Information Inequalities for Joint Distributions, With Interpretations and Applications

Mokshay Madiman, *Member, IEEE*, and Prasad Tetali

***Abstract*—Upper and lower bounds are obtained for the joint entropy of a collection of random variables in terms of an arbitrary collection of subset joint entropies. These inequalities generalize Shannon's chain rule for entropy as well as inequalities of Han, Fujishige, and Shearer. A duality between the upper and lower bounds for joint entropy is developed. All of these results are shown to be special cases of general, new results for submodular functions—thus, the inequalities presented constitute a richly structured class of Shannon-type inequalities. The new inequalities are applied to obtain new results in combinatorics, such as bounds on the number of independent sets in an arbitrary graph and the number of zero-error source-channel codes, as well as determinantal inequalities in matrix theory. A general inequality for relative entropies is also developed. Finally, revealing connections of the results to literature in economics, computer science, and physics are explored.**

***Index Terms*—Entropy-based counting, entropy inequality, inequality for minors, submodularity.**

I. Introduction

Let $X_1, X_2, \ldots, X_n$ be a collection of random variables. There are the familiar two canonical cases: (a) the random variables are real-valued and possess a probability density function, in which case h represents the differential entropy, or (b) they are discrete, in which case H represents the discrete entropy. More generally, if the joint distribution has a density f with respect to some reference product measure, the joint entropy may be defined by $-E[\log f(X_1, X_2, \ldots, X_n)]$; with this definition, H corresponds to counting measure and h to Lebesgue measure. The only assumption we will implicitly make throughout is that the joint entropy is finite, i.e., neither $-\infty$ nor $+\infty$.

We wish to discuss the relationship between the joint entropies of various subsets of the random variables $X_1, X_2, \ldots, X_n$. Thus we are motivated to consider an arbitrary collection $\mathcal{C}$ of subsets of $\{1, 2, \ldots, n\}$. The following conventions are useful:

- $[n]$ is the index set $\{1, 2, \ldots, n\}$. We equip this set with its natural (increasing) order, so that $1 < 2 < \cdots < n$. (Any other total order would do equally well, and indeed we use this flexibility later, but it is convenient to fix a default order).
- For any set $s \subset [n]$, X_s stands for the collection of random variables $(X_i : i \in s)$, with the indices taken in their increasing order.
- For any index i in $[n]$, define the *degree* of i in $\mathcal{C}$ as $r(i) = |\{t \in \mathcal{C} : i \in t\}|$. Let $r_-(s) = \min_{i \in s} r(i)$ denote the *minimal degree* in s, and $r_+(s) = \max_{i \in s} r(i)$ denote the *maximal degree* in s.

First, we present a weak form of our main inequality.

*Proposition I: (*Weak degree form*):* Let $X_1, \ldots, X_n$ be arbitrary random variables jointly distributed on some discrete sets. For any collection $\mathcal{C}$ such that each index i has nonzero degree

$$\sum_{s \in \mathcal{C}} \frac{H(X_s | X_{s^c})}{r_+(s)} \leq H\left(X_{[n]}\right) \leq \sum_{s \in \mathcal{C}} \frac{H(X_s)}{r_-(s)} \tag{1}$$

where $r_+(s)$ and $r_-(s)$ are the maximal and minimal degrees in s. If $\mathcal{C}$ satisfies $r_-(s) = r_+(s)$ for each s in $\mathcal{C}$, then (1) also holds for h in the setting of continuous random variables.

Proposition I unifies a large number of inequalities in the literature. Indeed,

1) Applying to the class $\mathcal{C}_1$ of singletons

$$\sum_{i=1}^{n} H\left(X_i | X_{[n]\setminus i}\right) \leq H\left(X_{[n]}\right) \leq \sum_{i=1}^{n} H(X_i). \tag{2}$$

The upper bound represents the subadditivity of entropy noticed by Shannon. The lower bound may be interpreted as the fact that the erasure entropy of a collection of random variables is not greater than their entropy; see Section VI for further comments.

2) Applying to the class $\mathcal{C}_{n-1}$ of all sets of $n-1$ elements

$$\frac{1}{n-1} \sum_{i=1}^{n} H(X_{[n]\setminus i} | X_i) \leq H\left(X_{[n]}\right) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_{[n]\setminus i}). \tag{3}$$

This is Han's inequality [23], [10], in its prototypical form.

Manuscript received May 22, 2007; revised August 14, 2009. Current version published May 19, 2010. The material in this paper was presented at the Information Theory and Applications Workshop, San Diego, CA, January 2007, and at the IEEE Symposium on Information Theory, Nice, France, June 2007. This work was supported in part by the NSF by Grants DMS-0401239 and DMS-0701043.

M. Madiman is with the Department of Statistics, Yale University, James Dwight Dana House, New Haven, CT 06511 USA (e-mail: mokshay.madiman@yale.edu).

P. Tetali is with the School of Mathematics and College of Computing, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: tetali@math.gatech.edu).

Communicated by U. Mitra, Associate Editor at Large.

Digital Object Identifier 10.1109/TIT.2010.2046253

3) Let $r_+ = \min_{i\in[n]} r(i)$ and $r_- = \max_{i\in[n]} r(i)$ be the minimal and maximal degrees with respect to $\mathcal{C}$. Using $r_- \le r_-(s)$ and $r_+ \le r_+(s)$, we have

$$\frac{1}{r_+}\sum_{s\in\mathcal{C}} H(X_s|X_{s^c}) \le H\left(X_{[n]}\right) \le \frac{1}{r_-}\sum_{s\in\mathcal{C}} H(X_s).$$

The upper bound is Shearer's lemma [9], known in the combinatorics literature [43]. The lower bound is new.

The paper is organized as follows. First, in Section II, the notions of fractional coverings and packings using hypergraphs, which provide a useful language for the information inequalities we present, are developed. In Section III, we present the main technical result of this paper, which is a new inequality for submodular (not necessarily monotone) set functions. Section IV presents the main entropy inequality of this paper, which strengthens Proposition I, and gives a very simple proof as a corollary of the general result for submodular functions. This entropy inequality is developed in two forms, which we call the strong fractional form and the strong degree form; Proposition I may then be thought of as the weak degree form. A different manifestation of the upper bound in this weak degree form of the inequality was recently proved (in a more involved manner) by Friedgut [15]; the relationship with his result is also further discussed in Section IV using the preliminary concepts developed in Section II.

While independent sets in graphs have always been of combinatorial and graph-theoretical interest, counting independent sets in bipartite graphs received renewed attention due to Kahn's entropy approach [26] to Dedekind's problem. Dedekind's problem involves counting the number of antichains in the Boolean lattice, or equivalently, counting the number of Boolean functions on n variables that can be constructed using only AND and OR (and no NOT) gates. To handle this problem by induction on the number of levels in the lattice, Kahn first derived a tight bound on the (logarithm of the) number of independent sets in a regular bipartite graph. In Section V, we build on Kahn's work to obtain a bound on number of independent sets in an arbitrary graph. We also generalize this to counting graph homomorphisms, with applications to graph coloring and zero-error source-channel codes.

In Section VI, we use our entropy inequalities for *continuous* random variables to prove a new family of determinantal inequalities that provide generalizations of the classical determinantal inequalities of Hadamard, Szasz, and Fischer.

Having presented two applications of our main inequalities, we move on to studying the structure of the inequalities more closely. In Section VII, we present a duality between our upper and lower bounds that generalizes a theorem of Fujishige [17]. In particular, we show that the collection of upper bounds on $H\left(X_{[n]}\right)$ for all collections $\mathcal{C}$ is equivalent to the collection of lower bounds. There we also discuss interpretations of the inequality relating to sensor networks and erasure entropy, and generalize the monotonicity property for special collections of subsets discovered by Han [23].

Section VIII presents some new entropy power inequalities for joint distributions, and points out an intriguing analogy between them and the recent subset sum entropy power inequalities of [33]. In Section IX, we prove inequalities for relative entropy between joint distributions. Interpretations of the relative entropy inequality through hypothesis testing and concentration of measure are also given there.

In Section X, we note that weaker versions of our main inequality for submodular functions follow from results developed in various communities (economics, computer science, physics); this history and the consequent connections do not seem to be well known or much tapped in information theory. Finally in Section XI, we conclude with some final remarks and brief discussion of other applications, including to multiuser information theory.

II. On Hypergraphs and Related Concepts

It is appropriate here to recall some terminology from discrete mathematics. A collection $\mathcal{C}$ of subsets of $[n]$ is called a *hypergraph*, and each set s in $\mathcal{C}$ is called a *hyperedge*. When each hyperedge has cardinality 2, then $\mathcal{C}$ can be thought of as the set of edges of an undirected graph on n labelled vertices. Thus all the statements made above can be translated into the language of hypergraphs. In the rest of this paper, we interchangeably use "hypergraph" and "collection" for $\mathcal{C}$, "hyperedge" and "set" for s in $\mathcal{C}$, and "vertex" and "index" for i in $[n]$.

We have the following standard definitions.

Definition I: The collection $\mathcal{C}$ is said to be r*-regular* if each index i in $[n]$ has the same degree r, i.e., if each vertex i appears in exactly r hyperedges of $\mathcal{C}$.

The following definitions extend the familiar notion of packings, coverings, and partitions of sets by allowing fractional counts. The history of these notions is unclear to us, but some references can be found in [44].

Definition II: Given a collection $\mathcal{C}$ of subsets of $[n]$, a function $\alpha : \mathcal{C} \to \mathbf{R}^+$, is called a *fractional covering*, if for each $i \in [n]$, we have $\sum_{s\in\mathcal{C}: i\in s} \alpha(s) \ge 1$.

Given $\mathcal{C}$, a function $\beta : \mathcal{C} \to \mathbf{R}^+$ is a *fractional packing*, if for each $i \in [n]$, we have $\sum_{s\in\mathcal{C}: i\in S} \beta(s) \le 1$.

If $\gamma : \mathcal{C} \to \mathbf{R}^+$ is both a fractional covering and a fractional packing, we call γ a *fractional partition*.

Note that the standard definition of a fractional packing of $[n]$ using $\mathcal{C}$ (as in [44]), would assign weights β_i to the elements, (rather than sets) $i \in [n]$, and require that, for each $s \in \mathcal{C}$, we have $\sum_{i\in s} \beta_i \le 1$. Our terminology can be justified, if one considers the "dual hypergraph," obtained by interchanging the role of elements and sets—consider the 0–1 incidence matrix (with rows indexed by the elements and columns by the sets) of the set system, and simply switch the roles of the elements and the sets.

The following simple lemmas are useful.

*Lemma I: (*Fractional Additivity*):* Let $\{a_i : i \in [n]\}$ be an arbitrary collection of real numbers. For any $s \subset [n]$, define $a_s = \sum_{j\in s} a_j$. For any fractional partition γ using any hypergraph $\mathcal{C}$, $a_{[n]} = \sum_{s\in\mathcal{C}} \gamma(s) a_s$. Furthermore, if each $a_i \ge 0$, then

$$\sum_{s\in\mathcal{C}} \beta(s) a_s \le a_{[n]} \le \sum_{s\in\mathcal{C}} \alpha(s) a_s \tag{4}$$

for any fractional packing β and any fractional covering α using $\mathcal{C}$.

