Combinatorial Applications of Shearer Inequalities in Graph Theory and Boolean Functions

Igal Sason, Technion - Israel Institute of Technology

November 19, 2024

Workshop on
Information Theory, Boolean Functions, and Lattice Problems
Hausdorff Research Institute for Mathematics (HIM), Bonn, Germany

Shearer's Lemma

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

Proposition 1.1 (Shearer's Lemma)

Let

- $n, m, k \in \mathbb{N}$,
- $X_1, \ldots, X_n$ be **discrete** random variables,
- $[n] \triangleq \{1, \ldots, n\}$,
- $\mathcal{S}_1, \ldots, \mathcal{S}_m \subseteq [n]$ be subsets such that each element $i \in [n]$ belongs to **at least** $k \geq 1$ of these subsets.
- $X^n \triangleq (X_1, \ldots, X_n)$, and $X_{\mathcal{S}_j} \triangleq (X_i)_{i \in \mathcal{S}_j}$ for all $j \in [m]$.

Then,

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

Special case: Subadditivity of the Shannon entropy

Let $n = m$ with $n \in \mathbb{N}$, and $\mathcal{S}_i = \{i\}$ (singletons) for all $i \in [n]$ $\Rightarrow$ every element $i \in [n]$ belongs to a single set among $\mathcal{S}_1, \ldots, \mathcal{S}_n$ (i.e., $k = 1$). By Shearer's Lemma, it follows that

$$\mathrm{H}(X^n) \leq \sum_{j=1}^{n} \mathrm{H}(X_j),$$

which is the subadditivity property of the Shannon entropy for discrete random variables.

Special case: Subadditivity of the Shannon entropy

Let $n = m$ with $n \in \mathbb{N}$, and $\mathcal{S}_i = \{i\}$ (singletons) for all $i \in [n]$
$\Rightarrow$ every element $i \in [n]$ belongs to a single set among $\mathcal{S}_1, \ldots, \mathcal{S}_n$ (i.e., $k = 1$). By Shearer's Lemma, it follows that

$$\mathrm{H}(X^n) \leq \sum_{j=1}^{n} \mathrm{H}(X_j),$$

which is the subadditivity property of the Shannon entropy for discrete random variables.

We will see shortly that, if every element $i \in [n]$ belongs to **exactly** k of the subsets $\mathcal{S}_j$ $(j \in [m])$, then Shearer's lemma also applies to continuous random variables $X_1, \ldots, X_n$, with entropy replaced by the differential entropy. Hence, Shearer's lemma yields the subadditivity property of the Shannon entropy for discrete and continuous random variables.

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

Proof of Shearer's Lemma (Cont.)

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

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 ?

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

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 $\mathsf{P}_{\mathcal{S}}$. If there exists $\theta > 0$ such that

$$\Pr[i \in \mathcal{S}] \geq \theta, \quad \forall\, i \in [n], \tag{1.9}$$

then,

$$\mathbb{E}_{\mathcal{S}}\big[\mathrm{H}(X_{\mathcal{S}})\big] \geq \theta\, \mathrm{H}(X^n). \tag{1.10}$$

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.)

$$
\begin{aligned}
\Longrightarrow \mathbb{E}_{\mathcal{S}}\big[\mathrm{H}(X_{\mathcal{S}})\big] &= \sum_{\mathcal{S} \subseteq [n]} \mathsf{P}_{\mathcal{S}}(\mathcal{S})\, \mathrm{H}(X_{\mathcal{S}}) \\
&\geq \sum_{\mathcal{S} \subseteq [n]} \left\{ \mathsf{P}_{\mathcal{S}}(\mathcal{S}) \sum_{i=1}^{n} \Big\{ \mathbb{1}\{i \in \mathcal{S}\}\, \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \Big\} \right\} \\
&= \sum_{i=1}^{n} \left\{ \sum_{\mathcal{S} \subseteq [n]} \Big\{ \mathsf{P}_{\mathcal{S}}(\mathcal{S})\, \mathbb{1}\{i \in \mathcal{S}\} \Big\}\, \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \right\} \\
&= \sum_{i=1}^{n} \Pr[i \in \mathcal{S}]\, \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \\
&\geq \theta \sum_{i=1}^{n} \mathrm{H}(X_i | X_1, \ldots, X_{i-1}) \qquad (1.11) \\
&= \theta\, \mathrm{H}(X^n).
\end{aligned}
$$

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 (1.11) holds with equality. Hence, if the condition in (1.9) is satisfied with equality for all $i \in [n]$, then (1.10) 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 a graph where every two vertices 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 self loops or parallel edges.

