

# Information Inequalities via Ideas from Additive Combinatorics

Chin Wa (Ken) Lau and Chandra Nair  
 Department of Information Engineering  
 The Chinese University of Hong Kong  
 Sha Tin, NT, Hong Kong  
 {kenlau,chandra}@ie.cuhk.edu.hk

## Abstract

Ruzsa's equivalence theorem provided a framework for converting certain families of inequalities in additive combinatorics to entropic inequalities (which sometimes did not possess stand-alone entropic proofs). In this work, we first establish formal equivalences between some families (different from Ruzsa) of inequalities in additive combinatorics and entropic ones. Secondly, we provide stand-alone entropic proofs for some previously known entropic inequalities that we established via Ruzsa's equivalence theorem. As a first step to further these equivalences, we provide an information theoretic characterization of the magnification ratio that is also of independent interest.

## I. INTRODUCTION

### A. Background

We seek to investigate and build upon the analogies and equivalence theorems between sumset inequalities in additive combinatorics and entropic inequalities in information theory. We are directly motivated by the work of Ruzsa [1] where a formal equivalence theory was proposed and established for certain families of sumset inequalities. Ruzsa categorized the inequalities into three scenarios [1]:

Scenario *A*: There exists an equivalence form (see Theorem 3) and explicit implication between a combinatorial inequality and an associated entropic inequality.

Scenario *B*: There exists a structural analog between a combinatorial inequality and an entropic inequality, but no direct equivalence is known. Sometimes, one directional implication could be established.

Scenario *C*: There is a combinatorial/entropic inequality, but the correctness of counterpart (analogous) inequality is unknown.

Most of the subsequent work has been done along the lines of the second scenario, i.e. analogous entropic inequalities without there being a formal equivalence. Tao established entropic analogs of the Plünnecke-Ruzsa-Frieman sumset and inverse sumset theory in 2010 [2]. Madiman, Marcus and Tetali established some entropic analogs and equivalence theorems based on partition-determined functions of random variables in 2012 [3]. Also, Kontoyiannis and Madiman explored the connection between sumsets and differential entropies [4]. We refer readers to [5], [6], [7] for more details. One can also a summary of the connection between combinatorial and entropic inequalities in [8].

The main contributions of this work are the following:

- a) we establish formal equivalence theorems (Theorem 1) between some combinatorial inequalities and entropic inequalities that Ruzsa had classified into Scenarios B and C. The entropic inequalities take a slightly different form than the analogous ones studied earlier. In some cases, the analogous entropic inequalities are stronger (Remark 3) while in some other cases the analogous (sometimes conjectured) ones does not imply the equivalent entropic inequalities (Remark 12).
- b) we use information theory based arguments to establish some entropic inequalities established by Ruzsa in Scenario A, and some other analogous ones. As a result, we are able to relax some assumptions about the ambient group structure (see Theorem 5). A key idea here is an entropic equality (Lemma 3), motivated from an analogous combinatorial lemma.
- c) we prove an information-theoretic characterization of the magnification ratio (Theorem 7), which serves as a primitive (as evidenced from Ruzsa's lecture notes [9]) to larger families of sumset inequalities.

A completely independent motivation for this study stems from an attempt to prove the subadditivity of certain entropic functionals related to establishing capacity regions in network information theory. In one particular but fundamental instance - the Gaussian interference channel - it appears that the additive structure of the channel should play a key role in the proof of the requisite sub-additive inequality.

### B. Notation

We will use  $(\mathbb{G}, +)$  to denote an abelian group and  $(\mathbb{T}, +)$  to denote a finitely generated torsion-free abelian group. If  $A$  is a finite set, we use  $|A|$  to denote the cardinality of  $A$ .

## II. MAIN

### A. Equivalence between sumset inequalities and entropic inequalities

We state a simple fact below. There is a trivial equivalence between cardinality inequalities and entropy inequalities via the observation that  $\log |A + B| = \max_{P_{XY}} H(X + Y)$ , where  $X$  takes values in  $A$  and  $Y$  takes values in  $B$ . The equality is clearly obtained by taking a uniform distribution on the support of  $|A + B|$ . However, we are seeking non-trivial versions of equivalence theorems.

**Theorem 1.** (*Generalized Ruzsa-type Equivalence Theorem*) *Let  $(\mathbb{T}, +)$  be a finitely generated torsion-free abelian group. Let  $f_1, \dots, f_k$  and  $g_1, \dots, g_\ell$  be linear functions on  $\mathbb{T}^n$  with integer coefficients, and let  $\alpha_1, \dots, \alpha_k, \beta_1, \dots, \beta_\ell$  be positive real numbers. For the linear function  $f_i$ , let  $S_i \subseteq [1 : n]$  denote the index set of non-zero coefficients. Similarly, for  $g_i$  let  $T_i \subseteq [1 : n]$  denote the corresponding index set of non-zero coefficients. (So, effectively,  $f_i$  and  $g_i$  are linear functions on  $\mathbb{T}_{S_i}$  and  $\mathbb{T}_{T_i}$  respectively). Further let us assume that  $S_i$  is a pairwise disjoint collection of sets. Then following statements are equivalent:*

a) *For any  $A_1, A_2, \dots, A_n$  that are finite subsets of  $\mathbb{T}$ , we have*

$$\prod_{i=1}^k |f_i(A_{S_i})|^{\alpha_i} \leq \prod_{i=1}^{\ell} |g_i(A_{T_i})|^{\beta_i},$$

where  $A_S = \otimes_{i \in S} A_i$ .

b) *For any  $m \in \mathbb{N}$ , and for any  $\hat{A}_1, \hat{A}_2, \dots, \hat{A}_n$  that are finite subsets of  $\mathbb{T}^m$ , we have*

$$\prod_{i=1}^k |\hat{f}_i(\hat{A}_{S_i})|^{\alpha_i} \leq \prod_{i=1}^{\ell} |\hat{g}_i(\hat{A}_{T_i})|^{\beta_i},$$

where  $\hat{A}_S = \otimes_{i \in S} \hat{A}_i$ , and  $\hat{f}_i$  (and  $\hat{g}_i$ ) are the natural coordinatewise extensions of  $f_i$  (and  $g_i$ ) respectively, mapping points in

$$\underbrace{\mathbb{T}^m \times \mathbb{T}^m \times \dots \times \mathbb{T}^m}_{n \text{ times}} \mapsto \mathbb{T}^m.$$

c) *For every sequence of random variables  $(X_1, \dots, X_n)$ , with fixed marginals  $P_{X_i}$  and having finite support in  $\mathbb{T}$ , we have*

$$\sum_{i=1}^k \alpha_i \max_{\Pi(X_{S_i})} H(f_i(X_{S_i})) \leq \sum_{i=1}^{\ell} \beta_i \max_{\Pi(X_{T_i})} H(g_i(X_{T_i})),$$

where  $\Pi(X_S)$  is collection of joint distributions  $P_{X_S}$  that are consistent with the marginals  $P_{X_i}, i \in S$ .

*Proof.* We will show that a)  $\implies$  b), b)  $\implies$  c), and c)  $\implies$  a). We make a brief remark on the three implications. That a)  $\implies$  b) has been used by Ruzsa in [1] and this is where the requirements that the functions be linear and that the ambient group be finitely generated and torsion-free play a crucial role. Now b)  $\implies$  c) is a rather standard argument in information theory community using the method of types (see Chapter 2 of [10]), and Sanov's theorem. Finally c)  $\implies$  a) is quite immediate by taking specific marginal distributions that induce uniform distributions on the support of  $f_i(X_{S_i})$  and is where the requirement that  $S_i$  be pairwise disjoint plays a role.

a)  $\implies$  b): We outline the method used by Ruzsa in [1]. By the classification theorem of finitely generated abelian groups, we know that a torsion-free finitely generated abelian group is isomorphic to  $\mathbb{Z}^d$ , for a finite  $d$ . We denote  $t$  to be a generic element in  $\mathbb{T}$ , (or equivalently  $\mathbb{Z}^d$ ). Let a linear function with integer coefficients  $f : \mathbb{T}^n \mapsto \mathbb{T}$ , be defined by  $f(t_1, \dots, t_n) = \sum_{i=1}^n a_i t_i$ . (In the context of our discussion, the locations of the non-zero values of  $a_i$  determine the support of  $f$ ). Similarly we denote  $\mathbf{t} = (t_1, \dots, t_m)$  to be a generic element in  $\mathbb{T}^m$ . Therefore, we have  $\hat{f}(\mathbf{t}_1, \dots, \mathbf{t}_n) = \sum_{i=1}^n a_i \mathbf{t}_i$ . Let  $\psi_q$  be a linear mapping from  $\mathbb{T}^m$  to  $\mathbb{T}$  defined as

$$\psi_q(\mathbf{t}) := t_1 + t_2 q + \dots + t_m q^{m-1}.$$

Observe that, by linearity,

$$\psi_q(\hat{f}(\mathbf{t}_1, \dots, \mathbf{t}_n)) = \psi_q\left(\sum_{i=1}^n a_i \mathbf{t}_i\right) = f(\psi_q(\mathbf{t}_1), \dots, \psi_q(\mathbf{t}_n)). \quad (1)$$

