Read all chapters assigned from IT-I (EE514, Winter 2012).
Read chapter 8 in the book.
Read chapter 9 in the book.
Read chapter 10 in the book (chapter on rate distortion theory).
Read chapter 14 in the book (Kolmogorov complexity)
Read chapter 13, section on Lempel Ziv compression, in the book.
Announcements, Assignments, and Reminders

- Please do use our discussion board (https://catalyst.uw.edu/gopost/board/bilmes/27386/) for all questions, comments, so that all will benefit from them being answered.
On Final Presentations

Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The papers must have been published in the last 10 years (so no old or classic papers).
**On Final Presentations**

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is.
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is.
- This is a real challenge and will require significant work! Many of the papers are complex. To get a good grade, you will need to work very hard to present very complex ideas in an extremely simple yet still precise way.
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is.
- This is a real challenge and will require significant work! Many of the papers are complex. To get a good grade, you will need to work very hard to present very complex ideas in an extremely simple yet still precise way.
- Again, don’t expect this to be easy, you might need to try a few topics until you find one that is suitable.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- **Wed, May 2nd**: Candidate proposed papers submitted. Include short at most 1-page write up: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page write up: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.

- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated write up (noting progress)

- Friday, May 25th: updated short write up on more details of how you will present the ideas in a simple fashion.

- Final presentations: Monday, June 4th in the afternoon late/evening (currently scheduled for 8:30am but that is too early). What to turn in: your slides and a short at most 4 page summary of the papers.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page write up: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.

- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated write up (noting progress)

- Friday, May 18th: short write up on more details of how you will present the ideas in a simple fashion.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page write up: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.
- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated write up (noting progress)
- Friday, May 18th: short write up on more details of how you will present the ideas in a simple fashion.
- Friday, May 25th: updated short write up on more details of how you will present the ideas in a simple fashion.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page write up: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.
- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated write up (noting progress)
- Friday, May 18th: short write up on more details of how you will present the ideas in a simple fashion.
- Friday, May 25th: updated short write up on more details of how you will present the ideas in a simple fashion.
- Final presentations: Monday, June 4th in the afternoon late/evening (currently scheduled for 8:30am but that is too early). What to turn in: your slides and a short at most 4 page summary of the papers.
All problem sets must be turned in via PDF files via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Problem set 1, due tonight at 11:00pm, see the problems listed on pdf page 161 of http://j.ee.washington.edu/~bilmes/classes/ee515a_spring_2012/lecture28_presented.pdf.

- Problem set 2, now due next Monday, May 21st, 11:45pm: Do book problems: 8.1, 8.8, 9.1, 9.2, 9.6, 10.5, 10.6, 13.5, 13.6, 14.3, 14.4, 14.5
Definitions: field, atom

- We have a set of random variables $X_1, X_2, \ldots, X_n$.
- For each random variable we associate a set $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.
- A field $\mathcal{F}_n$ can be generated by sets $\tilde{X}_{1:n}$ by taking unions ($\bigcup$), intersections ($\bigcap$), complementation ($\tilde{X}_i^c$), set subtractions/difference ($\setminus$) on combinations of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.
- An atom of $\mathcal{F}_n$ are sets of the form

$$\text{an atom } = \bigcap_{i=1}^{n} Y_i \text{ where } Y_i = \begin{cases} \tilde{X}_i \\ \tilde{X}_i^c \end{cases} \text{ or }$$

(1)

Ex: $n = 2$, 4 such atoms.

Ex: $n = 3$, 8 atoms
Definitions: signed measure

- We will be measuring these sets using a signed measure (meaning it might be positive or negative). In particular, a real-valued function $\mu$ defined on $\mathcal{F}_n$ is called a signed measure if it is set-additive, i.e., for disjoint $A$ and $B$, we have

$$\mu(A \cup B) = \mu(A) + \mu(B)$$

(2)

- for a signed measure, we must have $\mu(\emptyset) = 0$ since

$$\mu(A) = \mu(A + \emptyset) = \mu(A) + \mu(\emptyset).$$

- Note: For sets $A, B$, set-difference is $A \setminus B \equiv A \cap B^c$.

