Combinatorics, Probability and Computing (2010) 19, 603–639. © Cambridge University Press 2010
doi:10.1017/S0963548309990642

Sumset and Inverse Sumset Theory for Shannon Entropy

TERENCE TAO[†]

Department of Mathematics, UCLA, Los Angeles CA 90095-1555, USA
(e-mail: tao@math.ucla.edu)

Received 24 June 2009; revised 14 November 2009; first published online 22 January 2010

Let $G = (G, +)$ be an additive group. The sumset theory of Plünnecke and Ruzsa gives several relations between the size of sumsets $A + B$ of finite sets A, B, and related objects such as iterated sumsets kA and difference sets $A - B$, while the inverse sumset theory of Freiman, Ruzsa, and others characterizes those finite sets A for which $A + A$ is small. In this paper we establish analogous results in which the finite set $A \subset G$ is replaced by a discrete random variable X taking values in G, and the cardinality $|A|$ is replaced by the Shannon entropy $\mathbf{H}(X)$. In particular, we classify those random variables X which have small doubling in the sense that $\mathbf{H}(X_1 + X_2) = \mathbf{H}(X) + O(1)$ when X_1, X_2 are independent copies of X, by showing that they factorize as $X = U + Z$, where U is uniformly distributed on a coset progression of bounded rank, and $\mathbf{H}(Z) = O(1)$.

When G is torsion-free, we also establish the sharp lower bound $\mathbf{H}(X + X) \geqslant \mathbf{H}(X) + \frac{1}{2}\log 2 - o(1)$, where $o(1)$ goes to zero as $\mathbf{H}(X) \to \infty$.

1. Introduction

The purpose of this paper is to establish analogues of the Plünnecke–Ruzsa–Freiman sumset and inverse sumset theory for finite subsets of discrete additive groups, in the setting of discrete random variables in such groups.

1.1. Sumset and inverse sumset theory: a quick review

To motivate our results we begin by recalling some of the key results in sumset and inverse sumset theory. Let $G = (G, +)$ be an additive group. For any finite non-empty sets A, B in G, we define the sumset

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

[†] Supported by NSF Research Award DMS-0649473, the NSF Waterman award and a grant from the MacArthur Foundation.

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

the difference set

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

and the iterated sumsets $2A = A + A$, $3A = A + A + A$, *etc.* We use $|A|$ to denote the cardinality of a finite set A.

We have the trivial bounds

$$|A|, |B| \leqslant |A + B| \leqslant |A||B|, \tag{1.1}$$

and similarly for $A - B$. In particular, we see that the *doubling constant*

$$\sigma[A] := \frac{|A + A|}{|A|}$$

is at least one. It is easy to see that this doubling constant is precisely one if and only if A is the translate of a finite subgroup of G. Intuitively, one thus expects that if the doubling constant of A is bounded, then A should in some sense *behave like* a translate of a finite subgroup; this is one of the main objectives of the Plünnecke–Ruzsa sumset theory. One is furthermore interested in *classifying* those sets A of small doubling constant; this is the main objective of the Freiman–Ruzsa inverse sumset theory.

We now give some sample results in this theory. One of the simplest is the *Ruzsa triangle inequality*

$$|A - C| \leqslant \frac{|A - B||B - C|}{|B|}, \tag{1.2}$$

valid for all non-empty finite subsets A, B, C of G (see, *e.g.*, [9] and [12, Lemma 2.6]). In a similar spirit,

$$|A + B| \leqslant \frac{|A - B|^3}{|A||B|} \tag{1.3}$$

(see, *e.g.*, [9] and [12, Corollary 2.12]). If $\sigma[A] \leqslant K$, then one has the *Plünnecke–Ruzsa inequalities*

$$|nA - mA| \leqslant K^{n+m}|A| \tag{1.4}$$

for all $n, m \geqslant 1$ (see, *e.g.*, [9] and [12, Corollary 6.28]). We refer the reader to [9] or [12] for further details of these and related estimates.

Another basic result is the *Balog–Szemerédi–Gowers lemma* [3, 4], which involves *partial sumsets*

$$A \overset{E}{+} B := \{a + b : (a, b) \in E\}$$

for any subset E of $A \times B$:

Lemma 1.1 (Balog–Szemerédi–Gowers lemma). *Suppose that A, B are non-empty finite subsets of an additive group G, and let $E \subset A \times B$ be such that $|E| \geqslant |A||B|/K$ and $|A \overset{E}{+} B| \leqslant K|A|^{1/2}|B|^{1/2}$ for some $K \geqslant 1$. Then there exist subsets $A' \subset A, B' \subset B$ with $|A'| \gg |A|/K$, $|B'| \gg |B|/K$ such that $|A' + B'| \ll K^7|A'|^{1/2}|B|^{1/2}$.*

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Here and below, we use $X \ll Y$ or $X = O(Y)$ to denote the estimate $|X| \leqslant CY$ for some absolute constant Y, and $X \asymp Y$ as shorthand for $X \ll Y \ll X$. If we need the implied constant C to depend on a parameter, we will indicate this by subscripts; thus, for instance, $O_K(1)$ denotes a quantity bounded in magnitude by C_K for some C_K depending only on K.

Proof. See [12, Theorem 2.29]. □

Now we turn to inverse theorems. A basic concept here is that of a *coset progression*, which unifies the concept of a coset and of an arithmetic progression.

Definition 1.2 (Coset progression [5]). A *coset progression* in an additive group is any set of the form $H + P$, where H is a finite subgroup of G, and P is a generalized arithmetic progression, *i.e.*, a set of the form

$$P := \{x + n_1 r_1 + \cdots + n_d r_d : n_1 \in [0, N_1), \ldots, n_d \in [0, N_d)\},$$

where $d \geqslant 0$ is an integer, $x, r_1, \ldots, r_d$ lie in G, $N_1, \ldots, N_d \geqslant 1$ are integers, and $[0, N) := \{0, \ldots, N-1\}$. We call d the *rank* of the progression. We say that the coset progression is *t-proper* for some $t > 0$ if the sums $h + x + n_1 r_1 + \cdots + n_d r_d$ for $h \in H$ and $n_i \in [0, tN_i)$ are distinct, and *proper* if it is 1-proper.

It is easy to see that a coset progression of rank d has doubling constant at most 2^d. More generally, if A is a subset of a coset progression $H + P$ with $|A| \geqslant |H + P|/K$, then A has doubling constant at most $2^d K$. The following Freiman-type theorem, first proved in [5], establishes a partial converse to this claim.

Theorem 1.3 (Green–Ruzsa Freiman theorem). *Let G be an additive group, and let $A \subset G$ be a finite non-empty set with $\sigma[A] \leqslant K$ for some $K \geqslant 1$. Then there exists a coset progression $H + P$ of rank $O(K)$ and size $|H + P| \leqslant \exp(O(K^{O(1)}))|A|$ such that $A \subset H + P$.*

Proof. See [12, Theorem 5.44]. □

1.2. Shannon entropy

We now turn to the concept of Shannon entropy.

Definition 1.4 (Shannon entropy). Let A be a (discrete) set. Let $\mathrm{Pr}_c(A)$ denote the set of all probability measures on A with compact (*i.e.*, finite) support, or equivalently a function $p : A \to [0, 1]$ which is non-zero for only finitely many values, and adds up to one. Define an *A-random variable* to be a random variable X taking values in a finite subset $\mathrm{range}(X) := \{x \in A : \mathbf{P}(x \in X) \neq 0\}$, thus the distribution function $p_X(x) := \mathbf{P}(x \in X)$ of X lies in $\mathrm{Pr}_c(A)$. We write $X \equiv Y$ if $p_X = p_Y$, *i.e.*, if X, Y have the same distribution. We refer to random variables taking values in a finite set as *discrete random variables*.

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

The *Shannon entropy* $\mathbf{H}(p)$ of a probability distribution $p \in \Pr_c(A)$ is given by the formula

$$\mathbf{H}(p) := \sum_{x \in A} F(p(x)), \tag{1.5}$$

where $F : \mathbf{R}^+ \to \mathbf{R}^+$ is the function

$$F(x) := x \log \frac{1}{x}, \tag{1.6}$$

with the convention that $F(0) = 0$. Given an A-random variable X, we then define $\mathbf{H}(X) := \mathbf{H}(p_X)$.

The basic theory of Shannon entropy is reviewed in the Appendix. For now, we just remark that

$$0 \leqslant \mathbf{H}(X) \leqslant \log |\mathrm{range}(X)| \tag{1.7}$$

for any discrete random variable X, with equality in the former inequality if and only if X is deterministic (*i.e.*, it only takes on one value), and equality in the latter inequality if and only if it is uniformly distributed in $\mathrm{range}(X)$; see Lemma A.1 for a more precise statement. In particular, a *boolean* random variable (*i.e.*, one that takes values in $\{0,1\}$) has entropy at most $\log 2$ with our choice of normalization of entropy. One can view G-random variables as a generalization of the concept of a finite non-empty subset of G, in which the weight (or probability) assigned to each element in the range is not necessarily uniform.

Given two G-random variables X, Y (not necessarily independent), their sum $X + Y$ and $X - Y$ are also G-random variables, and $\mathrm{range}(X \pm Y) \subset \mathrm{range}(X) \pm \mathrm{range}(Y)$. From standard entropy inequalities, one has the trivial upper bound

$$\mathbf{H}(X \pm Y) \leqslant \mathbf{H}(X) + \mathbf{H}(Y), \tag{1.8}$$

while if X and Y are independent, one also has the trivial lower bound

$$\mathbf{H}(X), \mathbf{H}(Y) \leqslant \mathbf{H}(X \pm Y); \tag{1.9}$$

see Lemma 2.1. The lower bound (1.9) can of course fail if the independence hypothesis is dropped; for instance, one clearly has $\mathbf{H}(X - X) = 0$.

We can define the *doubling constant* $\sigma[X]$ of a G-random variable by the formula

$$\sigma[X] := \exp(\mathbf{H}(X_1 + X_2) - \mathbf{H}(X)),$$

where X_1, X_2 are independent copies of X, and thus $\sigma[X] \geqslant 1$ by (1.9). This quantity is related, but not identical, to the doubling constant $\sigma[A]$ of a set; indeed, from (1.7) we see that

$$\sigma[X] \leqslant \sigma[A] \tag{1.10}$$

whenever X is uniformly distributed on a finite non-empty subset A of G. However, the doubling constant of a random variable can be significantly smaller than that of its range. For instance, let A be an interval $[0, N)$ together with $\sqrt{N}$ (say) other integers in general

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

position, where N is large. Then the doubling constant of A is about $\sqrt{N}$, but the uniform distribution on A has doubling constant $O(1)$. Thus we see that a small amount of 'noise' (such as the $\sqrt{N}$ integers in general position) can significantly increase the doubling constant of a set, but have only a negligible impact on the doubling constant of a random variable. Heuristically, one can thus think of entropy sumset theory as a 'noise-tolerant' analogue of combinatorial sumset theory.

The analogue of a partial sumset $A \overset{E}{+} B$ here is the concept of a sum $X + Y$ of *non-independent* random variables X, Y. For instance, if $E \subset A \times B$ is a non-empty set, and (X, Y) is the random variable chosen uniformly at random from E, then $X + Y$ is a random variable ranging in $A \overset{E}{+} B$.

There are several ways to define the *distance* between two G-random variables X, Y (or their associated distributions p_X, p_Y). For instance, we can define their *total variation distance*

$$\mathrm{dist}_{TV}(X, Y) = \mathrm{dist}_{TV}(p_X, p_Y) := \sum_{x \in G} |p_X(x) - p_Y(x)|; \tag{1.11}$$

this is clearly a metric on $\mathrm{Pr}_c(G)$. Another useful distance is the *Rusza distance*

$$\mathrm{dist}_R(X, Y) = \mathrm{dist}_R(p_X, p_Y) := \mathbf{H}(X' - Y') - \frac{1}{2}\mathbf{H}(X') - \frac{1}{2}\mathbf{H}(Y'), \tag{1.12}$$

where X', Y' are independent copies of X, Y respectively; this is not quite a metric (in particular, $\mathrm{dist}_R(X, X) > 0$ in general), but does obey the triangle inequality, as we will see in Theorem 1.7 below. A third distance of importance to us is the following *transport distance*.

Definition 1.5 (Transport metric). Let G be an additive group, and let X, Y be G-random variables. We define the *entropy transport distance* $\mathrm{dist}_{\mathrm{tr}}(X, Y)$ from X to Y to be the infimum of $\mathbf{H}(Z)$, where Z ranges over all G-random variables (not necessarily independent of X) such that $X + Z \equiv Y$.

Observe that $\mathrm{dist}_{\mathrm{tr}}(X, Y) = 0$ if and only if Y has the distribution of a translate $X + c$ of X. Up to this equivalence, it is easy to see that the transport distance is indeed a metric. The notion of transport metric depends only on the distribution, so by abuse of notation we may define $\mathrm{dist}_{\mathrm{tr}}(p_X, p_Y) := \mathrm{dist}_{\mathrm{tr}}(X, Y)$. The notion of two random variables being close in transport metric is roughly analogous to the notion of (mutual) K*-control* of one set by another, introduced in [11].

Example 1.6. Let N be a large even integer, let X be the uniform distribution on $[0, N)$, and let Y be the uniform distribution on the even numbers in $[0, N)$. Then the total variation distance $\mathrm{dist}_{TV}(X, Y)$ is quite large (comparable to its maximal value of 2). On the other hand, the Ruzsa distance is quite small (of size $O(1)$). The transport distance is also of size $O(1)$; indeed, one can transport Y to X by adding a uniform boolean variable $Z \in \{0, 1\}$ which is independent of Y; conversely, one can transport X to Y by subtracting off the parity bit Z of X (which is *not* independent of X). In fact, the uniform

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

distribution on any dense subset of $[0, N)$ lies within $O(1)$ of X in the transport distance, although this is not as obvious to see; see Corollary 4.6 below.

