Sandwich bounds for joint entropy

Mokshay Madiman
Department of Statistics
Yale University
New Haven CT, USA
Email: mokshay.madiman@yale.edu

Prasad Tetali
School of Mathematics and College of Computing
Georgia Institute of Technology
Atlanta GA, USA
Email: tetali@math.gatech.edu

***Abstract*— New upper and lower bounds are given for joint entropy of a collection of random variables, in both discrete and continuous settings. These bounds generalize well-known information theoretic inequalities due to Han. A number of applications are suggested, including a new bound on the number of independent sets of a graph that is of interest in discrete mathematics, and a bound on the number of zero-error codes.**

I. Introduction

Let $X_1, X_2, \ldots, X_n$ be a collection of random variables. We assume that the joint distribution has a density f with respect to some reference measure, and define the joint entropy $H(X_1, X_2, \ldots, X_n) = -E[\log f(X_1, X_2, \ldots, X_n)]$. There are the familiar two canonical cases: (a) the random variables are real-valued and possess a probability density function, or (b) they are discrete. In the former case, H represents the differential entropy, and in the latter case, H represents the discrete entropy. Such distinctions do not matter in what follows, and we simply call H the entropy in all cases.

Shannon's chain rule says that

$$H(X, Y) = H(X) + H(Y|X) \tag{1}$$

where $H(Y|X) = E[-\log p(Y|X)]$ is the conditional entropy of Y given X. This rule, which is just an expression of the factorization of the joint distribution into a marginal and a conditional, extends to the consideration of n variables; indeed,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{<i}) \tag{2}$$

where $X_{<i}$ is used to denote $(X_j : j < i)$.

The main inequality we wish to present can be seen as a generalization of Shannon's chain rule. Since we wish to consider various subsets of random variables, the following notational conventions will be useful. Let $[n]$ be the index set $\{1, 2, \ldots, n\}$. We are interested in a collection $\mathcal{S}$ of subsets of $[n]$. For any set $\mathbf{s} \subset [n]$, $X_{\mathbf{s}}$ stands for the collection of random variables $(X_i : i \in \mathbf{s})$, with the indices taken in their increasing order. For any index i in $[n]$, define the *degree* of i in $\mathcal{S}$ as $r(i) = |\{\mathbf{t} \in \mathcal{S} : i \in \mathbf{t}\}|$.

First we present a weak form of our main inequality.

Proposition I:[Weak form] For any collection $\mathcal{S}$ such that every index i appears in at least one element of $\mathcal{S}$,

$$\sum_{\mathbf{s}\in\mathcal{S}} \frac{H(X_{\mathbf{s}}|X_{\mathbf{s}^c})}{r_+(\mathbf{s})} \le H(X_{[n]}) \le \sum_{\mathbf{s}\in\mathcal{S}} \frac{H(X_{\mathbf{s}})}{r_-(\mathbf{s})},$$

where $r_+(\mathbf{s}) = \max_{i\in\mathbf{s}} r(i)$ is the maximum degree within $\mathbf{s}$, and $r_-(\mathbf{s}) = \min_{i\in\mathbf{s}} r(i)$ is the minimum degree within $\mathbf{s}$.