- Any signed measure on $\mathcal{F}_n$ is defined by its value on the atoms. I.e., any $\tilde{X} \in \mathcal{F}_n$ can be represented as $\tilde{X} = \bigcup_i Y_i$ where $Y_i$ are appropriately chosen atoms.
Random variables

- We said earlier that $\tilde{X}_i$ was associated with a random variable $X_i$.
- Loosely speaking, the set $\tilde{X}_i$ represents the “uncertainty” or “information” contained within $X_i$.
- Ex: Two random variables $X_1, X_2$, define the universal set $\Omega = \tilde{X}_1 \cup \tilde{X}_2$ which is the set of everything (for the current $n = 2$) $\Rightarrow X^c \equiv \Omega \setminus X$ for any $X \in \mathcal{F}_n$.
- One of the atoms is always empty, namely $\tilde{X}_1^c \cap \tilde{X}_2^c = (\tilde{X}_1 \cup \tilde{X}_2)^c = \emptyset$, so this atom has no area and is not shown in the Venn diagram previously seen.
- For these random variables, we have the Shannon information measures: $H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2)$ $I(X_1; X_2)$.
- Lets associate these with $\mu$. 

Signed Measures and Shannon Measures

- We can make the following associations/definitions with signed measure $\mu^*$:

\[
\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2) \tag{3}
\]
\[
\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \tag{4}
\]
\[
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \tag{5}
\]
\[
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0 \tag{6}
\]

- We have instantiated the measures of the four atoms with values (could be arbitrary values, but we chose to use entropic quantities).

- Given these definitions, what would $\mu^*(\tilde{X}_1)$, $\mu^*(\tilde{X}_2)$, and $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$ be?

\[
\mu^*(\tilde{X}_1) = \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \tag{7}
\]
\[
= I(X_1; X_2) + H(X_1|X_2) = H(X_1) \tag{8}
\]
Unions of sets

- What if we define $\mu^*$ only on the unions of sets. I.e., we make the following definitions:

$$\mu^*(\emptyset) = 0$$  \hspace{1cm} (9)

$$\mu^*(\tilde{X}_1) = H(X_1)$$ \hspace{1cm} (10)

$$\mu^*(\tilde{X}_2) = H(X_2)$$ \hspace{1cm} (11)

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2)$$ \hspace{1cm} (12)

- Then from this, we can (using set theory) get the rest of the values, $I(X_1; X_2)$, $H(X_1|X_2)$, $H(X_2|X_1)$.

- E.g., we get:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2) = \mu(\tilde{X}_1) + \mu(\tilde{X}_2) - \mu(\tilde{X}_1 \cup \tilde{X}_2)$$ \hspace{1cm} (13)

$$= H(X_1) + H(X_2) - H(X_1, X_2) = I(X_1; X_2)$$ \hspace{1cm} (14)
Recovering Shannon

So we have recovered Shannon’s information measures with the following correspondence:

\[ H/I \leftrightarrow \mu^* \]  \hspace{1cm} (15)

, \leftrightarrow \cup \hspace{1cm} (16)

; \leftrightarrow \cap \hspace{1cm} (17)

| \leftrightarrow \setminus \hspace{1cm} //set minus \hspace{1cm} (18)

Note: with measure notation, no distinction between \( H \) and \( I \), we could call this \( H(X; Y) = I(X; Y) \) distinguished only by a semicolon rather than a comma.