Given the finite subsets  $\hat{A}_1, \dots, \hat{A}_n$  of  $\mathbb{T}^m$ , and the linear functions  $f_1, \dots, f_k$  and  $g_1, \dots, g_\ell$ , we can choose a  $q$  large enough that  $\psi_q(\hat{f}_i(\hat{A}_{S_i}))$  and  $\psi_q(\hat{g}_i(\hat{A}_{T_i}))$  are injections. Now set  $A_i = \psi_q(\hat{A}_i)$ . Therefore we have

$$|\hat{f}_i(\hat{A}_{S_i})| = |\psi_q(\hat{f}_i(\hat{A}_{S_i}))| \stackrel{(a)}{=} |f_i(\{\psi_q(\hat{A}_k)\}_{k \in S_i})| = |f_i(A_{S_i})|,$$

where (a) follows from (1). A similar equality holds for  $g$ 's as well. With these equalities, we have that a)  $\implies$  b).

b)  $\implies$  c): We are given a set of marginal distributions  $P_{X_1}, \dots, P_{X_n}$  whose supports are finite subsets of  $\mathbb{T}$ , say  $\mathcal{X}_1, \dots, \mathcal{X}_n$ . Consider a non-negative sequence  $\{\delta_m\}$ , where  $\delta_m \rightarrow 0$  and  $\sqrt{m} \cdot \delta_m \rightarrow \infty$  as  $m \rightarrow \infty$ . For every  $m$ , we construct the strongly typical sets  $\mathbb{T}_{(m, P_{X_i}, \delta_m)}$ , for  $1 \leq i \leq n$ , where

$$\mathbb{T}_{(m, P_{X_i}, \delta_m)} := \left\{ \mathbf{x} \in \mathcal{X}_i^m : \left| \frac{1}{m} N(a|\mathbf{x}) - P_{X_i}(a) \right| \leq \delta_m \cdot P_{X_i}(a) \text{ for any } a \in \mathcal{X}_i \right\}.$$

Suppressing dependence on other variables, let  $\hat{A}_i = \mathbb{T}_{(m, P_{X_i}, \delta_m)}$  for  $1 \leq i \leq n$ . Now consider a linear function  $f: \mathbb{T}_S \mapsto \mathbb{T}$  and let  $\hat{f}$  be the coordinate-wise extension of it to  $(\mathbb{T}^m)_S$ . Define  $Y = f(X_S)$ ,  $S \subseteq [1 : n]$ , and let  $\mathcal{M}_Y$  denote the set of probability distributions of  $Y$  induced by all couplings  $\Pi(X_S)$  that are consistent with the marginals  $P_{X_i}$  for  $i \in S$ . Let  $Q_Y$  be the uniform distribution on  $\mathcal{Y}$ , and by a routine application of Sanov's theorem we obtain that

$$\lim_{m \rightarrow \infty} \frac{1}{m} \log \frac{|\hat{f}(\hat{A}_S)|}{|\mathcal{Y}|^m} = \max_{P_Y \in \mathcal{M}_Y} H(Y) - \log |\mathcal{Y}| = \max_{\Pi(X_S)} H(f(X_S)) - \log |\mathcal{Y}|.$$

Therefore, we have

$$\lim_{m \rightarrow \infty} \frac{1}{m} |\hat{f}(\hat{A}_S)| = \max_{\Pi(X_S)} H(f(X_S)).$$

Thus the implication b)  $\implies$  c) is established.

c)  $\implies$  a): This is rather immediate. Since  $S_i$ 's are pairwise disjoint, let  $P_{X_{S_i}}$  induce a uniform distribution on  $f(A_{S_i})$  and let  $P_{X_i}$  be the induced marginals. Then it is clear that  $\max_{\Pi(X_{S_i})} H(f_i(X_{S_i})) = \log |f(A_{S_i})|$  and  $\max_{\Pi(X_{T_i})} H(g_i(X_{T_i})) \leq \log |g(A_{T_i})|$  and this completes the proof.  $\square$

The following corollaries to Theorem 1 lead to some entropic inequalities. Some of the sumset inequalities in literature are stated using Ruzsa-distance, and the equivalent entropic inequalities can be stated using a similar distance between distributions.

**Definition 1.** (Ruzsa Distance between Finite Sets, [11]) *The Ruzsa distance between two finite subsets  $A, B$  on an abelian group  $(\mathbb{G}, +)$  is defined as*

$$d_R(A, B) := \log \frac{|A - B|}{|A|^{1/2} |B|^{1/2}}.$$

**Remark 1.** *It is clear that  $d_R(A, B) = d_R(B, A)$  and that  $d_R(A, A) \geq 0$ .*

**Definition 2.** (Entropic Ruzsa Distance) *The entropic-Ruzsa "distance" between two distributions  $P_X, P_Y$  taking values in  $(\mathbb{G}, +)$  is defined as*

$$d_{HR}(X, Y) := \max_{P_{XY} \in \Pi(P_X, P_Y)} H(X - Y) - \frac{1}{2} H(X) - \frac{1}{2} H(Y),$$

where  $\Pi(P_X, P_Y)$  is the set of all coupling with the given marginals.

**Remark 2.** *The following remarks are worth noting with regards to the entropic Ruzsa-distance:*

- 1) *As with the abuse of notations in information theory  $d_{HR}(X, Y)$  is a function of  $P_X, P_Y$  and not of  $X$  and  $Y$ .*
- 2) *Just like the original Ruzsa distance between two sets, we have  $d_{HR}(X, Y) \geq 0$  (this follows by observing that when  $P_{XY} = P_X P_Y$ , we have  $H(X - Y) \geq \max\{H(X), H(Y)\}$  as  $0 \leq I(X; X - Y) = H(X - Y) - H(Y)$ ). Further it is immediate that  $d_{HR}(X, Y) = d_{HR}(Y, X)$ .*
- 3) *There is no relationship between  $d_{HR}(X, Y)$  and  $d_R(A, B)$  where  $A$  is the support of  $p_X$  and  $B$  is the support of  $p_Y$ .*
  - *Consider  $P_X$  and  $P_Y$  such that it is uniform on sets  $A$  and  $B$  respectively. Thus for any  $P_{XY} \in \Pi(P_X, P_Y)$  we have  $H(X - Y) \leq \log |A - B|$  and consequently  $d_{HR}(X, Y) \leq d_R(A, B)$  (and the inequality can be strict).*
  - *Consider a joint  $P_{XY}$  that uniform on  $A - B$  and let  $P_X$  and  $P_Y$  be its induced marginal distributions on sets  $A$  and  $B$  respectively. then as  $H(X) \leq \log |A|$  and  $H(Y) \leq \log |B|$ , we have  $d_{HR}(X, Y) \geq d_R(A, B)$  (and the inequality can be strict).*
- 4) *This definition is different from that of Tao [2], where he defines the similar quantity using independent coupling of  $P_X$  and  $P_Y$ . An advantage of our definition is that we have a formal equivalence between the two inequalities (one in sumset and one in entropy).*

Theorem 1 immediately implies the following entropic inequalities from the corresponding sumset inequalities.

**Corollary 1.** *For any distributions  $P_X, P_Y, P_Z$  with finite support on a finitely generated torsion-free group  $(\mathbb{T}, +)$ , we have*

$$d_{HR}(X, Z) \leq d_{HR}(X, Y) + d_{HR}(Y, Z),$$

or equivalently :  $H(Y) + \max_{\Pi(X, Z)} H(X - Z) \leq \max_{\Pi(X, Y)} H(X - Y) + \max_{\Pi(Y, Z)} H(Y - Z).$  (2)

*Proof.* In [11], Ruzsa showed that for any finite  $A, B, C$  on a finitely generated torsion-free abelian group  $(\mathbb{T}, +)$ , we have  $d_R(A, C) \leq d_R(A, B) + d_R(B, C)$ , or equivalently  $|B||A - C| \leq |A - B||B - C|$ . By applying Theorem 1, we will obtain the desired inequality.

**Remark 3.** *The entropic inequality in (2) can also be obtained as a direct consequence of a stronger entropic inequality that was established in [3]. There it was established that, if  $Y$  and  $(X, Z)$  are independent and taking values in an ambient abelian group  $(\mathbb{G}, +)$ , then one has  $H(Y) + H(X - Z) \leq H(X - Y) + H(Y - Z)$ . To see this observe that  $H(Y, X - Z) = H(X - Y, Y - Z) - I(X; Y - Z|X - Z)$ , and the requisite inequality is immediate.*

□

**Corollary 2.** *For distributions  $P_X, P_Y, P_Z$  with finite support on a finitely generated torsion-free group  $(\mathbb{T}, +)$ , we have*

$$H(X) + \max_{\Pi(Y, Z)} H(Y + Z) \leq \max_{\Pi(X, Y)} H(X + Y) + \max_{\Pi(X, Z)} H(X + Z). \quad (3)$$

*Proof.* In [11], Ruzsa showed that for any finite  $A, B, C$  on a finitely generated torsion-free abelian group  $(\mathbb{T}, +)$ , we have

$$|A||B + C| \leq |A + B||A + C|. \quad (4)$$

By applying Theorem 1, we obtain the desired entropic inequality.

