Combinatorial Applications of Shearer's Lemma & Generalizations

Igal Sason, Technion - Israel Institute of Technology

November 1, 2024

ETH, Zurich, Switzerland

Shearer's Lemma

Shearer's lemma extends the subadditivity property of Shannon entropy.

Proposition 1.1 (Shearer's Lemma)

Let $X_1, \ldots, X_n$ be discrete random variables, and let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ be subsets such that each element $i \in [n]$ belongs to at least $k \geq 1$ of these subsets. Then,

$$k \, \mathrm{H}(X^n) \leq \sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}), \tag{1.1}$$

with $X^n \triangleq (X_1, \ldots, X_n)$, and $X_{\mathcal{S}_j} \triangleq (X_i)_{i \in \mathcal{S}_j}$ for all $j \in [m]$.

Proof of Shearer's Lemma (Proposition 1.1)

- By assumption, $d(i) \geq k$ for all $i \in [n]$, where

$$d(i) \triangleq \left|\left\{j \in [m] : \, i \in \mathcal{S}_j\right\}\right|. \tag{1.2}$$

- Let $\mathcal{S} = \{i_1, \ldots, i_\ell\}$, $1 \leq i_1 < \ldots < i_\ell \leq n \implies |\mathcal{S}| = \ell$, $\mathcal{S} \subseteq [n]$.
- Let $X_{\mathcal{S}} \triangleq (X_{i_1}, \ldots, X_{i_\ell})$.
- By the chain rule and the fact that conditioning reduces entropy,

$$\begin{aligned}
\mathrm{H}(X_{\mathcal{S}}) &= \mathrm{H}(X_{i_1}) + \mathrm{H}(X_{i_2}|X_{i_1}) + \ldots + \mathrm{H}(X_{i_\ell}|X_{i_1}, \ldots, X_{i_{\ell-1}}) \\
&\geq \sum_{i \in \mathcal{S}} \mathrm{H}(X_i|X_1, \ldots, X_{i-1}) \\
&= \sum_{i=1}^{n} \Big\{ \mathbb{1}\{i \in \mathcal{S}\} \, \mathrm{H}(X_i|X_1, \ldots, X_{i-1}) \Big\}. \qquad (1.3)
\end{aligned}$$

Proof of Shearer's Lemma (Cont.)

$$\begin{aligned}
\sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}) &\geq \sum_{j=1}^{m} \sum_{i=1}^{n} \Big\{ \mathbb{1}\{i \in \mathcal{S}_j\} \, \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \Big\} \\
&= \sum_{i=1}^{n} \Big\{ \sum_{j=1}^{m} \mathbb{1}\{i \in \mathcal{S}_j\} \, \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \Big\} \\
&= \sum_{i=1}^{n} \Big\{ d(i) \, \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \Big\} \\
&\geq k \sum_{i=1}^{n} \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \qquad (1.4) \\
&= k \, \mathrm{H}(X^n),
\end{aligned}$$

where inequality (1.4) holds due to the nonnegativity of the conditional entropies of discrete random variables, and under the assumption that $d(i) \geq k$ for all $i \in [n]$.

Remark 1

1. Proposition 1.1 does not extend to continuous random variables, with entropies replaced by differential entropies, as the differential entropy of a continuous random variable may be negative, thereby invalidating inequality (1.4) under the assumption that $d(i) \geq k$ for all $i \in [n]$.
2. If each element $i \in [n]$ belongs to exactly k of the sets $\{\mathcal{S}_j\}_{j=1}^{m}$, then inequality (1.4) becomes an equality, irrespective of the nonnegativity issue of the conditional entropies.
   $\Longrightarrow$ If $d(i) = k$ for all $i \in [n]$, then Shearer's lemma extends to continuous random variables, with entropies replaced by differential entropies on both sides of inequality (1.1), as conditioning reduces the entropy for both discrete and continuous random variables.

A Geometric Application of Shearer's Lemma

Example 1.1

Let $\mathcal{P} \subseteq \mathbb{R}^3$ be a set of points that has at most r distinct projections on each of the XY, XZ and YZ planes. How large can this set be ?

A Geometric Application of Shearer's Lemma

Example 1.1

Let $\mathcal{P} \subseteq \mathbb{R}^3$ be a set of points that has at most r distinct projections on each of the XY, XZ and YZ planes. How large can this set be ?

As we shall see in the next slide,

$$|\mathcal{P}| \leq r^{\frac{3}{2}}.$$

Furthermore, that bound on the cardinality of the set $\mathcal{P}$ is achieved by a grid of $\sqrt{r} \times \sqrt{r} \times \sqrt{r}$ points, provided that r is a square of an integer.

Example 1.1 (cont.)

- By Shearer's lemma,

$$2\,\mathrm{H}(X,Y,Z) \le \mathrm{H}(X,Y) + \mathrm{H}(X,Z) + \mathrm{H}(Y,Z). \tag{1.5}$$

- Let $(X,Y,Z) \in \mathcal{P}$ be selected uniformly at random in $\mathcal{P}$. Then,

$$\mathrm{H}(X,Y,Z) = \log|\mathcal{P}|. \tag{1.6}$$

- By assumption, the set $\mathcal{P}$ has at most r distinct projections on each of the XY, XZ, and YZ planes. Hence,

$$\mathrm{H}(X,Y) \le \log r, \quad \mathrm{H}(X,Z) \le \log r, \quad \mathrm{H}(Y,Z) \le \log r. \tag{1.7}$$

- Combining (1.5)–(1.7) gives

$$2\log|\mathcal{P}| \le 3\log r, \tag{1.8}$$

and then exponentiating both sides of (1.8) gives $|\mathcal{P}| \le r^{\frac{3}{2}}$.

Direct proof of (1.5)

By the chain rule for the Shannon entropy,

$$\mathrm{H}(X,Y,Z) = \mathrm{H}(X) + \mathrm{H}(Y|X) + \mathrm{H}(Z|X,Y), \tag{1.9}$$

and as conditioning reduces entropy,

$$\begin{aligned}
\mathrm{H}(X,Y) &= \mathrm{H}(X) + \mathrm{H}(Y|X), && (1.10)\\
\mathrm{H}(X,Z) &= \mathrm{H}(X) + \mathrm{H}(Z|X) \\
&\geq \mathrm{H}(X) + \mathrm{H}(Z|X,Y), && (1.11)\\
\mathrm{H}(Y,Z) &= \mathrm{H}(Y) + \mathrm{H}(Z|Y), \\
&\geq \mathrm{H}(Y|X) + \mathrm{H}(Z|X,Y). && (1.12)
\end{aligned}$$

Combining (1.10)–(1.12) yields

$$\begin{aligned}
\mathrm{H}(X,Y) + \mathrm{H}(X,Z) + \mathrm{H}(Y,Z) &\geq 2\big(\mathrm{H}(X) + \mathrm{H}(Y|X) + \mathrm{H}(Z|X,Y)\big) \\
&= 2\,\mathrm{H}(X,Y,Z), && (1.13)
\end{aligned}$$

which proves inequality (1.5).