The Ruzsa distance, doubling constant, and transport distance interact well with each other. For instance, we have the identity

$$\sigma[X] = \exp(\mathrm{dist}_R(X, -X)) \tag{1.13}$$

and the Lipschitz-type properties

$$|\mathrm{dist}_R(X', Y') - \mathrm{dist}_R(X, Y)| \leqslant \frac{3}{2}(\mathrm{dist}_{\mathrm{tr}}(X, X') + \mathrm{dist}_{\mathrm{tr}}(Y, Y')) \tag{1.14}$$

for any G-random variables X, Y, X', Y', as can be seen by several applications of (1.8). In particular,

$$|\log \sigma[X] - \log \sigma[X']| \leqslant 3\,\mathrm{dist}_{\mathrm{tr}}(X, X'). \tag{1.15}$$

Thus we see that random variables which are close in transport distance are essentially equivalent from the perspective of their sumset theory.

1.3. Main results

We can now state our main results. We begin with some sumset estimates, analogous to (1.2), (1.3) and (1.4).

Theorem 1.7 (Entropy sumset estimates). *Let G be an additive group, and let X, Y, Z be G-random variables.*

- (*Ruzsa triangle inequality.*) *We have*

$$\mathrm{dist}_R(X, Z) \leqslant \mathrm{dist}_R(X, Y) + \mathrm{dist}_R(Y, Z). \tag{1.16}$$

- (*Sum-difference inequality.*) *We have*

$$\mathrm{dist}_R(X, -Y) \leqslant 3\,\mathrm{dist}_R(X, Y). \tag{1.17}$$

- (*Weak Plünnecke–Ruzsa inequality.*) *If $X_1, \ldots, X_n, X'_1, \ldots, X'_m$ are independent copies of X for some integers $n, m \geqslant 0$, then*

$$\mathbf{H}(X_1 + \cdots + X_n - X'_1 - \cdots - X'_m) \leqslant \mathbf{H}(X) + O((n + m)\log \sigma[X]). \tag{1.18}$$

We prove these estimates in Section 2. The estimate (1.18) loses an absolute constant in comparison to (the logarithm of) (1.4). This is because we do not know how to adapt the graph-theoretic Plünnecke inequality [8] to the entropy setting, and so must rely instead on some weaker but less deep arguments in [12] to establish results analogous to (1.4) instead.

The analogue of the Balog–Szemerédi theorem is a little more technical to state, requiring the concept of *conditional entropy* and *conditionally independent trials*, and will be deferred to Section 3.

We now turn to an inverse theorem for entropy in the spirit of Theorem 1.3.

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Theorem 1.8 (Inverse sumset theorem). *Let G be an additive group, and let X be a G-random variable.*

(i) $\sigma[X] = 1$ *if and only if X is the uniform distribution on a coset of a finite subgroup of G.*
(ii) *If $\sigma[X] \leqslant K$, then there exists a coset progression $H + P$ of rank $O_K(1)$ such that $\mathrm{dist}_{\mathrm{tr}}(X, U) \ll_K 1$, where U is the uniform distribution on $H + P$.*
(iii) *If $\mathrm{dist}_R(X, Y) \leqslant K$, then $\mathrm{dist}_{\mathrm{tr}}(X, Y) \ll_K 1$ and $\sigma[X] \leqslant K^4$.*

Note that the uniform distribution on a coset progression $H + P$ of rank d has doubling constant at most 2^d, by (1.10), so from (1.15) we obtain a partial converse to (ii): if $\mathrm{dist}_{\mathrm{tr}}(X, U) \leqslant K$, where U is the uniform distribution on a coset progression of rank at most K, then $\sigma[X] \ll_K 1$. Similarly, (1.14) gives a partial converse to (iii). Thus, up to constants, Theorem 1.8 gives a satisfactory description of random variables with small doubling constant or Ruzsa distance.

The implied constants in Theorem 1.8 can be explicitly computed from the proof, but are rather poor (being triple exponential in K). We will not attempt to optimize these constants here.

We prove Theorem 1.8 in Section 5.

1.4. The torsion-free case

When G is a torsion-free group (thus $nx \neq 0$ for all $x \neq 0$ in G and all integers $n > 0$), then the trivial doubling estimate $\sigma[A] \geqslant 1$ can be improved. Indeed,

$$|A + A| \geqslant 2|A| - 1$$

for any finite non-empty subset A of a torsion-free group G, since A can be mapped onto the integers by a Freiman isomorphism (see [12, Lemma 5.25]). In other words,

$$\sigma[A] \geqslant 2 - \frac{1}{|A|}. \tag{1.19}$$

The example of an arithmetic progression (*e.g.*, $A = [0, n)$) shows that this estimate is sharp.

One can ask whether the same statement holds for entropy. The following example shows that this is not quite the case. Let n be a large integer, and let X_n be the sum of n independent Bernoulli variables $\epsilon_1, \ldots, \epsilon_n \in \{-1, +1\}$ with an equal probability of each. From the central limit theorem (or Stirling's formula), we know that X_n is approximately distributed like a Gaussian of mean zero and variance n, thus

$$p_{X_n}(m) \approx \frac{1}{\sqrt{2\pi n}} e^{-m^2/2n}.$$

Approximating the Riemann sum by an integral, we then expect

$$\mathbf{H}(X_n) \approx \int_{\mathbf{R}} F\left(\frac{1}{\sqrt{2\pi n}} e^{-x^2/2n}\right) dx = \log \sqrt{2\pi n} + \frac{1}{2}.$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

It is not hard to make this heuristic precise, and obtain the asymptotic

$$\mathbf{H}(X_n) = \log \sqrt{2\pi n} + \frac{1}{2} + o(1).$$

In particular, since $X_n + X'_n \equiv X_{2n}$ if X'_n is an independent copy of X_n, we see that

$$\sigma[X_n] = \sqrt{2} - o(1),$$

which is less than one might have predicted from (1.19) and (1.10). The point is that in the entropy setting one can construct 'approximate Gaussian' counterexamples whose closest analogue in the combinatorial setting, namely the arithmetic progressions, are less efficient by a constant factor.

It should not be surprising to experts in information theory that this Gaussian-type bound is best possible.

Theorem 1.9. *If $\varepsilon > 0$, G is torsion-free, and X is a G-random variable, then*

$$\sigma(X) \geqslant \sqrt{2} - \varepsilon,$$

provided $\mathbf{H}(X)$ is sufficiently large depending on ε.

In asymptotic notation, Theorem 1.9 asserts that

$$\sigma(X) \geqslant \sqrt{2} - o_{\mathbf{H}(X)\to\infty}(1).$$

We prove Theorem 1.9 in Section 6 below. This result combines the inverse theorem in Theorem 1.8 with an analogous inequality concerning the Shannon entropy

$$\mathbf{H}_{\mathbf{R}}(X) := \int_{\mathbf{R}} F(p_X(x))\, dx$$

of a *continuous* random variable X taking values in $\mathbf{R}$, namely

$$\mathbf{H}_{\mathbf{R}}(S+T) \geqslant \frac{1}{2}(\mathbf{H}_{\mathbf{R}}(S) + \mathbf{H}_{\mathbf{R}}(T)) + \frac{1}{2}\log 2 \tag{1.20}$$

for independent continuous random variables S, T (see [1, Theorem 2]). The inverse sumset theory is necessary in order to approximate the discrete random variable by a continuous one in a certain sense.

In [1], the continuous entropy inequality

$$\mathbf{H}_{\mathbf{R}}(X_1 + \cdots + X_{n+1}) \geqslant \mathbf{H}_{\mathbf{R}}(X_1 + \cdots + X_n) + \log \frac{\sqrt{n+1}}{\sqrt{n}}$$

was established, where $X_1, \ldots, X_{n+1}$ were independent copies of the same continuous random variable. In view of Theorem 1.9, it is thus natural to conjecture that

$$\mathbf{H}(X_1 + \cdots + X_{n+1}) \geqslant \mathbf{H}(X_1 + \cdots + X_n) + \log \frac{\sqrt{n+1}}{\sqrt{n}} - \varepsilon \tag{1.21}$$

for any $\varepsilon > 0$ and any G-random variable X, if G is torsion-free and $\mathbf{H}(X)$ is sufficiently large depending on n, ε. Unfortunately we were not able to establish this because the inverse theorem is not applicable in this setting, nevertheless we believe (1.21) to be true.

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Finally, we remark that a number of additional entropy sumset inequalities were recently established in [7]. For instance, it was shown that

$$\mathbf{H}(X+Y+Z) \leqslant \frac{1}{2}(\mathbf{H}(X+Y)+\mathbf{H}(Y+Z)+\mathbf{H}(Z+X))$$

for independent G-random variables X, Y, Z, which is an entropy analogue of the inequality

$$|A+B+C| \leqslant |A+B|^{1/2}|B+C|^{1/2}|C+A|^{1/2}$$

(see, *e.g.*, [6] or [2] for a proof). However, these bounds are primarily of interest in the regime where the doubling constants of the sets involved are large, and so are not directly related to the ones presented here.

2. Sumset estimates

In this section we establish the various sumset estimates claimed in the Introduction, and in particular establish Theorem 1.7. The main tools will be entropy inequalities (in particular the submodularity inequality, Lemma A.2), elementary arithmetic identities, and independent and conditionally independent trials.

Readers who are familiar with the combinatorial analogues of these inequalities are invited to 'pretend' that all of the random variables below are uniformly distributed on various finite sets, and in particular on finite groups, in order to see the analogy between both the statements and the proofs of the combinatorial and the entropy estimates. Indeed, the arguments here were discovered by the reverse of this procedure, in which the author searched for the nearest entropy-theoretic analogue to each step in the combinatorial arguments. For instance, if the combinatorial argument required one to pick an object a from a finite set A, the entropy-based argument would instead consider an analogous random variable that was naturally associated to A; if the combinatorial argument required two objects to be related in some way, this usually manifested itself as a coupling of random variables (*e.g.*, by the use of conditionally independent trials); and so forth.

We begin with the trivial sum set estimates.

Lemma 2.1 (Trivial sumset estimate). *If X, Y are two G-random variables, and Z is a discrete random variable, then*

$$\mathbf{H}(X+Y|Z) \leqslant \mathbf{H}(X|Z)+\mathbf{H}(Y|Z).$$

If, furthermore, X, Y are conditionally independent relative to Z, then

$$\max(\mathbf{H}(X|Z), \mathbf{H}(Y|Z)) \leqslant \mathbf{H}(X+Y|Z).$$

In particular we have the inequalities (1.8), (1.9), *and* $\mathrm{dist}_R(X,Y) \geqslant 0$ *for all G-random variables X, Y.*

Proof. By conditioning on Z we may assume that Z is deterministic, thus the task reduces to showing (1.8) and (1.9). The former inequality follows from (A. 11) and (A. 12)

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

since (X, Y) determines $X + Y$. To prove the latter inequality, observe from (A. 10), (A. 16), and the independence of X, Y that

$$\mathbf{H}(X + Y) \geqslant \mathbf{H}(X + Y|Y) = \mathbf{H}(X|Y) = \mathbf{H}(X),$$

and similarly $\mathbf{H}(X + Y) \geqslant \mathbf{H}(Y)$, and the claim follows. □

Now we establish the Ruzsa triangle inequality (1.16), which we rewrite as

$$\mathbf{H}(X - Z) \leqslant \mathbf{H}(X - Y) + \mathbf{H}(Y - Z) - \mathbf{H}(Y),$$

where X, Y, Z are independent G-random variables. Observe that $(X - Y, Y - Z)$ and (X, Z) both determine $X - Z$, while $(X - Y, Y - Z)$ and (X, Z) jointly determine (X, Y, Z). By the submodularity inequality (Lemma A.2), we conclude that

$$\mathbf{H}(X, Y, Z) + \mathbf{H}(X - Z) \leqslant \mathbf{H}(X - Y, Y - Z) + \mathbf{H}(X, Z).$$

Applying (A. 11) and the independence hypotheses, we obtain the claim.

To prove (1.17), we introduce the idea of *conditionally independent trials*. Given two random variables X, Y (not necessarily independent), we can produce two conditionally independent trials X_1, X_2 of X relative to Y, defined by declaring $(X_1|Y = y)$ and $(X_2|Y = y)$ to be independent trials of $(X|Y = y)$ for all $y \in \text{range}(Y)$, thus in particular $X_1 \equiv X_2 \equiv X$, and X_1, X_2 are conditionally independent relative to Y. Observe from conditional independence that

$$\mathbf{H}(X_1, X_2|Y) = \mathbf{H}(X_1|Y) + \mathbf{H}(X_2|Y) = 2\mathbf{H}(X|Y),$$

and thus

$$\mathbf{H}(X_1, X_2, Y) = 2\mathbf{H}(X, Y) - \mathbf{H}(Y). \tag{2.1}$$

Let X, Y be independent G-random variables. Let $(X_1, Y_1), (X_2, Y_2)$ be conditionally independent trials of (X, Y) relative to $X - Y$; since (X, Y) determines $X - Y$, we conclude that $X_1 - Y_1 = X_2 - Y_2$. Let (X_3, Y_3) be another trial of (X, Y), independent of X_1, X_2, Y_1, Y_2; then we have the identity

$$X_3 + Y_3 = (X_3 - Y_2) - (X_1 - Y_3) + X_2 + Y_1.$$

Thus $(X_3 - Y_2, X_1 - Y_3, X_2, Y_1)$ and (X_3, Y_3) each determine $X_3 + Y_3$, while $(X_3 - Y_2, X_1 - Y_3, X_2, Y_1)$ and (X_3, Y_3) together determine $(X_1, X_2, X_3, Y_1, Y_2, Y_3)$; applying the submodularity inequality (Lemma A.2), we conclude

$$\mathbf{H}(X_1, X_2, X_3, Y_1, Y_2, Y_3) + \mathbf{H}(X_3 + Y_3) \leqslant \mathbf{H}(X_3 - Y_2, X_1 - Y_3, X_2, Y_1) + \mathbf{H}(X_3, Y_3).$$

But from (2.1), (A. 11), and the independence hypotheses, we have

$$\begin{aligned}
\mathbf{H}(X_1, X_2, X_3, Y_1, Y_2, Y_3) &= 2\mathbf{H}(X, Y) - \mathbf{H}(X - Y) + \mathbf{H}(X) + \mathbf{H}(Y), \\
\mathbf{H}(X_3 + Y_3) &= \mathbf{H}(X + Y), \\
\mathbf{H}(X_3 - Y_2, X_1 - Y_3, X_2, Y_1) &\leqslant 2\mathbf{H}(X - Y) + \mathbf{H}(X) + \mathbf{H}(Y), \\
\mathbf{H}(X_3, Y_3) &= \mathbf{H}(X) + \mathbf{H}(Y),
\end{aligned}$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

and thus

$$\mathbf{H}(X+Y) \leqslant 3\mathbf{H}(X-Y) - \mathbf{H}(X) - \mathbf{H}(Y), \tag{2.2}$$

which rearranges to form (1.17).