□

**Remark 4.** *The authors are not aware of a stand-alone information-theoretic proof of the above inequality and our results in Section III is a step at building an information-theoretic counterpart to the sumset arguments used to establish this. When  $X, Y$ , and  $Z$  are mutually independent, an entropic analogue has been established in [3], [5]. Note that in this case, by the data-processing inequality, we have  $I(Z; X + Y + Z) \leq I(Z; X + Z)$  implying*

$$H(X) + H(Y + Z) \leq H(X) + H(X + Y + Z) \leq H(X + Y) + H(X + Z)$$

*A relaxation of this proof to the case when  $X$  is independent of  $(Y, Z)$  would have yielded (3); however this relaxation does not seem immediate.*

### B. Katz-Tao Sum Difference Inequality

The following lemma was established by Katz and Tao [12] and used in the proof of certain sumset inequalities.

**Lemma 1.** [12, Lemma 2.1] *Let  $A$  and  $B_1, \dots, B_{n-1}$  be finite sets for some  $n$ . Let  $f_i : A \rightarrow B_i$  be a function for all  $i \in [1 : n - 1]$ . Then*

$$\begin{aligned} & \{(a_1, \dots, a_n) \in A^n : f_i(a_i) = f_i(a_{i+1}) \forall i \in [1 : n - 1]\} \\ & \geq \frac{|A|^n}{\prod_{i=1}^{n-1} |B_i|}. \end{aligned}$$

Motivated by this lemma, we will prove an information-theoretic version (which would imply the combinatorial version) and will turn out to be useful in several of our arguments. We will first present a lemma in a more general form than is used in this paper.

**Lemma 2.** *Suppose the following Markov chain holds:*

$$X_1 \rightarrow U_1 \rightarrow X_2 \rightarrow U_2 \rightarrow \dots \rightarrow X_{n-1} \rightarrow U_{n-1} \rightarrow X_n.$$

*Then,*

$$H(X_1, \dots, X_n, U_1, \dots, U_{n-1}) + \sum_{i=1}^{n-1} I(X_i; U_i) + \sum_{i=1}^{n-1} I(U_i; X_{i+1}) = \sum_{i=1}^n H(X_i) + \sum_{i=1}^{n-1} H(U_i).$$

*Proof.* This lemma is an immediate consequence of Chain Rule for entropy as follows. Note that the chain rule and the Markov Chain assumption yields

$$\begin{aligned} H(X_1, \dots, X_n, U_1, \dots, U_{n-1}) &= H(X_1) + \sum_{i=1}^{n-1} H(U_i|X_i) + \sum_{i=1}^{n-1} H(X_{i+1}|U_i) \\ &= H(X_1) + \sum_{i=1}^{n-1} (H(U_i) - I(U_i; X_i)) + \sum_{i=1}^{n-1} (H(X_{i+1}) - I(U_i; X_{i+1})). \end{aligned}$$

Now rearranging yields the desired equality.

□

As a special case of Lemma 2 we obtain the following version that turns out to be useful in this paper.

**Lemma 3.** Let  $(X_i)_{i=1}^n$  be a sequence of finite-valued random variables (defined on some common probability space) and  $(f_i, g_i)_{i=1}^{n-1}$  be a sequence of functions that take a finite set of values in some space  $\mathcal{S}$  such that:  $f_i(X_i) = g_i(X_{i+1})(=: U_i)$  and the following Markov chain holds,

$$X_1 \rightarrow U_1 \rightarrow X_2 \rightarrow U_2 \rightarrow \cdots \rightarrow X_{n-1} \rightarrow U_{n-1} \rightarrow X_n.$$

Then,

$$H(X_1, \dots, X_n) + \sum_{i=1}^{n-1} H(U_i) = \sum_{i=1}^n H(X_i).$$

*Proof.* Note that  $H(X_1, \dots, X_n) = H(X_1, \dots, X_n, U_1, \dots, U_{n-1})$  since  $U_i$  is determined by  $X_i$  (and also by  $X_{i+1}$ ). Further we also have  $I(U_i; X_i) = I(U_i; X_{i+1}) = H(U_i)$  for  $1 \leq i \leq n-1$ . Hence the desired consequence follows from Lemma 2.  $\square$

**Remark 5.** The following remarks are worth noting:

- Lemma 3 seems to play a similar role as the copy lemma [13] used in deriving several non-Shannon type inequalities.
- Note that Lemma 3 will imply Lemma 1 directly. Define

$$C = \{(a_1, \dots, a_n) \in A^n : f_i(a_i) = f_i(a_{i+1}) \forall i \in [1 : n-1]\}.$$

Suppose  $X_1, \dots, X_n$  have uniform marginals on  $A$ . Set  $f_i(X_i) = f_i(X_{i+1})(=: U_i)$  and construct a joint distribution such that the following Markov chain holds,

$$X_1 \rightarrow U_1 \rightarrow X_2 \rightarrow U_2 \rightarrow \cdots \rightarrow X_{n-1} \rightarrow U_{n-1} \rightarrow X_n.$$

Observe that  $(X_1, \dots, X_n)$  has a support on  $C$ . This implies, from Lemma 3, that

$$\begin{aligned} n \log |A| &= \sum_{i=1}^n H(X_i) = H(X_1, \dots, X_n) + \sum_{i=1}^{n-1} H(U_i) \\ &\leq \log |C| + \sum_{i=1}^{n-1} \log |B_i|. \end{aligned}$$

**Definition 3.** (*G*-restricted Sumset [1]) Suppose  $G$  is a subset of  $A \times B$ , where  $A, B$  are finite subsets of  $(\mathbb{G}, +)$ . We denote the *G*-restricted sumset and difference set of  $A$  and  $B$  as  $A \overset{G}{+} B$  and  $A \overset{G}{-} B$ .

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

$$A \overset{G}{-} B = \{a - b : a \in A, b \in B, (a, b) \in G\}.$$

**Theorem 2.** (*Katz-Tao Sum-Difference Inequality* [12]) For any  $G$ , a finite subset of  $\mathbb{T} \times \mathbb{T}$ , we have

$$|A \overset{G}{-} B| \leq |A|^{2/3} |B|^{2/3} |A \overset{G}{+} B|^{1/2}.$$

In [1], Ruzsa established the entropy version of Katz-Tao sum-difference inequality by using a formal equivalence theorem between *G*-restricted sumset inequalities and entropic inequalities.

**Theorem 3** (Equivalence Theorem 2, [1]). Let  $f, g_1, \dots, g_k$  be linear functions in two variables with integer coefficients, and let  $\alpha_1, \dots, \alpha_k$  be positive real numbers. The following statements are equivalent:

- 1) For every finite  $A \subseteq \mathbb{T} \times \mathbb{T}$  we have

$$|f(A)| \leq \prod |g_i(A)|^{\alpha_i},$$

where  $|f(A)|$  denotes the cardinality of the image  $f(A)$ .

- 2) For every pair  $X, Y$  of (not necessarily independent) random variables with values in  $(\mathbb{T}, +)$  such that the entropy of each  $g(X, Y)$  is finite, the entropy of  $f(X, Y)$  is also finite and it satisfies

$$H(f(X, Y)) \leq \sum \alpha_i H(g_i(X, Y)).$$

**Remark 6.** It may be worthwhile mentioning a key difference between Theorem 1 and Theorem 3. The equivalence in Theorem 3 follows when the sum-set inequalities hold for every *G*-restricted sumset. On the other hand, most of the inequalities in literature are established for the Minkowski sum of sets, and Theorem 1 holds under such a situation.

Consequently, Ruzsa obtained the following entropic inequality by applying Theorem 3 to Theorem 2.

**Theorem 4.** [1] Suppose  $X$  and  $Y$  are random variables with finite support on  $(\mathbb{T}, +)$ , we have

$$H(X - Y) \leq \frac{2}{3}H(X) + \frac{2}{3}H(Y) + \frac{1}{2}H(X + Y). \quad (5)$$

The main aim of this section is to give a stand-alone entropic proof (previously not known to the best of the knowledge of the authors) of (5). Further the entropic proof only necessitates that  $X$  and  $Y$  take values in some ambient abelian group  $\mathbb{G}$  (thus is a slight generalization of the result in the literature) and this relaxation extends back to the sumset inequality as well.