Proposition 1.2 (Shearer's Lemma: Second Version)

Let X^n be a discrete n-dimensional random vector, and let $\mathcal{S} \subseteq [n]$ be a random subset of $[n]$, independent of X^n, with an arbitrary probability mass function $\mathrm{P}_{\mathcal{S}}$. If there exists $\theta > 0$ such that

$$\Pr[i \in \mathcal{S}] \geq \theta, \quad \forall\, i \in [n], \tag{1.14}$$

then,

$$\mathbb{E}_{\mathcal{S}}\big[\mathrm{H}(X_{\mathcal{S}})\big] \geq \theta\, \mathrm{H}(X^n). \tag{1.15}$$

Proof of Proposition 1.2

By inequality (1.3), for any set $\mathcal{S} \subseteq [n]$,

$$\mathrm{H}(X_{\mathcal{S}}) \geq \sum_{i=1}^{n} \Big\{ \mathbb{1}\{i \in \mathcal{S}\}\ \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \Big\}.$$

Proof of Proposition 1.2 (cont.)

$$
\begin{aligned}
\implies \mathbb{E}_{\mathcal{S}}\big[\mathrm{H}(X_{\mathcal{S}})\big] &= \sum_{\mathcal{S}\subseteq[n]} \mathsf{P}_{\mathcal{S}}(\mathcal{S})\ \mathrm{H}(X_{\mathcal{S}}) \\
&\geq \sum_{\mathcal{S}\subseteq[n]} \Bigg\{ \mathsf{P}_{\mathcal{S}}(\mathcal{S}) \sum_{i=1}^{n} \Big\{ \mathbb{1}\{i\in\mathcal{S}\}\ \mathrm{H}(X_i|X_1,\ldots,X_{i-1}) \Big\} \Bigg\} \\
&= \sum_{i=1}^{n} \Bigg\{ \sum_{\mathcal{S}\subseteq[n]} \Big\{ \mathsf{P}_{\mathcal{S}}(\mathcal{S})\ \mathbb{1}\{i\in\mathcal{S}\} \Big\}\ \mathrm{H}(X_i|X_1,\ldots,X_{i-1}) \Bigg\} \\
&= \sum_{i=1}^{n} \mathrm{Pr}[i\in\mathcal{S}]\ \mathrm{H}(X_i|X_1,\ldots,X_{i-1}) \\
&\geq \theta \sum_{i=1}^{n} \mathrm{H}(X_i|X_1,\ldots,X_{i-1}) \qquad (1.16) \\
&= \theta\, \mathrm{H}(X^n).
\end{aligned}
$$

Shearer's Lemma: Second Version

Remark 2

Similarly to Remark 1, if $\Pr[i \in \mathcal{S}] = \theta$ for all $i \in [n]$, then inequality (1.16) holds with equality. Hence, if the condition in (1.14) is satisfied with equality for all $i \in [n]$, then (1.15) extends to continuous random variables, with entropies replaced by differential entropies.

Application of Proposition 1.2 to Graph Theory

Definition 1.2 (Complete and Simple Graphs)

A complete graph on n vertices, denoted by K_n, is defined by the property that every two vertices in the graph are adjacent (e.g., K_1 is an isolated vertex, K_2 is an edge, and K_3 is a triangle). A simple graph is a graph with no loops or edges with multiplicity.

Unless explicitly mentioned, all graphs are assumed to be undirected.

Proposition 1.3

Let G be a simple graph on n vertices, and let m_k be the number of the K_k induced subgraphs in G. Then, for all $\ell, r \in \mathbb{N}$ with $2 \leq \ell < r \leq n$,

$$m_r \leq \frac{(\ell!\, m_\ell)^{\frac{r}{\ell}}}{r!}. \tag{1.17}$$

Proof of Proposition 1.3

- Label the vertices of G by the elements of the set $[n]$, and let $\ell, r \in \mathbb{N}$ be arbitrary integers such that $2 \leq \ell \leq r \leq n$.
- Let $X_1, \ldots, X_r$ be random variables selected uniformly at random as the vertices of any complete induced subgraph K_r in G.
- Let m_r be the number of the induced subgraphs K_r in G. Then,

$$\mathrm{H}(X_1, \ldots, X_r) = \log(r!\, m_r), \tag{1.18}$$

since the r vertices of each complete induced subgraph K_r in G can be selected in $r!$ ways by arbitrarily permuting their order of selection.

- Let $\mathcal{S}$ be a uniformly selected subset of size ℓ from $[r]$. Then,

$$\Pr[i \in \mathcal{S}] = \frac{\ell}{r}, \quad \forall\, i \in [r]. \tag{1.19}$$

- By Proposition 1.2, it follows from (1.18) and (1.19) that

$$\mathbb{E}_{\mathcal{S}}\big[\mathrm{H}(X_{\mathcal{S}})\big] \geq \frac{\ell\, \log(r!\, m_r)}{r}. \tag{1.20}$$

Proof of Proposition 1.3 (cont.)

- $\implies \exists\, \mathcal{T} \in [r]$, with $|\mathcal{T}| = \ell$, for which

$$\mathrm{H}(X_{\mathcal{T}}) \geq \frac{\ell \, \log(r! \, m_r)}{r}. \tag{1.21}$$

- Furthermore, $X_{\mathcal{T}}$ is supported on a K_ℓ subgraph in G, so

$$\mathrm{H}(X_{\mathcal{T}}) \leq \log(\ell! \, m_\ell), \tag{1.22}$$

since, similarly, the ℓ vertices of each complete induced subgraph K_ℓ in G can be selected in $\ell!$ ways by arbitrarily permuting their order of selection.

- Combining (1.21) and (1.22) gives

$$\log(\ell! \, m_\ell) \geq \frac{\ell \, \log(r! \, m_r)}{r}, \tag{1.23}$$

and rearranging terms in (1.23) gives (1.17).

Example 1.3

Let G be a simple graph on n vertices with e edges and t triangles. Substituting $\ell = 2$ and $r = 3$ into (1.17), where $m_2 = e$ and $m_3 = t$, gives

$$t \leq \tfrac{1}{6}(2e)^{\frac{3}{2}}. \tag{1.24}$$

Inequality (1.24) can also be derived by using spectral graph theory. Let $\mathbf{A}$ be the adjacency matrix of G, with spectrum $\{\lambda_j\}_{j=1}^n$. Then,

$$\sum_{j=1}^{n} \lambda_j^2 = \mathrm{Tr}(\mathbf{A}^2) = 2e, \tag{1.25}$$

$$\sum_{j=1}^{n} \lambda_j^3 = \mathrm{Tr}(\mathbf{A}^3) = 6t, \tag{1.26}$$

$$\Longrightarrow 6t = \sum_{j=1}^{n} \lambda_j^3 \leq \left(\sum_{j=1}^{n} \lambda_j^2\right)^{\frac{3}{2}} = (2e)^{\frac{3}{2}}, \tag{1.27}$$

which coincides with (1.24).