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

1) Applying to the class $\mathcal{S}_1$ of singletons,
$$\sum_{i=1}^{n} H(X_i|X_{[n]\setminus i}) \le H(X_{[n]}) \le \sum_{i=1}^{n} H(X_i). \tag{3}$$
The upper bound represents the subadditivity of entropy noticed by Shannon. The lower bound may be interpreted as the fact that the "erasure entropy" [1] of a collection of random variables is not greater than their entropy.
2) Applying to the class $\mathcal{S}_{n-1}$ of all sets of $n-1$ elements,
$$\frac{1}{n-1}\sum_{i=1}^{n} H(X_{[n]\setminus i}|X_i) \le H(X_{[n]}) \le \frac{1}{n-1}\sum_{i=1}^{n} H(X_{[n]\setminus i}). \tag{4}$$
This is Han's inequality [2], [3], in its prototypical form.
3) Let $r_+ = \max_{i\in[n]} r(i)$ and $r_- = \min_{i\in[n]} r(i)$ be the minimal and maximal degrees with respect to $\mathcal{S}$. Using $r_- \le r_-(\mathbf{s})$ and $r_+ \ge r_+(\mathbf{s})$, we have
$$\frac{1}{r_+}\sum_{\mathbf{s}\in\mathcal{S}} H(X_{\mathbf{s}}|X_{\mathbf{s}^c}) \le H(X_{[n]}) \le \frac{1}{r_-}\sum_{\mathbf{s}\in\mathcal{S}} H(X_{\mathbf{s}}).$$
The upper bound is Shearer's lemma [4], known in the combinatorics literature [5]. The lower bound is new. A version of the general upper bound of Proposition I was first obtained by Friedgut [6].

In Section II, we present and prove our main inequality, which strengthens Proposition I. This inequality is developed in two forms, which we call the fractional form and the degree form. In Section III, we apply the degree form of the inequality to obtain a new upper bound on the number of independent sets of an arbitrary graph. Section IV generalizes this bound to estimate the number of zero-error codes. In Section V, a

duality between upper and lower bounds for joint entropy is presented. Section VI studies the special case of the collections $\mathcal{S}_m$ consisting of all sets of m elements, and recovers results of Han and Fujishige. Finally Section VII presents a version of our main inequality for relative entropy.

II. The Fractional and Degree Forms

The main inequality introduced in this work is the following generalization of Shannon's chain rule. Let $< \mathbf{s}$ denote the set of indices preceding every index in $\mathbf{s}$, and $> \mathbf{s}$ be defined analogously.

Theorem I:[Strong Fractional Form] Let $\mathcal{S}$ be any collection of subsets of $[n]$. A function $\alpha : \mathcal{S} \to \mathbb{R}_+$, is called a *fractional covering*, if for each $i \in [n]$, we have $\sum_{\mathbf{s}\in\mathcal{S}:i\in\mathbf{s}} \alpha(\mathbf{s}) \geq 1$. Similarly, $\beta : \mathcal{S} \to \mathbb{R}_+$ is a *fractional packing*, if for each $i \in [n]$, we have $\sum_{\mathbf{s}\in\mathcal{S}:i\in S} \beta(\mathbf{s}) \leq 1$. For any fractional packing β and any fractional covering α,

$$\sum_{S\in\mathcal{S}} \beta(\mathbf{s}) H(X_{\mathbf{s}}|X_{\mathbf{s}^c\setminus>\mathbf{s}}) \leq H(X_{[n]}) \leq \sum_{S\in\mathcal{S}} \alpha(\mathbf{s}) H(X_{\mathbf{s}}|X_{<\mathbf{s}}) .$$

Proof: By Shannon's chain rule for entropy,

$$H(X_{\mathbf{s}}|X_{<\mathbf{s}}) = \sum_{j\in\mathbf{s}} H(X_j|X_{<j\cap\mathbf{s}}, X_{<\mathbf{s}}). \tag{5}$$

Thus

$$\begin{aligned}
\sum_{\mathbf{s}\in\mathcal{S}} \alpha(\mathbf{s}) H(X_{\mathbf{s}}|X_{<\mathbf{s}}) &\overset{(a)}{=} \sum_{\mathbf{s}\in\mathcal{S}} \alpha(\mathbf{s}) \sum_{j\in\mathbf{s}} H(X_j|X_{<j\cap\mathbf{s}}, X_{<\mathbf{s}}) \\
&\overset{(b)}{\geq} \sum_{\mathbf{s}\in\mathcal{S}} \alpha(\mathbf{s}) \sum_{j\in\mathbf{s}} H(X_j|X_{<j}) \\
&\overset{(c)}{=} \sum_{j\in[n]} H(X_j|X_{<j}) \sum_{\mathbf{s}\in\mathcal{S}} \alpha(\mathbf{s}) \mathbf{1}_{\{j\in\mathbf{s}\}} \\
&\overset{(d)}{\geq} \sum_{j\in[n]} H(X_j|X_{<j}) \\
&\overset{(a)}{=} H(X_{[n]}),
\end{aligned}$$