Does this generalize? The hope is that if there is an information theoretic identity, it would occur iff there is a set theory identity.
General case, $n \geq 2$

- $n$ random variables $X_1, \ldots, X_n$ corresponding to sets $\tilde{X}_i$, and with $[n] = \{1, 2, \ldots, n\}$ the index set.
- $\Omega = \bigcup_i \tilde{X}_i$ is the universe, and empty atom again is:

$$A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset \quad (19)$$

- Non-empty atoms are $\mathcal{A} \triangleq \{\text{all atoms}\} \setminus \{A_0\}$ which are those that are not assuredly empty, so $|\mathcal{A}| = 2^n - 1$.
- When values of $\mu(\cdot)$ is given for all $\mathcal{A}$, then this defines $\mu(\cdot)$ on all $\mathcal{F}_n$. Why? Since the atoms $\mathcal{A}$ partition and hence fully cover $\Omega$.
- Notation: for $G \subseteq [n]$, $X_G = (X_i, i \in G)$ for index set $G$.
- Notation: for $G \subseteq [n]$, $\tilde{X}_G = \bigcup_{i \in G} \tilde{X}_i$.
- Definition, non-empty unions (note strictness on left side):

$$\mathcal{B} \triangleq \left\{ \tilde{X}_G : \emptyset \subset G \subseteq [n] \right\} \quad (20)$$
Inclusion/Exclusion

- We have \( n \) sets \( A_i \), for \( i = 1 \ldots n \) such that \( A_i \subseteq \Omega \). What we wish to prove is the form of the exclusion/exclusion formula.

\[
| \bigcap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \ldots
\]

(21)

\[
+ \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots
\]

(22)

\[
+ (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n|
\]

(23)

- Note the pattern: first we over count, then we undercount, and then overcount, etc. until the last term finally fixes things.

- Special case of sieve methods: general mathematical methods to count sizes of sets of integers.
The same exact argument can be used to show exclusion/exclusion formula for \( \mu \), i.e.,

\[
\mu(\cap_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cup A_j)
\]

\[+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cup A_j \cup A_k) + \ldots \quad (24)\]

\[+ (-1)^{n-1} \mu(A_1 \cup A_2 \cup \ldots \cup A_n) \quad (25)\]

A “dual” form has the form:

\[
\mu(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cap A_j)
\]

\[+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cap A_j \cap A_k) + \ldots \quad (27)\]

\[+ (-1)^{n-1} \mu(A_1 \cap A_2 \cap \ldots \cap A_n) \quad (28)\]
Inclusion/Exclusion

Another (easier?, shorter) way of writing these is as:

\[
\mu(\cap_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k+1} \left( \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \mu(A_{i_1} \cup \cdots \cup A_{i_k}) \right)
\]

(30)

and

\[
\mu(\cup_{i=1}^{n} A_i) = \sum_{k=1}^{n} (-1)^{k+1} \left( \sum_{1 \leq i_1 < i_2 < \cdots < i_k \leq n} \mu(A_{i_1} \cap \cdots \cap A_{i_k}) \right)
\]

(31)
What needs to be specified

Theorem 3.1

Signed measure $\mu$ on $\mathcal{F}_n$ is fully specified by $\{\mu(B) : B \in \mathcal{B}\}$ which can be any set of real numbers.

- So before, we defined $\mu$ on all atoms and noted that this defined $\mu$ everywhere else.
What needs to be specified

Theorem 3.1

Signed measure $\mu$ on $\mathcal{F}_n$ is fully specified by $\{\mu(B) : B \in \mathcal{B}\}$ which can be any set of real numbers.

- So before, we defined $\mu$ on all atoms and noted that this defined $\mu$ everywhere else.
- Here, in the above theorem, we are defining $\mu$ only on elements of $\mathcal{B}$ and are again saying that this defines the values of $\mu$ everywhere else.
What needs to be specified

**Theorem 3.1**

Signed measure $\mu$ on $\mathcal{F}_n$ is fully specified by $\{\mu(B) : B \in \mathcal{B}\}$ which can be any set of real numbers.

- So before, we defined $\mu$ on all atoms and noted that this defined $\mu$ everywhere else.
- Here, in the above theorem, we are defining $\mu$ only on elements of $\mathcal{B}$ and are again saying that this defines the values of $\mu$ everywhere else.
- We will see that this allows us to generate all standard mutual-information quantities, and some other less standard ones, using just entropy to fill out $\mu$. 
Proof of Theorem 3.1

Note that $|A| = |B| = 2^n - 1 \triangleq k$
Proof of Theorem 3.1

Note that $|\mathcal{A}| = |\mathcal{B}| = 2^n - 1 \triangleq k$

Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in \mathcal{A}$. $\text{length}(\vec{a}) = k$
Proof of Theorem 3.1

Note that $|\mathcal{A}| = |\mathcal{B}| = 2^n - 1 \triangleq k$

Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in \mathcal{A}$. $\text{length}(\vec{a}) = k$

Define $\vec{b} = [\ldots \mu(B) \ldots]^\top$ for all $B \in \mathcal{B}$. $\text{length}(\vec{b}) = k$
Proof of Theorem 3.1

proof of Theorem 3.1.