Finally, we establish (1.18). We let X and Y be independent G-random variables, and let $(X_0, Y_0), \dots, (X_n, Y_n)$ be independent trials of (X, Y). Set $S_i = X_i + Y_i$ for $0 \leqslant i \leqslant n$. We observe the identity

$$S_0 + \cdots + S_n = (Y_0 + X_1) + (Y_1 + X_2) + \cdots + (Y_{n-1} + X_n) + (Y_n + X_0).$$

In particular, we see that $(X_0, Y_0, S_1, \dots, S_n)$ and $(Y_0 + X_1, \dots, Y_{n-1} + X_n, Y_n + X_0)$ both determine $S_0 + \cdots + S_n$, while $(X_0, Y_0, S_1, \dots, S_n)$ and $(Y_0 + X_1, \dots, Y_{n-1} + X_n, Y_n + X_0)$ jointly determine $(X_0, \dots, X_n, Y_0, \dots, Y_n)$. Applying the submodularity inequality (Lemma A.2), we conclude

$$\begin{aligned}\mathbf{H}(X_0, \dots, X_n, Y_0, \dots, Y_n) + \mathbf{H}(S_0 + \cdots + S_n) \\ \leqslant \mathbf{H}(X_0, Y_0, S_1, \dots, S_n) + \mathbf{H}(Y_0 + X_1, \dots, Y_{n-1} + X_n, Y_n + X_0).\end{aligned}$$

But from (A. 11) and the independence hypotheses we have

$$\begin{aligned}\mathbf{H}(X_0, \dots, X_n, Y_0, \dots, Y_n) &= (n+1)(\mathbf{H}(X) + \mathbf{H}(Y)), \\ \mathbf{H}(X_0, Y_0, S_1, \dots, S_n) &= \mathbf{H}(X) + \mathbf{H}(Y) + n\mathbf{H}(X+Y), \\ \mathbf{H}(Y_0 + X_1, \dots, Y_{n-1} + X_n, Y_n + X_0) &\leqslant (n+1)\mathbf{H}(X+Y).\end{aligned}$$

Putting all this together, we obtain the inequality

$$\mathbf{H}(S_0 + \cdots + S_n) \leqslant (2n+1)\mathbf{H}(X+Y) - n\mathbf{H}(X) - n\mathbf{H}(Y).$$

In particular, if $X_1, \dots, X_{2n+2}$ are independent copies of X, then the above inequality (setting Y to be another independent copy of X) gives

$$\mathbf{H}(X_1 + \cdots + X_{2n+2}) \leqslant \mathbf{H}(X) + (2n+1)\log \sigma[X].$$

Applying (1.9), we conclude that

$$\mathbf{H}(X_1 + \cdots + X_n) = \mathbf{H}(X) + O(n \log \sigma[X])$$

for any $n \geqslant 1$. Applying (2.2), we then conclude that

$$\mathbf{H}(X_1 + \cdots + X_n - X'_1 - \cdots - X'_m) = \mathbf{H}(X) + O((n+m)\log \sigma[X])$$

for any $n, m \geqslant 1$, and the claim (1.18) follows. The proof of Theorem 1.7 is now complete. □

3. An entropy version of the Balog–Szemerédi–Gowers lemma

We now state an entropy analogue of the Balog–Szemerédi–Gowers lemma. In the combinatorial setting, one had the notion of a *refinement* A' of a set A, which was a subset A' of A which still had size comparable to A. In the entropy setting, the corresponding notion is that of a *conditioning* of a random variable X relative to some other related

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

random variable Y, such that $\mathbf{H}(X|Y)$ was still close to $\mathbf{H}(X)$. The entropy Balog–Szemerédi–Gowers lemma then asserts that if two weakly dependent random variables X, Y have a sum of small entropy, then there exist conditionings of X, Y (which capture most of the entropy) whose *independent* sum still has small entropy.

In fact, the conditioning can be given explicitly.

Theorem 3.1 (Entropy Balog–Szemerédi–Gowers lemma). *Let G be an additive group, and let X, Y be G-random variables that are weakly dependent in the sense that*

$$\mathbf{H}(X, Y) \geqslant \mathbf{H}(X) + \mathbf{H}(Y) - \log K \tag{3.1}$$

for some $K \geqslant 1$. Suppose also that

$$\mathbf{H}(X + Y) \leqslant \frac{1}{2}\mathbf{H}(X) + \frac{1}{2}\mathbf{H}(Y) + \log K. \tag{3.2}$$

Then, if we let $(X_1, Y), (X_2, Y)$ be conditionally independent trials of (X, Y) conditioning on Y, and then let (X_1, X_2, Y) and (X_1, Y') be conditionally independent trials of (X_1, X_2, Y) and (X_1, Y) conditioning on X_1, then X_2 and Y' are conditionally independent relative to X_1, Y, with

$$\mathbf{H}(X_2|X_1, Y) \geqslant \mathbf{H}(X) - \log K, \tag{3.3}$$

$$\mathbf{H}(Y'|X_1, Y) \geqslant \mathbf{H}(Y) - \log K, \tag{3.4}$$

$$\mathbf{H}(X_2 + Y'|X_1, Y) \leqslant \frac{1}{2}\mathbf{H}(X) + \frac{1}{2}\mathbf{H}(Y) + 7\log K. \tag{3.5}$$

This should be compared with Lemma 1.1. The appearance of the exponent 7 in both statements is not coincidental, as the proofs are fundamentally the same.

Remark. Let $E \subset A \times B$ be a regular bipartite graph between two finite non-empty sets A, B in G; thus the A-degree $|\{b \in B : (a, b) \in E\}|$ is independent of $a \in A$, and similarly the B-degree $|\{a \in A : (a, b) \in E\}|$ is independent of $b \in B$. Let (X, Y) be an element of E chosen uniformly at random. Then the random variables (Y', X_1, Y, X_2) defined above are drawn uniformly from the space of all paths (b, a, b', a') of length three in E; thus $(a, b), (a, b'), (a', b') \in E$. It may be helpful to keep this example in mind when going through the proof of Theorem 3.1. Not surprisingly, paths of length three also play a major role in the proof of Theorem 1.1.

We now establish the theorem. By construction, Y' and X_2, Y are conditionally independent relative to X_1, and thus X_2 and Y' are conditionally independent relative to X_1, Y, as claimed. Also, since X_1 is conditionally independent of X_2 relative to Y, then

$$\mathbf{H}(X_2|X_1, Y) = \mathbf{H}(X_2|Y) = \mathbf{H}(X|Y),$$

and (3.3) follows from (3.1). Similarly, since Y, Y' are conditionally independent relative to X_1,

$$\mathbf{H}(Y'|X_1, Y) = \mathbf{H}(Y'|X_1) = \mathbf{H}(Y|X),$$

and (3.4) follows from (3.1).

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

The only remaining claim to establish is (3.5). We need a preliminary lemma.

Lemma 3.2 (Weak Balog–Szemerédi–Gowers lemma). *We have*

$$\mathbf{H}(X_1 - X_2|Y) \leqslant \mathbf{H}(X) + 4\log K.$$

Proof. Let $(X_1, X_2, Y), (X_1, X_2, Y')$ be two conditionally independent copies of (X_1, X_2, Y) relative to (X_1, X_2). Observe that (X_1, X_2, Y), $(X_1 + Y', X_2 + Y', Y)$ both determine $(X_1 - X_2, Y)$, and that (X_1, X_2, Y) and $(X_1 + Y', X_2 + Y', Y)$ jointly determine (X_1, X_2, Y, Y'). Applying the submodularity inequality (Lemma A.2), we conclude that

$$\mathbf{H}(X_1, X_2, Y, Y') + \mathbf{H}(X_1 - X_2, Y) \leqslant \mathbf{H}(X_1, X_2, Y) + \mathbf{H}(X_1 + Y', X_2 + Y', Y).$$

But from (1.8), (2.1) and (A. 8),

$$\begin{aligned}
\mathbf{H}(X_1, X_2, Y, Y') &= 2\mathbf{H}(X_1, X_2, Y) - \mathbf{H}(X_1, X_2)\\
&\geqslant 4\mathbf{H}(X, Y) - 2\mathbf{H}(Y) - 2\mathbf{H}(X),\\
\mathbf{H}(X_1 - X_2, Y) &= \mathbf{H}(X_1 - X_2|Y) + \mathbf{H}(Y),\\
\mathbf{H}(X_1, X_2, Y) &= 2\mathbf{H}(X, Y) - \mathbf{H}(Y),\\
\mathbf{H}(X_1 + Y', X_2 + Y', Y) &\leqslant 2\mathbf{H}(X + Y) + \mathbf{H}(Y),
\end{aligned}$$

and thus

$$\mathbf{H}(X_1 - X_2|Y) \leqslant 2\mathbf{H}(X + Y) + \mathbf{H}(Y) + 2\mathbf{H}(X) - 2\mathbf{H}(X, Y),$$

and the claim then follows from (3.1) and (3.2). □

Now observe that (X_2, Y', Y) and $(X_1 - X_2, X_1 + Y', Y)$ both determine $(X_2 + Y', Y)$, and that (X_2, Y', Y) and $(X_1 - X_2, X_1 + Y', Y)$ jointly determine (X_1, X_2, Y, Y'). Applying the submodularity inequality (Lemma A.2), we conclude that

$$\mathbf{H}(X_1, X_2, Y, Y') + \mathbf{H}(X_2 + Y', Y) \leqslant \mathbf{H}(X_2, Y', Y) + \mathbf{H}(X_1 - X_2, X_1 + Y', Y).$$

But from (1.8), (A. 8), and (a generalization of) (2.1),

$$\begin{aligned}
\mathbf{H}(X_1, X_2, Y, Y') &= \mathbf{H}(X_1, X_2, Y) + \mathbf{H}(X_1, Y') - \mathbf{H}(X_1)\\
&= 2\mathbf{H}(X, Y) - \mathbf{H}(Y) + \mathbf{H}(X, Y) - \mathbf{H}(X),\\
\mathbf{H}(X_2 + Y', Y) &= \mathbf{H}(X_2 + Y'|Y) + \mathbf{H}(Y),\\
\mathbf{H}(X_2, Y', Y) &\leqslant \mathbf{H}(X_2, Y) + \mathbf{H}(Y')\\
&= \mathbf{H}(X, Y) + \mathbf{H}(Y),\\
\mathbf{H}(X_1 - X_2, X_1 + Y', Y) &\leqslant \mathbf{H}(X_1 - X_2|Y) + \mathbf{H}(Y) + \mathbf{H}(X_1 + Y')\\
&= \mathbf{H}(X_1 - X_2|Y) + \mathbf{H}(Y) + \mathbf{H}(X + Y).
\end{aligned}$$

Substituting these bounds, we obtain

$$\mathbf{H}(X_2 + Y'|Y) \leqslant \mathbf{H}(X_1 - X_2|Y) + 2\mathbf{H}(Y) + \mathbf{H}(X) + \mathbf{H}(X + Y) - 2\mathbf{H}(X, Y).$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Applying Lemma 3.2, (3.1) and (3.2), we conclude that

$$\mathbf{H}(X_2 + Y'|Y) \leqslant \frac{1}{2}\mathbf{H}(X) + \frac{1}{2}\mathbf{H}(Y) + 7\log K,$$

and the claim follows from (A. 12).

4. Uniformization

The main purpose of this section is to establish the following uniformization bound on groups, as well as an analogous result for coset progressions (see Corollary 4.6).

Theorem 4.1 (Uniformization on a group). *Let G be a finite group, let $p_U := \frac{1}{|G|}$ be the uniform distribution on G, and let $p \in \Pr_c(G)$ be another distribution, such that*

$$\mathbf{H}(p) \geqslant \log|G| - \log K$$

for some $K \geqslant 10$. Then

$$\mathrm{dist}_{\mathrm{tr}}(p, p_U) \ll \log K.$$

Since $\mathbf{H}(p_U) = \log|G|$, we see that this is sharp up to constants. One can view this theorem as a special case of Theorem 1.8, but with significantly better dependence on the constants.

We establish this theorem by a sequence of partial results. We first record a simple lemma that allows us to 'divide and conquer' the problem of estimating the transport distance between two random variables.