where (a) follows by the chain rule for entropy (5), (b) follows since extra conditioning reduces entropy, (c) follows by interchanging sums, and (d) follows by the definition of a fractional covering. The lower bound is proved in a similar fashion. ∎

We now use the fractional form of the main inequality, namely Theorem I, to prove the degree form of the inequality below.

Theorem II:[Strong Degree Form] Let $\mathcal{S}$ be any collection of subsets of $[n]$, such that every index i appears in at least one element of $\mathcal{S}$. Then

$$\sum_{\mathbf{s}\in\mathcal{S}} \frac{H(X_{\mathbf{s}}|X_{\mathbf{s}^c\setminus>\mathbf{s}})}{r_+(\mathbf{s})} \leq H(X_{[n]}) \leq \sum_{\mathbf{s}\in\mathcal{S}} \frac{H(X_{\mathbf{s}}|X_{<\mathbf{s}})}{r_-(\mathbf{s})}.$$

Proof: The numbers $\alpha(\mathbf{s}) = \frac{1}{r_-(\mathbf{s})}$ provide a fractional covering, which we call the *degree covering*. Indeed, as long as there is at least one set $\mathbf{s}$ in the collection $\mathcal{S}$ that contains i, we have

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

Similarly, the numbers $\beta(\mathbf{s}) = \frac{1}{r_+(\mathbf{s})}$ provide a fractional packing, which we call the *degree packing*. Applying Theorem I with the degree covering and degree packing for $\mathcal{S}$, we obtain Theorem II. ∎

Remark: 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 $< \mathbf{s}$ in the upper bound, and including conditioning on $> \mathbf{s}$ in the lower bound. Note that the upper bound of Proposition I was first proved, in a different form and in a more involved manner, by Friedgut [6].

Remark: Theorem II is actually equivalent to Theorem I. To see why Theorem II implies Theorem I, note that the sets in the collection $\mathcal{S}$ need not be distinct, and writing down Theorem II with arbitrary number of repetitions of the sets in $\mathcal{S}$ gives a version of Theorem I with rational coefficients. An approximation argument can then be used to complete the implication. This proof is similar to the one alluded to by Friedgut [6] for the version without ordering.

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

$$\sum_{i=1}^{n} H(X_i|X_{[n]\setminus\geq i}) \leq H(X_{[n]}) \leq \sum_{i=1}^{n} H(X_i|X_{<i}),$$

which is precisely Shannon's chain rule, since the upper and lower bounds are identical. Note in contrast the looseness of the upper and lower bounds in (3), which are tight if and only if the random variables X_i are independent.

III. Counting independent sets

A graph $G = (V, E)$ consists of a finite vertex set V and a collection E of two-element subsets of V called edges. 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.

Shearer's lemma, and more generally, entropy-based arguments, have proved very useful in combinatorics. Shearer's lemma was (implicitly) introduced in [4], and Kahn [7] stated an extension using the more familiar entropy notation. Recent applications of Shearer's lemma to difficult problems in combinatorics include [8], [7], [9], and [10]. Radhakrishnan [5] provides a nice survey of entropy ideas used for counting and various applications. The general strategy of entropy-based proofs in counting the number of objects in a class $\mathcal{C}$ of objects

is to consider a randomly drawn object X from the class, note that its entropy is $H(X) = \log|\mathcal{C}|$, and to estimate $H(X)$ using Shearer's lemma and further manipulation.