- Note that $|A| = |B| = 2^n - 1 \triangleq k$
- Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in A$. $\text{length}(\vec{a}) = k$
- Define $\vec{b} = [\ldots \mu(B) \ldots]^\top$ for all $B \in B$. $\text{length}(\vec{b}) = k$
- For any $B \in B$, we have $B = \bigcup_{\ell \in \mathcal{A}(B)} A_{\ell}$ with $A_{\ell} \in \mathcal{A}$ and where $\mathcal{A}(B)$ are the indices of the atoms that comprise $B$. 

...
Proof of Theorem 3.1

- Note that $|\mathcal{A}| = |\mathcal{B}| = 2^n - 1 \triangleq k$
- Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in \mathcal{A}$. $\text{length}(\vec{a}) = k$
- Define $\vec{b} = [\ldots \mu(B) \ldots]^\top$ for all $B \in \mathcal{B}$. $\text{length}(\vec{b}) = k$
- For any $B \in \mathcal{B}$, we have $B = \bigcup_{\ell \in \mathcal{A}(B)} A_\ell$ with $A_\ell \in \mathcal{A}$ and where $\mathcal{A}(B)$ are the indices of the atoms that comprise $B$.
- Therefore, there exists a unique $k \times k$ matrix $C_n$ such that $\vec{b} = C_n \vec{a}$
Proof of Theorem 3.1

Note that $|A| = |B| = 2^n - 1 \triangleq k$

Define $\vec{a} = [\ldots \mu(A) \ldots]^{\top}$ for all $A \in \mathcal{A}$. $\text{length}(\vec{a}) = k$

Define $\vec{b} = [\ldots \mu(B) \ldots]^{\top}$ for all $B \in \mathcal{B}$. $\text{length}(\vec{b}) = k$

For any $B \in \mathcal{B}$, we have $B = \bigcup_{\ell \in \mathcal{A}(B)} A_\ell$ with $A_\ell \in \mathcal{A}$ and where $\mathcal{A}(B)$ are the indices of the atoms that comprise $B$.

Therefore, there exists a unique $k \times k$ matrix $C_n$ such that $\vec{b} = C_n \vec{a}$

But also, we claim that any $\mu(A)$ for $A \in \mathcal{A}$ can be expressed as a linear combination of $\{\mu(B)\}_{B \in \mathcal{B}(A)}$ for the appropriate $\mathcal{B}(A) \subseteq \mathcal{B}$, and this can be done using inclusion/exclusion.
Proof of Theorem 3.1

Here, the inclusion/exclusion principle takes the form conditioned on (or excluding) $B$:

\[
\mu(\cap_{k=1}^{n} A_k \setminus B) = \sum_{1 \leq i \leq n} \mu(A_i \setminus B) - \sum_{1 \leq i < j \leq n} \mu((A_i \cup A_j) \setminus B) \\
+ \cdots + (-1)^{n+1} \mu((A_1 \cup A_2 \cup \cdots \cup A_n) \setminus B) \tag{32}
\]
Proof of Theorem 3.1

Here, the inclusion/exclusion principle takes the form conditioned on (or excluding) $B$:

$$
\mu(\cap_{k=1}^{n} A_k \setminus B) = \sum_{1 \leq i \leq n} \mu(A_i \setminus B) - \sum_{1 \leq i < j \leq n} \mu((A_i \cup A_j) \setminus B) + \cdots + (-1)^{n+1} \mu((A_1 \cup A_2 \cup \cdots \cup A_n) \setminus B) \quad (32)
$$

How can this help us? Note: this works for any number of sets $n$, not just all of them (so we are in some sense overloading the variable $n$).
proof of Theorem 3.1.