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_ℓ be the number of the K_ℓ induced subgraphs in G. Then, for all $\ell, r \in \mathbb{N}$ with $2 \leq \ell < r \leq n$,

$$m_r \leq \frac{(\ell!\, m_\ell)^{\frac{r}{\ell}}}{r!}. \tag{1.12}$$

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 $2 \leq \ell < r \leq n$.
- 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,

$$\mathrm{H}(X_1, \ldots, X_r) = \log(r!\, m_r), \tag{1.13}$$

since the r vertices of each complete induced subgraph K_r in G can be selected in $r!$ ways by permuting their order of selection.
- Let $\mathcal{S}$ be a uniformly selected subset of size ℓ from $[r]$. Then,

$$\Pr[i \in \mathcal{S}] = \frac{\ell}{r}, \quad \forall\, i \in [r]. \tag{1.14}$$

- By Proposition 1.2, it follows from (1.13) and (1.14) that

$$\mathbb{E}_{\mathcal{S}}\big[\mathrm{H}(X_{\mathcal{S}})\big] \geq \frac{\ell\, \log(r!\, m_r)}{r}. \tag{1.15}$$

Proof of Proposition 1.3 (cont.)

- $\implies \exists\, \mathcal{T} \in [r]$, with $|\mathcal{T}| = \ell$, for which
$$\mathrm{H}(X_{\mathcal{T}}) \geq \frac{\ell \, \log(r! \, m_r)}{r}. \tag{1.16}$$
- Furthermore, $X_{\mathcal{T}}$ is supported on a K_ℓ subgraph in G, so
$$\mathrm{H}(X_{\mathcal{T}}) \leq \log(\ell! \, m_\ell), \tag{1.17}$$
since, similarly, the ℓ vertices of each complete induced subgraph K_ℓ in G can be selected in $\ell!$ ways by permuting their order of selection.
- Combining (1.16) and (1.17) gives
$$\log(\ell! \, m_\ell) \geq \frac{\ell \, \log(r! \, m_r)}{r}, \tag{1.18}$$
and rearranging terms in (1.18) gives (1.12).

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 (1.12), where $m_2 = e$ and $m_3 = t$, gives

$$t \leq \tfrac{1}{6}(2e)^{\frac{3}{2}}. \tag{1.19}$$

Inequality (1.19) 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,

$$\sum_{j=1}^{n} \lambda_j^2 = \mathrm{Tr}(\mathbf{A}^2) = 2e, \tag{1.20}$$

$$\sum_{j=1}^{n} \lambda_j^3 = \mathrm{Tr}(\mathbf{A}^3) = 6t, \tag{1.21}$$

$$\Longrightarrow 6t = \sum_{j=1}^{n} \lambda_j^3 \leq \left(\sum_{j=1}^{n} \lambda_j^2 \right)^{\frac{3}{2}} \leq (2e)^{\frac{3}{2}}, \tag{1.22}$$

which coincides with (1.19).

Definition 1.4 (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

$$\left|\left\{j \in [m] : i \in \mathcal{S}_j\right\}\right| \leq k, \quad \forall\, i \in [n]. \tag{1.23}$$

Proposition 1.4 (A Probabilistic Result on Read-k Boolean Functions)