Below, we follow this approach, utilizing Theorem II instead of Shearer's lemma, to bound the number of independent sets of an arbitrary graph. Our bound extends the following recent result of Kahn, which he proved using Shearer's lemma. For a *d-regular* graph G (namely, one in which each vertex has the same degree d) on n vertices, the number of independent sets is bounded by

$$\prod_{v\in V} 2^{(p_a(v)+1)\frac{1}{d}} \leq 2^{\frac{N}{2}+\frac{N}{d}},$$

where $p_a(v)$ denotes the number of neighbors of v that precede it with respect to an arbitrary total order $\prec_a$ on V. We remove the assumption of regularity in the result below, at the cost of choosing a particular ordering.

Theorem III:[INDEPENDENT SETS] Let $G = (V, E)$ be an arbitrary graph on N vertices. 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

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

Proof: Let X be an independent set of G, chosen uniformly at random from $\mathcal{I}(G)$. The random independent set X can be represented by n indicator variables corresponding to the vertices, i.e., $X = (X(1), X(2), \ldots, X(n)) = (X_1, X_2, \ldots, X_n)$, where

$$X_i = \begin{cases} 1 & \text{if } i \in X \\ 0 & \text{otherwise} \end{cases}.$$

Let $\prec$ denote an ordering on vertices according to the decreasing order of their degrees. 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 $\mathcal{S}$ 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 $\mathcal{S}$, i.e., that the degree of i in the collection $\mathcal{S}$ 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 $\mathcal{S}$, we have

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

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

Let q denote the probability mass function of $X_{P(i)}$. In other words, $q(x_{P(i)}) = \Pr\{X_{P(i)} = x_{P(i)}\}$ for each $x_{P(i)} \in \mathcal{X}_i$, where $\mathcal{X}_i = \{x_{P(i)} : x_{[n]} \text{ represents an independent set}\}$. Then

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

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 $\log R(x_{P(i)})$. There are 2 cases to consider: (i) if $x_{P(i)} \neq 0_{P(i)}$, then X_i must be 0 since X is an independent set and vertex i is adjacent to each vertex in $P(i)$, whereas (ii) if $x_{P(i)} = 0_{P(i)}$, then X_i can be either 0 or 1. Thus, setting $q_0 = q(0_{P(i)})$,

$$\begin{aligned} H(X) &\leq \sum_{i\in V} \frac{1}{d(i)} \Bigg[q_0 \log \frac{2^{p(i)}}{q_0} \\ &\quad + \sum_{x_{P(i)}\in\mathcal{X}_i\setminus\{0_{P(i)}\}} q(x_{P(i)}) \log \frac{1}{q(x_{P(i)})}\Bigg] \\ &\leq \sum_{i\in V} \frac{1}{d(i)} \log(2^{p(i)} + 2^{p(i)}) \end{aligned}$$

using Jensen's inequality in the last step. The proof is completed by noting that $H(X) = \log|\mathcal{I}(G)|$. ∎

IV. AN APPLICATION TO ZERO-ERROR CODES

Given a graph $F = (V(F), E(F))$, possibly with self-loops, the set $Hom(G, F)$ of homomorphisms from G to F is defined as

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

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

The proof for the bound on independent sets given above extends to also provide an upper bound on the number of homomorphisms from an arbitrary graph G to an arbitrary graph F, as stated in Theorem IV below. The proof details are in [11].

Theorem IV:[GRAPH HOMOMORPHISMS] For any graph $G = (V, E)$ and any graph F,

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

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

By choosing appropriate graphs F, various corollaries can be obtained, including the independent set result of Theorem III, and a bound on the number of k-colorings of an arbitrary graph.

In [12], it is noted 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.