Shearer's Lemma on Hypergraphs

Definition 1.4 (Hypergraph)

A hypergraph H is a pair $H = (V, E)$ where V is a set of vertices, and E is a collection of subsets of V, called hyperedges. The cardinality of the vertex set V is called the order of the hypergraph, and the cardinality of the set E of hyperedges is called the size of the hypergraph.

A graph is a special case of a hypergraph, in which all hyperedges have cardinality 2.

Shearer's Lemma on Hypergraphs

Definition 1.4 (Hypergraph)

A hypergraph H is a pair $H = (V, E)$ where V is a set of vertices, and E is a collection of subsets of V, called hyperedges. The cardinality of the vertex set V is called the order of the hypergraph, and the cardinality of the set E of hyperedges is called the size of the hypergraph.

A graph is a special case of a hypergraph, in which all hyperedges have cardinality 2.

We next formulate Shearer's lemma in the context of hypergraphs.

Proposition 1.4 (Shearer's Lemma on Hypergraphs, Friedgut 2004)

Let $\mathsf{H} = (\mathsf{V}, \mathsf{E})$ be a hypergraph, and let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq \mathsf{V}$ be nonempty subsets of the vertex set such that every vertex $v \in \mathsf{V}$ belongs to at least k of the subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m$. For all $j \in [m]$, let $\mathsf{H}_j = (\mathcal{S}_j, \mathsf{E}_j)$ be the projection hypergraph, whose set of hyperedges is given by

$$\mathsf{E}_j = \big\{e \cap \mathcal{S}_j : e \in \mathsf{E}\big\}. \tag{1.28}$$

Then,

$$|\,\mathsf{E}\,|^k \leq \prod_{j=1}^{m} |\,\mathsf{E}_j\,|. \tag{1.29}$$

Inequality (1.29) establishes an upper bound on the size of a hypergraph, based on the sizes of its projection hypergraphs onto subsets, where each vertex is contained in at least $k \geq 1$ of these subsets.

Proof of Proposition 1.4

Without any loss of generality, let $\mathsf{V} = [n]$, where n is the order of H. Let $e \in \mathsf{E}$ be selected uniformly at random, and define the binary random variables $X_1, \ldots, X_n$ as the indicators of the n vertices in the selected hyperedge e, i.e., $X_i = \mathbb{1}\{i \in e\} \in \{0, 1\}$ for every $i \in [n]$. Then,

$$\mathrm{H}(X_1, \ldots, X_n) = \log |\, \mathsf{E} \,|. \tag{1.30}$$

For $j \in [m]$, the binary random vector $X_{\mathcal{S}_j}$ is one of $|\, \mathsf{E}_j \,|$ possibilities, so

$$\mathrm{H}(X_{\mathcal{S}_j}) \leq \log |\, \mathsf{E}_j \,|, \quad j \in [m]. \tag{1.31}$$

By Shearer's lemma (see Proposition 1.1),

$$k\, \mathrm{H}(X_1, \ldots, X_n) \leq \sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}), \tag{1.32}$$

$$\implies k \log |\, \mathsf{E} \,| \leq \sum_{j=1}^{m} \log |\, \mathsf{E}_j \,|. \tag{1.33}$$

Exponentiating both sides of (1.33) gives (1.29).

Proposition 1.5 (Weighted Shearer's Lemma on Hypergraphs, Friedgut 2004)

In the setting of Proposition 1.4, for each $j \in [m]$, let $w_j \colon \mathsf{E}_j \to \mathbb{R}_+$ be a nonnegative real-valued function. Then,

$$\left(\sum_{e \in \mathsf{E}} \prod_{j=1}^{m} w_j(e \cap \mathcal{S}_j) \right)^k \leq \prod_{j=1}^{m} \sum_{e_j \in \mathsf{E}_j} w_j(e_j)^k. \tag{1.34}$$

Proposition 1.5 reduces to Proposition 1.4 by setting all the nonnegative functions $w_j \colon \mathsf{E}_j \to \mathbb{R}_+$, with $j \in [m]$, to 1.

Application of Proposition 1.5 to Matrix Inequalities

Proposition 1.6

Let $\mathbf{A}_1, \ldots, \mathbf{A}_m$ be real matrices such that the product $\mathbf{A}_j \mathbf{A}_{j+1}$ is defined for all $j \in [m-1]$, and the product $\mathbf{A}_1 \ldots \mathbf{A}_m$ is a square matrix. Then,

$$\left| \operatorname{Tr}\left(\prod_{j=1}^{m} \mathbf{A}_j \right) \right| \leq \sqrt{\prod_{j=1}^{m} \operatorname{Tr}\left(\mathbf{A}_j \, \mathbf{A}_j^{\mathrm{T}}\right)}. \tag{1.35}$$

Corollary 1.5

Let $\mathbf{A} = (a_{i,\ell})$, $\mathbf{B} = (b_{\ell,k})$, and $\mathbf{C} = (c_{k,i})$ be real matrices. Then,

$$\left(\sum_{i,\ell,k} a_{i,\ell} \, b_{\ell,k} \, c_{k,i} \right)^2 \leq \sum_{i,\ell} a_{i,\ell}^2 \sum_{\ell,k} b_{\ell,k}^2 \sum_{k,i} c_{k,i}^2. \tag{1.36}$$

Proof of Proposition 1.6

Let $\mathbf{A}_1 = (a_1(i_1, i_2))$, $\mathbf{A}_2 = (a_2(i_2, i_3)), \ldots, \mathbf{A}_m = (a_m(i_m, i_1))$, which is a necessary and sufficient condition for the product $\mathbf{A}_j \mathbf{A}_{j+1}$ to be defined for all $j \in [m-1]$, and for the product $\mathbf{A}_1 \ldots \mathbf{A}_m$ to be a square matrix. Let $\mathcal{I}_j$ be the set of the i_j-values, for all $j \in [m]$.

The left and right-hand sides of (1.35) are, respectively, equal to

$$\left| \mathrm{Tr}\left(\prod_{j=1}^{m} \mathbf{A}_j \right) \right| = \left| \sum_{i_1, \ldots, i_m} a_1(i_1, i_2)\, a_2(i_2, i_3) \ldots a_m(i_m, i_1) \right|,$$

$$\sqrt{\prod_{j=1}^{m} \mathrm{Tr}\big(\mathbf{A}_j\, \mathbf{A}_j^{\mathrm{T}}\big)} = \sqrt{\sum_{i_1, i_2} a_1(i_1, i_2)^2 \sum_{i_2, i_3} a_2(i_2, i_3)^2 \ldots \sum_{i_1, i_m} a_m(i_m, i_1)^2},$$

so, without any loss of generality, it is assumed that all the entries of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are nonnegative. Otherwise, the left-hand side of (1.35) may only decrease by negating some of the entries of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_m$, whereas the right-hand side of (1.35) stays unaffected.