- Let $m, n, k \in \mathbb{N}$, with $k \leq m$.
- Let $\mathcal{S}_1, \dots, \mathcal{S}_m \subseteq [n]$, where every $i \in [n]$ belongs to at most k of the subsets $\{\mathcal{S}_j\}_{j=1}^m$.
- Let $\{f_j\}_{j=1}^m$ be a set of read-k Boolean functions, where

$$f_j \colon \{0,1\}^{|\mathcal{S}_j|} \to \{0,1\}.$$

- Let X^n be a binary random vector, uniformly distributed on $\{0,1\}^n$.
- 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,

$$\Pr[Y_1 = \ldots = Y_m = 1] \leq \left(\prod_{j=1}^m p_j \right)^{\frac{1}{k}}. \tag{1.24}$$

On Proposition 1.4

If $\{\mathcal{S}_j\}_{j=1}^m$ are disjoint sets, then $\{Y_j\}_{j=1}^m$ are statistically independent, so

$$\Pr[Y_1 = \ldots = Y_m = 1] = \prod_{j=1}^m p_j, \tag{1.25}$$

and inequality (1.24) holds with equality in that case.

On Proposition 1.4

If $\{\mathcal{S}_j\}_{j=1}^m$ are disjoint sets, then $\{Y_j\}_{j=1}^m$ are statistically independent, so

$$\Pr[Y_1 = \ldots = Y_m = 1] = \prod_{j=1}^{m} p_j, \tag{1.25}$$

and inequality (1.24) holds with equality in that case.

Proposition 1.4 extends equality (1.25) to an inequality that holds for random variables defined by an arbitrary set of read-k Boolean functions.

On Proposition 1.4

If $\{\mathcal{S}_j\}_{j=1}^m$ are disjoint sets, then $\{Y_j\}_{j=1}^m$ are statistically independent, so

$$\Pr[Y_1 = \ldots = Y_m = 1] = \prod_{j=1}^m p_j, \tag{1.25}$$

and inequality (1.24) holds with equality in that case.

Proposition 1.4 extends equality (1.25) to an inequality that holds for random variables defined by an arbitrary set of read-k Boolean functions.

Corollary 1.5

Under the setup in Proposition 1.4, let $\mathcal{S} \subseteq \{0,1\}^m$. Then,

$$\Pr[Y^m \in \mathcal{S}] \leq \sum_{\underline{c} \in \mathcal{S}} \prod_{j=1}^m \bigl(1 - p_j - c_j(1 - 2p_j)\bigr)^{\frac{1}{k}}. \tag{1.26}$$

Proof of Proposition 1.4

- Let
$$\begin{aligned}
Y_j &\triangleq f_j(X_{\mathcal{S}_j}), && (1.27)\\
y_j &\triangleq f_j(x_{\mathcal{S}_j}), \quad \forall\, x^n \in \{0,1\}^n,\ j \in [m], && (1.28)\\
q &\triangleq \Pr[Y_1 = \ldots = Y_m = 1], && (1.29)\\
p_j &\triangleq \Pr[Y_j = 1],\ \forall\, j \in [m], && (1.30)\\
\mathcal{A} &\triangleq \bigl\{x^n \in \{0,1\}^n : y_1 = \ldots = y_m = 1\bigr\}, && (1.31)\\
\mathcal{A}_j &\triangleq \bigl\{x_{\mathcal{S}_j} \in \{0,1\}^{|\mathcal{S}_j|} : y_j = 1\bigr\},\ \forall\, j \in [m]. && (1.32)
\end{aligned}$$

- X^n is uniformly distributed on $\{0,1\}^n$, so from (1.27)–(1.32),
$$\begin{aligned}
|\mathcal{A}| &= 2^n\, q, && (1.33)\\
|\mathcal{A}_j| &= 2^{|\mathcal{S}_j|}\, p_j,\ \forall\, j \in [m]. && (1.34)
\end{aligned}$$

- Let Z^n be uniformly distributed on $\mathcal{A}$. By (1.1), (1.23), and (1.33)
$$\begin{aligned}
k\, \mathrm{H}(Z^n) &\leq \sum_{j=1}^{m} \mathrm{H}(Z_{\mathcal{S}_j}), && (1.35)\\
\mathrm{H}(Z^n) &= \log|\mathcal{A}| = n + \log q. && (1.36)
\end{aligned}$$

Proof of Proposition 1.4 (cont.)