V. Duality

Consider the main entropy inequality, Theorem I, in its weaker version ignoring $< \mathbf{s}$ and $> \mathbf{s}$, for simplicity. That is,

$$\sum_{\mathbf{S}\in\mathcal{S}} \beta(\mathbf{s}) H(X_{\mathbf{S}}|X_{\mathbf{S}^c}) \le H(X_{[n]}) \le \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{s}) H(X_{\mathbf{S}}). \quad (6)$$

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

For a collection $\mathcal{S}$, with α (and β) denoting an arbitrary fractional covering (and packing, respectively) of $\mathcal{S}$, let

$$\text{Gap}_L(\mathcal{S},\beta) = H(X_{[n]}) - \sum_{\mathbf{S}\in\mathcal{S}} \beta(\mathbf{s}) H(X_{\mathbf{S}}|X_{\mathbf{S}^c})$$

and $$\text{Gap}_U(\mathcal{S},\alpha) = \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{s}) H(X_{\mathbf{S}}) - H(X_{[n]}).$$

Theorem IV:[Duality of Gaps] Define the complimentary collection to $\mathcal{S}$ as $\bar{\mathcal{S}} = \{\mathbf{s}^c : \mathbf{s} \in \mathcal{S}\}$. Then

$$\text{Gap}_U(\mathcal{S},\alpha) = \Big(\sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{s}) - 1\Big) \text{Gap}_L(\bar{\mathcal{S}},\bar{\beta}),$$

where $\bar{\beta}$ is a fractional packing using the complementary collection $\bar{\mathcal{S}}$ defined as, $\bar{\beta}(\mathbf{s}^c) = \frac{\alpha(\mathbf{S})}{\sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S})-1}$.

Proof: This is a straightforward computation involving the various log likelihoods. Indeed,

$$\begin{aligned}
\text{Gap}_U(\mathcal{S},\alpha) &= E\left[\log P(X_{[n]}) - \sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{s}) \log P(X_{\mathbf{S}})\right] \\
&= E\left[\log \frac{P(X_{[n]})}{\prod_{\mathbf{S}\in\mathcal{S}} P(X_{\mathbf{S}})^{\alpha(\mathbf{S})}}\right] \\
&= E\left[\log \prod_{\mathbf{S}\in\mathcal{S}} \left\{\frac{P(X_{[n]})}{P(X_{\mathbf{S}})}\right\}^{\alpha(\mathbf{S})} - \log P(X_{[n]})^{(\sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{S})-1)}\right] \\
&= \left(\sum_{\mathbf{S}\in\mathcal{S}} \alpha(\mathbf{s}) - 1\right)\left[H(X_{[n]}) - \sum_{\mathbf{S}\in\mathcal{S}} \bar{\beta}(\mathbf{s}^c) H(X_{\mathbf{S}^c}|X_{\mathbf{S}})\right],
\end{aligned}$$

which proves the claim. It is easy to check that $\bar{\beta}$ is indeed a fractional packing of $\bar{\mathcal{S}}$, using the fact that α is a fractional covering of $\mathcal{S}$. ∎

In particular, an upper bound for $H(X_{[n]})$ with respect to a collection $\mathcal{S}$ is equivalent to a lower bound for $H(X_{[n]})$ with respect to the complimentary collection $\bar{\mathcal{S}}$, implying that the *collection* of upper bounds for all collections and all fractional coverings is equivalent to the *collection* of lower bounds for all collections and all fractional packings!

VI. The Collections of k-sets

The gaps in the inequalities are particularly interesting when they are considered in the degree form of Proposition I. For simplicity, we restrict our attention to r-regular collections $\mathcal{S}$, i.e., collections in which each index has the same degree r. Suppose

$$g_L(\mathcal{S}) = H(X_{[n]}) - \frac{1}{r}\sum_{\mathbf{S}\in\mathcal{S}} H(X_{\mathbf{S}}|X_{\mathbf{S}^c})$$

and $$g_U(\mathcal{S}) = \frac{1}{r}\sum_{\mathbf{S}\in\mathcal{S}} H(X_{\mathbf{S}}) - H(X_{[n]})$$

are the gaps associated with Proposition I applied to $\mathcal{S}$.