Proof of Proposition 1.6 (cont.)

For proving inequality (1.35), let H be a complete m-partite hypergraph, whose vertex set is given by

$$\mathsf{V} = \bigcup_{j=1}^{m} \mathcal{I}_j, \tag{1.37}$$

and whose set of hyperedges is given by

$$\mathsf{E} = \Big\{ \{i_1, \ldots, i_m\} : (i_1, \ldots, i_m) \in \mathcal{I}_1 \times \ldots \times \mathcal{I}_m \Big\}. \tag{1.38}$$

Define, furthermore, the following m subsets of V:

$$\mathcal{S}_1 = \mathcal{I}_1 \cup \mathcal{I}_2, \quad \mathcal{S}_2 = \mathcal{I}_2 \cup \mathcal{I}_3, \ \ldots \ , \mathcal{S}_m = \mathcal{I}_m \cup \mathcal{I}_1, \tag{1.39}$$

so, every element in V is included in at least two of the m sets $\mathcal{S}_1, \ldots, \mathcal{S}_m$.

Let $e \in \mathsf{E}$, and $e_j = e \cap \mathcal{S}_j \in \mathsf{E}_j$ for all $j \in [m]$.
Assign the weight $w_j(e_j) = a_j(i_j, i_{j+1}) \geq 0$, for each $j \in [m-1]$, and assign the weight $w_m(e_m) = a_m(i_m, i_1) \geq 0$.

Proof of Proposition 1.6 (cont.)

An application of Proposition 1.6, with $k = 2$, gives that the left and right-hand sides of (1.34) are, respectively, equal to

$$\left(\sum_{e \in \mathsf{E}} \prod_{j=1}^{m} w_j(e \cap \mathcal{S}_j) \right)^k = \left(\sum_{i_1, \ldots, i_m} a_1(i_1, i_2)\, a_2(i_2, i_3) \ldots a_m(i_m, i_1) \right)^2$$

$$= \left(\mathrm{Tr} \left(\prod_{j=1}^{m} \mathbf{A}_j \right) \right)^2, \tag{1.40}$$

and,

$$\prod_{j=1}^{m} \sum_{e_j \in \mathsf{E}_j} w_j(e_j)^k = \sum_{i_1, i_2} a_1(i_1, i_2)^2 \ \ldots \ \sum_{i_1, i_m} a_m(i_m, i_1)^2$$

$$= \prod_{j=1}^{m} \mathrm{Tr}\big(\mathbf{A}_j\, \mathbf{A}_j^{\mathrm{T}}\big). \tag{1.41}$$

Substituting (1.40) and (1.41) into (1.34) gives (1.35).

Proposition 1.7 (Shearer's Lemma for the Relative Entropy, Gavinsky *et al.*, 2015)

- Let $X_1, \ldots, X_n$ be discrete random variables.
- Let $U_1, \ldots, U_n$ be independent random variables, where U_i has an equiprobable distribution over a set containing the support of X_i.
- Let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ be subsets such that each element $i \in [n]$ is contained in at most $k \geq 1$ of these subsets.

Then,

$$k \, \mathsf{D}\big(\mathsf{P}_{X^n} \| \mathsf{P}_{U^n}\big) \geq \sum_{j=1}^{m} \mathsf{D}\big(\mathsf{P}_{X_{\mathcal{S}_j}} \| \mathsf{P}_{U_{\mathcal{S}_j}}\big). \tag{1.42}$$

Proof of Proposition 1.7

Without any loss of generality, one can assume that every element $i \in [n]$ is included in exactly k subsets among $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$. This holds since

- By the chain rule for the relative entropy, adding some elements to a set $\mathcal{S}_j$ cannot decrease the relative entropy $\mathsf{D}\big(\mathsf{P}_{X_{\mathcal{S}_j}} \| \mathsf{P}_{U_{\mathcal{S}_j}}\big)$.
  $\Longrightarrow$ The right-hand side of (1.42) cannot be decreased as a result of adding elements to some subsets $\{\mathcal{S}_j\}_{j=1}^m$.
- The left-hand side of (1.42) stays, however, unaffected as a result of these additions of elements.

The following equality holds:

$$\mathsf{D}\big(\mathsf{P}_{X^n} \| \mathsf{P}_{U^n}\big) = \mathrm{H}(U^n) - \mathrm{H}(X^n), \tag{1.43}$$

since $U_1, \ldots, U_n$ are independent random variables, and U_i is equiprobable over a set containing the support of X_i for all $i \in [n]$.

Proof of Proposition 1.7 (Cont.)

- Given that $\{U_i\}_{i=1}^n$ are independent random variables, and by the assumption that $d(i) = k$ for all $i \in [n]$ (see (1.2)), meaning each element in $[n]$ belongs to exactly k subsets among $\mathcal{S}_1, \ldots, \mathcal{S}_m$,

$$\begin{aligned} k\,\mathrm{H}(U^n) &= k \sum_{i=1}^{n} \mathrm{H}(U_i) \\ &= \sum_{i=1}^{n} d(i)\,\mathrm{H}(U_i) \\ &= \sum_{j=1}^{m} \mathrm{H}(U_{\mathcal{S}_j}). \end{aligned} \tag{1.44}$$

- By Shearer's lemma (Proposition 1.1),

$$k\,\mathrm{H}(X^n) \leq \sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}). \tag{1.45}$$

Proof of Proposition 1.7 (Cont.)

Combining (1.43), (1.44), and (1.45) gives

$$\begin{aligned} k\,\mathsf{D}\big(\mathsf{P}_{X^n}\,\|\,\mathsf{P}_{U^n}\big) &= k\,\mathrm{H}(U^n) - k\,\mathrm{H}(X^n) \\ &= \sum_{j=1}^{m} \mathrm{H}(U_{\mathcal{S}_j}) - k\,\mathrm{H}(X^n) \\ &\geq \sum_{j=1}^{m} \mathrm{H}(U_{\mathcal{S}_j}) - \sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}) \\ &= \sum_{j=1}^{m} \big(\mathrm{H}(U_{\mathcal{S}_j}) - \mathrm{H}(X_{\mathcal{S}_j})\big) \\ &= \sum_{j=1}^{m} \mathsf{D}\big(\mathsf{P}_{X_{\mathcal{S}_j}}\,\|\,\mathsf{P}_{U_{\mathcal{S}_j}}\big). \end{aligned}$$

Q.E.D.