Proof: Interchanging sums implies

$$\sum_{s\in\mathcal{C}} \alpha(s) \sum_{i\in s} a_i = \sum_{i\in[n]} a_i \sum_{s\in\mathcal{C}} \alpha(s)\mathbf{1}_{\{i\in s\}} \geq \sum_{i\in[n]} a_i$$

using the definition of a fractional covering. The other statements are similarly obvious. ∎

We introduce the notion of quasi-regular hypergraphs.

Definition III: The hypergraph $\mathcal{C}$ is *quasi-regular* if the degree function $r : [n] \to \mathbb{Z}_+$ defined by $r(i) = |\{s \in \mathcal{C} : s \ni i\}|$ is constant on s, for each s in $\mathcal{C}$.

Example: One can construct simple examples of quasi-regular hypergraphs using what are called biregular graphs in the graph theory literature. Consider a bipartite graph on vertex sets V_1 and V_2 (i.e., all edges go between V_1 and V_2), such that every vertex in V_1 has degree r_1 and every vertex in V_2 has degree r_2. Such a graph always exists if $|V_1|r_1 = |V_2|r_2$. Now consider the hypergraph on $V_1 \cup V_2$ with hyperedges being the neighborhoods of vertices in the bipartite graph. This hypergraph is quasi-regular (with degrees being r_1 and r_2), and it is not regular if r_1 is different from r_2.

There is a sense in which all quasi-regular hypergraphs are similar to the example above; specifically, any quasi-regular hypergraph has a canonical decomposition as a disjoint union of regular subhypergraphs.

Lemma II: Suppose the hypergraph $\mathcal{C}$ on the vertex set $[n]$ is quasi-regular. Then one can partition $[n]$ into disjoint subsets $\{V_m\}$, and $\mathcal{C}$ into disjoint subhypergraphs $\{\mathcal{C}_m\}$ such that each $\mathcal{C}_m$ is a regular hypergraph on vertex set V_m.

Proof: Consider the equivalence relation on $[n]$ induced by the degree, i.e., i and j are related if $r(i) = r(j)$. This relation decomposes $[n]$ into disjoint equivalence classes $\{V_m\}$. Since $\mathcal{C}$ is quasi-regular, all indices in s have the same degree for each set $s \in \mathcal{C}$, and hence each $s \in \mathcal{C}$ is a subset of exactly one equivalence class V_m. Q.E.D. ∎

The notion of quasi-regularity is related to a useful fractional covering/packing pair. As long as there is at least one set s in the hypergraph $\mathcal{C}$ that contains i, we have

$$\sum_{s\in\mathcal{C}, s\ni i} \frac{1}{r_-(s)} = \sum_{s\in\mathcal{C}} \frac{\mathbf{1}_{\{i\in s\}}}{r_-(s)} \geq \sum_{s\in\mathcal{C}} \frac{\mathbf{1}_{\{i\in s\}}}{r(i)} = 1$$

so that $\alpha(s) = \frac{1}{r_-(s)}$ provide a fractional covering. Similarly, the numbers $\beta(s) = \frac{1}{r_+(s)}$ provide a fractional packing.

Definition IV: Let $\mathcal{C}$ be any hypergraph on $[n]$ such that every index appears in at least one hyperedge. The fractional covering given by $\alpha(s) = \frac{1}{r_-(s)}$ is called the *degree covering*, and the fractional packing given by $\beta(s) = \frac{1}{r_+(s)}$ is called the *degree packing*.

The following lemma is a trivial consequence of the definitions.

Lemma III: If $\mathcal{C}$ is quasi-regular, the degree packing and degree covering coincide and provide a fractional partition of $[n]$ using $\mathcal{C}$. In particular, $a_{[n]} = \sum_{s\in\mathcal{C}} a_s / r_-(s)$.

One may define the weight of a fractional partition as follows.

Definition V: Let γ be a fractional partition (or a fractional covering or packing). Then the weight of γ is $w(\gamma) = \sum_{s\in\mathcal{C}} \gamma(s)$.

There are natural optimization problems associated with the weight function. The problem of minimizing the weight of α over all fractional coverings α is the called the optimal fractional covering problem, and that of maximizing the weight of β over all fractional packings β is the called the optimal fractional packing problem. These are linear programming relaxations of the integer programs associated with optimal covering and optimal packing, which are of course important in many applications. Much work has been done on these problems, including studies of the integrality gap (see, e.g., [44]).

One may also define a notion of duality for fractional partitions.

Definition VI: For any hypergraph $\mathcal{C}$, define the complimentary hypergraph as $\bar{\mathcal{C}} = \{s^c : s \in \mathcal{C}\}$. If α is a fractional covering (or packing) using $\mathcal{C}$, the dual fractional packing (respectively, covering) using $\bar{\mathcal{C}}$ is defined by

$$\bar{\alpha}(s^c) = \frac{\alpha(s)}{w(\alpha) - 1}.$$

To see that this definition makes sense (say for the case of a fractional covering α), note that for each $i \in [n]$

$$\begin{aligned}\sum_{s^c\in\bar{\mathcal{C}}, s^c\ni i} \bar{\alpha}(s^c) &= \sum_{s\in\mathcal{C}, i\notin s} \frac{\alpha(s)}{w(\alpha)-1} \\ &= \frac{\sum_{s\in\mathcal{C}} \alpha(s) - \sum_{s\in\mathcal{C}, i\in s} \alpha(s)}{w(\alpha)-1} \\ &\leq \frac{w(\alpha)-1}{w(\alpha)-1} = 1.\end{aligned}$$

III. New Inequality for Submodular Functions

The following definitions are necessary in order to state the main technical result of this paper.

Definition VII: The set function $f : 2^{[n]} \to \mathbb{R}$ is *submodular* if

$$f(s) + f(t) \geq f(s \cup t) + f(s \cap t)$$

for every $s, t \subset [n]$. If $-f$ is submodular, we say that f is *supermodular*.

Definition VIII: For any disjoint subsets s and t of $[n]$, define $f(s|t) = f(s \cup t) - f(t)$. For a fixed subset $t \subsetneq [n]$, the function $f_t : 2^{[n]\setminus t} \to \mathbb{R}$ defined by $f_t(s) = f(s|t)$ is called f *conditional on* t.

For any $s \subset [n]$, denote by $< s$ the set of indices less than every index in s. Similarly, $> s$ is the set of indices greater than every index in s. Also, the index i is identified with the set $\{i\}$;

thus, for instance, $< i$ is well-defined. We also write $[i : i+k]$ for $\{i, i+1, \ldots, i+k-1, i+k\}$. Note that $[n] = [1 : n]$.

Lemma IV: Let $f : 2^{[n]} \to \mathbb{R}$ be any submodular function with $f(\phi) = 0$.

1) If s, t, u are disjoint sets

$$f(s|t,u) \leq f(s|t). \tag{5}$$

2) The following "chain rule" expression holds for $f([n])$:

$$f([n]) = \sum_{i \in [n]} f(i| < i).$$

Proof: First note that if s, t, u are disjoint sets, then submodularity implies

$$f(s \cup t \cup u) + f(t) \leq f(s \cup t) + f(t \cup u)$$

which is equivalent to $f(s|t,u) \leq f(s|t)$.

The "chain rule" expression for $f([n])$ is obtained by induction. Note that $f([2]) = f(1) + f(2|1) = f(1|\phi) + f(2|1)$ since $f(\phi) = 0$. Now assume the chain rule holds for $[n]$, and observe that

$$f([n+1]) = f([n]) + f(n+1|[n]) = \sum_{i \in [n+1]} f(i| < i)$$

where we used the induction hypothesis for the second equality. ∎

Theorem I: Let $f : 2^{[n]} \to \mathbb{R}$ be any submodular function with $f(\phi) = 0$. Let γ be any fractional partition with respect to any collection $\mathcal{C}$ of subsets of $[n]$. Then

$$\sum_{s \in \mathcal{C}} \gamma(s) f(s|s^c \setminus > s) \leq f([n]) \leq \sum_{s \in \mathcal{C}} \gamma(s) f(s| < s).$$

Proof: The chain rule (actually a slightly extended version of it with additional conditioning in all terms that can be proved in exactly the same way) implies

$$f(s| < s) = \sum_{j \in s} f(j| < j \cap s, < s). \tag{6}$$

Thus

$$\begin{aligned}
\sum_{s \in \mathcal{C}} \gamma(s) f(s| < s) &\overset{(a)}{=} \sum_{s \in \mathcal{C}} \gamma(s) \sum_{j \in s} f(j| < j \cap s, < s) \\
&\overset{(b)}{\geq} \sum_{s \in \mathcal{C}} \gamma(s) \sum_{j \in s} f(j| < j) \\
&\overset{(c)}{=} \sum_{j \in [n]} f(j| < j) \sum_{s \in \mathcal{C}} \gamma(s) \mathbf{1}_{\{j \in s\}} \\
&\overset{(d)}{=} \sum_{j \in [n]} f(j| < j) \\
&\overset{(a)}{=} f\left(X_{[n]}\right)
\end{aligned}$$

where (a) follows by the chain rule (6), (b) follows from(5), (c) follows by interchanging sums, and (d) follows by the definition of a fractional partition.

The lower bound may be proved in a similar fashion by a chain of inequalities. Indeed

$$\begin{aligned}
&\sum_{s \in \mathcal{C}} \gamma(s) f(s|s^c \setminus > s) \\
&\quad \overset{(a)}{=} \sum_{s \in \mathcal{C}} \gamma(s) \sum_{j \in s} f(j| < j \cap s, s^c \setminus > s) \\
&\quad \overset{(b)}{\leq} \sum_{s \in \mathcal{C}} \gamma(s) \sum_{j \in s} f(j| < j) \\
&\quad \overset{(c)}{=} \sum_{j \in [n]} f(j| < j) \sum_{s \in \mathcal{C}} \mathbf{1}_{\{j \in s\}} \gamma(s) \\
&\quad \overset{(d)}{=} \sum_{j \in [n]} f(j| < j) \\
&\quad \overset{(a)}{=} f([n])
\end{aligned}$$

where (a), (b), (c), and (d) follow as above. ∎

Remark 1: The key new element in this result is the fact that one can use, for any ordering on the ground set $[n]$, the conditional values of f that appear in the upper and lower bounds for $f([n])$. Because of (5), this is an improvement over simply using "unconditional" values of f. The latter weaker inequality has been implicit in the cooperative game theory literature; various historical remarks explicating these connections are given in Section X.

The fact that Theorem I does not require f to be monotone is important for many applications (including for differential entropy). However, if indeed the set function of interest is monotone, then more can be said—specifically, Theorem I can be extended to allow fractional coverings and packings as coefficients in the bounds.

Corollary I: Let $f : 2^{[n]} \to \mathbb{R}$ be any submodular function with $f(\phi) = 0$, such that $f([j])$ is nondecreasing in j for $j \in [n]$. Then, for any collection $\mathcal{C}$ of subsets of $[n]$

$$\sum_{s \in \mathcal{C}} \beta(s) f(s|s^c \setminus > s) \leq f([n]) \leq \sum_{s \in \mathcal{C}} \alpha(s) f(s| < s)$$

where β is any fractional packing and α is any fractional covering of $\mathcal{C}$.