Corollary I:[Duality for Regular Collections] For a r-regular collection $\mathcal{S}$,

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

The special collections $\mathcal{S}_k$, $k = 1, 2, \ldots, n$, consisting of all k-sets or sets of size k, are of particular interest. Indeed, Han's theorem [2] implies Proposition I for these collections. We note that Fujishige [13], building on terminology of Han, called the quantity $g_U(\mathcal{S}_k)$ a "total correlation", and $g_L(\mathcal{S}_k)$ a "dual total correlation". Applied to the collection $\mathcal{S}_k$, Corollary I implies that

$$\frac{g_L(\mathcal{S}_{n-k})}{g_U(\mathcal{S}_k)} = \frac{k}{n-k}.$$

This recovers an observation made in [13]. Further connections with [13] are explored in [11].

Han also demonstrated a monotonicity property of the total correlations (and the dual total correlations). Since this complements the duality result, we state it below and note that it follows from Han's inequality (4) (see, e.g., [14]).

Theorem V:[Han's Monotonicity] Both $g_L(\mathcal{S}_k)$ and $g_U(\mathcal{S}_k)$ are monotonically decreasing in k.

Remark: [14] gives a nice interpretation of this fact. Let

$$h_k^{(U)} = \frac{1}{\binom{n}{k}} \sum_{\mathbf{S}:|\mathbf{S}|=k} \frac{H(X_{\mathbf{S}})}{k}$$

denote the joint entropy per element for subsets of size k averaged over all k-element subsets. Since $g_k^{(U)} = n h_k^{(U)} - H(X_{[n]})$, Theorem V asserts that $h_k^{(U)}$ is decreasing in k. Suppose we have n sensors collecting data relevant to the task at hand. Suppose due to experimental conditions, at any time, we only have access to a random subset of m sensor measurements out of n. On average, are we getting more information as m increases? [14] notes that the answer to this and related questions is contained in Han's theorem.

Remark: [14] also uses Han's theorem to demonstrate determinantal inequalities including the Hadamard and Szasz

inequalities. In a similar manner, the more general entropy inequality of Theorem I implies the following more general determinantal inequality by considering multivariate normal distributions. Let K be a positive definite $n \times n$ matrix and let $\mathcal{S}$ be a collection of subsets of $[n]$. Let $K(\mathbf{s})$ denote the submatrix corresponding to the rows and columns indexed by elements of $\mathbf{s}$, and $|\cdot|$ denote determinant. Then for any α^* that is both a fractional packing and a fractional covering,

$$\prod_{\mathbf{s}\in\mathcal{S}} \Big(\frac{|K|}{|K(\mathbf{s}^c)|}\Big)^{\alpha^*(\mathbf{s})} \leq |K| \leq \prod_{\mathbf{s}\in\mathcal{S}} |K(\mathbf{s})|^{\alpha^*(\mathbf{s})}.$$

Further details are in [11].

VII. An Inequality for Relative Entropy

Now suppose P is a joint distribution on n random variables taking values in chosen spaces, such that P is absolutely continuous with respect to the product measure Q. Let p and q denote the densities of these measures with respect to a common reference measure (which may be taken to be Q itself). Let us use $p_{X_\mathbf{S}}$ to denote the marginal density of $X_\mathbf{S}$ when $X_{[n]}$ is distributed according to P. Recall also the definition of the relative entropy

$$D(p_{X_\mathbf{S}} \| q_{X_\mathbf{S}}) = E_P\left[\log \frac{p(X_\mathbf{S})}{q(X_\mathbf{S})}\right]$$

and the conditional relative entropy

$$D(p_{X_\mathbf{S}|X_\mathbf{T}} \| q_{X_\mathbf{S}|X_\mathbf{T}} | P) = E_{p_{X_\mathbf{T}}} D(p_{X_\mathbf{S}|X_\mathbf{T}} \| q_{X_\mathbf{S}|X_\mathbf{T}}).$$

We have the following result, the proof of which we omit for brevity and may be found in [11].