Application of Shearer's Lemma to Families of Read-k Functions

Let $m, n \in \mathbb{N}$, and let

- $X_1, \ldots, X_n$ be independent random variables, and $X_i \in \mathcal{X}_i$, $\forall\, i \in [n]$,
- $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$, and $X_{\mathcal{S}_j} \triangleq (X_i)_{i \in \mathcal{S}_j}$ for all $j \in [m]$,
- $\{f_j\}_{j=1}^m$ be a family of functions, where $f_j \colon \prod\limits_{i \in \mathcal{S}_j} \mathcal{X}_i \to \mathbb{R}$, $\forall\, j \in [m]$,
- Y^m be a random vector, where $Y_j \triangleq f_j(X_{\mathcal{S}_j})$ for all $j \in [m]$,
- $p_j \triangleq \mathbb{E}\big[Y_j\big]$, for all $j \in [m]$, exist and be finite.

Consider first the following setup:

- If $\{\mathcal{S}_j\}_{j=1}^m$ are disjoint sets, then $\{Y_j\}_{j=1}^m$ are statistically independent.
- If also $Y^m \in \{0,1\}^m$ ($\implies p_j = \Pr[Y_j = 1]$, $\forall\, j \in [m]$), then

$$\Pr[Y_1 = \ldots = Y_m = 1] = \prod_{j=1}^m p_j. \tag{1.46}$$

Definition 1.6 (A Set of Read-k Functions)

A set of read-k functions is a set of functions where each input variable appears in the arguments of at most k different functions within that set. In other words, in the context of such a set, each variable can only be read or accessed by at most k functions.

A set of functions $\{f_j\}_{j=1}^m$, whose arguments are $x_1, \ldots, x_n$, is read-k if there exist subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ such that f_j depends on the vector $x_{\mathcal{S}_j} \triangleq (x_i)_{i \in \mathcal{S}_j}$, for each $j \in [m]$, and

$$\left|\left\{j \in [m] : \, i \in \mathcal{S}_j\right\}\right| \leq k, \quad \forall \, i \in [n]. \tag{1.47}$$

Definition 1.6 (A Set of Read-k Functions)

A set of read-k functions is a set of functions where each input variable appears in the arguments of at most k different functions within that set. In other words, in the context of such a set, each variable can only be read or accessed by at most k functions.

A set of functions $\{f_j\}_{j=1}^m$, whose arguments are $x_1, \ldots, x_n$, is read-k if there exist subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ such that f_j depends on the vector $x_{\mathcal{S}_j} \triangleq (x_i)_{i \in \mathcal{S}_j}$, for each $j \in [m]$, and

$$\left|\left\{j \in [m] : \, i \in \mathcal{S}_j\right\}\right| \leq k, \quad \forall \, i \in [n]. \tag{1.47}$$

Shearer's lemma (Proposition 1.1) enables the extension of equality (1.46) into an inequality that holds for random variables defined by an arbitrary set of read-k Boolean functions.

Proposition 1.8 (A Probabilistic Result on Read-k Boolean Functions)

- Let $m, n, k \in \mathbb{N}$, with $k \leq m$,
- Let $\{f_j\}_{j=1}^{m}$ be a set of read-k Boolean functions,
- Let X^n be a binary random vector, uniformly distributed on $\{0,1\}^n$,
- Let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$, where every $i \in [n]$ belongs to at most k of the subsets $\{\mathcal{S}_j\}_{j=1}^{m}$,
- Let $\{Y_j\}_{j=1}^{m}$ be defined as $Y_j \triangleq f_j(X_{\mathcal{S}_j})$, $\forall\, j \in [m]$,
- Let $p_j \triangleq \Pr[Y_j = 1]$ for all $j \in [m]$.

Then,

$$\Pr[Y_1 = \ldots = Y_m = 1] \leq \left(\prod_{j=1}^{m} p_j \right)^{\frac{1}{k}}. \tag{1.48}$$

Proposition 1.8 (A Probabilistic Result on Read-k Boolean Functions)

- Let $m, n, k \in \mathbb{N}$, with $k \leq m$,
- Let $\{f_j\}_{j=1}^m$ be a set of read-k Boolean functions,
- Let X^n be a binary random vector, uniformly distributed on $\{0,1\}^n$,
- Let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$, where every $i \in [n]$ belongs to at most k of the subsets $\{\mathcal{S}_j\}_{j=1}^m$,
- Let $\{Y_j\}_{j=1}^m$ be defined as $Y_j \triangleq f_j(X_{\mathcal{S}_j})$, $\forall\, j \in [m]$,
- Let $p_j \triangleq \Pr[Y_j = 1]$ for all $j \in [m]$.

Then,

$$\Pr[Y_1 = \ldots = Y_m = 1] \leq \left(\prod_{j=1}^m p_j \right)^{\frac{1}{k}}, \tag{1.48}$$

with equality in (1.48) if $k = 1$, or if the following two conditions are satisfied under the given setup:
$(X_1 = \ldots = X_n = 1 \iff Y_1 = \ldots = Y_m = 1)$ and $p_1 \ldots p_m = 2^{-nk}$.

Example 1.7

- Let $\mathsf{G} = \mathsf{G}(n, p)$ be a random graph on n vertices, where any two vertices are independently adjacent with probability p.
- Let E_v be an event which depends on the edges that are incident to the vertex $v \in \mathsf{V}(\mathsf{G})$.
- An edge $e \in \mathsf{E}(\mathsf{G})$ can only affect the two events E_{v_1} and E_{v_2}, where v_1 and v_2 are the endpoints of e.
- By construction, the edges in $\mathsf{E}(\mathsf{G})$ are statistically independent.
- Let $Y_v \triangleq \mathbb{1}\{\mathrm{E}_v\}$ for all $v \in \mathsf{V}(\mathsf{G})$. Then, every edge $e \in \mathsf{E}(\mathsf{G})$ affects at most $k = 2$ values among the binary random variables $\{Y_v\}_{v \in \mathsf{V}(\mathsf{G})}$.
- By Proposition 1.8, it follows that

$$\Pr\left(\bigcap_{v \in \mathsf{V}(\mathsf{G})} \mathrm{E}_v\right) \leq \sqrt{\prod_{v \in \mathsf{V}(\mathsf{G})} \Pr(\mathrm{E}_v)}. \tag{1.49}$$

Binary Relative Entropy

Let $p, q \in [0,1]$. The binary relative entropy, $\mathsf{D_b}(p\,\|\,q)$, is defined to be the relative entropy from the Bernoulli distribution $(p, 1-p)$ to the Bernoulli distribution $(q, 1-q)$, i.e.,

$$\mathsf{D_b}(p\,\|\,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \tag{1.50}$$

with the convention that $0 \log 0 \triangleq \lim_{x \to 0^+} x \log x = 0$. In particular,

$$\mathsf{D_b}(p\,\|\,\tfrac{1}{2}) = 1 - \mathsf{H_b}(p), \quad \forall\, p \in [0,1]. \tag{1.51}$$