Proof: The proof is almost exactly the same as that of Theorem I. The only difference is that the equalities (d) there in the upper and lower bounds are replaced by appropriate inequalities, using Definition II and the nonnegativity of $f(j| < j)$. ∎

Observe that if f defines a polymatroid (i.e., f is not only submodular but also nondecreasing in the sense that $f(s) \leq f(t)$ if $s \subset t$), then the condition of Corollary I is automatically satisfied.

IV. Entropy Inequalities

A. Strong Fractional Form

The main entropy inequality introduced in this work is the following generalization of Shannon's chain rule.

*Theorem I': (*STRONG FRACTIONAL FORM*):*
For any collection $\mathcal{C}$ of subsets of $[n]$

$$\sum_{s\in\mathcal{C}}\beta(s)H(X_s|X_{s^c\setminus>s}) \le H\left(X_{[n]}\right) \le \sum_{s\in\mathcal{C}}\alpha(s)H(X_s|X_{<s})$$

and

$$\sum_{s\in\mathcal{C}}\gamma(s)h(X_s|X_{s^c\setminus>s}) \le h\left(X_{[n]}\right) \le \sum_{s\in\mathcal{C}}\gamma(s)h(X_s|X_{<s})$$

where β is any fractional packing, α is any fractional covering, and γ is any fractional partition of $\mathcal{C}$.

To see that Theorem I' follows from Theorem I, we need to check that the joint entropy set function $f(s) = H(X_s)$ is a submodular function with $f(\phi) = 0$. The submodularity of f is a well known result that to our knowledge was first explicitly mentioned by Fujishige [17], although he appears to partially attribute the result to a 1960 paper of Watanabe that we have been unable to find. It follows from the fact that $H(X_s) + H(X_t) - H(X_{s\cup t}) - H(X_{s\cap t}) = I(X_{s\setminus t}; X_{t\setminus s}|X_{s\cap t})$ is a conditional mutual information (see, e.g., [10]), which is guaranteed to be nonnegative by Jensen's inequality. To see that the "correct" definition of $f(\phi) = 0$, note that the "unconditional" entropy $H(X_s)$ should be equal to $H(X_s|X_\phi)$, but the latter is $H(X_s) - H(X_\phi)$ by definition, which suggests that $H(X_\phi) = 0$.

There are two points deserving of emphasis here. First, although the statements of Theorem I' (or of Theorem I) above appear to be new, we do not claim that the proof of it is particularly novel; indeed, it is a careful refinement of the proof given by Llewellyn and Radhakrishnan for Shearer's lemma in the discrete setting (see [43]). Also, fractional partitions have been well known in combinatorics, and an anonymous referee mentioned that he or she had known of (but not published) the fractional version of Shearer's lemma for polymatroids. In this sense, our main technical contribution is to isolate the elements actually needed in the proof, observing for instance that the result is not particular to entropy and furthermore, even the full strength of the polymatroid structure is not needed. Second, we would again like to stress the freedom given by Theorem I' in terms of choice of ordering, as this can be useful in applications. For convenience of notation, we simply labelled the indices using the natural numbers in Theorem I' and used the ordering $1 < 2 < \cdots < n$, but one may equally well use another labeling or ordering.

Remark 2: It is natural to ask what choices of fractional packing and covering optimize the lower and upper bounds respectively. For a given collection of subset entropies, the optimal choices are clearly the solution of a linear program. Indeed, the best upper bound is obtained, for $w_s = H(X_s|X_{<s})$, by solving

$$\begin{aligned} &\text{Minimize} \quad \textstyle\sum_{s\in\mathcal{C}}\alpha(s)w_s \\ &\text{subject to} \quad \alpha(s)\ge 0 \text{ and } \textstyle\sum_{s\in\mathcal{C},s\ni i}\alpha(s)\ge 1. \end{aligned}$$

When the subset entropies are all equal, this is just the problem of optimal fractional covering discussed in Section II.

B. Strong Degree Form

The choice of α as the degree covering and β as the degree packing in Theorem I' gives the strong degree form of the inequality.

*Theorem II: (*STRONG DEGREE FORM*):* Let $\mathcal{C}$ be any collection of subsets of $[n]$, such that every index i appears in at least one element of $\mathcal{C}$. Then

$$\sum_{s\in\mathcal{C}}\frac{H\left(X_s|X_{s^c\setminus>s}\right)}{r_+(s)} \le H\left(X_{[n]}\right) \le \sum_{s\in\mathcal{C}}\frac{H(X_s|X_{<s})}{r_-(s)}.$$

If $\mathcal{C}$ is quasi-regular, then the above inequality also holds for h in place of H.

Remark 3: This also proves Proposition I. Indeed, since conditioning reduces entropy, Proposition I is just the loose form of Theorem II obtained by dropping the conditioning on $<s$ in the upper bound, and including conditioning on $>s$ in the lower bound.

Remark 4: The collections $\mathcal{C}$ for which the results in this paper hold need not consist of distinct sets. That is, one may have multiple copies of a particular $s \subset [n]$ contained in $\mathcal{C}$, and as long as this is taken into account in counting the degrees of the indices (or checking that a set of coefficients forms a fractional packing or covering), the statements extend. We will make use of this feature when developing applications to combinatorics in Section V.

Remark 5: Using the previous remark, one may write down Theorem II with arbitrary numbers of repetitions of each set in $\mathcal{C}$. This gives a version of Theorem I' with rational coefficients, following which an approximation argument can be used to obtain Theorem I'. This proof is similar to the one alluded to by Friedgut [15] for the version without ordering. Thus Theorem II is actually equivalent to Theorem I'.

The strong degree form of the inequality generalizes Shannon's chain rule. In order to see this, simply choose the collection $\mathcal{C}$ to be $\mathcal{C}_1$, the collection of all singletons. For this collection, Theorem II says

$$\sum_{i=1}^{n} H(X_i|X_{[n]\setminus\ge i}) \le H\left(X_{[n]}\right) \le \sum_{i=1}^{n} H(X_i|X_{<i})$$

which is precisely Shannon's chain rule (see, e.g., Shannon [45] and Cover and Thomas [10]), since the upper and lower bounds are identical. Note in contrast the looseness of the upper and lower bounds in (2), which are tight if and only if the random variables X_i are independent.

Application of Theorem II to nonsymmetric collections is also of interest. For instance, choosing $\mathcal{C}$ to be the class of all sets of k consecutive integers yields $r_- = 1$ and $r_+ = k$. Thus

$$\frac{H\left(X_{[n]}\right)}{\sum_{j\in[n]} H(X_{[j:l(j)]|X_{<j}})} \in \left[\frac{1}{k}, 1\right] \tag{7}$$

where $l(j) = \min\{j + k - 1, n\}$.

C. Weak Fractional Form

Theorems I' and II can be weakened by removing the conditioning in the upper bound, and adding conditioning in the

lower bound; from the latter, one obtains the weak degree form of Proposition I, and from the former, one obtains the weak fractional form of our main inequality.

*Proposition II: (*WEAK FRACTIONAL FORM*):* For any hypergraph $\mathcal{C}$ on $[n]$

$$\sum_{s\in\mathcal{C}}\beta(s)H\left(X_s|X_{s^c}\right)\leq H\left(X_{[n]}\right)\leq\sum_{s\in\mathcal{C}}\alpha(s)H(X_s) \quad (8)$$

and

$$\sum_{s\in\mathcal{C}}\gamma(s)h\left(X_s|X_{s^c}\right)\leq h\left(X_{[n]}\right)\leq\sum_{s\in\mathcal{C}}\gamma(s)h(X_s) \quad (9)$$

where β is any fractional packing, α is any fractional covering, and γ is any fractional partition of $\mathcal{C}$.

Remark 6: While the main inequality as stated in both its degree form (Theorem II) and its fractional form (Theorem I') seems novel, the bounds have been known to various levels of generality, as pointed out in the Introduction. In the discrete mathematics community, particular forms of the upper bound have been well known ever since the introduction of Shearer's lemma by Chung, Graham, Frankl, and Shearer [9] (see also [43] and [25]). In the level of generality of Proposition II, the fractional form was demonstrated by Friedgut [15] in terms of hypergraph projections. Friedgut's proof of the upper bound is perhaps not as transparent as the one we give. In the information theory community, both the upper and lower bounds of Proposition II have been known for the special case of the hypergraphs $\mathcal{C}_k$ (consisting of all sets of k elements out of n), since the work of Han [23] and Fujishige [17].

Remark 7: In the case of independent random variables, the joint entropy $H(X_s)=H\left(X_s|X_{s^c}\right)=\sum_{i\in s}H(X_i)$ is additive. Thus in that case, for any quasi-regular hypergraph $\mathcal{C}$, Proposition I holds with equality, and this is just Lemma III with $a_i=H(X_i)$. Similarly, thanks to Lemma I, Proposition II holds with equality for independent random variables when $\alpha=\beta$ is a fractional partition.

We believe that both the degree formulations of Proposition I and Theorem II, and the fractional formulations of Theorem I' and Proposition II are useful ways to think about these inequalities, and that they pave the way to the discovery of new applications. We illustrate this by using the degree formulation to count independent sets in graphs in Section V, and by using the fractional formulation to obtain new determinantal inequalities in Section VI.

V. AN APPLICATION TO COUNTING

A. Entropy and Counting

It is necessary to recall some terminology from graph theory. For our purposes, a graph $G=(V,E)$ consists of a finite vertex set V and a collection E of two-element subsets of V called edges (allowing repetition, i.e., self-loops). Thus G is a special case of a hypergraph, each hyperedge having cardinality 2. Two vertices are said to be adjacent, if there is an edge containing both of them. An independent set of G is a subset V_I of V such that no two vertices in V_I are adjacent.

Given a graph $F=(V(F),E(F))$, the set $\mathrm{Hom}(G,F)$ of homomorphisms from G to F is defined as

$$\mathrm{Hom}(G,F)=\{x:V\to V(F)\,\text{s.t.}\,uv\in E\Rightarrow x(u)x(v)\in E(F)\}.$$

Let $K_{a,b}$ denote the complete bipartite graph between parts of sizes a and b, respectively.

Shearer's lemma, and more generally, entropy-based arguments, have proved very useful in combinatorics. Shearer's lemma was (implicitly) introduced by Chung, Graham, Frankl and Shearer [9], and Kahn [25] stated an extension using the more familiar entropy notation. Recent applications of Shearer's lemma to difficult problems (where counting bounds are a key step in obtaining the results) include [19], [16], [26], [25], [6], and [21]. Radhakrishnan [43] provides a nice survey of entropy ideas used for counting and various applications; see also [1].

The general strategy of entropy-based proofs in counting is as follows:

- To count the number of objects in a certain class $\mathcal{C}$ of objects, consider a randomly drawn object X from the class and note that its entropy is $H(X)=\log|\mathcal{C}|$.
- Represent X using a collection of discrete random variables, and apply a Shearer-type lemma to bound $H(X)$ using certain subset entropies for a clever choice of hypergraph dictated by the problem.
- Perform an estimation of the resulting bound, using Jensen's inequality if necessary.

Below, we follow this direction of work and demonstrate a counting application of the new inequality. In particular, we use Theorem I' to bound the number of independent sets of an arbitrary graph, the number of proper graph colorings with a fixed number of colors, and more generally the number of graph homomorphisms.