**Theorem 5.** Suppose  $X$  and  $Y$  are random variables with finite support on an ambient abelian group  $\mathbb{G}$ , we have

$$H(X - Y) \leq \frac{2}{3}H(X) + \frac{2}{3}H(Y) + \frac{1}{2}H(X + Y). \quad (6)$$

Before we prove this theorem, we make the following observation.

**Lemma 4.** To prove (6), it suffices to consider  $P_{X,Y}$  such that  $X - Y$  implies  $(X, Y)$  with probability one.

*Proof.* Define  $f(P_{X,Y}) := \frac{2}{3}H(X) + \frac{2}{3}H(Y) + \frac{1}{2}H(X + Y) - H(X - Y)$ . Consider the closed convex set of probability distributions  $P_{X,Y}$  that have a support on  $\text{supp}(X) \times \text{supp}(Y)$  and have a fixed  $P_{X-Y}$ . Note that  $f(P_{X,Y})$  is concave on this convex set and hence the minimum occurs at the extreme points of this set. Since the extreme points of this set are  $P_{X,Y}$  such that  $X - Y$  implies  $(X, Y)$ , the lemma is established.  $\square$

*Proof of Theorem 5.* Suppose  $(X, Y)$  are random variables such that  $(X, Y)$  is determined by  $(X - Y)$ . Then consider a joint distribution  $(X, Y, Y^\dagger)$  such that  $Y \rightarrow X \rightarrow Y^\dagger$  forms a Markov chain and  $(X, Y)$  shares the same marginal as  $(X, Y^\dagger)$ . From Lemma 3 (considering  $(X, Y) - X - (X, Y^\dagger)$ ) we have

$$H(X, Y, Y^\dagger) = H(X, Y) + H(X, Y^\dagger) - H(X) = 2H(X - Y) - H(X). \quad (7)$$

Here, the last equality comes by combining the assumptions that  $(X, Y) \stackrel{(d)}{=} (X, Y^\dagger)$  and that  $(X, Y)$  is determined by  $(X - Y)$ .

The main idea is that by inducing couplings between the copies  $(X_i, Y_i, Y_i^\dagger)$  of  $(X, Y, Y^\dagger)$  via functions  $f_k$ , one can express the joint entropy of the copies using a smaller subset of the variables. To this end, define three functions:

$$\begin{aligned} f_1(x, y, y^\dagger) &= (x + y, x + y^\dagger), \\ f_2(x, y, y^\dagger) &= (y, y^\dagger), \\ f_3(x, y, y^\dagger) &= (x + y, y^\dagger). \end{aligned}$$

Consider a joint distribution of  $(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger)$  such that the following three conditions are satisfied:

- 1)  $(X_i, Y_i, Y_i^\dagger)$  shares the same marginal as  $(X, Y, Y^\dagger)$  for  $1 \leq i \leq 4$ . (they are copies)
- 2)  $f_i(X_i, Y_i, Y_i^\dagger) = f_i(X_{i+1}, Y_{i+1}, Y_{i+1}^\dagger)$  for  $1 \leq i \leq 3$ . (these induce the couplings)
- 3)  $(X_1, Y_1, Y_1^\dagger) \rightarrow f_1(X_1, Y_1, Y_1^\dagger) \rightarrow (X_2, Y_2, Y_2^\dagger) \rightarrow f_2(X_2, Y_2, Y_2^\dagger) \rightarrow (X_3, Y_3, Y_3^\dagger) \rightarrow f_3(X_3, Y_3, Y_3^\dagger) \rightarrow (X_4, Y_4, Y_4^\dagger)$  forms a Markov chain.

Now by Lemma 3, we have

$$H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger) = 4H(X, Y, Y^\dagger) - H(X + Y, X + Y^\dagger) - H(Y, Y^\dagger) - H(X + Y, Y^\dagger). \quad (8)$$

From condition 2) and the definition of  $f_1, f_2, f_3$  construction, we have the following equalities:

$$\begin{aligned} X_1 + Y_1 &= X_2 + Y_2, & X_1 + Y_1^\dagger &= X_2 + Y_2^\dagger, \\ Y_2 &= Y_3, & Y_2^\dagger &= Y_3^\dagger, & X_3 + Y_3 &= X_4 + Y_4, & Y_3^\dagger &= Y_4^\dagger. \end{aligned}$$

From this, we obtain the following:

$$Y_1 - Y_1^\dagger = Y_2 - Y_2^\dagger = Y_3 - Y_3^\dagger.$$

Consequently, we have

$$\begin{aligned} X_4 - Y_4^\dagger &= (X_4 + Y_4) - Y_4 - Y_4^\dagger \\ &= (X_3 + Y_3) - Y_4^\dagger - Y_4 = X_3 + (Y_3 - Y_3^\dagger) - Y_4 \\ &= X_3 + Y_1 - Y_1^\dagger - Y_4. \end{aligned}$$

Therefore  $X_4 - Y_4^\dagger$  is a function of  $(X_1, Y_1, Y_1^\dagger, X_3, Y_4)$  and since  $X - Y$  implies  $(X, Y)$  (from Lemma 4) we see that  $(X_4, Y_4^\dagger)$  is a function of  $(X_1, Y_1, Y_1^\dagger, X_3, Y_4)$ . To complete the argument, observe that  $Y_2 = Y_3 = X_4 + Y_4 - X_3$ ,  $Y_2^\dagger = Y_3^\dagger = Y_4^\dagger$ , and

$X_2 = X_1 + Y_1 - Y_2$ . This implies that  $(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger)$  is a function of  $(X_1, Y_1, Y_1^\dagger, X_3, Y_4)$  and hence

$$H(X_1, Y_1, Y_1^\dagger, X_2, Y_2, Y_2^\dagger, X_3, Y_3, Y_3^\dagger, X_4, Y_4, Y_4^\dagger) = H(X_1, Y_1, Y_1^\dagger, X_3, Y_4). \quad (9)$$

By using (8) and (9), we have

$$\begin{aligned} 0 &= 4H(X, Y, Y^\dagger) - H(X + Y, X + Y^\dagger) - H(Y, Y^\dagger) - H(X + Y, Y^\dagger) - H(X_1, Y_1, Y_1^\dagger, X_3, Y_4) \\ &= 3H(X, Y, Y^\dagger) - H(X + Y, X + Y^\dagger) - H(Y, Y^\dagger) - H(X + Y, Y^\dagger) - H(X_3, Y_4 | X_1, Y_1, Y_1^\dagger). \end{aligned}$$

Now using (7) to replace  $H(X, Y, Y^\dagger)$  we have

$$\begin{aligned} 0 &= 6H(X - Y) - 3H(X) - H(X + Y, X + Y^\dagger) - H(Y, Y^\dagger) - H(X + Y, Y^\dagger) - H(X_3, Y_4 | X_1, Y_1, Y_1^\dagger) \\ &\geq 6H(X - Y) - 3H(X) - H(X + Y) - H(X + Y^\dagger) - H(Y) - H(Y^\dagger) - H(X + Y) - H(Y^\dagger) - H(X_3) - H(Y_4) \\ &= 6H(X - Y) - 4H(X) - 4H(Y) - 3H(X + Y). \end{aligned}$$

This completes the proof of the theorem.  $\square$

**Remark 7.** *The crux of the argument presented here is not new. The ideas are borrowed from similar arguments in [12]. The purpose is mainly to illustrate that certain arguments in sumset literature have an almost verbatim counterpart in the entropic language.*

### C. Sum-difference Inequality

In this section we give some generalization of analogous entropic inequalities and this leads, in the reverse direction, to a sumset inequality that we had not seen in literature.

**Theorem 6.** (Sum-difference Inequality) [11, Theorem 5.3] *The Ruzsa distance between two finite subsets  $A, B$  on an abelian group  $(\mathbb{G}, +)$  satisfies*

$$\begin{aligned} d_R(A, -B) &\leq 3d_R(A, B), \\ \text{or equivalently } |A + B||A||B| &\leq |A - B|^3. \end{aligned} \quad (10)$$

**Proposition 1.** (Entropic Sum-difference Inequality) *Let  $X_1, Y_1, X_2, Y_2, X_3, Y_3$  be random variables (on a common probability space) with finite support on an abelian group  $(\mathbb{G}, +)$  such that  $X_1 - Y_1 = X_2 - Y_2$  ( $=: U$ ) and also satisfies that  $(X_1, Y_1) \rightarrow U \rightarrow (X_2, Y_2)$  forms a Markov chain. Further, suppose  $(X_1, Y_1, X_2, Y_2)$  and  $(X_3, Y_3)$  are independent. Then the following inequality holds:*

$$H(X_1, Y_1) + H(X_2, Y_2) + H(X_3 + Y_3) \leq H(X_1 - Y_1) + H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1). \quad (11)$$

*Proof.* Since  $U = X_1 - Y_1 = X_2 - Y_2$  and  $(X_1, Y_1) \rightarrow U \rightarrow (X_2, Y_2)$  forms a Markov chain, from Lemma 3 we have

$$H(X_1, Y_1, X_2, Y_2) + H(U) = H(X_1, Y_1) + H(X_2, Y_2) \quad (12)$$

We now decompose  $H(X_1, Y_1, X_2, Y_2, X_3, Y_3 | X_3 + Y_3)$  in two ways. Firstly, since  $(X_1, Y_1, X_2, Y_2)$  and  $(X_3, Y_3)$  are independent, we have

$$\begin{aligned} &H(X_1, Y_1, X_2, Y_2, X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_1, X_2, Y_2) + H(X_3, Y_3 | X_3 + Y_3) \\ &\stackrel{(a)}{=} H(X_1, Y_1) + H(X_2, Y_2) - H(U) + H(X_3, Y_3 | X_3 + Y_3), \end{aligned}$$

where (a) follows by (12).