Binary Relative Entropy

Let $p, q \in [0,1]$. The binary relative entropy, $\mathsf{D_b}(p\,\|\,q)$, is defined to be the relative entropy from the Bernoulli distribution $(p, 1-p)$ to the Bernoulli distribution $(q, 1-q)$, i.e.,

$$\mathsf{D_b}(p\,\|\,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \tag{1.50}$$

with the convention that $0 \log 0 \triangleq \lim_{x \to 0^+} x \log x = 0$. In particular,

$$\mathsf{D_b}(p\,\|\,\tfrac{1}{2}) = 1 - \mathsf{H_b}(p), \quad \forall\, p \in [0,1]. \tag{1.51}$$

An Application of Proposition 1.7: Chernoff-Like Bounds

Building on Proposition 1.7, the following result establishes Chernoff-like bounds for the one-sided tail probabilities of sums of dependent random variables.

Proposition 1.9 (Chernoff-Like Bounds for Sums of Read-k Functions, Gavinsky *et al.*, 2015)

- Let $m, n, k \in \mathbb{N}$, with $k \leq m$,
- Let $X_1, \ldots, X_n$ be independent discrete random variables,
- Let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$, containing every $i \in [n]$ in at most k subsets,
- Let $\{f_j\}_{j=1}^m$ be a set of read-k functions with range in $[0, 1]$,
- Let $p_j \triangleq \mathbb{E}[Y_j]$, where $Y_j \triangleq f_j(X_{\mathcal{S}_j})$ for all $j \in [m]$,
- Let $p \triangleq \frac{1}{m} \sum_{j=1}^m p_j$.

Then, the following Chernoff-like bounds hold for every $\varepsilon > 0$:

$$\Pr\left[\sum_{j=1}^m Y_j \geq m\,(p+\varepsilon)\right] \leq \exp\left(-\frac{m}{k} \cdot \mathsf{D}_{\mathsf{b}}\big((p+\varepsilon)\,\|\,p\big)\right) \leq \mathrm{e}^{-\frac{2m\varepsilon^2}{k}}, \tag{1.52}$$

$$\Pr\left[\sum_{j=1}^m Y_j \leq m\,(p-\varepsilon)\right] \leq \exp\left(-\frac{m}{k} \cdot \mathsf{D}_{\mathsf{b}}\big((p-\varepsilon)\,\|\,p\big)\right) \leq \mathrm{e}^{-\frac{2m\varepsilon^2}{k}}. \tag{1.53}$$

Application: On the Number of Length-r Cycles in a Random Graph

Let $\mathsf{G} = \mathsf{G}(n,p)$ be a random graph on n vertices, where each pair of vertices is adjacent with probability p, independently of every other pair. Let $N_r(\mathsf{G})$ be the number of length-r cycles in a randomly selected graph $\mathsf{G} = \mathsf{G}(n,p)$. By Proposition 1.9 it can be shown that, for all $\varepsilon > 0$,

$$\Pr\Big[\big|N_r(\mathsf{G}) - \mathbb{E}[N_r(\mathsf{G})]\big| \geq \varepsilon\,\mathbb{E}[N_r(\mathsf{G})]\Big] \leq 2\,\mathrm{e}^{-\frac{n(n-1)}{r}\cdot\varepsilon^2 p^{2r}}, \tag{1.54}$$

where

$$\mathbb{E}[N_r(\mathsf{G})] = \tfrac{1}{2}(r-1)!\binom{n}{r}p^r. \tag{1.55}$$

Example 1.8 (On the Number of Triangles in a Random Graph)

Since C_3 is a triangle, substituting $r = 3$ into (1.54) and (1.55) specializes to a result in the paper by Gavinsky *et al.* (*Random Structures and Algorithms*, 2015).

Generalizations of Shearer Inequalities

A Generalized Version of Shearer's Lemma

We next provide a generalized version of Shearer's Lemma. To that end, let Ω be a finite and non-empty set, and let $f\colon 2^{\Omega} \to \mathbb{R}$ be a real-valued set function (i.e., f is defined for all subsets of Ω).

Definition 2.1 (Sub/Supermodular function)

The set function $f\colon 2^{\Omega} \to \mathbb{R}$ is submodular if

$$f(\mathcal{T}) + f(\mathcal{S}) \geq f(\mathcal{T} \cup \mathcal{S}) + f(\mathcal{T} \cap \mathcal{S}), \qquad \forall\, \mathcal{S}, \mathcal{T} \subseteq \Omega \tag{2.1}$$

Likewise, f is supermodular if $-f$ is submodular.

Equivalent Condition for Submodularity

An identical characterization of submodularity is the diminishing return property, which is stated as follows.

Proposition 2.1

A set function $f : 2^{\Omega} \to \mathbb{R}$ is submodular if and only if whenever

$$\mathcal{S} \subset \mathcal{T} \subset \Omega, \;\; \omega \in \mathcal{T}^{\mathrm{c}} \implies f(\mathcal{S} \cup \{\omega\}) - f(\mathcal{S}) \geq f(\mathcal{T} \cup \{\omega\}) - f(\mathcal{T}). \quad (2.2)$$

Equivalent Condition for Submodularity

An identical characterization of submodularity is the diminishing return property, which is stated as follows.

Proposition 2.1

A set function $f : 2^{\Omega} \to \mathbb{R}$ is submodular if and only if whenever

$$\mathcal{S} \subset \mathcal{T} \subset \Omega, \;\; \omega \in \mathcal{T}^{\mathrm{c}} \implies f(\mathcal{S} \cup \{\omega\}) - f(\mathcal{S}) \geq f(\mathcal{T} \cup \{\omega\}) - f(\mathcal{T}). \quad (2.2)$$

The equivalent condition for the submodularity of f in (2.2) means that the larger is the set, the smaller is the increase in f when a new element is added.

Definition 2.2 (Monotonic set function)

The set function $f\colon 2^{\Omega} \to \mathbb{R}$ is *monotonically increasing* if

$$\mathcal{S} \subseteq \mathcal{T} \subseteq \Omega \implies f(\mathcal{S}) \leq f(\mathcal{T}). \tag{2.3}$$

Likewise, f is *monotonically decreasing* if $-f$ is monotonically increasing.

Definition 2.2 (Monotonic set function)

The set function $f\colon 2^{\Omega} \to \mathbb{R}$ is *monotonically increasing* if

$$\mathcal{S} \subseteq \mathcal{T} \subseteq \Omega \implies f(\mathcal{S}) \leq f(\mathcal{T}). \tag{2.3}$$

Likewise, f is *monotonically decreasing* if $-f$ is monotonically increasing.