B. Counting Graph Homomorphisms

Using Shearer's entropy inequality as a key ingredient, Kahn [27] recently showed a bound on the number of independent sets of a regular graph, building on his earlier result [25] for bipartite, regular graphs. Kahn's proof extends in a straightforward way, as observed by Galvin [20], to also bound from above the number of homomorphisms from a d-regular graph to an arbitrary graph F. Theorem IV below extends the observations of Kahn and Galvin to bound the number of graph homomorphisms from an arbitrary graph G to an arbitrary graph F.

*Theorem III: (*GRAPH HOMOMORPHISMS*):* For any N-vertex graph G and any graph F

$$|\mathrm{Hom}(G,F)|\leq\prod_{v\in V}|\mathrm{Hom}(K_{p(v),p(v)},F)|^{\frac{1}{d(v)}} \quad (10)$$

where $p(v)$ denotes the number of vertices preceding v in any ordering induced by decreasing degrees.

Proof: Let X be chosen uniformly at random from $\mathrm{Hom}(G,F)$. The random homomorphism X can be represented by the values it assigns to each $i\in V$, i.e., $X=(X(1),X(2),\ldots,X(n))=(X_1,X_2,\ldots,X_n)$, where

$X_i \in V_F$. By definition, X_i and X_j are connected in F if i and j are connected in G. We aim to bound $H(X)$ from above.

Let $\prec$ denote an ordering on vertices according to the decreasing order of their degrees (ties may be broken, for instance, by using an underlying lexicographic ordering of V). For each $i \in V$, let

$$P(i) = \{j \in V : \{i,j\} \in E \text{ and } j \prec i\}$$

and define $p(i) = |P(i)|$. Consider the collection $\mathcal{C}$ to be the collection of $P(i)$, and in addition, $p(i)$ copies of singleton sets $\{i\}$, for each i. Then observe that each i is covered by $d(i)$ sets in $\mathcal{C}$, i.e., that the degree of i in the collection $\mathcal{C}$ is $r(i) = d(i)$. Indeed, each i appears in $d(i) - p(i)$ sets of the form, $P(j)$, corresponding to each j such that $i \prec j$ and $\{i,j\} \in E$, and once in each of the $p(i)$ singleton sets $\{i\}$.

By the upper bound in Theorem II applied to this collection $\mathcal{C}$, we have

$$\begin{aligned} H(X) &\le \sum_{i\in V} \frac{1}{\min_{j\in P(i)} d(j)} H\Big(X_{P(i)} | X_{\prec P(i)}\Big) \\ &\quad + \sum_{i\in V} \frac{p(i)}{d(i)} H(X_i | X_{\prec i}) \\ &\le \sum_{i\in V} \Big(\frac{1}{d(i)} H(X_{P(i)}) + \frac{p(i)}{d(i)} H(X_i | X_{P(i)})\Big) \end{aligned}$$

by relaxing the conditioning and by the fact that the chosen ordering makes $j \in P(i)$ imply $d(j) \ge d(i)$.

Let q_i denote the probability mass function of $X_{P(i)}$, which takes its values in $\mathcal{X}_i = \{x_{P(i)} : x \in \mathrm{Hom}(G,F)\}$. In other words, $q(x_{P(i)})$ is the probability that $X_{P(i)} = x_{P(i)}$, under the uniform distribution on X. Finally, let $R(x_{P(i)})$ be the number of values that X_i can take given that $X_{P(i)} = x_{P(i)}$, i.e., the support size of the conditional distribution of X_i given $X_{P(i)} = x_{P(i)}$. Note that this is also the number of possible extensions of the partial homomorphism on $P(i)$ to a partial homomorphism on $P(i) \cup \{i\}$.

Then

$$\begin{aligned} &H(X_{P(i)}) + p(i) H(X_i | X_{P(i)}) \\ &\quad \le \sum_{x_{P(i)} \in \mathcal{X}_i} \Big(q(x_{P(i)}) \log \frac{1}{q(x_{P(i)})} \\ &\qquad\qquad + p(i) q(x_{P(i)}) H(X_i | X_{P(i)} = x_{P(i)})\Big) \\ &\quad \le \sum_{x_{P(i)} \in \mathcal{X}_i} q(x_{P(i)}) \log \frac{R(x_{P(i)}))^{p(i)}}{q(x_{P(i)})} \\ &\quad \le \log \sum_{x_{P(i)} \in \mathcal{X}_i} R(x_{P(i)}))^{p(i)} \end{aligned}$$

where $R(x_{P(i)})$ is the cardinality of the range of X_i given that $X_{P(i)} = x_{P(i)}$, and we have bounded $H(X_i | X_{P(i)} = x_{P(i)})$ by $\log R(x_{P(i)})$, and the last inequality follows by Jensen's inequality. Thus

$$H(X) \le \sum_{i\in V} \frac{1}{d(i)} \log\Big(\sum_{x_{P(i)} \in \mathcal{X}_i} R_i(x_{P(i)})^{p(i)}\Big).$$

The proof is completed by observing that, for any $i \in V$

$$\sum_{x_{P(i)} \in \mathcal{X}_i} R_i(x_{P(i)})^{p(i)} \le |\mathrm{Hom}(K_{p(i),p(i)}, F)|. \tag{11}$$

Indeed, first note that every (partial) homomorphism $x_{P(i)}$ of $P(i)$ for any graph G (regardless of the ordering $\prec$) is trivially a valid (partial) homomorphism of one side of $K_{p(i),p(i)}$, since each side of this bipartite graph has no edges and $|P(i)| = p(i)$. Furthermore, for a valid $x_{P(i)}$, the number of extensions $R_i(x_{P(i)})$ to i is the same whether the graph is G or $K_{p(i),p(i)}$, since it only depends on F. This proves (11). Note that the inequality(11) can be strict, since there can be partial homomorphisms of one side of $K_{p(i),p(i)}$ to a given F which are not necessarily valid while considering (partial) homomorphisms from G to F, since the induced graph on $P(i)$, for a given i, might have some edges. (This corrects the claim in [21] that(11) holds with equality.) ■

Nayak, Tuncel, and Rose [42] note that zero-error source-channel codes are precisely graph homomorphisms from a "source confusability graph" G_U to a "channel characteristic graph" G_X. Thus, Theorem IV may also be interpreted as giving a bound on the number of zero-error source channel codes that exist for a given source-channel pair.

C. Counting Independent Sets

By choosing appropriate graphs F, various corollaries can be obtained. In particular, it is well known that the problem of counting independent sets in a graph can be cast in the language of graph homomorphisms. Choose F to be the graph on two vertices joined by an edge, and with a self-loop on one of the vertices. Then, by considering the set of vertices of G that are mapped to the unlooped vertex in F, it is easy to see that each homomorphism from G to F corresponds to an independent set of G. This yields the following corollary.

*Corollary II:(*INDEPENDENT SETS*):* Let $G = (V,E)$ be an arbitrary graph on N vertices, and let $\mathcal{I}(G)$ denote the set of independent sets of G. Let $\prec$ denote an ordering on V according to decreasing order of degrees of the vertices, breaking ties arbitrarily. Let $p(v)$ denote the number of neighbors of v which precede v, under the $\prec$ ordering. Then

$$|\mathcal{I}(G)| \le \prod_{v\in V} 2^{(p(v)+1)\frac{1}{d(v)}}.$$

Specializing to the case of d-regular graphs $G = (V,E)$ on n vertices, it is clear that

$$|\mathcal{I}(G)| \le \prod_{v\in V} 2^{(p_a(v)+1)\frac{1}{d}} \le 2^{\frac{n}{2}+\frac{n}{d}}.$$

where $\prec_a$ is an arbitrary total order on V, and $p_a(v)$ is the number of vertices preceding v in this order, which are neighbors of v. This recovers Kahn's unpublished result [27] for d-regular graphs, which generalized his earlier result [25] for the d-regular, bipartite case. Note that we removed the assumption of regularity in Kahn's result by making a choice of ordering.

There is another way to view this result that is useful in computational geometry. Namely, if one considers a region (of, say, Euclidean space) and a finite family of subsets $\mathcal{F} = \{A_v : v \in V\}$ of this region, then one can define the intersection graph $G_{\mathcal{F}}$ of this family by connecting i and j in V if and only if $A_i \cap A_j \neq \phi$. Then the independent sets of $G_{\mathcal{F}}$ are in one-to-one correspondence with packings of the region using sets in the family $\mathcal{F}$. Thus Corollary II also gives a bound on the number of packings of a region using a given family of sets.

Another easy corollary of Theorem III is to graph colorings. Recall that a (proper) r-coloring of the vertices of G is a mapping $f : V \to [r]$ so that $u, v \in V$ and $uv \in E$ implies that $f(u) \neq f(v)$. Consider the constraint graph $F = K_r$ the complete graph on r vertices, for $r \geq 2$. Then $\mathrm{Hom}(G, K_r)$ corresponds to the number of (proper) r-colorings of the vertices of G. Thus the above theorem yields a corresponding upper bound on the number of r-colorings of a graph G, by replacing $\mathrm{Hom}(K_{p(v),p(v)}, F)$ in (10) with the number of r-colorings of the complete bipartite graph $K_{p(v),p(v)}$.

VI. Application to Determinantal Inequalities

The connection between determinants of positive definite matrices and multivariate normal distributions is classical. For example, Bellman's text [3] on matrix analysis makes extensive use of an "integral representation" of determinants in terms of an integrand of the form $e^{-<x,Ax>}$, which is essentially the Gaussian density. The classical determinantal inequalities of Hadamard and Fischer then follow from the subadditivity of entropy. This approach seems to have been first cast in probabilistic language by Dembo, Cover, and Thomas [11], who further showed that an inequality of Szasz can be derived (and generalized) using Han's inequality. Following this well-trodden path, Proposition II yields the following general determinantal inequality.

*Corollary III: (*Determinantal Inequalities*):* Let K be a positive definite $n \times n$ matrix and let $\mathcal{C}$ be a hypergraph on $[n]$. Let $K(s)$ denote the submatrix corresponding to the rows and columns indexed by elements of s. Then, using $|M|$ denote the determinant of M, we have for any fractional partition α^*

$$\prod_{s \in \mathcal{C}} \left(\frac{|K|}{|K(s^c)|} \right)^{\alpha^*(s)} \leq |K| \leq \prod_{s \in \mathcal{C}} |K(s)|^{\alpha^*(s)}.$$

The proof follows from Proposition II via the fact that any positive definite $n \times n$ matrix K can be realized as the covariance matrix of a multivariate normal distribution $N(0, K)$, whose entropy is

$$H\left(X_{[n]}\right) = \tfrac{1}{2} \log[(2\pi e)^n |K|]$$

and furthermore, that if $X_{[n]} \sim N(0, K)$, then $X_s \sim N(0, K(s))$. Note that an alternative approach to proving Corollary III would be to directly apply Theorem I to the known fact (called the Koteljanskii or sometimes the Hadamard-Fischer inequality) that the set function $f(s) = \log |K(s)|$ is submodular.

For an r-regular hypergraph $\mathcal{C}$, using the degree partition in Corollary III implies that

$$|K|^r \leq \prod_{s \in \mathcal{C}} |K(s)|.$$

Considering the hypergraphs $\mathcal{C}_1$ and $\mathcal{C}_{n-1}$ then yields the Hadamard and prototypical Szasz inequality, while the Fischer inequality follows by considering $\mathcal{C} = \{s, s^c\}$, for an arbitrary $s \subset [n]$.