Lemma 4.2 (Transport splitting lemma). *Let G be a group, let X, Y be G-random variables, and let S be another discrete random variable; we do not assume X, Y, S to be independent. Then*

$$\mathrm{dist}_{\mathrm{tr}}(X, Y) \leqslant \mathbf{H}(S) + \sum_{s \in \mathrm{range}(S)} p_S(s)\,\mathrm{dist}_{\mathrm{tr}}((X|S=s), (Y|S=s)).$$

Proof. Let $\varepsilon > 0$. For each $s \in \mathrm{range}(S)$, we can use Definition 1.5 to select a random variable Z_s conditioned to the event $S = s$ of entropy $\mathbf{H}(Z_s) \leqslant \mathrm{dist}_{\mathrm{tr}}((X|S=s), (Y|S=s)) + \varepsilon$ such that $(X + Z_s|S=s) \equiv (Y|S=s)$. If we then let Z be the random variable whose conditioning to $S = s$ equals Z_s, then $X + Z \equiv Y$, and (by (A. 10))

$$\mathbf{H}(Z) \leqslant \mathbf{H}(S) + \mathbf{H}(Z|S) = \mathbf{H}(S) + \sum_{s \in \mathrm{range}(S)} p_S(s)\mathbf{H}(Z|S=s).$$

Putting all this together, we conclude that

$$\mathrm{dist}_{\mathrm{tr}}(X, Y) \leqslant \mathbf{H}(S) + \sum_{s \in \mathrm{range}(S)} p_S(s)\,\mathrm{dist}_{\mathrm{tr}}((X|S=s), (Y|S=s)) + |\,\mathrm{range}(S)|\varepsilon.$$

Since ε was arbitrary, the claim follows. □

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Next, we show that one can converge exponentially fast to the uniform distribution in the L^2 sense.

Lemma 4.3 (L^2 flattening lemma). *Let G be a finite group, let $p_U := \frac{1}{|G|}$ be the uniform distribution on G, and let $p \in \Pr_c(G)$ be another distribution. Then, for any integer $k \geqslant 1$, one can find a distribution $p_k \in \Pr_c(G)$ such that*

$$\mathrm{dist}_{\mathrm{tr}}(p, p_k) \leqslant k \log 2$$

and

$$\|p_k - p_U\|_{\ell^2(G)} \leqslant 2^{-k/2} \|p - p_U\|_{\ell^2(G)}.$$

Proof. By induction it suffices to verify the case $k = 1$. We use the first moment method. Let h be chosen uniformly at random from G, and let

$$p_1(x) := \frac{1}{2}(p(x) + p(x - h)).$$

Clearly p_1 is the convolution of p with a Bernoulli variable of entropy $\log 2$, and so $\mathrm{dist}_{\mathrm{tr}}(p, p_1) \leqslant \log 2$. On the other hand, a straightforward calculation using $\sum_{x \in G} p(x) = 1$ reveals the identity

$$\mathbf{E}_h \|p_1 - p_U\|^2_{\ell^2(G)} = \frac{1}{2} \|p - p_U\|^2_{\ell^2(G)},$$

and the claim follows. □

We now combine these lemmas to pass to ℓ^2-bounded random variables.

Lemma 4.4 (Entropy-uniform to ℓ^2-bounded). *Let G be a finite group, and let $p \in \Pr_c(G)$ be such that*

$$\mathbf{H}(p) \geqslant \log |G| - \log K$$

for some $K \geqslant 10$. Then there exists $q \in \Pr_c(G)$ with

$$\mathrm{dist}_{\mathrm{tr}}(p, q) \ll \log K$$

and

$$\|q\|_{\ell^2(G)} \ll 1/|G|^{1/2}.$$

Proof. The basic idea here is to flatten all the regions of G in which X has an abnormally high probability density.

By Lemma A.1, we have

$$\sum_{k=1}^{\infty} 2^k \mathbf{P}(X \in A_k) \ll \log K, \tag{4.1}$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

where A_k are the sets

$$A_k := \left\{ x \in G : \frac{2^{2^{k-1}}}{|G|} \leqslant p(x) < \frac{2^{2^k}}{|G|} \right\}$$

for $k \geqslant 1$, and then set $A_0 := G \backslash \bigcup_{k=1}^{\infty} A_k$; thus the $A_0, A_1, \ldots$ partition G (and thus only finitely many are non-empty).

Let X be a random variable with distribution p. For each $k \geqslant 0$, let E_k be the event that $X \in A_k$, and thus the E_k partition probability space. From (4.1) we have

$$\sum_{k=1}^{\infty} 2^k \mathbf{P}(E_k) \ll \log K. \tag{4.2}$$

Suppose that $k \geqslant 1$ is such that E_k has positive probability. Then we can define X_k to be the random variable $X_k := (X|E_k)$. Observe that p_{X_k} is bounded above by $\frac{2^{2^k}}{\mathbf{P}(E_k)|G|}$, so we have the crude bound

$$\|p_{X_k}\|_{\ell^2(G)} \leqslant \frac{2^{2^k}}{\mathbf{P}(E_k)|G|^{1/2}}.$$

Applying Lemma 4.3, with k replaced by, say, $2 \times (2^k + \log \frac{1}{\mathbf{P}(E_k)})$, one can thus find $q_k \in \mathrm{Pr}_c(G)$ such that

$$\mathrm{dist}_{\mathrm{tr}}(p_{X_k}, q_k) \ll 2^k + \log \frac{1}{\mathbf{P}(E_k)} \tag{4.3}$$

and

$$\|q_k - p_U\|_{\ell^2(G)} \leqslant 1/|G|^{1/2}. \tag{4.4}$$

(Indeed, one could even gain a factor of 2^{-2^k} on the right-hand side of (4.4), though this turns out to be unnecessary for the current argument.)

Now set $q \in \mathrm{Pr}_c(G)$ to be the probability distribution

$$q = 1_{E_0} p_X + \sum_{k=1}^{\infty} \mathbf{P}(E_k) q_k. \tag{4.5}$$

Observe that p_X is bounded by $2/|G|$ on E_0. From (4.5), (4.4) and the triangle inequality, we conclude that

$$\|q\|_{\ell^2(G)} \ll 1/|G|^{1/2}.$$

From Lemma 4.2 (setting S to be the random variable induced by the partition E_k), we see that

$$\mathrm{dist}_{\mathrm{tr}}(p, q) \leqslant \sum_{k=0}^{\infty} \mathbf{P}(E_k) \log \frac{1}{\mathbf{P}(E_k)} + \sum_{k=1}^{\infty} \mathbf{P}(E_k)\, \mathrm{dist}_{\mathrm{tr}}(p_{X_k}, q_k). \tag{4.6}$$

From (4.6), (4.3) and (4.2), we conclude that

$$\mathrm{dist}_{\mathrm{tr}}(p, q) \ll \log K + \sum_{k=0}^{\infty} \mathbf{P}(E_k) \log \frac{1}{\mathbf{P}(E_k)}.$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

But from (4.2), $\mathbf{P}(E_k) \ll 2^{-k} \log K$, and so

$$\mathbf{P}(E_k) \log \frac{1}{\mathbf{P}(E_k)} \ll (1+k) 2^{-k} \log K,$$

and thus

$$\sum_{k=0}^{\infty} \mathbf{P}(E_k) \log \frac{1}{\mathbf{P}(E_k)} \ll \log K$$

and the claim follows. □

From the triangle inequality, it is now clear that Theorem 4.1 follows from Lemma 4.4, Lemma 4.3, and the next lemma.

Lemma 4.5 (ℓ^2-bounded to uniform). *Let G be a finite group, and let $p \in \mathrm{Pr}_c(G)$ be such that $\|p - p_U\|_{\ell^2(G)} \leqslant 1/|G|^{1/2}$. Then* $\mathrm{dist}_{\mathrm{tr}}(p, p_U) \ll 1$.

Proof. The idea is to manually transport away the most severe irregularities in the distribution of p to obtain a new distribution that is significantly closer to uniform in the ℓ^2 norm, and then iterate.

Let C_G be the supremum of $\mathrm{dist}_{\mathrm{tr}}(p, p_U)$ for all $p \in \mathrm{Pr}_c(G)$ with $\|p - p_U\|_{\ell^2(G)} \leqslant 1/|G|^{1/2}$. It is easy to see that C_G is finite for any fixed finite G; our task is to show that $C_G \ll 1$ (uniformly in G).

Let $k \geqslant 1$ be a large integer to be chosen later. Let $p \in \mathrm{Pr}_c(G)$ be such that $\|p - p_U\|_{\ell^2(G)} \leqslant 1/|G|^{1/2}$. Then by Lemma 4.3 one can find $q \in \mathrm{Pr}_c(G)$ with $\mathrm{dist}_{\mathrm{tr}}(p, q) \leqslant k \log 2$ and

$$\|q - p_U\|_{\ell^2(G)} \leqslant 2^{-k/2}/|G|^{1/2}. \tag{4.7}$$

If $q \equiv p_U$, then we have $\mathrm{dist}_{\mathrm{tr}}(p, p_U) \leqslant k \log 2$, so suppose instead that q is not identically equal to p_U. Then the quantity

$$\sigma := \sum_{x \in G : q(x) > p_U(x)} q(x) - p_U(x) = \sum_{x \in G : q(x) < p_U(x)} p_U(x) - q(x)$$

is non-zero; from (4.7) and the Cauchy–Schwarz inequality we also have

$$\sigma < 2^{-k/2}. \tag{4.8}$$

Let $q_+, q_- \in \mathrm{Pr}_c(G)$ be the probability distributions defined by

$$q_+(x) := \frac{1}{\sigma} 1_{q(x) > p_U(x)} (q(x) - p_U(x))$$

and

$$q_-(x) := \frac{1}{\sigma} 1_{q(x) < p_U(x)} (p_U(x) - q(x)),$$

and thus

$$q = p_U + \sigma q_+ - \sigma q_-.$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

We can thus build a random variable X with distribution q by creating a boolean random variable $S \in \{0,1\}$ with $p_S(1) = \sigma$, then setting $(X|S=1)$ to have distribution q_+ and $(X|S=0)$ to have distribution $\frac{1}{1-\sigma}(p_U - \sigma q_-)$. If we let Y be a random variable with $(Y|S=1)$ having distribution q_- and $(Y|S=0)$ having distribution $\frac{1}{1-\sigma}(p_U - \sigma q_-)$, we see that $Y \equiv p_U$. From Lemma 4.2 we conclude that

$$\mathrm{dist}_{\mathrm{tr}}(q, p_U) \leqslant \mathbf{H}(S) + \sigma\, \mathrm{dist}_{\mathrm{tr}}(q_+, q_-) \ll \sigma \log \frac{1}{\sigma} + \sigma\, \mathrm{dist}_{\mathrm{tr}}(q_+, q_-).$$

Now we estimate $\mathrm{dist}_{\mathrm{tr}}(q_+, q_-)$. From (4.7) we see that

$$\|q_+\|_{\ell^2(G)}, \|q_-\|_{\ell^2(G)} \ll \frac{2^{-k/2}}{\sigma |G|^{1/2}}.$$

Applying Lemma 4.3, one can find $r_+, r_- \in \mathrm{Pr}_c(G)$ with $\|r_\pm - p_U\|_{\ell^2(G)} \leqslant 1/|G|^{1/2}$ such that

$$\mathrm{dist}_{\mathrm{tr}}(q_\pm, r_\pm) \ll 1 + \log \frac{2^{-k/2}}{\sigma}.$$

By definition of C_G, we then have

$$\mathrm{dist}_{\mathrm{tr}}(r_\pm, p_U) \ll C_G.$$

Putting all this together using the triangle inequality, we see that

$$\mathrm{dist}_{\mathrm{tr}}(p, p_U) \ll k + \sigma \log \frac{1}{\sigma} + \sigma \left(C_G + \log \frac{2^{-k/2}}{\sigma} \right).$$

Taking the worst-case value of σ using (4.8), we conclude

$$\mathrm{dist}_{\mathrm{tr}}(p, p_U) \ll k + 2^{-k/2} C_G,$$

and thus, taking suprema in p,

$$C_G \ll k + 2^{-k/2} C_G.$$

Setting k sufficiently large, we conclude

$$C_G \leqslant \frac{1}{2} C_G + O(1),$$

and the claim follows. □

Corollary 4.6 (Uniformization on coset progressions). *Let $H + P$ be a proper coset progression of rank d in some additive group G, and let $p \in \mathrm{Pr}_c(H + P)$ be such that*

$$\mathbf{H}(p) \geqslant \log |H + P| - \log K$$

for some $K \geqslant 10$. Let p_U be the uniform distribution on $H + P$. Then

$$\mathrm{dist}_{\mathrm{tr}}(p, p_U) \ll \log K + d. \tag{4.9}$$

Proof. We can view $H + P$ as the homomorphic image of $B := H \times [0, N_1) \times \cdots \times [0, N_d)$ for some integers $N_1, \ldots, N_d \geqslant 1$. Let $\tilde{p} \in \mathrm{Pr}_c(B)$ be the pullback of p to B, and

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

similarly define $\tilde{p}_U$. We can then embed B in the finite group $G := H \times \mathbf{Z}/(2N_1\mathbf{Z}) \times \cdots \times \mathbf{Z}/(2N_d\mathbf{Z})$. Observe that

$$\mathbf{H}(\tilde{p}), \mathbf{H}(\tilde{p}_U) \geqslant \log|G| - \log K - O(d),$$

and so, by Theorem 4.1,

$$\mathrm{dist}_{\mathrm{tr}}(\tilde{p}, p_G), \mathrm{dist}_{\mathrm{tr}}(\tilde{p}_U, p_G) \ll \log K + d,$$

where p_G is the uniform distribution on G. Thus, by the triangle inequality,

$$\mathrm{dist}_{\mathrm{tr}}(\tilde{p}, \tilde{p}_U) \ll \log K + d.$$

Observe that as $\tilde{p}, \tilde{p}_U$ both range in B, the shifts needed to transport $\tilde{p}$ to $\tilde{p}_U$ range in $B - B$, and so do not encounter the 'wrap-around' effects of the cyclic groups $\mathbf{Z}/(2N_1\mathbf{Z}), \ldots, \mathbf{Z}/(2N_d\mathbf{Z})$. Thus we can push this transport bound back to $H + P$ and establish (4.9) as desired. □

5. The inverse entropy theorem

We now prove Theorem 1.8. We begin with the easy claim (i). If X is the uniform distribution on a coset of a finite group then $X + X$ is uniformly distributed on another coset of this group, and so $\sigma[X] = 1$ as claimed. Now suppose that $\sigma[X] = 1$, thus $\mathbf{H}(X_1 + X_2) = \mathbf{H}(X)$, where X_1, X_2 are independent copies of X. Inspecting the proof of Lemma 2.1 we conclude that $\mathbf{H}(X_1 + X_2) = \mathbf{H}(X_1 + X_2|X_2)$, which by the discussion after (A. 10) implies that $X_1 + X_2$ and X_2 are independent, or equivalently that the distribution of $(X_1 + X_2|X_2 = x)$ is independent of $x \in \mathrm{range}(X)$. This implies that the probability distribution of X is invariant under translations in $\mathrm{range}(X) - \mathrm{range}(X)$, which quickly implies that $\mathrm{range}(X) - \mathrm{range}(X)$ is a finite subgroup of G, and that X is uniformly distributed on a coset of this subgroup, as desired.