Definition 2.3 (Polymatroid, ground set and rank function)

Let $f\colon 2^{\Omega} \to \mathbb{R}$ be submodular and monotonically increasing set function with $f(\varnothing) = 0$. The pair (Ω, f) is called a polymatroid, Ω is called a ground set, and f is called a rank function.

Proposition 2.2 (Information-Theoretic Set Functions)

Let Ω be a finite and non-empty set, and let $\{X_\omega\}_{\omega\in\Omega}$ be a collection of discrete random variables. Then, the following holds:

1. The set function $f\colon 2^\Omega \to \mathbb{R}$, given by

$$f(\mathcal{T}) \triangleq \mathrm{H}(X_\mathcal{T}), \quad \mathcal{T} \subseteq \Omega, \tag{2.4}$$

   is a rank function.

2. The set function $f\colon 2^\Omega \to \mathbb{R}$, given by

$$f(\mathcal{T}) \triangleq \mathrm{H}(X_\mathcal{T}|X_{\mathcal{T}^\mathrm{c}}), \quad \mathcal{T} \subseteq \Omega, \tag{2.5}$$

   is supermodular, monotonically increasing, and $f(\varnothing) = 0$.

Proposition 2.2 (cont.)

3. The set function $f\colon 2^{\Omega} \to \mathbb{R}$, given by

$$f(\mathcal{T}) \triangleq \mathrm{I}(X_{\mathcal{T}}; X_{\mathcal{T}^{\mathrm{c}}}), \quad \mathcal{T} \subseteq \Omega, \tag{2.6}$$

is submodular, $f(\varnothing) = 0$, but f is not a rank function. The latter holds since the equality $f(\mathcal{T}) = f(\mathcal{T}^{\mathrm{c}})$, for all $\mathcal{T} \subseteq \Omega$, implies that f is not a monotonic function.

4. Let $\mathcal{U}, \mathcal{V} \subseteq \Omega$ be disjoint subsets, and let the entries of the random vector $X_{\mathcal{V}}$ be conditionally independent given $X_{\mathcal{U}}$. Then, the set function $f\colon 2^{\mathcal{V}} \to \mathbb{R}$ given by

$$f(\mathcal{T}) \triangleq \mathrm{I}(X_{\mathcal{U}}; X_{\mathcal{T}}), \quad \mathcal{T} \subseteq \mathcal{V}, \tag{2.7}$$

is a rank function.

Proposition 2.2 (cont.)

5. Let $X_\Omega = \{X_\omega\}_{\omega\in\Omega}$ be independent random variables, and let the set function $f\colon 2^\Omega \to \mathbb{R}$ be given by

$$f(\mathcal{T}) \triangleq \mathrm{H}\left(\sum_{\omega\in\mathcal{T}} X_\omega\right), \quad \mathcal{T}\subseteq\Omega. \tag{2.8}$$

Then, f is a rank function.

Proof.

We prove Item (a), in regard to the entropy as a set function $f\colon 2^{\Omega} \to \mathbb{R}$, given in (2.4). It is clear that $f(\varnothing) = 0$, and also f is monotonically increasing. The submodularity of f is next verified. Let $\mathcal{S} \subset \mathcal{T} \subset \Omega$ and $\omega \in \mathcal{T}^{\mathrm{c}} \triangleq \Omega \setminus \mathcal{T}$. Then,

$$\begin{aligned} f(\mathcal{T} \cup \{\omega\}) - f(\mathcal{T}) &= \mathrm{H}(X_{\mathcal{T} \cup \{\omega\}}) - \mathrm{H}(X_{\mathcal{T}}) \\ &= \mathrm{H}(X_{\omega} | X_{\mathcal{T}}) \\ &= \mathrm{H}(X_{\omega} | X_{\mathcal{S}}, X_{\mathcal{T} \setminus \mathcal{S}}) \\ &\leq \mathrm{H}(X_{\omega} | X_{\mathcal{S}}) \qquad (2.9) \\ &= \mathrm{H}(X_{\mathcal{S} \cup \{\omega\}}) - \mathrm{H}(X_{\mathcal{S}}) \\ &= f(\mathcal{S} \cup \{\omega\}) - f(\mathcal{S}), \end{aligned}$$

which asserts the submodularity of $f \implies f$ is a rank function.
The proofs for the set functions in (2.5)–(2.8) are left as exercises. □

Proposition 2.3 (Generalized Version of Shearer's Lemma: I.S, 2022)

Let Ω be a finite set, let $\{\mathcal{S}_j\}_{j=1}^{M}$ be a finite collection of subsets of Ω (with $M \in \mathbb{N}$), and let $f\colon 2^{\Omega} \to \mathbb{R}$ be a set function.

1. If f is non-negative and submodular, and every element in Ω is included in at least $d \geq 1$ of the subsets $\{\mathcal{S}_j\}_{j=1}^{M}$, then

$$\sum_{j=1}^{M} f(\mathcal{S}_j) \geq d\, f(\Omega). \tag{2.10}$$

2. If f is a rank function, $\mathcal{A} \subset \Omega$, and every element in $\mathcal{A}$ is included in at least $d \geq 1$ of the subsets $\{\mathcal{S}_j\}_{j=1}^{M}$, then

$$\sum_{j=1}^{M} f(\mathcal{S}_j) \geq d\, f(\mathcal{A}). \tag{2.11}$$

Proposition 2.3 $\Longrightarrow$ Sherarer's Lemma in Proposition 1.1

Item 1 of Proposition 2.3 yields Sherarer's Lemma in Proposition 1.1 since the set function given in (2.4) is submodular, and it is also nonnegative for discrete random variables (in light of Item 1 of Proposition 2.2).

Proposition 2.4 (Madiman and Tetali, 2010)

Let $X_1, \ldots, X_n$ be discrete random variables, and let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ be arbitrary subsets of $[n]$, with $m, n \in \mathbb{N}$. For every $i \in [n]$, let

$$d(i) = \big|\{j \in [m] : i \in \mathcal{S}_j\}\big|, \tag{2.12}$$

and, for an arbitrary subset $\mathcal{A} \subseteq [n]$, let

$$d_-(\mathcal{A}) = \min_{i \in \mathcal{A}} d(i), \tag{2.13a}$$

$$d_+(\mathcal{A}) = \max_{i \in \mathcal{A}} d(i). \tag{2.13b}$$

If $d(i) > 0$ for all $i \in [n]$ (i.e., each element in $[n]$ belongs to at least one of the subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m$), then

$$\sum_{j=1}^{m} \frac{\mathrm{H}(X_{\mathcal{S}_j} | X_{\mathcal{S}_j^{\mathrm{c}}})}{d_+(\mathcal{S}_j)} \leq \mathrm{H}(X^n) \leq \sum_{j=1}^{m} \frac{\mathrm{H}(X_{\mathcal{S}_j})}{d_-(\mathcal{S}_j)}. \tag{2.14}$$

By the proof of Proposition 2.4, the two inequalities extend to continuous random variables under the following condition.

