set of integers of positive Schirelmann density that contains 0 is a basis for the positive integers.

5.1.7 [312] Let A, B be sets of integers such that $1 \in A$ and $0 \in B$. Show that $\sigma(A+B) \geq \sigma(A) + \sigma(B) - \sigma(A)\sigma(B)$, where $\sigma()$ is the Schnirelmann density from Exercise 5.1.6. (Hint: order the positive elements of A as $a_1 < a_2 < \cdots$, and observe that $|(A+B) \cap [a_n, a_{n+1})| \geq 1 + |B \cap [1, a_{n+1} - a_n - 1]|$.)

5.1.8 [311], [201], [202] Let A and B be additive sets in an ambient group Z. Prove that

$$|A+B| \geq |A| + |B| - \min_{c \in A+B} |\{(a,b) \in A+B : a+b=c\}|.$$

(This can be done either by Kneser's theorem, or more directly via the e-transform method.)

5.1.9 Let p be a prime, let $N \geq 1$, and let $A_1, \ldots, A_N$ be additive sets in $\mathbf{Z}_p$ such that $|A_1| + \cdots + |A_N| = p + N - 1$. Use the Cauchy–Davenport inequality to show that $A_1 + \cdots + A_N = p$. Conversely, show that this statement can be used to imply the Cauchy–Davenport inequality.

5.1.10 What happens if one extends Theorem 5.9 to cover the cases $|A| = 1$, $|B| = 1$, or $|A+B| = p-1$? (The case $|A+B| = p$ is much more difficult to analyze and does not have as simple a characterization.)

5.1.11 [201] Let A, B be additive sets in ambient group Z such that $|A|, |B| > 1$, $|\mathrm{Sym}_1(A+B)| = 1$, and $|A+B| < |A| + |B|$. By analyzing the proof of Kneser's theorem (and Vosper's theorem) carefully, show that $A+B$ is either equal to an arithmetic progression, or there exists a finite subgroup G of Z such that $A+B$ consists of one or more cosets of G, and possibly a subset of one other coset of G. (Compare with Exercise 5.1.5 and Exercise 3.2.7.)

5.1.12 [242] Let A, B be open subsets of the torus $(\mathbf{R}/\mathbf{Z})^d$. Prove the *Mann–Kneser–Macbeath inequality* $\mathrm{mes}(A+B) \geq \min(\mathrm{mes}(A) + \mathrm{mes}(B), 1)$, where mes() denotes the usual Haar measure on the torus. (Hint: discretize the torus to $(\mathbf{Z}/p\mathbf{Z})^d$ for some large prime p, apply Kneser's theorem, and then take limits.) Give examples to show that this inequality cannot be improved. One can extend this result to arbitrary measurable subsets of the torus with some additional analytic arguments. See [27] for some recent developments concerning this inequality. This inequality should be contrasted with the Brunn–Minkowski inequality (Theorem 3.16), and shows that sum sets in $(\mathbf{R}/\mathbf{Z})^d$ and sum sets in $\mathbf{R}^d$ behave slightly differently.

5.1.13 [116] Let $N \geq 0$ be an integer, and let A, B be non-empty subsets of $[0, N]$ such that $0, N \in A$ and $|A| + |B| \geq N + 3$. Prove that $|A+B| \geq |B| + N$.