On the other hand, we have

$$\begin{aligned} &H(X_1, Y_1, X_2, Y_2, X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1, X_3, Y_3 | X_3 + Y_3) \\ &\leq H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1 | X_3 + Y_3) + H(X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1, X_3 + Y_3) - H(X_3 + Y_3) \\ &\quad + H(X_3, Y_3 | X_3 + Y_3) \\ &= H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1) - H(X_3 + Y_3) \\ &\quad + H(X_3, Y_3 | X_3 + Y_3). \end{aligned}$$

The last equality is a consequence of the observation that  $(X_1, Y_2, X_2 - Y_3, X_3 - Y_1)$  implies  $(X_1, Y_2, X_2 + Y_1 - (X_3 + Y_3))$ . However as  $X_1 + Y_2 = X_2 + Y_1$  by assumption, we observe that  $H(X_3 + Y_3 | X_1, Y_2, X_2 - Y_3, X_3 - Y_1) = 0$  and thus justifying the equality.

By combining these two decompositions, we obtain

$$\begin{aligned} & H(X_1, Y_1) + H(X_2, Y_2) + H(X_3 + Y_3) \\ & \leq H(U) + H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1). \end{aligned}$$

□

**Remark 8.** *The arguments here are motivated by similar arguments in the sunset literature [14], as well as in Tao's work on a similar inequality in [2].*

**Corollary 3.** *In addition to the assumptions on  $X_1, Y_1, X_2, Y_2, X_3, Y_3$  imposed in Proposition 1, let us assume that  $X_1$  is independent of  $Y_1$  and  $X_2$  independent of  $Y_2$ . Then we have*

$$H(X_2) + H(Y_1) + H(X_3 + Y_3) \leq H(X_1 - Y_1) + H(X_3 - Y_1) + H(X_2 - Y_3).$$

*Proof.* The proof is immediate from Proposition 1 along with the observation that the assumptions imply  $H(X_1, Y_1) = H(X_1) + H(Y_1)$ ,  $H(X_2, Y_2) = H(X_2) + H(Y_2)$ , and using the sub-additivity of entropy applied to  $H(X_1, Y_2, X_2 - Y_3, X_3 - Y_1)$ . □

**Remark 9.** *Suppose  $X$  and  $Y$  are independent random variables having finite support on  $\mathbb{G}$ , and random variables  $X_3, Y_3$  also have finite support on  $\mathbb{G}$ , then observe that we can always construct a coupling  $(X_1, Y_1, X_2, Y_2, X_3, Y_3)$  satisfying the assumptions of Corollary 3, so that  $(X_1, Y_1)$  and  $(X_2, Y_2)$  are distributed as  $(X, Y)$ .*

**Corollary 4** (Generalized Ruzsa sum-difference inequality). *Let  $A, B, C, D$  be finite subsets of an abelian group  $(\mathbb{G}, +)$ . Then the following sunset inequality holds:*

$$|A||B||C + D| \leq |A - B||C - B||A - D|,$$

or equivalently

$$d_R(C, -D) \leq d_R(C, B) + d_R(B, A) + d_R(A, D).$$

*Proof.* Suppose  $X$  be a uniform distribution on  $A$  and  $Y$  be a uniform distribution on  $B$ . Further let  $X_3, Y_3$  be taking values on  $C, D$  (respectively) such that  $X_3 + Y_3$  is uniform on  $C + D$ . Let  $(X_1, Y_2, X_2, Y_2, X_3, Y_3)$  be the coupling according to Remark 9 and observe that Corollary 3 implies that

$$\begin{aligned} & \log |A| + \log |B| + \log |C + D| \\ & \leq H(U) + H(X_3 - Y_1) + H(X_2 - Y_3) \\ & \leq \log |A - B| + \log |C - B| + \log |A - D|. \end{aligned}$$

Here, the second inequality used that the entropy of a finite valued random variable is upper bounded by the logarithm of its support size. □

**Remark 10.** *Setting  $C = A$  and  $D = B$ , we can see that the above is a generalization of Theorem 6.*

**Corollary 5.** *For any distributions  $P_U, P_V, P_X, P_Y$  with finite support on a finitely generated torsion-free group  $(\mathbb{T}, +)$ , we have*

$$\begin{aligned} & H(X) + H(Y) + \max_{\Pi(U, V)} H(U + V) \leq \\ & \max_{\Pi(X, Y)} H(X - Y) + \max_{\Pi(X, U)} H(X - U) + \max_{\Pi(V, Y)} H(V - Y). \end{aligned}$$

*Proof.* From Corollary 3, for any finite  $A, B, C, D$  on a finitely generated torsion-free abelian group  $(\mathbb{T}, +)$ , we have

$$|A||B||C + D| \leq |A - B||A - D||C - B|.$$

By applying Theorem 1, we will obtain the desired inequality. □

**Remark 11.** *Setting  $U = Y$  and  $V = X$  from the above result. We will obtain an entropic analog of sum-difference inequality*

$$\begin{aligned} & d_{HR}(X, -Y) \leq 3d_{HR}(X, Y), \\ & \text{or equivalently } H(X) + H(Y) + \max_{\Pi(X, Y)} H(X + Y) \leq 3 \max_{\Pi(X, Y)} H(X - Y). \end{aligned}$$

**Remark 12.** *There seems to be no direct implication between these two statements:*



- Suppose  $X$  and  $Y$  are independent, we have  $H(X) + H(Y) + H(X + Y) \leq 3H(X - Y)$ . This was the previously considered analogous form of the sum-difference inequality (10), established in [2].
- For any  $P_X, P_Y$ , we have

$$H(X) + H(Y) + \max_{\Pi(X,Y)} H(X + Y) \leq 3 \max_{\Pi(X,Y)} H(X - Y).$$

This is the formally established equivalent form of the sum-difference inequality (10).

### III. ENTROPIC FORMULATION OF MAGNIFICATION RATIO

There are a large number on sumset inequalities that do not have entropic equivalences yet, such as Plünnecke–Ruzsa inequality (even though some entropic analogs have been established in [2], [4]). A combinatorial primitive that frequently occurs in the combinatorial proofs is the notion of a magnification ratio (see the lecture notes: [9]). In this section we establish an entropic characterization of the magnification ratio and in addition to this result being potentially useful in deriving new entropic equivalences (future research), it may also be of independent interest to the combinatorics community.

Let  $G \subseteq A \times B$  be a finite bipartite graph such that there are no isolated vertices in  $A$  or  $B$ . For every  $S \subseteq A$ , let  $\mathcal{N}(S) \subseteq B$  denote the set of neighbours of  $S$ .

**Definition 4.** The magnification ratio of  $G$  from  $A$  to  $B$  is defined as

$$\mu_{A \rightarrow B}(G) = \min_{S \subseteq A, S \neq \emptyset} \frac{|\mathcal{N}(S)|}{|S|}.$$

**Definition 5.** (Channel Consistent with a Bipartite Graph) Let  $\mathcal{W}$  be the set of all possible channels (or probability transition matrices) from  $A$  to  $B$ . Given a bipartite graph  $G \subseteq A \times B$ , we define

$$\mathcal{W}(G) := \{W \in \mathcal{W} : W(Y = b|X = a) = 0 \text{ if } (a, b) \notin G\},$$

to be the set of all channels consistent with the bipartite graph  $G$ . Note that  $\mathcal{W}(G)$  is a closed and compact set.

In the above, we think of  $X$  (taking values in  $A$ ) as the input and  $Y$  (taking values in  $B$ ) as the output of a channel  $W_{Y|X}$ . Given an input distribution  $P_X$ , we define

$$\lambda_{A \rightarrow B}(G; P_X) := \max_{W \in \mathcal{W}(G)} (H(Y) - H(X)).$$

Given a fixed  $P_X$ , it is rather immediate that  $H(Y)$  is concave in  $W_{Y|X}$ . Let  $W^*(G; P_X) \in \mathcal{W}(G)$  denote a corresponding optimizer, i.e.

$$W^*(G; P_X) := \arg \max_{W \in \mathcal{W}(G)} (H(Y) - H(X)).$$

In the event that the optimizer is a convex set, we just define it to be an arbitrary element of this set.