Now we prove the more difficult claims (ii) and (iii). We begin with a special case of (iii), in which Y is already uniform.

Proposition 5.1. *Let X be a G-random variable, and let $H + P$ be a coset progression of rank d. Let U be the uniform distribution on $H + P$, and suppose that* $\mathrm{dist}_R(X, U) \leqslant \log K$. *Then* $\mathrm{dist}_{\mathrm{tr}}(X, U) \ll_{K,d} 1$. □

Proof. By translating $H + P$ if necessary, we may assume $0 \in H + P$. We allow all implied constants to depend on K, d. The basic idea here is to treat $H + P$ as an approximate group, and somehow pass to a 'quotient' of G by $H + P$. The reader is encouraged to consider the special case $P = \{0\}$, in which this quotienting idea can be made precise.

Let S be a maximal subset of G with the property that the translates $s + H + P$ of $H + P$ for $s \in S$ are all disjoint. (For G infinite, the existence of such an S is guaranteed by Zorn's lemma.) Clearly the translates $s + (H + P) - (H + P)$ cover G. From this, the disjointness of the $s + H + P$, and the greedy algorithm, we can thus partition G into sets A_s for $s \in S$, where

$$s + (H + P) \subset A_s \subset s + (H + P) - (H + P).$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

One should view the partition A_s as a crude approximation of the (non-existent) quotient of G by $H+P$.

Take U, X to be independent. From (1.12) and the hypothesis $\mathrm{dist}_R(X,U) \leqslant \log K$ (and the fact that $-U$ is equivalent to a translate of U), one has

$$\mathbf{H}(X+U) \leqslant \frac{1}{2}\mathbf{H}(X) + \frac{1}{2}\mathbf{H}(U) + O(1).$$

Of course, $\mathbf{H}(U) = \log|H+P|$. Applying (1.9), we conclude that

$$\mathbf{H}(X) = \log|H+P| + O(1)$$

and

$$\mathbf{H}(X+U) = \log|H+P| + O(1),$$

and thus

$$\sum_{s\in S}\sum_{x\in A_s} p_{X+U}(x)\log\frac{1}{p_{X+U}(x)} = \log|H+P| + O(1).$$

On the other hand, if $s \in S$ and $x \in A_s$, one clearly has

$$\begin{aligned} p_{X+U}(x) &\leqslant \frac{1}{|H+P|}\mathbf{P}(X \in s + (H+P) - 2(H+P)) \\ &\leqslant \frac{1}{|H+P|}\mathbf{P}(X+U \in s + 2(H+P) - 2(H+P)), \end{aligned}$$

and thus

$$\log\frac{1}{p_{X+U}(x)} \geqslant \log|H+P| + \log\frac{1}{\mathbf{P}(X+U \in s+2(H+P)-2(H+P))}.$$

Since $\sum_{s\in S}\sum_{x\in A_x} p_{X+U}(x) = 1$, we conclude that

$$\sum_{s\in S}\sum_{x\in A_s} p_{X+U}(x)\log\frac{1}{\mathbf{P}(X+U \in s+2(H+P)-2(H+P))} \leqslant O(1),$$

or equivalently

$$\sum_{s\in S} c_s \log\frac{1}{\mathbf{P}(X+U \in s+2(H+P)-2(H+P))} \leqslant O(1),$$

where $c_s := \mathbf{P}(X+U \in A_s)$. Observe that $s+2(H+P)-2(H+P)$ can be covered by at most $O(1)$ sets $A_{s'}$, where $s' \in s+3(H+P)-3(H+P)$. (Indeed, all such $A_{s'}$ are disjoint, have cardinality comparable to $|H+P|$, and are contained in $s+4(H+P)-4(H+P)$, which has cardinality $O(|H+P|)$.) Thus, by the pigeonhole principle, for every $s \in S$ there exists $s'(s) \in S \cap (s+3(H+P)-3(H+P))$ such that

$$\mathbf{P}(X+U \in s+2(H+P)-2(H+P)) \ll c_{s'(s)},$$

and thus

$$\sum_{s\in S} c_s \log\frac{1}{c_{s'(s)}} \leqslant O(1). \tag{5.1}$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Let Y be the random variable $Y := s$, where s is the unique $s \in S$ such that $X + U \in A_s$. Then $X - Y$ takes values in $(H+P) - 2(H+P)$. We now claim that

$$\mathbf{H}(Y) \leqslant O(1), \tag{5.2}$$

or in other words that

$$\sum_{s \in S} F(c_s) \leqslant O(1). \tag{5.3}$$

This is almost (5.1), but we have to replace $s'(s)$ by s. To do this, we let $C > e$ be a large quantity to be chosen later, and split the sum in (5.3) into three terms: one where $c_{s'(s)} \geqslant 1/e$, one where $1/e > c_{s'(s)} \geqslant Cc_s$ and one where $c_{s'(s)} \leqslant Cc_s$.

In the first case, observe that there are only $O(1)$ possible values of $s'(s)$; each one of these is associated to $O(1)$ possible values of s (since $s \in s'(s) + 3(H+P) - 3(H+P)$ and the $s + H + P$ are disjoint), so the net contribution to (5.3) here is $O(1)$.

In the second case, we observe from (A. 6) that

$$F(c_s) \leqslant 2F(1/C)F(Cc_s) \leqslant 2F(1/C)F(c_{s'(s)}),$$

so the contribution of this term to (5.3) is at most

$$2F(1/C) \sum_{s \in S} F(c_{s'(s)}).$$

But each s' can arise from at most $O(1)$ choices of S, so we can bound this contribution by at most

$$\frac{1}{2} \sum_{s' \in S} F(c_{s'}) = \frac{1}{2}\mathbf{H}(Y)$$

if $C = O(1)$ is chosen appropriately.

For the third case, we see that

$$c_s \log \frac{1}{c_s} \leqslant c_s \log \frac{1}{c_{s'(s)}} + c_s \log C,$$

and so by (5.3) the net contribution of this case is

$$\leqslant O(1) + \log C = O(1).$$

Thus $\mathbf{H}(Y) \leqslant \frac{1}{2}\mathbf{H}(Y) + O(1)$, and the claim (5.2) follows. In particular

$$\mathrm{dist}_{\mathrm{tr}}(X, X - Y) \ll 1.$$

But by (1.8),

$$\mathbf{H}(X - Y) \geqslant \mathbf{H}(X) - O(1) \geqslant \log |H + P| - O(1).$$

The random variable $X - Y$ ranges in $(H+P) - 2(H+P)$, which is a coset progression of rank d and cardinality $O(|H+P|)$. Applying Corollary 4.6 to $X - Y$, we conclude that

$$\mathrm{dist}_{\mathrm{tr}}(X - Y, U_{(H+P)-2(H+P)}) \ll 1,$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

where $U_{(H+P)-2(H+P)}$ is the uniform distribution on $(H+P)-2(H+P)$. Direct computation shows that

$$\mathrm{dist}_{\mathrm{tr}}(U_{(H+P)-2(H+P)}, U) \ll 1,$$

and the claim follows from another application of the triangle inequality. □

In view of the above proposition, it suffices to show the following.

Proposition 5.2. *If $\sigma[X] \leqslant K$, then there exists a coset progression $H+P$ of rank $O_K(1)$ such that $\mathrm{dist}_R(X,U) \ll_K 1$, where U is the uniform distribution on $H+P$.* □

Indeed, (ii) follows immediately from Propositions 5.2 and 5.1, while if we are in the situation of (iii), then from Theorem 1.7 we have

$$\begin{aligned}\mathrm{dist}_R(X,-X) &\leqslant \mathrm{dist}_R(X,Y)+\mathrm{dist}_R(X,-Y)\\ &\leqslant 4\,\mathrm{dist}_R(X,Y)\\ &\leqslant 4\log K,\end{aligned}$$

and thus from (1.13) we have $\sigma[X] \leqslant K^4$. By Proposition 5.2, one can then find a uniform distribution U on a coset progression $H+P$ on rank $O_K(1)$ such that $\mathrm{dist}_R(X,U) \ll_K 1$, hence $\mathrm{dist}_{\mathrm{tr}}(X,U) \ll_K 1$ by Proposition 5.1. Meanwhile, from the Ruzsa triangle inequality (1.16), we have $\mathrm{dist}_R(Y,U) \ll_K 1$ and so $\mathrm{dist}_{\mathrm{tr}}(Y,U) \ll_K 1$, and so by the triangle inequality we have $\mathrm{dist}_{\mathrm{tr}}(X,Y) \ll_K 1$, as claimed.

It remains to establish Proposition 5.2. We begin with an approximate formula for $\mathbf{H}(X+Y)-\mathbf{H}(X)$ when X, Y are independent.

Lemma 5.3 (Sumset entropy increase formula). *Let X, Y be independent. Then*

$$\sum_{y\in\mathrm{range}(Y)} p_Y(y) \sum_{z\in\mathrm{range}(X+Y)} p_{X+y}(z)\log_+\frac{p_{X+y}(z)}{p_{X+Y}(z)} = \mathbf{H}(X+Y)-\mathbf{H}(X)+O(1),$$

where $\log_+ x := \max(\log x, 0)$.

Proof. To simplify the summation notation, it will be understood throughout that $y \in \mathrm{range}(Y)$ and $z \in \mathrm{range}(X+Y)$. We have

$$\begin{aligned}\mathbf{H}(X+Y)-\mathbf{H}(X) &= \sum_y p_Y(y)(\mathbf{H}(X+Y)-\mathbf{H}(X+y))\\ &= \sum_y p_Y(y)\sum_z (F(p_{X+Y}(z))-F(p_{X+y}(z))).\end{aligned}$$

Since

$$\sum_y p_Y(y)(p_{X+Y}(z)-p_{X+y}(z)) = 0$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

for all z, we thus have

$$\begin{aligned}&\mathbf{H}(X+Y)-\mathbf{H}(X)\\&\quad=\sum_y p_Y(y)\sum_z F(p_{X+Y}(z))+F'(p_{X+Y}(z))(p_{X+y}(z)-p_{X+Y}(z))-F(p_{X+y}(z)).\end{aligned}$$

From (A. 5), the summand is equal to

$$p_{X+y}(z)\log_+\frac{p_{X+y}(z)}{p_{X+Y}(z)}+O(p_{X+y}(z))+O(p_{X+Y}(z)).$$

Since

$$\sum_y p_Y(y)\sum_z p_{X+y}(z)=\sum_y p_Y(y)\sum_z p_{X+Y}(z)=1,$$

the desired claim follows. □

This leads us to our first structural result on random variables of bounded doubling, namely that they are approximately uniformly distributed in a set that captures the bulk of the entropy.

Proposition 5.4 (X is approximately uniformly distributed). *If $\sigma[X]\leqslant K$, then there exists a non-empty set A of cardinality*

$$|A|\asymp_K \exp(\mathbf{H}(X)), \tag{5.4}$$

such that

$$p_X(x)\asymp_K \exp(-\mathbf{H}(X)) \tag{5.5}$$

for all $x\in A$.

Proof. We allow implied constants to depend on K. Write $Z:=X_1+X_2$ for the sum of two independent copies of X. From Lemma 5.3 we have

$$\sum_y p_X(y)\sum_z p_{X+y}(z)\log_+\frac{p_{X+y}(z)}{p_Z(z)}\ll 1, \tag{5.6}$$

where it is understood that $y\in\mathrm{range}(X)$ and $z\in\mathrm{range}(Z)$.