We remark that one can interpret Corollary III using the all-minors matrix-tree theorem (see, e.g., Chaiken [7] or Lewin [32]). This is a generalization of the matrix tree theorem of Kirchhoff [29], which states that the determinant of any cofactor of the Laplacian matrix of a graph is the total number of distinct spanning trees in the graph, and interprets all minors of this matrix in terms of combinatorial properties of the graph.

VII. Duality and Monotonicity of Gaps

Consider the weak fractional form of Theorem I, namely

$$\sum_{s \in \mathcal{C}} \gamma(s) f(s|s^c) \leq f([n]) \leq \sum_{s \in \mathcal{C}} \gamma(s) f(s).$$

We observe that there is a duality between the upper and lower bounds, relating the gaps in this inequality.

*Theorem IV: (*Duality of Gaps*):* Let $f : 2^{[n]} \to \mathbb{R}$ be a submodular function with $f(\phi) = 0$. Let γ be an arbitrary fractional partition using some hypergraph $\mathcal{C}$ on $[n]$. Define the lower and upper gaps by

$$\mathrm{Gap}_L(f, \mathcal{C}, \gamma) = f([n]) - \sum_{s \in \mathcal{C}} \gamma(s) f(s|s^c)$$

and

$$\mathrm{Gap}_U(f, \mathcal{C}, \gamma) = \sum_{s \in \mathcal{C}} \gamma(s) f(s) - f([n]). \tag{12}$$

Then

$$\frac{\mathrm{Gap}_U(f, \mathcal{C}, \gamma)}{w(\gamma)} = \frac{\mathrm{Gap}_L(f, \bar{\mathcal{C}}, \bar{\gamma})}{w(\bar{\gamma})} \tag{13}$$

where w is the weight function and $\bar{\gamma}$ is the dual fractional partition defined in Section II.

Proof: This follows easily from the definitions. Indeed

$$\begin{aligned} f([n]) &- \sum_{s^c \in \bar{\mathcal{C}}} \bar{\gamma}(s^c) f(s^c|s) \\ &= f([n]) - \sum_{s \in \mathcal{C}} \frac{\gamma(s)}{w(\gamma) - 1} [f([n]) - f(s)] \\ &= \frac{\sum_{s \in \mathcal{C}} \gamma(s) f(s)}{w(\gamma) - 1} - \left[\frac{w(\gamma)}{w(\gamma) - 1} - 1 \right] f([n]) \\ &= \frac{1}{w(\gamma) - 1} \left[\sum_{s \in \mathcal{C}} \gamma(s) f(s) - f([n]) \right], \end{aligned}$$

and

$$w(\bar{\gamma}) = \sum_{s^c \in \bar{\mathcal{C}}} \bar{\gamma}(s^c) = \frac{\sum_{s \in \mathcal{C}} \gamma(s)}{w(\gamma) - 1} = \frac{w(\gamma)}{w(\gamma) - 1}.$$

Dividing the first expression by the second yields the result. ■

Note that the upper bound for $f([n])$ with respect to $(\mathcal{C},\gamma)$ is equivalent to the lower bound for $f([n])$ with respect to the dual $(\bar{\mathcal{C}},\bar{\gamma})$, implying that the *collection* of upper bounds for all hypergraphs and all fractional coverings is equivalent to the *collection* of lower bounds for all hypergraphs and all fractional packings. Also, it is clear that under the assumptions of Corollary I, one can state a duality result extending Theorem IV by replacing γ by any fractional covering α, and $\bar{\gamma}$ by the dual fractional packing $\bar{\alpha}$. From Theorem IV, it is clear by symmetry that also

$$\frac{\mathrm{Gap}_L(f,\mathcal{C},\gamma)}{w(\gamma)} = \frac{\mathrm{Gap}_U(f,\bar{\mathcal{C}},\bar{\gamma})}{w(\bar{\gamma})}. \tag{14}$$

However, (13) and(14) do not imply any relation between $\mathrm{Gap}_U(f,\mathcal{C},\gamma)$ and $\mathrm{Gap}_L(f,\mathcal{C},\gamma)$.

The gaps in the inequalities have especially nice structure when they are considered in the weak degree form, i.e., for the fractional partition using a r-regular hypergraph $\mathcal{C}$, all of whose coefficients are $1/r$. The associated gaps are

$$g_L(f,\mathcal{C}) = f([n]) - \frac{1}{r}\sum_{s\in\mathcal{C}} f(s|s^c)$$
$$\text{and } g_U(f,\mathcal{C}) = \frac{1}{r}\sum_{s\in\mathcal{C}} f(s) - f([n]). \tag{15}$$

*Corollary IV: (*DUALITY FOR REGULAR COLLECTIONS*):* Let $f : 2^{[n]} \to \mathbb{R}$ be a submodular function with $f(\phi) = 0$. For a r-regular collection $\mathcal{C}$

$$\frac{g_L(f,\bar{\mathcal{C}})}{g_U(f,\mathcal{C})} = \frac{r}{|\mathcal{C}| - r}.$$

Let us now specialize to the entropy set function $\mathsf{e}(s)$—we use this to mean either $H(X_s)$ (if the random variables X_i are discrete) or $h(X_s)$ (if the random variables X_i are continuous). The special hypergraphs $\mathcal{C}_k$, $k = 1,2,\ldots,n$, consisting of all k-sets or sets of size k, are of particular interest, and a lot is already known about the gaps for these collections. For instance, Han's inequality [23] already implies Proposition I for these hypergraphs, and Corollary IV applied to these hypergraphs implies that

$$\frac{g_L(\mathsf{e},\mathcal{C}_{n-k})}{g_U(\mathsf{e},\mathcal{C}_k)} = \frac{k}{n-k}$$

recovering an observation made by Fujishige [17]. Indeed, Theorem IV and Corollary IV generalize what [17] interpreted using the duality of polymatroids, since our assumptions are weaker and the assertions broader. Fujishige [17] considered these gaps important enough to merit a name: building on terminology of Han [23], he called the quantity $g_U(\mathsf{e},\mathcal{C}_k)$ a "total correlation," and $g_L(\mathsf{e},\mathcal{C}_k)$ a "dual total correlation."

In two particular cases, the gaps have simple expressions as relative entropies (see Section IX for definitions). First, note that the lower gap in Han's inequality (3) is related to the dependence measure that generalizes the mutual information

$$\begin{aligned}(n-1)g_L(\mathsf{e},\mathcal{C}_{n-1}) &= g_U(\mathsf{e},\mathcal{C}_1)\\ &= \sum_{i\in[n]} \mathsf{e}(\{i\}) - \mathsf{e}([n])\\ &= D(P_{X_{[n]}}\|P_{X_1}\times\cdots\times P_{X_n}).\end{aligned} \tag{16}$$

It is trivial to see that the gap is zero if and only if the random variables are independent.

Second, the lower gap in Proposition I with respect to the singleton class $\mathcal{C}_1$ is related to the upper gap in the prototypical form (3) of Han's inequality

$$\begin{aligned}g_L(\mathsf{e},\mathcal{C}_1) &= (n-1)g_U(\mathsf{e},\mathcal{C}_{n-1})\\ &= \sum_{i\in[n]} D(P_{X_i|X_{[n]\setminus i}}\|P_{X_i|X_{<i}}|P).\end{aligned} \tag{17}$$

(Here, the last equality comes from simple manipulation of the pointwise log likelihoods.) Note that for the gap to be zero, each of the relative entropies on the right must be zero. In particular, $D(P_{X_1|X_{[2:n]}}\|P_{X_1}) = 0$, which implies that X_1 is independent of the remaining random variables. By applying the same fact to the collection of random variables under different orderings, one sees that $X_{[n]}$ must be an independent collection of random variables.

The latter observation is relevant to the study of the *erasure entropy* of a collection of random variables, defined by Verdú and Weissman [50] to be

$$H^-\left(X_{[n]}\right) = \sum_{i=1}^{n} H(X_i|X_{[n]\setminus i}).$$

They give several motivations for defining these quantities; most significantly, the erasure entropy has an operational significance as the number of bits required to reconstruct a symbol erased by an erasure channel. Theorem 1 in [50] states that $H^-\left(X_{[n]}\right) \le H\left(X_{[n]}\right)$ with equality if and only if the X_i are independent. The inequality here is simply the lower bound of Proposition I applied to the singleton class $\mathcal{C}_1$, and is thus a special case of our results. The difference between the joint entropy of $X_{[n]}$ and its erasure entropy is just $g_L(\mathsf{e},\mathcal{C}_1)$, and the characterization of equality in terms of independence follows from the remarks above. It would be interesting to see if the more general bounds on joint entropy developed here can also be given an operational meaning using appropriate erasure-type channels.

Apart from the eponymous duality between the total and dual total correlations discussed above, these quantities also satisfy a monotonicity property, sometimes called Han's theorem (cf., [23]). Since this complements the duality result, we state it below in the more general submodular function setting.

*Corollary V: (*MONOTONICITY OF GAPS*):* Let $f : 2^{[n]} \to \mathbb{R}$ be a submodular function with $f(\phi) = 0$, and let $g_L(f,\mathcal{C}_k)$ and $g_U(f,\mathcal{C}_k)$ be defined by (15). Then both $g_L(f,\mathcal{C}_k)$ and $g_U(f,\mathcal{C}_k)$ are monotonically decreasing in k.

Proof: Proposition I, applied to the collection $\mathcal{C}_k$, immediately implies that $0 = g_U(f,\mathcal{C}_n) \le g_U(f,\mathcal{C}_k)$, for $k \in [n]$, on observing that $r_-(s) = r_+(s) = \binom{n-1}{k-1}$. To obtain the full chain of inequalities, first note that for any s in $\mathcal{C}_{k+1}$

$$f(s) \le \frac{1}{k}\sum_{i\in s} f(s\setminus i).$$

Thus

$$\begin{aligned}&g_U(f,\mathcal{C}_k) - g_U(f,\mathcal{C}_{k+1})\\ &\quad= \frac{1}{\binom{n-1}{k-1}}\sum_{s\in\mathcal{C}_k} f(s) - \frac{1}{\binom{n-1}{k}}\sum_{s\in\mathcal{C}_{k+1}} f(s)\\ &\quad\ge \frac{1}{\binom{n-1}{k-1}}\left[\sum_{s\in\mathcal{C}_k} f(s) - \frac{1}{n-k}\sum_{s\in\mathcal{C}_{k+1}}\sum_{i\in s} f(s\setminus i)\right].\end{aligned}$$

To complete the proof, note that

$$\begin{aligned}\sum_{s\in\mathcal{C}_{k+1}}\sum_{i\in s} f(s\setminus i) &= \sum_{i\in[n]}\sum_{s\in\mathcal{C}_{k+1},s\ni i} f(s\setminus i)\\ &= \sum_{i\in[n]}\sum_{s\in\mathcal{C}_k,i\notin s} f(s) = \sum_{s\in\mathcal{C}_k}\sum_{i\notin s} f(s)\\ &= (n-k)\sum_{s\in\mathcal{C}_k} f(s).\end{aligned}$$

■

Again specializing to the joint entropy function, let

$$\mathsf{e}_k^{(U)} = \frac{1}{\binom{n}{k}}\sum_{s:|s|=k}\frac{\mathsf{e}(s)}{k}$$

denote the joint entropy per element for subsets of size k averaged over all k-element subsets, and

$$\mathsf{e}_k^{(L)} = \frac{1}{\binom{n}{k}}\sum_{s:|s|=k}\frac{\mathsf{e}(s|s^c)}{k}$$

denote the corresponding average of conditional entropy per element. Since $g_U(\mathsf{e},\mathcal{C}_k) = n\mathsf{e}_k^{(U)} - \mathsf{e}([n])$ and $g_L(\mathsf{e},\mathcal{C}_k) = \mathsf{e}([n]) - n\mathsf{e}_k^{(L)}$, Corollary V asserts that $\mathsf{e}_k^{(U)}$ is decreasing in k, while $\mathsf{e}_k^{(L)}$ is increasing in k. Dembo, Cover and Thomas [11] give a nice interpretation of this fact, briefly outlined below.