Finally, we define the quantity

$$\lambda_{A \rightarrow B}(G) := \min_{P_X} \lambda_{A \rightarrow B}(G; P_X) \tag{13}$$

$$= \min_{P_X} \max_{W \in \mathcal{W}(G)} (H(Y) - H(X)). \tag{14}$$

The main result of this section is the following result.

**Theorem 7** (Entropic characterization of magnification ratio).

$$\begin{aligned} \log \mu_{A \rightarrow B}(G) &= \lambda_{A \rightarrow B}(G), \text{ or equivalently,} \\ \log \mu_{A \rightarrow B}(G) &= \min_{P_X} \max_{W \in \mathcal{W}(G)} (H(Y) - H(X)). \end{aligned}$$

*Proof.* We first establish that  $\lambda_{A \rightarrow B}(G) \leq \log \mu_{A \rightarrow B}(G)$ . This direction is rather immediate. Let

$$A^* := \arg \min_{S \subseteq A, S \neq \emptyset} \frac{|\mathcal{N}(S)|}{|S|}.$$

So we have  $\mu_{A \rightarrow B}(G) = \frac{|\mathcal{N}(A^*)|}{|A^*|}$ . Let  $P_X$  be the uniform distribution on  $A^*$ . Then note that

$$\begin{aligned} \lambda_{A \rightarrow B}(G) &\leq \lambda_{A \rightarrow B}(G; P_X) \\ &= \max_{W \in \mathcal{W}(G)} (H(Y) - H(X)) \\ &= \max_{W \in \mathcal{W}(G)} (H(Y) - \log |A^*|) \\ &\leq \log |\mathcal{N}(A^*)| - \log |A^*| = \log \mu_{A \rightarrow B}(G). \end{aligned}$$

This completes this direction.

We next establish that  $\mu_{A \rightarrow B}(G) \leq \log \lambda_{A \rightarrow B}(G)$ . This direction is comparatively rather involved whose main ingredient is the following lemma:

**Lemma 5.** *There exists a  $P_X^*$ , an optimizer of the outer minimization problem in*

$$\min_{P_X} \max_{W \in \mathcal{W}(G)} (H(Y) - H(X)),$$

such that the inner optimizer  $W^*(G; P_X^*)$  induces a uniform output distribution on  $\mathcal{N}(S^*)$ .

Now, let  $S^*$  be the support of  $P_X^*$ . If so, one would have

$$\lambda_{A \rightarrow B}(G) = H(Y) - H(X) = \log |\mathcal{N}(S^*)| - H(X) \geq \log \frac{|\mathcal{N}(S^*)|}{|S^*|} \geq \min_{S \subseteq A, S \neq \emptyset} \frac{|\mathcal{N}(S)|}{|S|} = \mu_{A \rightarrow B}(G),$$

and the proof is complete.  $\square$

We will now develop some preliminaries needed to establish Lemma 5.

**Definition 6.** *Given an input distribution  $P_X$  and a bipartite graph  $G$ , we define an edge  $(a, b) \in G$  to be active under  $W^*(G; P_X)$  if  $W^*(b|a) > 0$ . Otherwise, it is said to be inactive.*

**Lemma 6.** *Let  $S$  be the support of  $P_X$ .*

- 1) *Any maximizer  $W^*(G; P_X)$  induces an output distribution,  $P_Y$ , such that the support of  $P_Y$  is  $\mathcal{N}(S)$ .*
- 2) *Let  $a_1 \in S$  and  $(a_1, b_1), (a_1, b_2)$  be edges in  $G$ .*
  - a) *If the edges  $(a_1, b_1)$  and  $(a_1, b_2)$  are active under  $W^*(G; P_X)$ , then  $P_Y(b_1) = P_Y(b_2)$ .*
  - b) *If  $(a_1, b_1)$  is active and  $(a_1, b_2)$  is inactive under  $W^*(G; P_X)$ , then  $P_Y(b_1) \geq P_Y(b_2)$ .*

*Proof.* The proof of part 1) proceeds by contradiction. Assume that there exists  $b_1 \in \mathcal{N}(S)$  such that  $P_Y(b_1) = 0$ . This implies that there exists  $a_1 \in S$ , such that  $(a_1, b_1) \in G$  and  $W_{Y|X}^*(b_1|a_1) = 0$  as  $P_Y(b_1) = 0$ . Further since  $P_X(a_1) > 0$ , there exists  $b_2 \in \mathcal{N}(S)$  with  $(a_1, b_2) \in G$  and  $W_{Y|X}^*(b_2|a_1) > 0$ . For  $\alpha \geq 0$  and sufficiently small, define  $W_\alpha$  as follows:

$$W_{Y|X, \alpha}(b|a) = \begin{cases} W_{Y|X}^*(b|a) + \alpha = \alpha, & (a, b) = (a_1, b_1) \\ W_{Y|X}^*(b|a) - \alpha, & (a, b) = (a_1, b_2) \\ W_{Y|X}^*(b|a), & \text{otherwise} \end{cases}$$

Define  $f(\alpha) := H(Y_\alpha) - H(X)$ , where  $P_{Y_\alpha}$  is the output distribution of  $P_X$  under  $W_\alpha$ . Note that

$$f'(\alpha) = P_X(a_1) \log \left( \frac{P_Y(b_2) - \alpha P_X(a_1)}{\alpha P_X(a_1)} \right).$$

By assumption,  $W_0 = W^*$  is a maximizer of  $f(\alpha)$ . However,  $f'(\alpha) \rightarrow +\infty$  as  $\alpha \rightarrow 0^+$ , yielding the requisite contradiction.

We now establish part 2). Note that  $H(Y)$  is concave in  $\mathcal{W}(G)$  and all constraints in  $\mathcal{W}(G)$  is linear under  $\mathcal{W}$ . Therefore, Karush–Kuhn–Tucker(KKT) conditions are the necessary and sufficient conditions for optimality for  $W_{Y|X}$ . We rewrite the optimization problem as follows,

$$\begin{aligned} & \min_{W \in \mathcal{W}(G)} (H(Y) - H(X)) \\ & \text{subject to } W(b|a) \geq 0, a \in S, (a, b) \in G \cdot \\ & \quad \sum_b W(b|a) = 1, a \in S \end{aligned}$$

Define the Lagrangian as follows,

$$\mathcal{L}(W) := H(Y) - \mu_{a,b} W(a|b) + \sum_a \lambda_a \left( \sum_b W(b|a) - 1 \right).$$

The KKT condition for optimality is  $W \in \mathcal{W}$  and for any  $a \in S$  and  $(a, b) \in G$ , we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W(b|a)} &= -P_X(a)(\log P_Y(b) + 1) - \mu_{a,b} + \lambda_a = 0, \\ \mu_{a,b} W(b|a) &= 0, \\ \mu_{a,b} &\geq 0. \end{aligned}$$

By solving the above conditions, we have

$$P_Y(b) = \exp \left( - \frac{(\tilde{\lambda}_a + \mu_{a,b})}{P_X(a)} \right),$$

where  $\tilde{\lambda}_a = P_X(a) - \lambda_a$ .

a) Suppose  $(a_1, b_1)$  and  $(a_1, b_2)$  are active. This implies that  $\mu_{a_1, b_1} = \mu_{a_1, b_2} = 0$ , and forces  $P_Y(b_1) = P_Y(b_2)$ .

b) Suppose  $(a_1, b_1)$  is active and  $(a_1, b_2)$  is inactive. We have  $\mu_{a_1, b_1} = 0$  and  $\mu_{a_1, b_2} \geq 0$ , this implies  $P_Y(b_1) \geq P_Y(b_2)$ .

This establishes part 2) of the lemma.  $\square$

Based on  $P_X$  (with support  $S$ ) and the properties of the maximizer  $W^*(G; P_X)$ , we induce equivalence relationships between elements in  $\mathcal{N}(S)$  and between elements in  $S$ . Let  $P_Y$  be the distribution on  $\mathcal{N}(S)$  induced by  $P_X$  and  $W^*(G; P_X)$ . For  $b_1, b_2 \in \mathcal{N}(S)$ , we say that  $b_1 \sim b_2$  if  $P_Y(b_1) = P_Y(b_2)$ . We use the above to induce an equivalence relationship on  $S$  as follows: For  $a_1, a_2 \in S$ , we say that  $a_1 \sim a_2$  if there exists  $b_1, b_2 \in \mathcal{N}(S)$  such that the edges  $(a_1, b_1)$  and  $(a_2, b_2)$  are active (see Definition 6) and  $b_1 \sim b_2$ .