- Since by assumption $Z^n \in \mathcal{A}$, it follows from (1.31) and (1.32) that $Z_{\mathcal{S}_j} \in \mathcal{A}_j$, $\forall\, j \in [m]$. By (1.34),

$$\mathrm{H}(Z_{\mathcal{S}_j}) \leq \log |\mathcal{A}_j| = |\mathcal{S}_j| + \log p_j. \tag{1.37}$$

$$\begin{aligned}
\Longrightarrow \; k\,(n + \log q) &= k\; \mathrm{H}(Z^n) \quad \text{(by (1.36))} \\
&\leq \sum_{j=1}^{m} \mathrm{H}(Z_{\mathcal{S}_j}) \quad \text{(by (1.35))} \\
&\leq \sum_{j=1}^{m} \bigl(|\mathcal{S}_j| + \log p_j\bigr) \quad \text{(by (1.37))} \\
&= \sum_{j=1}^{m} |\mathcal{S}_j| + \log\left(\prod_{j=1}^{m} p_j\right) \\
&\leq kn + \log\left(\prod_{j=1}^{m} p_j\right) \quad \text{(by (1.23))}.
\end{aligned} \tag{1.38}$$

Rearranging terms in (1.38) gives (1.24), as required.

Example 1.6

- Let $\mathsf{G} = \mathsf{G}(n, \frac{1}{2})$ be a random graph on n vertices, where any two vertices are independently adjacent with probability one-half.
- 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 $e \in \mathsf{E}(\mathsf{G})$ can influence at most $k = 2$ of the binary random variables $\{Y_v\}_{v \in \mathsf{V}(\mathsf{G})}$.
- By Proposition 1.4, it follows that

$$\Pr\left(\bigcap_{v \in \mathsf{V}(\mathsf{G})} \mathrm{E}_v\right) \leq \sqrt{\prod_{v \in \mathsf{V}(\mathsf{G})} \Pr(\mathrm{E}_v)}. \tag{1.39}$$

Proposition 1.5 (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,

$$k\, D\big(P_{X^n} \,\|\, P_{U^n}\big) \geq \sum_{j=1}^{m} D\big(P_{X_{\mathcal{S}_j}} \,\|\, P_{U_{\mathcal{S}_j}}\big). \tag{1.40}$$

Proof Outline of Proposition 1.5

Without any loss of generality, one can assume that every element $i \in [n]$ is included in exactly k subsets among $\mathcal{S}_1, \dots, \mathcal{S}_m \subseteq [n]$. 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 $\mathsf{D}\big(\mathsf{P}_{X_{\mathcal{S}_j}} \,\|\, \mathsf{P}_{U_{\mathcal{S}_j}}\big)$.
- The left-hand side of (1.40) stays, however, unaffected.

The following equality holds:

$$\mathsf{D}\big(\mathsf{P}_{X^n} \,\|\, \mathsf{P}_{U^n}\big) = \mathrm{H}(U^n) - \mathrm{H}(X^n), \tag{1.41}$$

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

$$k\,\mathrm{H}(U^n) = k \sum_{i=1}^{n} \mathrm{H}(U_i) = \sum_{j=1}^{m} \mathrm{H}(U_{\mathcal{S}_j}),$$

$$k\,\mathrm{H}(X^n) \leq \sum_{j=1}^{m} \mathrm{H}(X_{\mathcal{S}_j}) \qquad \text{(by Shearer's lemma (Proposition 1.1)).}$$

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.,

$$\mathsf{D_b}(p \| q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \tag{1.42}$$

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

$$\mathsf{D_b}(p \| \tfrac{1}{2}) = 1 - \mathsf{H_b}(p), \quad \forall\, p \in [0,1]. \tag{1.43}$$

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.,

$$\mathsf{D_b}(p \| q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}, \tag{1.42}$$

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

$$\mathsf{D_b}(p \| \tfrac{1}{2}) = 1 - \mathsf{H_b}(p), \quad \forall\, p \in [0,1]. \tag{1.43}$$

An Application of Proposition 1.5: Chernoff-Like Bounds

Building on Proposition 1.5, the following result establishes Chernoff-like bounds for the one-sided tail probabilities of sums of dependent random variables.