Suppose we have n sensors collecting data relevant to the task at hand. For instance, the sensors might be measuring the temperature of the ocean at various points, or they might be evaluating the probability that a human face is in a collection of camera images taken along the boundary of a high-security site, or they might be taking measurements of neurons in a monkey's brain. Suppose due to experimental conditions, at any time, we only have access to a random subset of m sensor measurements out of n. Then Han's monotonicity theorem implies that, on average, we are getting more information as m increases, etc.

VIII. Entropy Power Inequalities

Theorem I' implies similar inequalities for entropy powers. Recall that the *entropy power* of the random vector X_s is

$$\mathcal{N}(X_s) = e^{\frac{2h(X_s)}{|s|}}.$$

This is sometimes standardized by a constant $(2\pi e)$, which is convenient in the continuous case as it allows for a comparison with a multivariate normal distribution. For discrete random variables, one can replace h by H in the above definition.

Corollary VI: Let γ be any fractional partition of $[n]$ using the hypergraph $\mathcal{C}$. Then

$$\mathcal{N}\left(X_{[n]}\right) \le \sum_{s\in\mathcal{C}} w_s\mathcal{N}(X_s)$$

where $w_s = \frac{\gamma(s)|s|}{n}$ are weights that sum to 1 over $s\in\mathcal{C}$.

Proof: First note that

$$\sum_{s\in\mathcal{C}} w_s = \sum_{s\in\mathcal{C}}\frac{\gamma(s)}{n}\sum_{i\in[n]}\mathbf{1}_{i\in s} = \sum_{i\in[n]}\frac{1}{n}\sum_{s\in\mathcal{C},s\ni i}\gamma(s) = 1$$

since γ is a fractional partition. Thus

$$\begin{aligned}\exp\left\{\frac{2h\left(X_{[n]}\right)}{n}\right\} &\le \exp\left\{\frac{2}{n}\sum_{s\in\mathcal{C}}\gamma(s)h(X_s)\right\}\\ &= \exp\left\{\sum_{s\in\mathcal{C}} w_s\frac{2h(X_s)}{|s|}\right\}\\ &\le \sum_{s\in\mathcal{C}} w_s\mathcal{N}(X_s)\end{aligned}$$

where the first inequality follows from Proposition II, and the last inequality follows by Jensen's inequality. ■

Remark 8: Corollary VI generalizes an implication of [10, Th. 16.5.2], which looks at the collections of k-sets. Note that, as in the special case covered in [10], Corollary VI continues to hold with the entropy power $\mathcal{N} = \mathcal{N}_2$ replaced throughout by any of the quantities $\mathcal{N}_c(X_s) = \exp\{ch(X_s)/|s|\}$ for any $c > 0$. As in the case of entropy, the bounds on the entropy powers associated with the hypergraphs $\mathcal{C}_m$ and the degree covering satisfy a monotonicity property. Indeed, by [10, Th. 16.5.2]

$$\frac{1}{\binom{n}{m}}\sum_{s\in\mathcal{C}_{n-m}}\mathcal{N}_c(X_s)$$

is a decreasing sequence in m. More interesting than entropy power inequalities for joint distributions, however, are entropy power inequalities for sums of independent random variables with densities. Introduced by [45] and [48] in seminal contributions, they have proved to be extremely useful and surprisingly deep—with connections to functional analysis, central limit theorems, and to the determination of capacity and rate regions for problems in information theory. Recently the first author showed (building on work in [2] and [33]) the following generalized entropy power inequality. For independent real-valued random variables X_i with densities and finite variances

$$\mathcal{N}\left(\sum_{i\in[n]} X_i\right) \ge \sum_{s\in\mathcal{C}}\gamma(s)\mathcal{N}\left(\sum_{i\in s} X_i\right) \tag{18}$$

for any fractional partition γ with respect to any hypergraph $\mathcal{C}$ on $[n]$. Inequality (18) shares an intriguing similarity of form to the inequalities of this paper, although it is much harder to prove.

The formal similarity between results for joint entropy and for entropy power of sums extends further. For instance, the fact that

$$\frac{1}{\binom{n}{m}}\sum_{s\in\mathcal{C}_{n-m}}\mathcal{N}\left(\sum_{i\in s} X_i\right)$$

is an increasing sequence in m, can be thought of as a formal dual of Han's theorem. It is an open question whether upper bounds for entropy power of sums can be obtained that are analogous to the lower bound in Theorem I'.

IX. Inequality for Relative Entropy, and Interpretations

Let A be either a countable set, or a Polish (i.e., complete separable metric) space equipped as usual with its Borel σ-algebra of measurable sets. Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on

the Polish product space A^n. For any nonempty subset s of $[n]$, write $\mathbb{P}_s$ for the marginal probability measure corresponding to the coordinates in s. Recall the definition of the relative entropy:

$$D(\mathbb{P}_s||\mathbb{Q}_s) = E_{\mathbb{P}}\left[\log \frac{d\mathbb{P}_s}{d\mathbb{Q}_s}\right] \in [0,\infty]$$

when $\mathbb{P}_s$ is absolutely continuous with respect to $\mathbb{Q}_s$, and $D(\mathbb{P}_s||\mathbb{Q}_s) = +\infty$ otherwise.

One may also define the conditional relative entropy by

$$D(\mathbb{P}_{s|t}||\mathbb{Q}_{s|t}|\mathbb{P}) = E_{\mathbb{P}_t} D(\mathbb{P}_{s|t}||\mathbb{Q}_{s|t}) \tag{19}$$

where $\mathbb{P}_{s|t}$ is understood to mean the conditional distribution (under $\mathbb{P}$) of the random variables corresponding to s given particular values of the random variables corresponding to t; then $E_{\mathbb{P}_t}$ denotes the averaging using $\mathbb{P}_t$ over the values that are conditioned on. We have freely used (regular) conditional distributions in these definitions; the existence of these is justified by the fact that we are working with Polish spaces.

For compactness of notation, let us set

$$d(s) = D(\mathbb{P}_s||\mathbb{Q}_s).$$

From the definitions above, it is easy to verify the chain rule

$$d(s\cup t) = D(\mathbb{P}_{s|t}||\mathbb{Q}_{s|t}|\mathbb{P}) + d(t)$$

for disjoint s and t, so that following the terminology developed in Section III, one should set

$$d(s|t) = D(\mathbb{P}_{s|t}||\mathbb{Q}_{s|t}|\mathbb{P}).$$

Theorem V: Let $\mathbb{Q}$ be a product probability measure on A^n, where A is a Polish space as above. Suppose $\mathbb{P}$ is a probability measure on A^n such that the set function $d : 2^{[n]} \to [0,\infty]$ given by

$$d(s) = D(\mathbb{P}_s||\mathbb{Q}_s)$$

does not take the value $+\infty$ for any $s \subset [n]$. Then $d(s)$ is supermodular.

Proof: For any nonempty $s, t \subset [n]$, we have

$$\begin{aligned} &d(s\cup t) + d(s\cap t) - d(s) - d(t) \\ &\quad = [d(s\cup t) - d(t)] - [d(s) - d(s\cap t)] \\ &\quad = d(s\cup t\setminus t\,|\,t) - d(s\setminus s\cap t|s\cap t). \end{aligned}$$

Since $s\cup t\setminus t = s\setminus s\cap t$, it would suffice to prove for disjoint sets s' and t that

$$d(s'|t) \geq d(s'|t') \tag{20}$$

for any $t' \subset t$.

However observe that, since $\mathbb{Q}$ is a product probability measure

$$d(s'|t) = E_{\mathbb{P}_t} D\left(\mathbb{P}_{s'|t}||\mathbb{Q}_{s'}\right) = E_{\mathbb{P}_{t'}} E_{\mathbb{P}_{t\setminus t'}} D\left(\mathbb{P}_{s'|t}||\mathbb{Q}_{s'}\right)$$

and

$$d(s'|t') = E_{\mathbb{P}_{t'}} D\left(\mathbb{P}_{s'|t'}||\mathbb{Q}_{s'}\right) = E_{\mathbb{P}_{t'}} D\left(E_{\mathbb{P}_{t\setminus t'}}\mathbb{P}_{s'|t}||\mathbb{Q}_{s'}\right)$$

so that (20) is an immediate consequence of the convexity of relative entropy (see, e.g., [10]). ■

Based on the supermodularity proved in Theorem V, Theorem I applied to $-d(s)$ immediately implies the following corollary.

Corollary VII: Under the assumptions of Theorem V

$$\begin{aligned} &\sum_{s\in\mathcal{C}} \gamma(s) D\left(\mathbb{P}_{s|s^c\setminus > s}||\mathbb{Q}_s|\mathbb{P}\right) \\ &\quad \geq D(\mathbb{P}_{[n]}||\mathbb{Q}_{[n]}) \\ &\quad \geq \sum_{s\in\mathcal{C}} \gamma(s) D(\mathbb{P}_{s|<s}||\mathbb{Q}_s|\mathbb{P}) \end{aligned} \tag{21}$$

where γ is any fractional partition using any hypergraph $\mathcal{C}$ on $[n]$.

Remark 9: We mention a hypothesis testing interpretation for the following easier-to-parse corollary of Corollary VII: for r-regular hypergraphs $\mathcal{C}$ on $[n]$

$$D(\mathbb{P}_{[n]}||\mathbb{Q}_{[n]}) \geq \frac{1}{r}\sum_{s\in\mathcal{C}} D(\mathbb{P}_s||\mathbb{Q}_s). \tag{22}$$

Suppose $\mathbb{P}$ and $\mathbb{Q}$ are two competing hypotheses for the joint distribution of $X_{[n]}$. Then it is a classical fact due to Chernoff (see, e.g., Cover and Thomas [10], where it is called Stein's lemma) that the best error exponent for a hypothesis test between $\mathbb{P}$ and $\mathbb{Q}$ based on a large number of i.i.d. observations of the random vector $X_{[n]}$ is given by $D(\mathbb{P}_{[n]}||\mathbb{Q}_{[n]})$. One may ask the following question: If one has partial access to all observations (for instance, one observes only X_s out of each $X_{[n]}$), then how much is our capacity to distinguish between the two hypotheses $\mathbb{P}$ and $\mathbb{Q}$ worsened? Corollary VII can be interpreted as giving us estimates that relate our capacity to distinguish between the two hypotheses given all the data to our capacity to distinguish between the two hypotheses given various subsets of the data.