**Remark 13.** *The main observation is that the active edges in  $W^*(G; P_X)$  partitions the graph into disconnected components and further there is a one-to-one correspondence between the equivalence classes in  $\mathcal{N}(S)$  and the equivalence classes in  $S$ . To see this: consider an equivalence class  $T \subset \mathcal{N}(S)$  and let  $\hat{S} = \{a \in S : (a, b) \text{ is active for some } b \in T\}$ . From Lemma 6, we see that all elements in  $\hat{S}$  are equivalent to each other and there is no active edge  $(a, b)$  where  $a \in \hat{S}$  and  $b \notin T$ . Further if  $a_1 \in S \setminus \hat{S}$ , then observe that  $a_1$  is not equivalent to any element in  $\hat{S}$ .*

Let  $T_1, \dots, T_k$  be the partition of  $\mathcal{N}(S)$  into equivalence classes and let  $S_1, \dots, S_k$  be the corresponding partition of  $S$  into equivalence classes. We can define a total order on the equivalence classes of  $\mathcal{N}(S)$  as follows: we say  $T_{i_1} \geq T_{i_2}$  if  $P_Y(b_{i_1}) \geq P_Y(b_{i_2})$ . This also induces a total order on the equivalence classes on  $S$ . Further, without loss of generality, let us assume that  $T_1, \dots, T_k$  (and correspondingly  $S_1, \dots, S_k$ ) be monotonically decreasing according to the order defined above.

#### A. Proof of Lemma 5

*Proof.* Let  $P_X^*$  be an optimizer of the outer minimization problem in (13) and let  $S^*$  be its support. Further, let  $S_1, \dots, S_k$  be the equivalence classes (that form a partition of  $S$ ) induced by  $W^*(G; P_X^*)$ . If  $k = 1$ , i.e. there is only one equivalence class, then Lemma 6 implies that  $P_X^*$  and  $W^*(G; P_X^*)$  induces a uniform output distribution on  $\mathcal{N}(S^*)$ . Therefore, our goal is to show the existence of an optimizer  $P_X^*$  that induces exactly one equivalence class.

Let  $S_1$  and  $S_2$  be the largest and second largest element under the total ordering mentioned previously. Let  $m_\ell = |S_\ell|$ ,  $n_\ell = |T_\ell|$ , and for  $1 \leq i \leq k$ , let  $s_{i,j}, 1 \leq j \leq m_i$  be an enumeration of the elements of  $S_i$  and  $t_{i,j}, 1 \leq j \leq n_i$  be an enumeration of the elements of  $T_i$ . Further let  $p_{i,j} = P_X^*(s_{i,j})$  and  $p_i = \sum_{j=1}^{m_i} p_{i,j}$ . Since the induced output probabilities on the elements of  $T_i$  is uniform (by the definition of equivalence class), observe that  $q_{i,j} := P_Y^*(t_{i,j}) = \frac{p_i}{n_i}$  for all  $1 \leq j \leq n_i$ .

By the grouping property of entropy, we have

$$\begin{aligned} H(X) &= H(p_{1,1}, \dots, p_{1,m_1}, p_{2,1}, \dots, p_{2,m_2}, p_{3,1}, \dots, p_{k,m_k}) \\ &= p_1 H\left(\frac{p_{1,1}}{p_1}, \dots, \frac{p_{1,m_1}}{p_1}\right) + p_2 H\left(\frac{p_{2,1}}{p_2}, \dots, \frac{p_{2,m_2}}{p_2}\right) \\ &\quad + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \\ &\quad + H(p_1 + p_2, p_{3,1}, \dots, p_{k,m_k}). \end{aligned}$$

Similarly,

$$\begin{aligned} H(Y) &= p_1 H\left(\frac{1}{n_1}, \dots, \frac{1}{n_1}\right) + p_2 H\left(\frac{1}{n_2}, \dots, \frac{1}{n_2}\right) \\ &\quad + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) \\ &\quad + H(p_1 + p_2, q_{3,1}, \dots, q_{k,n_k}). \end{aligned}$$

Define a parameterized family of input distributions  $\tilde{P}_{X(\alpha)}$  as follows:

$$\tilde{P}_{X(\alpha)}(s_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) p_{i,j}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) p_{i,j}, & i = 2 \\ p_{i,j}, & \text{otherwise.} \end{cases}$$

By Lemma 7 we know that for  $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ , where

$$\alpha_{\max} := \frac{p_1 n_2 - p_2 n_1}{n_1 + n_2} \geq 0 \geq n_2 \left(\frac{p_3}{n_3} - \frac{p_2}{n_2}\right) =: \alpha_{\min},$$

$W^*(G; P_X^*)$  remain the optimal channel. Observe that the induced output distribution is

$$\tilde{P}_{Y(\alpha)}(t_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) q_{i,j} = \frac{p_i}{n_i} - \frac{\alpha}{n_i}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) q_{i,j} = \frac{p_i}{n_i} + \frac{\alpha}{n_i}, & i = 2 \\ q_{i,j}, & \text{otherwise.} \end{cases}$$

This implies  $\lambda_{A \rightarrow B}(G; \tilde{P}_{X(\alpha)}) = H(\tilde{Y}(\alpha)) - H(\tilde{X}(\alpha))$ . Note that

$$\begin{aligned} \lambda_{A \rightarrow B}(G; \tilde{P}_{X(\alpha)}) &:= H(\tilde{Y}(\alpha)) - H(\tilde{X}(\alpha)) \\ &= (p_1 - \alpha) \left( H\left(\frac{1}{n_1}, \dots, \frac{1}{n_1}\right) - H\left(\frac{p_{1,1}}{p_1}, \dots, \frac{p_{1,m_1}}{p_1}\right) \right) \\ &\quad + (p_2 + \alpha) \left( H\left(\frac{1}{n_2}, \dots, \frac{1}{n_2}\right) - H\left(\frac{p_{2,1}}{p_2}, \dots, \frac{p_{2,m_2}}{p_2}\right) \right) \\ &\quad + H(p_1 + p_2, q_{3,1}, \dots, q_{k,n_k}) - H(p_1 + p_2, p_{3,1}, \dots, p_{k,m_k}) \\ &= (p_1 - \alpha) f_1 + (p_2 + \alpha) f_2 + H(p_1 + p_2, q_{3,1}, \dots, q_{k,n_k}) \\ &\quad - H(p_1 + p_2, p_{3,1}, \dots, p_{k,m_k}), \end{aligned}$$

where

$$\begin{aligned} f_1 &= H\left(\frac{1}{n_1}, \dots, \frac{1}{n_1}\right) - H\left(\frac{p_{1,1}}{p_1}, \dots, \frac{p_{1,m_1}}{p_1}\right), \\ f_2 &= H\left(\frac{1}{n_2}, \dots, \frac{1}{n_2}\right) - H\left(\frac{p_{2,1}}{p_2}, \dots, \frac{p_{2,m_2}}{p_2}\right). \end{aligned}$$

Thus,  $\lambda_{A \rightarrow B}(G; \tilde{P}_{X(\alpha)})$  is linear in  $\alpha$ . At  $\alpha = 0$ , note that  $\tilde{P}_{X(\alpha)} = P_X^*$ , and hence is a minimizer of  $\lambda_{A \rightarrow B}(G; \tilde{P}_{X(\alpha)})$ . Therefore, this necessitates that  $f_1 = f_2$ , and for  $\alpha \in [\alpha_{\min}, \alpha_{\max}]$  we have that  $\lambda_{A \rightarrow B}(G; \tilde{P}_{X(\alpha)})$  is a constant. Consequently, both  $\tilde{P}_{X(\alpha_{\min})}$  and  $\tilde{P}_{X(\alpha_{\max})}$  are minimizers of the outer minimization problem.

If we consider  $\tilde{P}_{X(\alpha_{\max})}$  observe that we have  $\tilde{P}_{Y(\alpha_{\max})}(t_{1,j}) = \tilde{P}_{Y(\alpha_{\max})}(t_{2,j})$ . Therefore  $t_{1,j} \sim t_{2,j}$  and this causes  $T_1$  and  $T_2$  to merge into a new equivalence class. Therefore, we have a minimizer of the outer minimization problem with  $k - 1$  equivalence classes. We can proceed by induction till we get a single equivalence class. Note that the output elements in an equivalent class has the same probability, and the support of the induced output distribution is the neighbourhood of the support of  $P_X^*$  (see Lemma 6). Therefore establishing that  $\tilde{P}_{X(\alpha_{\max})}$  induces a single equivalence class establishes Lemma 5.