Proposition 1.6 (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 $\varepsilon > 0$:

$$\Pr\left[\sum_{j=1}^m Y_j \geq m\,(p+\varepsilon)\right] \leq \exp\left(-\frac{m}{k} \cdot \mathsf{D_b}\big((p+\varepsilon)\,\|\,p\big)\right) \leq \mathrm{e}^{-\frac{2m\varepsilon^2}{k}}, \tag{1.44}$$

$$\Pr\left[\sum_{j=1}^m Y_j \leq m\,(p-\varepsilon)\right] \leq \exp\left(-\frac{m}{k} \cdot \mathsf{D_b}\big((p-\varepsilon)\,\|\,p\big)\right) \leq \mathrm{e}^{-\frac{2m\varepsilon^2}{k}}. \tag{1.45}$$

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 1.6 it can be shown that, for all $\varepsilon > 0$,

$$\Pr\Big[\big|N_r(\mathsf{G}) - \mathbb{E}[N_r(\mathsf{G})]\big| \geq \varepsilon\, \mathbb{E}[N_r(\mathsf{G})]\Big] \leq 2\, \mathrm{e}^{-\frac{n(n-1)}{r} \cdot \varepsilon^2 p^{2r}}, \tag{1.46}$$

where

$$\mathbb{E}[N_r(\mathsf{G})] = \tfrac{1}{2}(r-1)! \binom{n}{r} p^r. \tag{1.47}$$

Example 1.7 (On the Number of Triangles in a Random Graph)

Since C_3 is a triangle, substituting $r = 3$ into (1.46) and (1.47) specializes to a result in the paper by Gavinsky *et al.* (*Random Structures and Algorithms*, 2015).

Some Generalizations of Shearer's Lemma

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

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

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 $f\colon 2^{\Omega} \to \mathbb{R}$ 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
$$f(\mathcal{T}) \triangleq \mathrm{H}(X_{\mathcal{T}} | X_{\mathcal{T}^c}), \quad \mathcal{T} \subseteq \Omega, \tag{2.5}$$
is supermodular, monotonically increasing, and $f(\varnothing) = 0$.

Proposition 2.2 (cont.)

3. The set function $f: 2^{\Omega} \to \mathbb{R}$, 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 $f: 2^{\mathcal{V}} \to \mathbb{R}$ 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 $\omega \in \mathcal{T}^{\mathsf{c}} \triangleq \Omega \setminus \mathcal{T}$. Then,

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

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)

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

I. Sason, "Information inequalities via submodularity, and a problem in extremal graph theory," *Entropy*, vol. 24, paper 597, pp. 1–31, April 2022.

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,

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

□

Special Cases

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,

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

and, for continuous random variables,

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

where $\widetilde{X}^{(i)} \triangleq (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) = X_{[n]\setminus\{i\}} = (X^{i-1}, X_{i+1}^n)$ 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).

Summary

- Shearer's Lemmata (Propositions 1.1 and 1.2).
- Applications:
  - Geometry (Example 1.1).
  - Graph theory (Proposition 1.3).
  - Read-k Boolean functions (Proposition 1.4).
  - Probabilistic results in graph theory (Example 1.6).
  - Version of Shearer's lemma for the relative entropy (Proposition 1.5).
  - Chernoff-like bounds for sums of read-k functions (Proposition 1.6).
  - It can be also applied to hypergraphs, with consequences in linear algebra (Friedgut, 2004). Not covered in this talk.
- Generalizations of Shearer and Han Inequalities:
  - Extension of Shearer inequalities to submodular functions (Proposition 2.3), with links to IT set functions (Proposition 2.2).
  - Extension of Shearer's lemma and Han's inequality, and a counterpart of these inequalities (Proposition 2.4).
  - Connection of the erasure divergence (Proposition 2.5).
  - Special cases (Examples 2.6, 2.7, and 2.8).

I am sorry for not being able to attend the workshop in person. For comments or questions, my e-mail address is *sason@ee.technion.ac.il*.