Interestingly, Corollary VII implies a tensorization property of the entropy functional $\mathrm{Ent}_{\mathbb{Q}}(f) = E_{\mathbb{Q}}[f\log f] - (E_{\mathbb{Q}}f)\log(E_{\mathbb{Q}}f)$, defined for positive functions f. From the special case of Corollary VII corresponding to Han's inequality (i.e., the hypergraph $\mathcal{C}_{n-1}$ of all subsets of size $n-1$), one obtains the classical tensorization property, as noticed in [38]. We present a generalized tensorization inequality for the entropy functional with respect to a product measure by utilizing the power of Corollary VII more fully.

Corollary VIII: Let $\mathcal{C}$ be an r-regular hypergraph on $[n]$. Then

$$\mathrm{Ent}_{\mathbb{Q}_{[n]}}(g) \leq \frac{1}{r} E_{\mathbb{Q}} \sum_{s\in\mathcal{C}} \mathrm{Ent}_{\mathbb{Q}_s}(g).$$

We omit the proof, which is based on the observation that $\mathrm{Ent}_{\mathbb{Q}}(f) = (E_{\mathbb{Q}}f)\, D(\mathbb{P}||\mathbb{Q})$, where $\mathbb{P}$ is the probability measure such that $\frac{d\mathbb{P}}{d\mathbb{Q}} = \frac{f}{E_{\mathbb{Q}}f}$, and follows the same line of argument as in [38].

The tensorization property of the entropy functional is of enormous utility in functional analysis, and the study of isoperimetry, concentration of measure, and convergence of

Markov processes to stationarity. For instance, see [22], [4], and [30], where the classical tensorization property is used to prove logarithmic Sobolev inequalities for Gaussian, Poisson, and compound Poisson distributions, respectively.

X. Historical Remarks

It turns out that the main technical result of this paper, Theorem I, is related to a wide body of work in a number of fields, including the study of combinatorial optimization of set functions in computer science, the study of cooperative games in economics, the study of capacities in probability theory, and of course the study of structural properties of entropy in information theory, which has been our present focus. In this section, we sketch these connections and place our work in context.

The following terminology is useful.

Definition IX: The set function f is *fractionally subadditive* if

$$f([n]) \leq \sum_{s \in \mathcal{C}} \gamma(s) f(s) \tag{23}$$

for any $\mathcal{C} \subset 2^{[n]}$, and for any fractional partition $\gamma : \mathcal{C} \to \mathbb{R}_+$ of $[n]$. If the inequality is reversed, we say f is fractionally superadditive.

Note that Theorem I has the following corollary (basically Proposition II for general submodular functions), obtained by using(5) to weaken the upper bound in Theorem I.

Corollary IX: If $f : 2^{[n]} \to \mathbb{R}$ is submodular and $f(\phi) = 0$, then it is fractionally subadditive.

It is pertinent to observe that there is no monotonicity assumption needed for Corollary IX. This result has a long history, and has rarely been explicitly stated in the literature although aspects of it have been rediscovered on multiple occasions in various fields. First we describe how it is implicit in the classical theory of cooperative games.

In cooperative game theory, a set function $f : 2^{[n]} \to \mathbb{R}_+$ is called a value function; it can be thought of as describing the payoff that can be obtained by arbitrary coalitions of n players, and it is canonical to take $f(\phi) = 0$. Different assumptions on the value function f correspond to different kinds of games. For instance, a *balanced game* is one for which the value function is fractionally superadditive, i.e.

$$f([n]) \geq \sum_{s \in \mathcal{C}} \gamma(s) f(s) \tag{24}$$

holds for every fractional partition γ. If the value function f is supermodular, the corresponding game is said to be a *convex game*.

One solution concept for cooperative games is the core, a subset of Euclidean space representing possible allocations of the payoff to players. (We do not bother to define it here; it suffices for our brief remarks here to know that such a notion exists.) The fundamental Bondareva–Shapley theorem [5], [46] states that the game with transferable utility associated with the value function f has a nonempty core if and only if it is balanced. Separately, it is known from even earlier work of Kelley [28] (see also Shapley [47] who rediscovered it in the language of games) that a convex game has a nonempty core. Putting these together, one sees that a convex game must be balanced. This yields a statement very similar to that of Corollary IX.

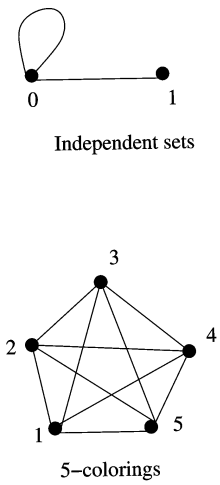

Fig. 1. Graphs F relevant for counting independent sets and number of 5-colorings.

Much more recently, yet another related approach to the relationship between submodularity and fractional subadditivity has come from the theory of combinatorial auctions. Lehmann *et al.* [31] showed that every submodular function is "XOS" (terminology that again we do not bother to explain here). Feige [14] showed that XOS and fractionally subadditive are identical. We refer the reader to the mentioned papers for definitions and details.

To summarize, the literature from cooperative game theory and combinatorial auction theory imply Corollary IX.

While we had expected direct proofs of Corollary IX to exist in the literature, we had initially been unable to find a reference. After the first version of this paper was written and presented at various venues, we were informed by Alan Sokal that it has indeed been explicitly stated and proved in the French statistical physics literature [40] (see also [49], where it is applied to entropy in a statistical physics context).

The above discussion is also related to the theory of polymatroids. A nondecreasing and submodular set function $f : 2^{[n]} \to \mathbb{R}_+$ with $f(\phi) = 0$ is sometimes called a β-function. This class of functions has been intensely studied ever since the pioneering work of Edmonds [13], who used them to define polymatroids. Note that the nondecreasing property (i.e., $f(t) \leq f(s)$ whenever $t \subset s \subset [n]$) implies that f is nonnegative. It is pertinent to note that the extra properties inherent in polymatroid theory are not required for Corollary IX and Theorem I (for instance, either a monotonicity or a nonnegativity requirement for f would rule out an application to the differential entropy); so Theorem I is really just a basic fact about submodular functions.

XI. Discussion

The inequalities presented in this note are contributions to a large body of work on the structural properties of the entropy function for joint distributions. While the origins of such work

lie in Shannon's foundational paper, let us again mention (see also the discussion after Theorem I') that the important observation of submodularity of the joint entropy function goes back at least to Fujishige [17]. There have also been interesting new developments in the last few years, namely the discovery of the so-called "non-Shannon inequalities". Motivated by the goal of characterizing the possible joint entropy set functions $\mathsf{e}(s) = H(X_s)$ for the discrete entropy as the underlying joint distribution is varied arbitrarily, Zhang and Yeung [51] revealed a fascinating phenomenon: if one thinks of each such e (corresponding to any joint distribution on n copies of a discrete alphabet) as being a vector of dimension 2^n, then the set of vectors one obtains in this manner is a strict subset of the set of vectors corresponding to polymatroidal functions for any $n \geq 4$. The constraints on joint entropy that are not automatic consequences of a polymatroid property were termed "non-Shannon inequalities" in [51]. For more recent developments on this subject, one may consult [24], [39], [12].

In the context of these works, it is pertinent to note that all of the inequalities in this paper are Shannon inequalities, in the sense that they follow from submodularity of an entropy function. Indeed, our study was based on the set function $\mathsf{e}(s) = H(X_s)$, from consideration of which our main entropy inequality (Theorem I') was derived. However, since we now know from the mentioned literature that entropy satisfies additional constraints beyond submodularity, a natural question arises. If it is true that the set function $\bar{\mathsf{e}}(s) = H(X_s|X_{<s})$ is itself submodular, so that Theorem I' then follows by an application of Corollary IX to $\bar{\mathsf{e}}$ rather than an application of Theorem I to e, then we would have a tighter outer bound on the space of joint entropy set functions. The following counterexample shows that this is not the case.

Proposition III: The set function $\bar{\mathsf{e}}(s)$ is not submodular.

Proof: We construct a counterexample with $n = 4$ random variables. Consider the sets $s = \{1,3\}$ and $t = \{3,4\}$. Then $s \cup t = \{1,3,4\}$ and $s \cap t = \{3\}$. If $\bar{\mathsf{e}}$ is submodular, then since s contains the first element

$$H(X_s) + H(X_t|X_{<t}) \geq H(X_{s\cup t}) + H(X_{s\cap t}|X_{<(s\cap t)})$$

which in our case becomes

$$H(X_{\{1,3\}}) + H\left(X_{\{3,4\}}|X_{\{1,2\}}\right) \geq H\left(X_{\{1,3,4\}}\right) + H\left(X_{\{3\}}|X_{\{1,2\}}\right). \quad (25)$$

By the chain rule

$$H(X_{\{1,3,4\}}) = H(X_{\{1,3\}}) + H(X_4|X_{\{1,3\}})$$

and

$$H\left(X_{\{3,4\}}|X_{\{1,2\}}\right) = H\left(X_4|X_{\{1,2,3\}}\right) + H\left(X_3|X_{\{1,2\}}\right)$$

so that (25) reduces to

$$H\left(X_{\{1,3\}}\right) + H\left(X_4|X_{\{1,2,3\}}\right) + H\left(X_3|X_{\{1,2\}}\right) \geq H\left(X_{\{1,3\}}\right) + H\left(X_4|X_{\{1,3\}}\right) + H\left(X_3|X_{\{1,2\}}\right)$$

and then simply to $H(X_4|X_{\{1,2,3\}}) \geq H(X_4|X_{\{1,3\}})$. However, this is in general not true since conditioning reduces entropy, and thus the hypothesis of submodularity is falsified. ■

Note, however, that such a counterexample is only possible when $s \cup t$ is strictly smaller than the index set $[n]$.

The relationship between the inequalities for discrete and continuous entropy in this paper is worth noting. Observe that a slightly more general class of inequalities holds for discrete entropy as compared to differential entropy (for instance, only fractional partitions are allowed in the differential entropy context in Theorem I'); however, this is not surprising and indeed follows from the equivalences explored by Chan [8].

The structural properties of entropy discussed in this work are not just of abstract interest. Some applications, to determinant inequalities and counting problems, have already been mentioned in earlier sections. The inequalities discussed also have close connections with several classical multiuser information theoretic problems, including the Slepian-Wolf data compression problem and the multiple access channel. In particular, for the Slepian-Wolf problem where data from n sources is to be losslessly compressed in a distributed fashion, it is the set function $H(X_s|X_{s^c})$ rather than $H(X_s)$ that plays the key role. Consequently, the *lower* bound in Theorem I' has a crucial significance: it is equivalent to the existence of a rate point whose sum rate is the same as the rate achievable for nondistributed compression (namely $H\left(X_{[n]}\right)$), and is one way of showing that no extra cost is paid in terms of asymptotic rate for the distributed nature of the task. These connections merit a separate and more detailed exploration, and are discussed along with several other applications of cooperative game theory to multiuser problems in [36].

Chain rules for entropy and relative entropy have played an important role in information theory since their recognition by Shannon. Here we have presented several inequalities for information in joint distributions that go beyond the chain rules but can also be thought of as deeper consequences of them. While these relate the information in projections of a random vector onto different subspaces, more general inequalities can be formulated that apply to a rich class of functions beyond projections (such as the sum), and these are described along with applications to additive combinatorics and matrix analysis in the follow-up works [34], [37]. We anticipate further extensions and applications of these inequalities in the future.