Corollary 2.4

Let $X_1, \ldots, X_n$ be discrete random variables, and let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ be arbitrary subsets of $[n]$, with $m, n \in \mathbb{N}$. If every element $i \in [n]$ belongs to exactly a fixed number $k > 0$ of these subsets, then

$$\sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j} | X_{\mathcal{S}_j^{\mathrm{c}}}) \leq k\, \mathrm{H}(X^n) \leq \sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}). \tag{2.15}$$

Furthermore, if $X_1, \ldots, X_n$ are continuous random variables then, under the above assumption on k,

$$\sum_{j=1}^{m} \mathsf{h}(X_{\mathcal{S}_j} | X_{\mathcal{S}_j^{\mathrm{c}}}) \leq k\, \mathsf{h}(X^n) \leq \sum_{j=1}^{m} \mathsf{h}(X_{\mathcal{S}_j}). \tag{2.16}$$

Definition 2.5 (Erasure Entropy)

The erasure entropy of a discrete random vector X^n is given by

$$\begin{aligned} \mathrm{H}^{-}(X^n) &= \sum_{i=1}^{n} \mathrm{H}(X_i | X_{[n]\setminus\{i\}}) \\ &= \sum_{i=1}^{n} \mathrm{H}(X_i | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n). \end{aligned} \tag{2.17}$$

For a continuous random vector, the conditional entropy on the right-hand side of (2.17) is replaced by the conditional differential entropy.

Reference

S. Verdú and T. Weissman, "The information lost in erasures," *IEEE Trans. on Information Theory*, vol. 54, no. 11, pp. 5030–5058, November 2008.

Proposition 2.5 (Verdú and Weissman, 2008)

The difference between the Shannon and erasure entropies of a random vector X^n is given by

$$\mathrm{H}(X^n) - \mathrm{H}^-(X^n) = \sum_{i=1}^{n} \mathrm{I}(X_i; X_{i+1}^n | X^{i-1}) \geq 0, \tag{2.18}$$

where

$$X_i^j \triangleq (X_i, X_{i+1}, \ldots, X_j), \quad 1 \leq i \leq j \leq n, \tag{2.19}$$

with the convention that it is void if $i > j$.

Proposition 2.5 (Verdú and Weissman, 2008)

The difference between the Shannon and erasure entropies of a random vector X^n is given by

$$\mathrm{H}(X^n) - \mathrm{H}^-(X^n) = \sum_{i=1}^{n} \mathrm{I}(X_i; X_{i+1}^n | X^{i-1}) \geq 0, \tag{2.18}$$

where

$$X_i^j \triangleq (X_i, X_{i+1}, \ldots, X_j), \quad 1 \leq i \leq j \leq n, \tag{2.19}$$

with the convention that it is void if $i > j$.

By Proposition 2.5, the erasure entropy is always less than or equal to the Shannon entropy, and the difference between these entropies is equal to the total conditional mutual information between the present and future given the past.

Proposition 2.5 (Verdú and Weissman, 2008)

The difference between the Shannon and erasure entropies of a random vector X^n is given by

$$\mathrm{H}(X^n) - \mathrm{H}^-(X^n) = \sum_{i=1}^{n} \mathrm{I}(X_i; X_{i+1}^n | X^{i-1}) \geq 0, \tag{2.18}$$

where

$$X_i^j \triangleq (X_i, X_{i+1}, \ldots, X_j), \quad 1 \leq i \leq j \leq n, \tag{2.19}$$

with the convention that it is void if $i > j$.

Proof.

By the chain rule of the Shannon entropy and by Definition 2.5,

$$\begin{aligned}\mathrm{H}(X^n) - \mathrm{H}^-(X^n) &= \sum_{i=1}^{n} \mathrm{H}(X_i | X^{i-1}) - \sum_{i=1}^{n} \mathrm{H}(X_i | X^{i-1}, X_{i+1}^n) \\ &= \sum_{i=1}^{n} \mathrm{I}(X_i; X_{i+1}^n | X^{i-1}) \geq 0.\end{aligned}$$

Example 2.6

Applying Corollary 2.4 to the singletons $\mathcal{S}_i = \{i\}$ for all $i \in [n]$ (so $m = n$) gives that, for discrete random variables $\{X_i\}_{i=1}^n$,

$$\sum_{i=1}^{n} \mathrm{H}(X_i | X_{[n]\setminus\{i\}}) \leq \mathrm{H}(X^n) \leq \sum_{i=1}^{n} \mathrm{H}(X_i), \tag{2.20}$$

and similarly, for continuous random variables,

$$\sum_{i=1}^{n} \mathsf{h}(X_i | X_{[n]\setminus\{i\}}) \leq \mathsf{h}(X^n) \leq \sum_{i=1}^{n} \mathsf{h}(X_i). \tag{2.21}$$

- The rightmost inequalities in (2.20) and (2.21) show the subadditivity of the Shannon entropy.
- The leftmost inequalities in (2.20) and (2.21) represent the fact that the erasure entropy cannot be larger than the Shannon entropy (see Proposition 2.5).

Special Cases (cont.)

Example 2.7

Applying Corollary 2.4 to the collection of all the n subsets of $[n]$ whose size is $n-1$, we get that every element in $[n]$ belongs to exactly $n-1$ of these subsets, so for discrete random variables,

$$\frac{1}{n-1}\sum_{i=1}^{n}\mathrm{H}(\widetilde{X}^{(i)}|X_i) \leq \mathrm{H}(X^n) \leq \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{H}(\widetilde{X}^{(i)}), \tag{2.22}$$

and, for continuous random variables,

$$\frac{1}{n-1}\sum_{i=1}^{n}\mathsf{h}(\widetilde{X}^{(i)}|X_i) \leq \mathsf{h}(X^n) \leq \frac{1}{n-1}\sum_{i=1}^{n}\mathsf{h}(\widetilde{X}^{(i)}), \tag{2.23}$$

where $\widetilde{X}^{(i)} \triangleq (X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) = X_{[n]\setminus\{i\}} = (X^{i-1},X_{i+1}^n)$ for $i \in [n]$. The rightmost inequality is Han's inequality.

Special Cases (cont.)

Example 2.8

Let $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ be arbitrary sets such that every element $i \in [n]$ belongs to at least k of these subsets of $[n]$. Then, $d_-(\mathcal{S}_j) \geq k$ for all $j \in [m]$. By the rightmost inequality in Proposition 2.4, it follows that for every discrete random vector X^n,

$$k \, \mathrm{H}(X^n) \leq \sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}), \tag{2.24}$$

which is Shearer's lemma (Proposition 1.1).