Theorem VI:[Bounding Relative Entropy] For any collection $\mathcal{S}$ of subsets of $[n]$, and any σ-finite product measure Q,

$$\sum_{\mathbf{s}\in\mathcal{S}} \beta(\mathbf{s}) D(p_{X_\mathbf{S}|X_{\mathbf{S}^c\setminus>\mathbf{S}}} \| q_{X_\mathbf{S}} | P) \geq D(p_{X_{[n]}} \| q_{X_{[n]}}) \geq \sum_{\mathbf{s}\in\mathcal{S}} \alpha(\mathbf{s}) D(p_{X_\mathbf{S}|X_{<\mathbf{S}}} \| q_{X_\mathbf{S}} | P).$$

Note that the discrete and continuous cases of Theorem I can be seen as special cases of Theorem VI by choosing Q to be counting measure and Lebesgue measure respectively.

To get a sense of what Theorem VI means, we note that it implies

$$D(p_{X_{[n]}} \| q_{X_{[n]}}) \leq \sum_{\mathbf{s}\in\mathcal{S}} \frac{D(p_{X_\mathbf{S}} \| q_{X_\mathbf{S}})}{r_-(\mathbf{s})}, \tag{7}$$

which has a hypothesis testing interpretation. Suppose P and Q are two competing hypotheses for the joint distribution of $X_{[n]}$. By Stein's lemma [15], [3], the best error exponent for a hypothesis test between P and Q based on a large number N of i.i.d. observations of the random vector $X_{[n]}$ is given by $D(p_{X_{[n]}} \| q_{X_{[n]}})$. One may ask the following question: If one has partial access to all observations (for instance, one observes only $X_\mathbf{S}$ out of each $X_{[n]}$), then how much is our capacity to distinguish between the two hypotheses P and Q worsened? Theorem VI 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.

VIII. Conclusions

The main entropy inequalities we present in Theorems I and II are of interest in their own right, but we also demonstrated their usefulness by applying them to obtain new combinatorial results. Further details can be found in the full paper [11], where we also explore connections to the new entropy power inequalities developed in [16], as well as to multi-user information theory, submodular function theory and game theory. We believe that the information inequalities developed here will continue to find applications in information theory and related fields.

Acknowledgment

We are indebted to Andrew Barron for many useful discussions, the beneficial influence of his joint work with one of us on entropy power inequalities, and for pointing out an error in an earlier draft of this paper. We thank the anonymous referees for constructive comments.

References

[1] S. Verdú and T. Weissman, "Erasure entropy," *Proc. IEEE Intl. Symp. Inform. Th., Seattle*, 2006.

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

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

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

[5] J. Radhakrishnan, "Entropy and counting," in *IIT Kharagpur Golden Jubilee Volume*, 2001. [Online]. Available: http://www.tcs.tifr.res.in/ jaikumar/mypage.html

[6] E. Friedgut, "Hypergraphs, entropy, and inequalities," *The American Mathematical Monthly*, vol. 111, no. 9, pp. 749–760, November 2004.

[7] 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.

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

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

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

[11] M. Madiman and P. Tetali, "Information inequalities for joint distributions, with interpretations and applications," *In preparation*, 2007.

[12] J. Nayak, E. Tuncel, and K. Rose, "Zero-error source–channel coding-with side information," *IEEE Trans. Inform. Th.*, vol. 52, pp. 4626–4629, 2006.

[13] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Information and Control*, vol. 39, pp. 55–72, 1978.

[14] A. Dembo, T. Cover, and J. Thomas, "Information-theoretic inequalities," *IEEE Trans. Inform. Theory*, vol. 37, no. 6, pp. 1501–1518, 1991.

[15] H. Chernoff, "Large-sample theory: Parametric case," *Ann. Math. Stat.*, vol. 27, pp. 1–22, 1956.

[16] M. Madiman and A. Barron, "Generalized entropy power inequalities and monotonicity properties of information," To appear in *IEEE Trans. Inform. Theory*, 2007.