Now let $0<\varepsilon<0.1$ be a small constant (depending on K) to be chosen later. From (5.6) we have

$$\sum_y p_X(y)\sum_{z:p_{X+y}(z)\geqslant e^{1/\varepsilon}p_Z(z)} p_{X+y}(z)/\varepsilon\ll 1,$$

and thus

$$\sum_z\Big[\sum_{y:p_{X+y}(z)\geqslant e^{1/\varepsilon}p_Z(z)} p_X(y)p_{X+y}(z)\Big]\ll\varepsilon.$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Swapping y and $z-y$ (using the identity $p_{X+y}(z) = p_X(z-y)$), we also have

$$\sum_z \sum_{y: p_X(y) \geqslant e^{1/\varepsilon} p_Z(z)} p_X(y) p_{X+y}(z) \ll \varepsilon.$$

Also, observe that

$$\sum_z \sum_{y: p_{X+y}(z) \leqslant \varepsilon p_Z(z)} p_X(y) p_{X+y}(z) \leqslant \sum_z \varepsilon p_Z(z) \sum_y p_X(y) = \varepsilon,$$

and similarly

$$\sum_z \sum_{y: p_X(y) \leqslant \varepsilon p_Z(z)} p_X(y) p_{X+y}(z) \leqslant \varepsilon.$$

Finally, we have

$$\sum_z \sum_y p_X(y) p_{X+y}(z) = 1. \tag{5.7}$$

Putting all these estimates together (with ε sufficiently small), we conclude that

$$\sum_z \sum_{y: p_X(y), p_{X+y}(z) \asymp p_Z(z)} p_X(y) p_{X+y}(z) > 1/2. \tag{5.8}$$

From (5.8), and the pigeonhole principle, there exists a $z_0 \in \mathrm{range}(Z)$ such that

$$\sum_{y: p_X(y), p_{X+y}(z_0) \asymp p_Z(z_0)} p_X(y) p_{X+y}(z_0) > p_Z(z_0)/2.$$

The left-hand side can be bounded crudely by

$$\ll |\{y : p_X(y) \asymp p_Z(z_0)\}| p_Z(z_0)^2.$$

Thus, if we let A denote the set

$$A := \{y : p_X(y) \asymp p_Z(z_0)\}, \tag{5.9}$$

then

$$|A| \gg 1/p_Z(z_0).$$

Since

$$1 \geqslant \sum_{y \in A} p_X(y) \asymp |A| p_Z(z_0),$$

we conclude that

$$|A| \asymp 1/p_Z(z_0),$$

and hence, by (5.9),

$$p_X(y) \asymp 1/|A| \tag{5.10}$$

for all $y \in A$. In particular, we have

$$\mathbf{P}(X \in A) \asymp 1. \tag{5.11}$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

To conclude the lemma, we need to show that

$$\log|A| = \mathbf{H}(X) + O(1).$$

We may assume by a limiting argument that the events $X \in A$ and $X \notin A$ have non-zero probability. Let X_1, X_2 be independent copies of X, and let Y be the indicator random variable $Y = 1_{X_1 \in A}$. Then, by (A. 10) and (A. 7),

$$\begin{aligned}\mathbf{H}(X_1 + X_2) &\geqslant \mathbf{H}(X_1 + X_2|Y) \\ &= \mathbf{P}(X_1 \in A)\mathbf{H}(X_1 + X_2|X_1 \in A) + \mathbf{P}(X_1 \notin A)\mathbf{H}(X_1 + X_2|X_1 \notin A). \quad (5.12)\end{aligned}$$

Now, from Lemma 2.1 we have

$$\mathbf{H}(X_1 + X_2|X_1 \notin A) \geqslant \mathbf{H}(X_1|X_1 \notin A) = \mathbf{H}(X|X \notin A)$$

and

$$\mathbf{H}(X_1 + X_2|X_1 \in A) \geqslant \mathbf{H}(X_2) = \mathbf{H}(X).$$

On the other hand, since $\sigma[X] \leqslant K$ by hypothesis, $\mathbf{H}(X_1 + X_2) \leqslant \mathbf{H}(X) + O(1)$. Putting all these estimates together, we obtain

$$\mathbf{H}(X) + O(1) \geqslant \mathbf{P}(X \in A)\mathbf{H}(X) + \mathbf{P}(X \notin A)\mathbf{H}(X|X \notin A),$$

and hence

$$\mathbf{H}(X|X \notin A) \leqslant \mathbf{H}(X) + O(1/\mathbf{P}(X \notin A)). \quad (5.13)$$

A similar argument (swapping the roles of A and its complement) gives

$$\mathbf{H}(X|X \in A) \leqslant \mathbf{H}(X) + O(1/\mathbf{P}(X \in A)). \quad (5.14)$$

But from (5.10) and (5.11),

$$\mathbf{H}(X|X \in A) = \log|A| + O(1). \quad (5.15)$$

Combining this with (5.14) and (5.11), we obtain the upper bound $\log|A| \leqslant \mathbf{H}(X) + O(1)$.

Now we establish the lower bound. From (5.15) and (5.14),

$$\mathbf{H}(X|Y) \leqslant \mathbf{P}(X \in A)\log|A| + \mathbf{P}(X \notin A)\mathbf{H}(X) + O(1).$$

Since Y is boolean, we have $\mathbf{H}(Y) \leqslant \log 2$. In particular

$$\mathbf{H}(X|Y) = \mathbf{H}(X, Y) - \mathbf{H}(Y) \geqslant \mathbf{H}(X) - \log 2.$$

Combining this with the previous bound and (5.11), we see that

$$\log|A| \geqslant \mathbf{H}(X) - O\left(\frac{1}{\mathbf{P}(X \in A)}\right) \geqslant \mathbf{H}(X) - O(1),$$

as desired. □

Now we show that A has large additive energy.

Proposition 5.5 (A has large energy). *Let $\sigma[X] \leqslant K$, and let A be the set in Proposition 5.4. Then $|\{a_1, a_2, a_3, a_4 \in A : a_1 + a_2 = a_3 + a_4\}| \gg_K |A|^3$.*

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Proof. Again, we allow implied constants to depend on K. Let X_1, X_2 be independent copies of X, and let Y_1, Y_2 be the indicators $Y_i = 1_{X_i \in A}$. We have

$$\begin{aligned}\mathbf{H}(X) + O(1) &\geqslant \mathbf{H}(X_1 + X_2)\\ &\geqslant \mathbf{H}(X_1 + X_2|Y_1, Y_2)\\ &= \mathbf{P}(X_1 \in A)\mathbf{P}(X_2 \in A)\mathbf{H}(X_1 + X_2|X_1, X_2 \in A)\\ &\quad + \mathbf{P}(X_1 \in A)\mathbf{P}(X_2 \notin A)\mathbf{H}(X_1 + X_2|X_1 \in A; X_2 \notin A)\\ &\quad + \mathbf{P}(X_1 \notin A)\mathbf{P}(X_2 \in A)\mathbf{H}(X_1 + X_2|X_1 \notin A; X_2 \in A)\\ &\quad + \mathbf{P}(X_1 \notin A)\mathbf{P}(X_2 \notin A)\mathbf{H}(X_1 + X_2|X_1, X_2 \notin A)\end{aligned}$$

and

$$\begin{aligned}\mathbf{H}(X) &\leqslant \frac{1}{2}\mathbf{H}(X_1, Y_1) + \frac{1}{2}\mathbf{H}(X_2, Y_2)\\ &\leqslant \frac{1}{2}\mathbf{H}(X_1|Y_1) + \frac{1}{2}\mathbf{H}(X_2|Y_2) + \log 2\\ &= \mathbf{P}(X_1 \in A)\mathbf{P}(X_2 \in A)\left(\frac{1}{2}\mathbf{H}(X_1|X_1 \in A) + \frac{1}{2}\mathbf{H}(X_2|X_2 \in A)\right)\\ &\quad + \mathbf{P}(X_1 \in A)\mathbf{P}(X_2 \notin A)\left(\frac{1}{2}\mathbf{H}(X_1|X_1 \in A) + \frac{1}{2}\mathbf{H}(X_2|X_2 \notin A)\right)\\ &\quad + \mathbf{P}(X_1 \notin A)\mathbf{P}(X_2 \in A)\left(\frac{1}{2}\mathbf{H}(X_1|X_1 \notin A) + \frac{1}{2}\mathbf{H}(X_2|X_2 \in A)\right)\\ &\quad + \mathbf{P}(X_1 \notin A)\mathbf{P}(X_2 \notin A)\left(\frac{1}{2}\mathbf{H}(X_1|X_1 \notin A) + \frac{1}{2}\mathbf{H}(X_2|X_2 \notin A)\right)\\ &\quad + \log 2.\end{aligned}$$

Now, applying Lemma 2.1 we have

$$\mathbf{H}(X_1 + X_2|X_1 \in A_1, X_2 \in A_2) \geqslant \frac{1}{2}\mathbf{H}(X_1|X_1 \in A_1) + \frac{1}{2}\mathbf{H}(X_2|X_2 \in A_2)$$

for any sets A_1, A_2. Inserting this into the first estimate and then subtracting from the second, we conclude in particular that

$$\mathbf{P}(X_1 \in A)\mathbf{P}(X_2 \in A)(\mathbf{H}(X_1 + X_2|X_1, X_2 \in A) - \frac{1}{2}\mathbf{H}(X_1|X_1 \in A) - \frac{1}{2}\mathbf{H}(X_2|X_2 \in A)) \leqslant O(1),$$

and hence (by (5.11))

$$\mathbf{H}(X_1 + X_2|X_1, X_2 \in A) \leqslant \mathbf{H}(X|X \in A) + O(1).$$

Let X' be the random variable X conditioned to the event $X \in A$. Then X' now obeys the hypotheses of Proposition 5.4 (with K replaced by a larger but still bounded quantity). Repeating the derivation of (5.8), we conclude

$$\sum_z \sum_{y : p_{X'}(y), p_{X'+y}(z) \asymp p_{Z'}(z)} p_{X'}(y) p_{X'+y}(z) \geqslant 1/2,$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

where Z' is the sum of two independent copies of X'. Observe that the summand here vanishes unless $y, z - y \in A$, in which case the summand is $O(1/|A|^2)$. Setting $x := z - y$, we conclude that

$$|\{x, y \in A : p_{Z'}(x + y) \asymp 1/|A|\}| \gg |A|^2. \tag{5.16}$$

Since $p_{Z'}(z) \asymp |\{a, a' \in A : a + a' = z\}|/|A|^2$, the claim follows. □

From Proposition 5.5 (or (5.16)), one can find $E \subset A \times A$ with $|E| \gg_K |A|^2$ such that $|A \overset{E}{+} A| \ll_K |A|$. Applying the Balog–Szemerédi–Gowers theorem (Lemma 1.1), we conclude that there exists a subset A' of A with $|A' + A'| \asymp_K |A'| \asymp_K |A|$. Applying Freiman's theorem in an arbitrary additive group (Theorem 1.3), we conclude the following.

Corollary 5.6 (Concentration on a coset progression). *If $\sigma[X] \leqslant K$, then there exists a coset progression $H + P$ of rank $O_K(1)$ and cardinality*

$$|H + P| \asymp_K \exp(\mathbf{H}(X)),$$

such that

$$p_X(x) \asymp_K \exp(-\mathbf{H}(X)) \tag{5.17}$$

for $\gg_K |H + P|$ elements x of $H + P$.

Now we are ready to prove Proposition 5.2. Let X_1, X_2 be independent copies of X, let Y_2 be the indicator of the event $X_2 \in H + P$, and let X_2' be the conditioning of X_2 to the event $X_2 \in H + P$. Let U be a uniform distribution on $H + P$, taken to be independent of X_1, X_2'. From Corollary 5.6 we have

$$\mathbf{P}(X_2 \in H + P) \asymp 1 \tag{5.18}$$

and

$$\mathbf{H}(X_2') = \mathbf{H}(X) + O(1) = \log |H + P| + O(1). \tag{5.19}$$

(The lower bound on $\mathbf{H}(X_2')$ follows from (5.17) and the definition of entropy; the upper bound follows from Jensen's inequality, Lemma A.1.)

Next, observe that

$$\begin{aligned}
\mathbf{H}(X_1 + X_2) &\geqslant \mathbf{H}(X_1 + X_2|Y_2) \\
&= \mathbf{P}(X_2 \in H + P)\mathbf{H}(X_1 + X_2|X_2 \in H + P) \\
&\quad + \mathbf{P}(X_2 \notin H + P)\mathbf{H}(X_1 + X_2|X_2 \notin H + P) \\
&\geqslant \mathbf{P}(X_2 \in H + P)\mathbf{H}(X_1 + X_2') + \mathbf{P}(X_2 \notin H + P)\mathbf{H}(X_1) \\
&= \mathbf{H}(X_1) + \mathbf{P}(X_2 \in H + P)(\mathbf{H}(X_1 + X_2') - \mathbf{H}(X_1)).
\end{aligned}$$

By the hypothesis $\sigma[X] \leqslant K$, we have $\mathbf{H}(X_1 + X_2) \leqslant \mathbf{H}(X_1) + O(1)$. Applying (5.18), we conclude that

$$\mathbf{H}(X_1 + X_2') \leqslant \mathbf{H}(X_1) + O(1).$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

From this and (5.19), we see that $\mathrm{dist}_R(X_1, -X_2') = O(1)$. Meanwhile, from Jensen's inequality we have

$$\begin{aligned}\mathbf{H}(X_2' - U) &\leqslant \log|(H+P)-(H+P)| \\ &\leqslant \log|H+P| + O(1),\end{aligned}$$

which implies that $\mathrm{dist}_R(X_2', U) = O(1)$. Applying the triangle inequality (1.16), we obtain the claim.

The proof of Proposition 5.2, and thus Theorem 1.8, is now complete.

6. Proof of Theorem 1.9

We now prove Theorem 1.9. The basic idea is to get enough control on X that one can find a 'smooth' direction in which to approximate the discrete random variable by a continuous one.

Fix ε, and assume X to be a G-random variable with $\mathbf{H}(X)$ sufficiently large depending on ε. We assume for contradiction that the claim failed; thus (after adjusting ε slightly)

$$\mathbf{H}(X_1 + X_2) < \mathbf{H}(X) + \frac{1}{2}\log 2 - \varepsilon.$$

We can then apply Theorem 1.8(ii) and express $X = U + Z$, where U is the uniform distribution in a coset progression $H + P$ of rank $O(1)$ and cardinality $O(\exp(\mathbf{H}(X)))$, and $\mathbf{H}(Z) = O(1)$. Since G is torsion-free, the H component of the coset progression is trivial; thus U is just the uniform distribution on P.