Alternately, if we consider  $\tilde{P}_{X(\alpha_{\min})}$  observe that we have  $\tilde{P}_{Y(\alpha_{\min})}(t_{2,j}) = \tilde{P}_{Y(\alpha_{\min})}(t_{3,j})$ . Therefore  $t_{2,j} \sim t_{3,j}$  and this causes  $T_2$  and  $T_3$  to merge into a new equivalence class. Therefore, again we have a minimizer of the outer minimization problem with  $k - 1$  equivalence classes. Proceeding we can again reduce to a single equivalence class and hence establish Lemma 5.  $\square$

**Remark 14.** The argument above can be used to infer (with minimal modifications) that any minimizer  $P_X^*$  of the outer minimization problem must have  $f_i = f_j$ , where

$$\begin{aligned} f_i &= H\left(\frac{1}{n_i}, \dots, \frac{1}{n_i}\right) - H\left(\frac{p_{i,1}}{p_i}, \dots, \frac{p_{i,m_i}}{p_i}\right), \\ f_j &= H\left(\frac{1}{n_j}, \dots, \frac{1}{n_j}\right) - H\left(\frac{p_{j,1}}{p_j}, \dots, \frac{p_{j,m_j}}{p_j}\right). \end{aligned}$$

Further  $\lambda_{A \rightarrow B}(G; \tilde{P}_{X^*}) = \sum_{i=1}^k p_i f_i$ . Since all  $f_i$ 's are identical, we have  $\lambda_{A \rightarrow B}(G) = f_1$ . Therefore the restriction of  $\tilde{P}_{X^*}$  to the first equivalence class is also a minimizer of the outer minimization problem, and observe that the induced output is uniform in  $T_1$ .

**Lemma 7** (Reweight input equivalence class probabilities preserves the optimality of the channel). *Let the partition  $S_1 \geq S_2 \geq \dots \geq S_k$  (of  $S$ , the support of  $P_X$ ) be the monotonically decreasing order of equivalence classes induced by  $W^*(G; P_X)$ . Define a parameterized family of input distributions  $\tilde{P}_{X(\alpha)}$  as follows*

$$\tilde{P}_{X(\alpha)}(s_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) p_{i,j}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) p_{i,j}, & i = 2 \\ p_{i,j}, & \text{otherwise.} \end{cases}$$

Then  $W^*(G; P_X)$  continues to be an optimal channel under  $\tilde{P}_{X(\alpha)}$  for  $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ , where

$$\alpha_{\max} := \frac{p_1 n_2 - p_2 n_1}{n_1 + n_2} \geq 0 \geq n_2 \left( \frac{p_3}{n_3} - \frac{p_2}{n_2} \right) =: \alpha_{\min}.$$

*Proof.* We recall the KKT conditions (from the proof of Lemma 6), which are necessary and sufficient for the inner optimization problem, to verify the optimality of  $W^*(G, P_X)$ . The KKT condition for optimality is that for any  $a \in S$  and  $(a, b) \in G$ , we have

$$\begin{aligned} -P_X(a)(\log P_Y(b) + 1) - \mu_{a,b} + \lambda_a &= 0, \\ \mu_{a,b} W(b|a) &= 0, \\ \mu_{a,b} &\geq 0. \end{aligned}$$

For  $a \in S$  and  $(a, b) \in G$ , let  $\lambda_a, \mu_{a,b}$  denote the dual parameters that certify the optimality of  $W^*(G, P_X)$  for  $P_X^*$ . Now define

$$\lambda_a(\alpha) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) \left(\lambda_a + P_X^*(a) \log \left(1 - \frac{\alpha}{p_1}\right)\right) & a \in S_1 \\ \left(1 + \frac{\alpha}{p_2}\right) \left(\lambda_a + P_X^*(a) \log \left(1 + \frac{\alpha}{p_2}\right)\right) & a \in S_2 \\ \lambda_a, & \text{otherwise.} \end{cases}$$

Using the channel  $W^*(G; P_X)$ , the induced output distribution of  $\tilde{P}_{X(\alpha)}$ , is given by

$$\tilde{P}_{Y(\alpha)}(t_{i,j}) = \begin{cases} \left(1 - \frac{\alpha}{p_1}\right) q_{i,j} = \frac{p_i}{n_i} - \frac{\alpha}{n_i}, & i = 1 \\ \left(1 + \frac{\alpha}{p_2}\right) q_{i,j} = \frac{p_i}{n_i} + \frac{\alpha}{n_i}, & i = 2 \\ q_{i,j} = \frac{p_i}{n_i}, & \text{otherwise.} \end{cases}$$

Observe that if  $(a, b_a)$  is an active edge under  $W^*(G; P_X)$ , then note that  $P_{Y(\alpha)}(b_a)$  only depends on  $a$ , or rather only on the equivalence class that  $a$  (or equivalently  $b_a$ ) belongs to. Define

$$\mu_{a,b}(\alpha) = P_{X(\alpha)}(a)(\log P_{Y(\alpha)}(b_a) - \log P_{Y(\alpha)}(b)).$$

Note that  $\mu_{a,b}(\alpha) \geq 0$  as long as

$$1 \geq \frac{p_1}{n_1} - \frac{\alpha}{n_1} \geq \frac{p_2}{n_2} + \frac{\alpha}{n_2} \geq \frac{p_3}{n_3},$$

or the ordering of equivalence classes remains unchanged. (Note that: if  $k = 2$ , i.e. there are only two partitions, then we set  $p_3 = 0$ .) This is equivalent to  $\alpha \geq \max\{n_2 \left(\frac{p_3}{n_3} - \frac{p_2}{n_2}\right), p_1 - n_1\}$  and  $\alpha \leq \frac{p_1 n_2 - p_2 n_1}{n_1 + n_2}$ . Since  $n_1 \geq 1$ , and by our ordering of equivalence classes, we have  $\frac{p_1}{n_1} \geq \frac{p_2}{n_2} \geq \frac{p_3}{n_3}$ ; a moments reflection implies the following:

$$\frac{p_1 n_2 - p_2 n_1}{n_1 + n_2} \geq 0 \geq n_2 \left( \frac{p_3}{n_3} - \frac{p_2}{n_2} \right) \geq p_1 - n_1.$$

Therefore  $\alpha \in [\alpha_{\min}, \alpha_{\max}]$  preserves the ordering of equivalence classes. A simple substitution shows that the dual variables  $\lambda_a(\alpha)$  and  $\mu_{a,b}(\alpha)$  defined above serve as witnesses for the optimality of  $W^*(G; P_X)$  for  $P_{X(\alpha)}$ . This completes the proof of the lemma.  $\square$

**Remark 15.** *The idea of the above proof is the following. The reweighting of the input classes preserves the uniformity of the output probabilities within each equivalent class, as well as the ordering between the output probabilities between equivalent classes. This happens to be the KKT conditions for the maximality of the channel. The limits are achieved with the output probability in an equivalence class equals the value in its adjacent class. At this point, there are potentially multiple optimizers for the inner problem, and there could be a rearrangement of the active and inactive edges as you change  $\alpha$  further.*

#### IV. CONCLUSION

In this paper we develop some information theoretic tools for proving information inequalities by borrowing from similar combinatorial tools developed in additive combinatorics. In reverse, the tools can also be used to generalize some results (or ambient group structure) in additive combinatorics.

## REFERENCES

- [1] I. Z. Ruzsa, "Sumsets and entropy," *Random Structures & Algorithms*, vol. 34, no. 1, pp. 1–10, 2009.
- [2] T. Tao, "Sumset and inverse sumset theory for Shannon entropy," *Combinatorics, Probability and Computing*, vol. 19, no. 4, pp. 603–639, 2010.
- [3] M. Madiman, A. W. Marcus, and P. Tetali, "Entropy and set cardinality inequalities for partition-determined functions," *Random Structures & Algorithms*, vol. 40, no. 4, pp. 399–424, 2012.
- [4] I. Kontoyiannis and M. Madiman, "Sumset and inverse sumset inequalities for differential entropy and mutual information," *IEEE transactions on information theory*, vol. 60, no. 8, pp. 4503–4514, 2014.
- [5] M. Madiman, "On the entropy of sums," in *2008 IEEE Information Theory Workshop*. IEEE, 2008, pp. 303–307.
- [6] A. Lapidoth and G. Pete, "On the entropy of the sum and of the difference of independent random variables," in *2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2008, pp. 623–625.
- [7] M. Madiman and I. Kontoyiannis, "The entropies of the sum and the difference of two iid random variables are not too different," in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 1369–1372.
- [8] A. Espuny Díaz, "Entropy methods for sumset inequalities," Master's thesis, Universitat Politècnica de Catalunya, 2016.
- [9] I. Z. Ruzsa, "Sumsets and structure," *Combinatorial number theory and additive group theory*, pp. 87–210, 2009.
- [10] I. Csiszár and J. Körner, *Information theory: Coding theorems for discrete memoryless systems*. Cambridge University Press, 1 2011.
- [11] I. Z. Ruzsa, "Sums of finite sets, number theory (new york, 1991–1995), 281–293," 1996.
- [12] N. H. Katz and T. Tao, "Bounds on arithmetic projections, and applications to the Kakeya conjecture," *Mathematical Research Letters*, vol. 6, no. 6, pp. 625–630, 1999.
- [13] Z. Zhang and R. Yeung, "On characterization of entropy function via information inequalities," *IEEE Transactions on Information Theory*, vol. 44, no. 4, pp. 1440–1452, July 1998.
- [14] B. Green, "Additive combinatorics - chapter 2," 2009. [Online]. Available: <https://people.maths.ox.ac.uk/greenbj/notes.html>