Just before this paper went to press, we became aware of some early references describing the use of information-theoretic arguments for counting. Please see: [N. Pippenger, "An information-theoretic method in combinatorial theory," *J. Comb. Theory, Ser. A*, vol. 23, pp. 99–104, 1977] and [N. Pippenger, "Entropy and Enumeration of Boolean Functions," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2096–2100, Sep. 1999] and references therein.

ACKNOWLEDGMENT

The authors would like to thank A. Barron for many useful discussions, and for the indirect influence of his joint work [33] with M. Madiman on entropy power inequalities. They would also like to thank the organizers of the IEEE International Symposium on Information Theory 2006 in Seattle, WA, where they met and initiated this work, and R. Bapat, S. Chaiken, U. Feige, G. Kalai, and A. Sokal for help with references. P. Tetali would like to thank the Theory Group at Microsoft Research for hosting him during the period this research was

carried out. They would also like to thank the three anonymous referees for very thorough feedback that eliminated an error and significantly improved the paper.

REFERENCES

[1] N. Alon and J. H. Spencer, *The Probabilistic Method*, second ed. New York: Wiley , 2000.

[2] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, "Solution of Shannon's problem on the monotonicity of entropy," *J. Amer. Math. Soc.*, vol. 17, no. 4, pp. 975–982, 2004.

[3] R. Bellman, *Introduction to Matrix Analysis*. New York: McGraw-Hill, 1960.

[4] S. G. Bobkov and M. Ledoux, "On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures," *J. Funct. Anal.*, vol. 156, no. 2, pp. 347–365, 1998.

[5] O. N. Bondareva, "Some applications of the methods of linear programming to the theory of cooperative games (in Russian)," *Problemy Kibernetiki*, vol. 10, pp. 119–139, 1963.

[6] G. Brightwell and P. Tetali, "The number of linear extensions of the boolean lattice," *Order*, vol. 20, pp. 333–345, 2003.

[7] S. Chaiken, "A combinatorial proof of the all minors matrix tree theorem," *SIAM J. Algebr. Discr. Methods*, vol. 3, no. 3, pp. 319–329, 1982.

[8] T. H. Chan, "Balanced information inequalities," *IEEE Trans. Inf. Theory*, vol. 49, pp. 3261–3267, 2003.

[9] F. R. K. Chung, R. L. Graham, P. Frankl, and J. B. Shearer, "Some intersection theorems for ordered sets and graphs," *J. Combinator. Theory, Ser. A*, vol. 43, pp. 23–37, 1986.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[11] A. Dembo, T. M. Cover, and J. A. Thomas, "Information-theoretic inequalities," *IEEE Trans. Inf. Theory*, vol. 37, pp. 1501–1518, 1991.

[12] R. Dougherty, C. Freiling, and K. Zeger, "Networks, matroids and non-Shannon information inequalities," *IEEE Trans. Inf. Theory*, vol. 53, no. 6, pp. 1949–1969, Jun. 2007.

[13] J. Edmonds, "Submodular functions, matroids and certain polyhedra," in *Proc. Int. Conf. Combinator. Struct. Appl.*, 1970.

[14] U. Feige, "On maximizing welfare when utility functions are subadditive," presented at the 38th ACM Symp. Theory Comput., Seattle, WA, May 2006.

[15] E. Friedgut, "Hypergraphs, entropy, and inequalities," *Amer. Math. Month.*, vol. 111, no. 9, pp. 749–760, Nov. 2004.

[16] E. Friedgut and J. Kahn, "On the number of copies of one hypergraph in another," *Israel J. Math.*, vol. 105, pp. 251–256, 1998.

[17] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Inf. Contr.*, vol. 39, pp. 55–72, 1978.

[18] S. Fujishige, *Submodular Functions and Optimization, Volume 58 of Annals of Discrete Mathematics*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2005.

[19] Z. Füredi, "Scrambling permutations and entropy of hypergraphs," *Random Structures Algorithms*, vol. 8, no. 2, pp. 97–104, 1996.

[20] D. Galvin, Pers. commun., 2006.

[21] D. Galvin and P. Tetali, "On weighted graph homomorphisms," *DIMACS-AMS Special Volume*, vol. 63, pp. 13–28, 2004.

[22] L. Gross, "Logarithmic Sobolev inequalities," *Amer. J. Math.*, vol. 97, no. 4, pp. 1061–1083, 1975.

[23] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inf. Contr.*, vol. 36, no. 2, pp. 133–156, 1978.

[24] B. Ibinson, N. Linden, and A. Winter, "All inequalities for the relative entropy," in *Proc. IEEE Intl. Symp. Inform. Theory, Seattle*, 2006, pp. 237–241.

[25] J. Kahn, "An entropy approach to the hard-core model on bipartite graphs," *Combinator., Probabil. Comput.*, vol. 10, pp. 219–237, 2001.

[26] J. Kahn, "Entropy, independent sets and antichains: A new approach to Dedekind's problem," *Proc. Amer. Math. Soc.*, vol. 130, no. 2, pp. 371–378, 2001.

[27] J. Kahn, Pers. Commun., 2006.

[28] J. L. Kelley, "Measures on Boolean algebras," *Pacific J. Math.*, vol. 9, pp. 1165–1177, 1959.

[29] G. Kirchhoff, "Über die auflösung der gleichungen, auf welche man bei der untersuchung der linearen verteilung galvanischer ströme geführt wird," *Ann. Phys. Chem.*, vol. 72, pp. 497–508.

[30] I. Kontoyiannis and M. Madiman, "Measure concentration for compound Poisson distributions," *Elect. Commun. Probab.*, vol. 11, pp. 45–57, 2006.

[31] B. Lehmann, D. Lehmann, and N. Nisan, "Combinatorial auctions with decreasing marginal utilities," in *Proc. 3rd ACM Conf. Electron. Commerce*, Tampa, FL, 2001, pp. 18–28.

[32] M. Lewin, "A generalization of the matrix-tree theorem," *Math. Z.*, vol. 181, no. 1, pp. 55–70, 1982.

[33] M. Madiman and A. R. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Trans. Inf. Theory*, vol. 53, pp. 2317–2329, Jul. 2007.

[34] M. Madiman, A. Marcus, and P. Tetali, Entropy and Set Cardinality Inequalities for Partition-Determined Functions, With Applications to Sumsets, 2008 [Online]. Available: http://arxiv.org/abs/0901.0055

[35] M. Madiman and P. Tetali, "Sandwich bounds for joint entropy," presented at the IEEE Int. Symp. Inf. Theory Nice, France, Jun. 2007.

[36] M. Madiman, "Cores of cooperative games in information theory," *EURASIP J. Wireless Commun. Netw.*, no. 318704, 2008.

[37] M. Madiman, *Determinant and Trace Inequalities for Sums of Positive-Definite Matrices* 2008, unpublished manuscript.

[38] P. Massart, "Some applications of concentration inequalities to statistics," *Ann. Faculté des Sci. Toulouse*, vol. IX, no. 2, pp. 245–303, 2000.

[39] F. Matúš, "Two constructions on limits of entropy functions," *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 320–330, Jan. 2007.

[40] J. M. Ollagnier and D. Pinchon, "Filtre moyennant et valeurs moyennes des capacités invariantes," *Bull. Soc. Math. France*, vol. 110, no. 3, pp. 259–277, 1982.

[41] H. Narayanan, "Submodular functions and electrical networks," in *Annals of Discrete Mathematics*. Amsterdam, The Netherlands: North-Holland, 1997, vol. 54.

[42] J. Nayak, E. Tuncel, and K. Rose, "Zero-error source-channel codingwith side information," *IEEE Trans. Inf. Theory*, vol. 52, pp. 4626–4629, 2006.

[43] J. Radhakrishnan, "Entropy and counting," in *Computational Mathematics, Modelling and Algorithms*, J. C. Misra, Ed. New Delhi, India: Narosa, 2003.

[44] E. R. Scheinerman and D. H. Ullman, *Fractional Graph Theory*. New York: Wiley, 1997.

[45] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[46] L. S. Shapley, "On balanced sets and cores," *Naval Res. Logist. Quarter.*, vol. 14, pp. 453–560, 1967.

[47] L. S. Shapley, "Cores of convex games," *Int. J. Game Theory*, vol. 1, no. 1, pp. 11–26, 1971.

[48] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Contr.*, vol. 2, pp. 101–112, 1959.

[49] A. C. D. van Enter, R. Fernández, and A. D. Sokal, "Regularity properties and pathologies of position-space renormalization-group transformations: Scope and limitations of Gibbsian theory," *J. Statist. Phys.*, vol. 72, no. 5-6, pp. 879–1167, 1993.

[50] S. Verdú and T. Weissman, "Erasure entropy," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, 2006.

[51] J. Zhang and R. W. Yeung, "On characterization of entropy function via information inequalities," *IEEE Trans. Inf. Theory*, vol. 44, pp. 1440–1452, 1998.

Mokshay Madiman (S'04–M'06) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, in 1999, and the Sc.M. and Ph.D. degrees in applied mathematics from Brown University, Providence, RI, in 2001 and 2005, respectively.

He spent the summer of 1998 as a Research Intern at the Institute of Mathematical Sciences, Chennai, India. He joined the Department of Statistics at Yale University, New Haven, CT, in 2005, as a Gibbs Assistant Professor. In the spring of 2007, he was a Visiting Fellow at the School of Technology and Computer Science at the Tata Institute of Fundamental Research (TIFR), Mumbai, India. Since July 2006, he has been an Assistant Professor of Statistics at Yale University, with a courtesy appointment in Applied Mathematics, as well. His current research interests span information theory and statistical inference, as well as aspects of probabilistic combinatorics, convex geometry, theoretical and applied probability, and the semantics of natural language.

Dr. Madiman was awarded a Laha Travel Award by the Institute of Mathematical Statistics (IMS) in order to present a paper at the IMS Annual Meeting in Rio de Janeiro, Brazil, in 2006. He also received a Junior Faculty Fellowship from Yale University in the spring of 2009.

Prasad Tetali received the B.E. degree from Andhra University, the M.S. degree from the Indian Institute of Science, Bangalore, and the Ph.D. degree in computer science in 1991 from the Courant Institute of Mathematical Sciences (NYU), New York.

After postdoctoral positions at DIMACS (Rutgers) and the AT&T Bell Labs, he joined the School of Mathematics at the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1994. Since 2000, he has also held a joint appointment with the School of Computer Science at Georgia Tech. He is currently a Professor at Georgia Tech, where he is affiliated with the Ph.D. program in Algorithms, Combinatorics and Optimization (ACO), and the Algorithms & Randomness Center (ARC) Thinktank. His research interests span the topics: probabilistic combinatorics, Markov chains, randomized algorithms, computational number theory, and statistical physics. He is an author (jointly with R. Montenegro) of a book titled *Mathematical Aspects of Mixing Times in Markov Chains* (NOW Publishers, 2006).

Dr. Tetali is the Editor-in-Chief of the SIAM *Journal on Discrete Mathematics* and also an Associate Editor of *Annals of Applied Probability, Random Structures & Algorithms,* and the *Journal of Combinatorics*. He was recognized as a SIAM Fellow in 2009 for his contributions to discrete mathematics and algorithms.