Since $\mathbf{H}(Z) = O(1)$, we have

$$\sum_z p_Z(z)\log\frac{1}{p_Z(z)} = O(1).$$

Let $0 < \delta < 1/2$ be a small number depending on ε to be chosen later. Let $A := \{z : p_Z(z) \geqslant \delta\}$; thus $|A| \leqslant 1/\delta$. Also, since

$$\sum_z p_Z(z)\log\frac{1}{p_Z(z)} \geqslant \left(\log\frac{1}{\delta}\right)\sum_{z\notin A} p_Z(z) = \left(\log\frac{1}{\delta}\right)\mathbf{P}(Z \notin A),$$

we see that

$$\mathbf{P}(Z \notin A) \ll \frac{1}{\log\frac{1}{\delta}}.$$

This implies that

$$\mathbf{H}(1_{Z\in A}) \ll \frac{\log\log\frac{1}{\delta}}{\log\frac{1}{\delta}}.$$

This implies from (A. 14) that

$$\mathbf{H}(X|1_{Z\in A}) \geqslant \mathbf{H}(X) - O\left(\frac{\log\log\frac{1}{\delta}}{\log\frac{1}{\delta}}\right),$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

and thus

$$\mathbf{H}(X|Z \in A)\mathbf{P}(Z \in A) + \mathbf{H}(X|Z \notin A)\mathbf{P}(Z \notin A) \geqslant \mathbf{H}(X) - O\left(\frac{\log\log \frac{1}{\delta}}{\log \frac{1}{\delta}}\right). \tag{6.1}$$

If we let X_1, Z_1 and X_2, Z_2 be independent copies of X, Z, then we have

$$\begin{aligned}\mathbf{H}(X_1 + X_2) &\geqslant \mathbf{H}(X_1 + X_2|1_{Z_1\in A}, 1_{Z_2\in A})\\ &\geqslant \mathbf{H}(X_1 + X_2|Z_1, Z_2 \in A)\mathbf{P}(Z \in A)^2\\ &\quad + \mathbf{H}(X_1 + X_2|Z_1 \in A; Z_2 \notin A)\mathbf{P}(Z \in A)(1 - \mathbf{P}(Z \in A))\\ &\quad + \mathbf{H}(X_1 + X_2|Z_2 \in A; Z_1 \notin A)\mathbf{P}(Z \in A)(1 - \mathbf{P}(Z \in A))\\ &\quad + \mathbf{H}(X_1 + X_2|Z_1, Z_2 \notin A)(1 - \mathbf{P}(Z \in A))^2.\end{aligned}$$

From Lemma 2.1 we have

$$\begin{aligned}\mathbf{H}(X_1 + X_2|Z_2 \in A; Z_1 \notin A) &\geqslant \mathbf{H}(X|Z \in A),\\ \mathbf{H}(X_1 + X_2|Z_1 \in A; Z_2 \notin A) &\geqslant \mathbf{H}(X|Z \notin A),\\ \mathbf{H}(X_1 + X_2|Z_1, Z_2 \notin A) &\geqslant \mathbf{H}(X|Z \notin A),\end{aligned}$$

so from (6.1) we see that

$$\begin{aligned}&\mathbf{H}(X_1 + X_2)\\ &\geqslant \mathbf{H}(X) - O\left(\frac{\log\log \frac{1}{\delta}}{\log \frac{1}{\delta}}\right) + (\mathbf{H}(X_1 + X_2|Z_1, Z_2 \in A) - \mathbf{H}(X|Z \in A))\mathbf{P}(Z \in A)^2.\end{aligned}$$

Thus, by taking δ small enough, it will suffice to show that

$$\mathbf{H}(X_1' + X_2') \geqslant \mathbf{H}(X') + \frac{1}{2}\log 2 - \varepsilon/2 \tag{6.2}$$

(say), where $X' := (X|Z \in A)$ and X_1', X_2' are independent copies of X'.

Observe that X' ranges in the set $A + P$; since $|A| \ll_\delta 1$, we may place $A + P$ inside a progression Q of rank $O_\delta(1)$ and size $O_\delta(\exp(\mathbf{H}(X))) = O_\delta(|P|)$; by [13, Theorem 1.9], we may assume that Q is 4-proper, and thus

$$Q = \{a + n_1v_1 + \cdots + n_dv_d : n_1 \in [0, N_1), \ldots, n_d \in [0, N_d)\}$$

for some $d = O_\delta(1)$ and integers $N_1, \ldots, N_d$, and the sums $a + n_1v_1 + \cdots + n_dv_d$ for $n_1 \in [0, 4N_1), \ldots, n_d \in [0, 4N_d)$ are all distinct. Using a Freiman isomorphism (see, *e.g.*, [12, Section 5.3]), we may thus identify Q with the box $B := [0, N_1) \times \cdots \times [0, N_d)$ in $\mathbf{Z}^d$.

Let X'' be the counterpart of X' in B; thus X'' is Freiman isomorphic to X'. Since $X' = (U + Z|Z \in A)$, with U the uniform distribution on P, we see that

$$p_{X''}(x) \ll_\delta 1/|P| \asymp_\delta 1/|B| \tag{6.3}$$

for all $x \in B$.

Now we establish some 'smoothness' in the probability distribution function $p_{X_1''+X_2''}$ in some short direction, as measured using the total variation metric (1.11).

Lemma 6.1 (Smoothness of $p_{X_1''+X_2''}$). *Let $0 < \mu < 1$. Then, if $\mathbf{H}(X)$ is sufficiently large depending on μ, δ, there exists a non-zero $r \in [0, N_1) \times \cdots \times [0, N_d)$ with $|r| \ll_\delta \mu^{-O_\delta(1)}$ such*

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

that

$$\mathrm{dist}_{TV}(X_1'' + X_2'' + r, X_1'' + X_2'') \ll_\delta \mu. \tag{6.4}$$

Proof. For this lemma it is convenient to embed B and X'' inside the finite group

$$G' := \mathbf{Z}/3N_1\mathbf{Z} \times \cdots \times \mathbf{Z}/3N_d\mathbf{Z}.$$

Thus $p_{X''}$ is now a function on G'. The left-hand side of (6.4) can thus be written as

$$\sum_{x \in G'} |p_{X''} * p_{X''}(x + r) - p_{X''} * p_{X''}(x)|,$$

where $p_{X''} * p_{X''}$ is the convolution

$$p_{X''} * p_{X''}(x) := \sum_{y \in G'} p_{X''}(y) p_{X''}(x - y).$$

We introduce the Fourier coefficients

$$\hat{p}_{X''}(\chi) := \sum_{x \in G'} p_{X''}(x)\overline{\chi(x)}$$

for all characters $\chi : G' \to S^1$ in the Pontryagin dual $\hat{G}'$ of G'. From Plancherel's theorem and (6.3),

$$\sum_{\chi \in \hat{G}'} |\hat{p}_{X''}(\chi)|^2 = |G'| \sum_{x \in G'} |p_{X''}(x)|^2 \ll_\delta 1. \tag{6.5}$$

Thus, if we set

$$\Lambda := \{\chi \in \hat{G}' : |\hat{p}_{X''}(\chi)| \geqslant \mu\}, \tag{6.6}$$

then

$$|\Lambda| \ll_\delta \mu^{-2}. \tag{6.7}$$

If $\mathbf{H}(X)$ is large, then $N_1 \cdots N_d$ is large. By the Kronecker approximation theorem (see, *e.g.*, [12, Corollary 3.25]), if $\mathbf{H}(X)$ is large enough depending on δ, μ, we may thus find a non-zero $r \in [0, N_1) \times \cdots \times [0, N_d)$ with $|r| \ll_\delta \mu^{-O_\delta(1)}$, such that

$$|\chi(r) - 1| \leqslant \mu^2 \tag{6.8}$$

for all $\chi \in \Lambda$.

Fix this r. From the Fourier inversion formula,

$$p_{X''} * p_{X''}(x) = \frac{1}{|G'|} \sum_{\chi \in \hat{G}'} \hat{p}_{X''}(\chi)^2 \chi(x),$$

and thus

$$p_{X''} * p_{X''}(x + r) - p_{X''} * p_{X''}(x) = \frac{1}{|G'|} \sum_{\chi \in \hat{G}'} \hat{p}_{X''}(\chi)^2 (\chi(r) - 1)\chi(x).$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

By Plancherel's theorem, we conclude

$$\sum_{x\in G'} |p_{X''} * p_{X''}(x+r) - p_{X''} * p_{X''}(x)|^2 = \frac{1}{|G'|}\sum_{\chi\in\hat{G}'} |\hat{p}_{X''}(\chi)|^4 |\chi(r)-1|^2.$$

From (6.7) and (6.8), the contribution of the terms with $\chi \in \Lambda$ are $O_\delta(\mu^2/|G'|)$; by (6.6) and (6.5), the contribution of the terms with $\chi \notin \Lambda$ are also $O_\delta(\mu^2/|G'|)$. We thus have

$$\sum_{x\in G'} |p_{X''} * p_{X''}(x+r) - p_{X''} * p_{X''}(x)|^2 \ll_\delta \mu^2/|G'|,$$

and the claim follows from the Cauchy–Schwarz inequality. □

Let $0 < \mu < 1$ be a small number (depending on δ, ε) to be chosen later, and let r be as in the above lemma. We can write $r = mr'$, where $m \geqslant 1$ is an integer with

$$m \ll_\delta \mu^{-O_\delta(1)}, \tag{6.9}$$

and r' is irreducible in $\mathbf{Z}^d$. Applying yet another Freiman isomorphism, we may then map B to the integers so that r maps to m. If X''' is the image of X'' under this isomorphism, then X''' is isomorphic to X', ranges over at most $|B|$ values, and

$$\mathrm{dist}_{TV}(X_1''' + X_2''' + m, X_1''' + X_2''') \ll_\delta \mu. \tag{6.10}$$

Our task is now to show that

$$\mathbf{H}(X_1''' + X_2''') \geqslant \mathbf{H}(X''') + \frac{1}{2}\log 2 - \varepsilon/2. \tag{6.11}$$

To motivate the general argument later, let us first consider the simpler *irreducible case* when $m = 1$; thus the distribution function $p_{X_1'''+X_2'''}(x)$ looks 'locally smooth'. To exploit this, let U be the *continuous* random variable uniformly distributed in $[0,1]$, independent of X'''. Recall that the continuous Shannon entropy $\mathbf{H}_\mathbf{R}(V)$ of a random variable on $\mathbf{R}$ with distribution $p_V(x)\,dx$ is given by

$$\mathbf{H}_\mathbf{R}(V) := \int_\mathbf{R} F(p_V(x))\,dx.$$

A short computation then relates the continuous Shannon entropy to the discrete entropy:

$$\mathbf{H}_\mathbf{R}(X''' + U) = \mathbf{H}(X''').$$

Now let us look at the continuous variable $V := X_1''' + U_1 + X_2''' + U_2$. We write

$$\mathbf{H}_\mathbf{R}(V) = \log|P| + \int_\mathbf{R} F(p_V(x)) - p_V(x)\log|P|\,dx,$$

where p_V is the density function of V.

Observe that for any $x \in [n, n+1]$, the density function $p_V(x)$ of V at x is equal to some average of $p_{X_1'''+X_2'''}(n)$, $p_{X_1'''+X_2'''}(n-1)$; thus

$$p_V(x) = p_{X_1'''+X_2'''}(n) + O(g(n)),$$

where

$$g(n) := |p_{X_1'''+X_2'''}(n) - p_{X_1'''+X_2'''}(n-1)|.$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

In particular, $p_V(x) \ll_\delta 1/|P|$, by (6.3). Using the elementary estimate

$$F(b) - b\log|P| = F(a) - a\log|P| + O_\delta\left(\left(\frac{\mu}{|P|} + |b-a|\right)\log\frac{1}{\mu}\right)$$

when $a, b \ll_\delta 1/|P|$ (which arises from the fact that $F'(c) = \log|P| + O_\delta(\log\frac{1}{\mu})$ for $\mu/|P| \leqslant c \ll_\delta 1/|P|$), we thus have

$$\begin{aligned}\mathbf{H}_{\mathbf{R}}(V) = \log|P| + \sum_{n\in \mathrm{range}(X_1'''+X_2''')+\{0,1\}} & F(p_{X_1'''+X_2'''}(n)) \\ & - p_{X_1'''+X_2'''}(n)\log|P| + O_\delta\left(\left(\frac{\mu}{|P|} + g(n)\right)\log\frac{1}{\mu}\right).\end{aligned}$$

From (6.10),

$$\sum_{n\in \mathrm{range}(X_1'''+X_2''')+\{0,1\}} \left(\frac{\mu}{|P|} + g(n)\right) \ll_\delta \mu,$$

and thus

$$\mathbf{H}_{\mathbf{R}}(V) = \mathbf{H}(X_1''' + X_2''') + O_\delta\left(\mu\log\frac{1}{\mu}\right).$$

On the other hand, from Shannon's inequality (1.20), we have

$$\mathbf{H}_{\mathbf{R}}(V) \geqslant \mathbf{H}_{\mathbf{R}}(X''' + U) + \frac{1}{2}\log 2.$$

Putting all this together, we obtain

$$\mathbf{H}(X_1''' + X_2''') \geqslant \mathbf{H}(X''') + \frac{1}{2}\log 2 - O_\delta\left(\mu\log\frac{1}{\mu}\right),$$

and the claim (6.11) follows by taking μ small enough.

Now we return to the general case, when m is not necessarily 1. We then introduce the random variable $W := X''' \bmod m \in \mathbf{Z}/m\mathbf{Z}$, and define W_1, W_2 analogously. Then

$$\mathbf{H}(X_1''' + X_2''') \geqslant \mathbf{H}(X_1''' + X_2'''|W_1).$$

Observe that $X_1''' + X_2'''$ and W_1 determine W_2, and thus

$$\mathbf{H}(X_1''' + X_2'''|W_1) = \mathbf{H}(W_2) + \mathbf{H}(X_1''' + X_2'''|W_1, W_2).$$

We can write

$$\mathbf{H}(X_1''' + X_2'''|W_1, W_2) = \sum_{w_1,w_2\in\mathbf{Z}/m\mathbf{Z}} p_{W_1}(w_1)p_{W_2}(w_2)\mathbf{H}(X_{1,w_1} + X_{2,w_2}),$$

where, for $i = 1, 2$, X_{i,w_i} is the $\mathbf{Z}$-random variable $(X_i - w_i)/m$ conditioned to the event $W_i = w_i$. Meanwhile,

$$\mathbf{H}(X''') = \mathbf{H}(W) + \sum_{w_1\in\mathbf{Z}/m\mathbf{Z}} p_{W_1}(w_1)\mathbf{H}(X_{1,w_1}),$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

and similarly with the 1 index replaced by 2, thus

$$\mathbf{H}(X''') = \mathbf{H}(W) + \sum_{w_1,w_2 \in \mathbf{Z}/m\mathbf{Z}} p_{W_1}(w_1) p_{W_2}(w_2) \frac{1}{2}(\mathbf{H}(X_{1,w_1}) + \mathbf{H}(X_{2,w_2})).$$

Putting all this together, we see that to show (6.11) it will suffice to prove that

$$\sum_{w_1,w_2 \in \mathbf{Z}/m\mathbf{Z}} p_{W_1}(w_1) p_{W_2}(w_2) [\mathbf{H}(X_{1,w_1} + X_{2,w_2}) - \frac{1}{2}(\mathbf{H}(X_{1,w_1}) + \mathbf{H}(X_{2,w_2}))] \geqslant \frac{1}{2} \log 2 - \varepsilon/2. \tag{6.12}$$

From Lemma 2.1, the expression in brackets is non-negative. Thus we may restrict the sum to a smaller range of w_1, w_2; more specifically, we will restrict to the range where

$$p_{W_1}(w_1), p_{W_2}(w_2) \geqslant \mu/m. \tag{6.13}$$

On this range, we have from (6.3) and (6.9) that

$$p_{X_{i,w_i}}(x) \ll_\delta \mu^{-O_\delta(1)}/|P|$$

for all $i = 1, 2$ and $x \in \mathbf{Z}$; also observe that $p_{X_{i,w_i}}$ takes on at most $|P|$ values. From (6.10) we have

$$\sum_{w_i \in \mathbf{Z}/m\mathbf{Z}} p_{W_i}(w_i) \operatorname{dist}_{TV}(X_{i,w_i} + 1, X_{i,w_i}) \ll_\delta \mu \tag{6.14}$$

for $i = 1, 2$. We will now restrict w_1, w_2 further, by imposing the additional restriction

$$\operatorname{dist}_{TV}(X_{i,w_i} + 1, X_{i,w_i}) \leqslant \mu^{1/2} \tag{6.15}$$

for $i = 1, 2$.

Now we repeat the arguments from the $m = 1$ case. Let U_1, U_2 be independent copies of the uniform distribution of $[0, 1]$. Then, as before, we have

$$\mathbf{H}_{\mathbf{R}}(X_{i,w_i} + U_i) = \mathbf{H}(X_{i,w_i})$$

and

$$\mathbf{H}_{\mathbf{R}}(X_{1,w_1} + U_1 + X_{2,w_2} + U_2) = \mathbf{H}(X_{1,w_1} + X_{2,w_2}) + O_\delta\left(\mu^{1/2} \log \frac{1}{\mu}\right).$$

Applying (1.20), we conclude

$$\mathbf{H}(X_{1,w_1} + X_{2,w_2}) - \frac{1}{2}(\mathbf{H}(X_{1,w_1}) + \mathbf{H}(X_{2,w_2})) \geqslant \frac{1}{2} \log 2 - O_\delta\left(\mu^{1/2} \log \frac{1}{\mu}\right).$$

To conclude the proof of (6.12), it thus suffices (taking μ small enough) to show that

$$\sum_{w_1,w_2} p_{W_1}(w_1) p_{W_2}(w_2) \geqslant 1 - \varepsilon/4$$

(say), where w_1, w_2 range over all pairs in $\mathbf{Z}/m\mathbf{Z}$ obeying (6.13) and (6.15). But the contribution of those w_1, w_2 that fail to obey (6.13) is $O(\mu)$, while from (6.14) the contribution of the w_1, w_2 that fail to obey (6.15) is $O(\mu^{1/2})$, and the claim follows by taking μ small enough.

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Acknowledgements

The material in Sections 2 and 3 is based on some unpublished notes of the author with Van Vu. We are indebted to the anonymous referee for a careful reading of the paper and many useful corrections and suggestions.

Appendix: Basic properties of entropy

The function $F(x) := x \log \frac{1}{x}$ defined in (1.6) has first derivative

$$F'(x) = \log \frac{1}{x} - 1$$

and second derivative

$$F''(x) = -\frac{1}{x}$$

for $x > 0$. From this one easily concludes that F is concave on $\mathbf{R}^+$, and increasing for $x < 1/e$. In particular, we have the upper bound

$$F(x) \leqslant F(1/e) = 1/e, \tag{A.1}$$

the inequality

$$F(y) \leqslant F(x) + F'(x)(y - x) \tag{A.2}$$

for all $y \geqslant 0$ and $x > 0$, and the subadditivity property

$$F(x + y) \leqslant F(x) + F(y) \tag{A.3}$$

for all $x, y \geqslant 0$. In particular, we have the triangle inequality

$$|F(a) - F(b)| \leqslant F(|a - b|) \tag{A.4}$$

for $0 \leqslant a, b \leqslant 1/e$. From the identity

$$F(x) + F'(x)(y - x) - F(y) = y\left(\frac{x}{y} - 1 - \log \frac{x}{y}\right)$$

and (A.2), we obtain the bound

$$F(x) + F'(x)(y - x) - F(y) = y \log_+ \frac{y}{x} + O(x) + O(y), \tag{A.5}$$

where $\log_+ x := \max(\log x, 0)$. Finally, from the identity

$$F(ax) = F(a)F(x)\left(\frac{1}{\log \frac{1}{a}} + \frac{1}{\log \frac{1}{x}}\right),$$

we see that

$$F(ax) \leqslant 2F(a)F(x) \tag{A.6}$$

whenever $0 \leqslant a, x \leqslant 1/e$.

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

Lemma A.1 (Jensen bound). *Let A be a finite set, and let X be an A-random variable. Then $\mathbf{H}(X) \leqslant \log|A|$. Furthermore, if*

$$\mathbf{H}(X) \geqslant \log|A| - \log K$$

for some $K \geqslant 1$, then

$$\sum_{k=1}^{\infty} 2^k \mathbf{P}(X \in A_k) \ll 1 + \log K,$$

where

$$A_k := \{x \in A : 2^{2^{k-1}} \leqslant p_X(x)|A| \leqslant 2^{2^k}\}.$$

Proof. For the first bound, we observe that

$$\begin{aligned}\mathbf{H}(X) &= \sum_{x\in A} F(p_X(x)) \\ &\leqslant \sum_{x\in A} F\left(\frac{1}{|A|}\right) + F'\left(\frac{1}{|A|}\right)\left(p_X(x) - \frac{1}{|A|}\right) \\ &= \log|A|,\end{aligned}$$

as required. Similarly, if $\mathbf{H}(X) \geqslant \log|A| - \log K$, then the above argument shows that

$$\sum_{x\in A} F\left(\frac{1}{|A|}\right) + F'\left(\frac{1}{|A|}\right)\left(p_X(x) - \frac{1}{|A|}\right) - F(p_X(x)) \leqslant \log K.$$

From (A. 2), the summand is non-negative. From (A. 5), the summand is $p_X(x)\log(|A|p_X(x)) + O(p_X(x))$ for $p_X(x) \geqslant 1/|A|$, and the claim follows by decomposing the x variable into the sets A_k. □

Let X be a discrete random variable, and let E be an event which occurs with positive probability. Then we can define the *conditioned random variable* $(X|E)$ by restricting the underlying probability measure to E (and then dividing out by $\mathbf{P}(E)$ to recover the normalization); thus

$$p_{(X|E)}(x) = \mathbf{P}(x \in X \wedge E)/\mathbf{P}(E).$$

In the special case where E is an event of the form $X \in A$ for some set A, we conclude that

$$p_{(X|X\in A)}(x) = \frac{1_A(x)p_X(x)}{\sum_{y\in A} p_X(y)}.$$

Given two random variables X, Y (not necessarily independent), we define the *conditional entropy* $\mathbf{H}(X|Y)$ by the formula

$$\mathbf{H}(X|Y) := \sum_{y\in \mathrm{range}(Y)} p_Y(y)\mathbf{H}(X|Y = y). \tag{A. 7}$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

A standard computation reveals the identity

$$\mathbf{H}(X|Y) = \mathbf{H}(X,Y) - \mathbf{H}(Y), \tag{A.8}$$

and in particular

$$\mathbf{H}(X|Y) = \mathbf{H}(X,Y|Y). \tag{A.9}$$

Meanwhile, we have the total probability formula

$$p_X(x) = \sum_{y \in \mathrm{range}(Y)} p_Y(y) p_{(X|Y=y)}(x).$$

Comparing this with (A. 7) and Jensen's inequality using the concavity of F, we conclude that

$$\mathbf{H}(X|Y) \leqslant \mathbf{H}(X), \tag{A.10}$$

with equality if and only if $(X|Y = y) \equiv X$ for all $y \in \mathrm{range}(Y)$, or in other words if X and Y are independent. From this and (A. 8), we conclude that

$$\mathbf{H}(X,Y) \leqslant \mathbf{H}(X) + \mathbf{H}(Y). \tag{A.11}$$

We say that a discrete random variable Y is *determined* by another discrete random variable X if $Y = f(X)$ for some function $f : \mathrm{range}(X) \to \mathrm{range}(Y)$. From the sub-additivity property (A. 3), we see that

$$\mathbf{H}(Y) \leqslant \mathbf{H}(X) \tag{A.12}$$

whenever X determines Y. For instance, since (X,Y) determines both X and Y,

$$\mathbf{H}(X), \mathbf{H}(Y) \leqslant \mathbf{H}(X,Y), \tag{A.13}$$

and hence, by (A. 8) and (A. 10),

$$\mathbf{H}(X) - \mathbf{H}(Y) \leqslant \mathbf{H}(X|Y) \leqslant \mathbf{H}(X). \tag{A.14}$$

If X determines Y, then X and (X,Y) determine each other, and so $\mathbf{H}(X,Y) = \mathbf{H}(X)$; in particular,

$$\mathbf{H}(X|Y) = \mathbf{H}(X) - \mathbf{H}(Y) \tag{A.15}$$

and $\mathbf{H}(Y|X) = 0$.

We have the following useful inequality.

Lemma A.2 (Submodularity inequality). *If X_0, X_1, X_2, X_{12} are random variables such that X_1 and X_2 each determine X_0, and (X_1, X_2) determine X_{12}, then*

$$\mathbf{H}(X_{12}) + \mathbf{H}(X_0) \leqslant \mathbf{H}(X_1) + \mathbf{H}(X_2).$$

Proof. By (A. 15) and (A. 14) it suffices to show that

$$\mathbf{H}(X_{12}|X_0) \leqslant \mathbf{H}(X_1|X_0) + \mathbf{H}(X_2|X_0).$$

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press

By (A. 7) it suffices to show that

$$\mathbf{H}(X_{12}|X_0 = x_0) \leqslant \mathbf{H}(X_1|X_0 = x_0) + \mathbf{H}(X_2|X_0 = x_0)$$

for all $x_0 \in \text{range}(X_0)$. But by hypothesis, $(X_1|X_0 = x_0)$ and $(X_2|X_0 = x_0)$ determine $(X_{12}|X_0 = x_0)$, and the claim then follows from (A. 11) and (A. 12). □

As a special case of Lemma A.2 (and (A. 9)), we see that

$$\mathbf{H}(Y|Z) \leqslant \mathbf{H}(X|Z) \tag{A. 16}$$

whenever (X, Z) determines Y. Similarly, we have

$$\mathbf{H}(X, Y|Z) \leqslant \mathbf{H}(X|Z) + \mathbf{H}(Y|Z) \tag{A. 17}$$

for any X, Y, Z, with equality if and only if $(X|Z = z)$ and $(Y|Z = z)$ are independent for all $z \in \text{range}(Z)$, *i.e.*, if X and Y are *conditionally independent* relative to Z.

References

[1] Artstein, S., Ball, K., Barthe, F. and Naor, A. (2004) Solution of Shannon's problem on the monotonicity of entropy. *J. Amer. Math. Soc.* **17** 975–982.

[2] Balister, P. and Bollobás, B. Projections, entropy and sumsets. *Combinatorica*, to appear.

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

[4] Gowers, T. (1998) A new proof of Szemerédi's theorem for arithmetic progressions of length four. *Geom. Funct. Anal.* **8** 529–551.

[5] Green, B. and Ruzsa, I. (2007) Freiman's theorem in an arbitrary abelian group. *J. London Math. Soc.* **75** 163–175.

[6] Gyarmati, K., Matolcsi, M. and Ruzsa, I. A superadditivity and submultiplicativity property for cardinalities of sumsets. *Combinatorica*, to appear.

[7] Madiman, M., Marcus, A. and Tetali, P. Entropy and set cardinality inequalities for partition-determined functions, with applications to sumsets. Preprint.

[8] Plünnecke, H. (1969) Eigenschaften und Abschätzungen von Wirkungsfunktionen. BMwF-GMD-22 Gesellschaft für Mathematik und Datenverarbeitung, Bonn.

[9] Ruzsa, I. (1996) Sums of finite sets. In *Number Theory: New York Seminar* (D. V. Chudnovsky, G. V. Chudnovsky and M. B. Nathanson, eds), Springer.

[10] Tao, T. (2008) Product set estimates for non-commutative groups. *Combinatorica* (5) **28** 547–594.

[11] Tao, T. Freiman's theorem for solvable groups. Preprint.

[12] Tao, T. and Vu, V. (2006) *Additive Combinatorics*, Cambridge University Press.

[13] Tao, T. and Vu, V. (2008) John-type theorems for generalized arithmetic progressions and iterated sumsets. *Adv. Math.* **219** 428–449.

https://doi.org/10.1017/S0963548309990642 Published online by Cambridge University Press