Since $A \setminus B = A \cap B^c$, every atom $A \in \mathcal{A}$ corresponds to:

$$A = \bigcap_{i=1}^{n} Y_i = \left( \bigcap_{j:Y_j=\tilde{X}_j} \tilde{X}_j \right) \bigcap \left( \bigcap_{j:Y_j=\tilde{X}_j^c} \tilde{X}_j^c \right)$$  \hspace{1cm} (33)

$$= \left( \bigcap_{j:Y_j=\tilde{X}_j} \tilde{X}_j \right) \bigcap \left( \bigcup_{j:Y_j=\tilde{X}_j} \tilde{X}_j \right)^c$$  \hspace{1cm} (34)

$$= \left( \bigcap_{j:Y_j=\tilde{X}_j} \tilde{X}_j \right) \setminus \left( \bigcup_{j:Y_j=\tilde{X}_j^c} \tilde{X}_j \right)$$  \hspace{1cm} (35)

$$\cdots$$  \hspace{1cm} (36)
Proof of Theorem 3.1

Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(32)) may take the form:

\[ \mu(A_i \cup A_j \cup \cdots \cup A_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_l) \]  

(37)

\[ = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_l) - \mu(\bigcup_{\ell} \tilde{X}_l) \]  

(38)

which is true since \( \mu(A \setminus B) = \mu(A \cup B) - \mu(B) \).
Proof of Theorem 3.1

Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(32)) may take the form:

\[ \mu(A_i \cup A_j \cup \cdots \cup A_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_l) \]  \hspace{1cm} (37)

\[ = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_l) - \mu(\bigcup_{\ell} \tilde{X}_l) \]  \hspace{1cm} (38)

which is true since \( \mu(A \setminus B) = \mu(A \cup B) - \mu(B) \).

Thus, the measure of any atom \( A \in \mathcal{A} \) is representable as a sum of weighted measures of the unions of the basic sets \( \mathcal{B} \)!

...
Proof of Theorem 3.1

Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(32)) may take the form:

\[
\mu(A_i \cup A_j \cup \cdots \cup A_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_l) \quad (37)
\]

\[
= \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_l) - \mu(\bigcup_{\ell} \tilde{X}_l) \quad (38)
\]

which is true since \(\mu(A \setminus B) = \mu(A \cup B) - \mu(B)\).

Thus, the measure of any atom \(A \in \mathcal{A}\) is representable as a sum of weighted measures of the unions of the basic sets \(\mathcal{B}\)!

Therefore, there exists a \(k \times k\) matrix \(D_n\) such that \(\vec{a} = D_n \vec{b}\) (before we had \(\vec{b} = C_n \vec{a}\)).
Proof of Theorem 3.1

Since $C_n$ is unique, so is $D_n$ with $D_n = C_n^{-1}$. 

So to summarize, we can define quantities only on $B$ and it defines the measures for all elements of $F_n$. For example, we can define just the values $H(X_G)$ for $G \subseteq [n]$ and this defines every other information theoretic value.
Proof of Theorem 3.1

Proof of Theorem 3.1.

- Since $C_n$ is unique, so is $D_n$ with $D_n = C_n^{-1}$.
- So to summarize, we can define quantities only on $B$ and it defines the measures for all elements of $F_n$. 
Proof of Theorem 3.1

Since $C_n$ is unique, so is $D_n$ with $D_n = C_n^{-1}$.

So to summarize, we can define quantities only on $\mathcal{B}$ and it defines the measures for all elements of $\mathcal{F}_n$.

For example, we can define just the values $H(X_G)$ for $G \subseteq [n]$ and this defines every other information theoretic value.
Examples

- Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.
Examples

- Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.
- Let $A, B, C$ be sets.
Examples

- Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.
- Let \( A, B, C \) be sets.
- **Lemma:**
  \[
  \mu((A \cap B) \setminus C) = \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C).
  \]
Examples

- Here are two examples of the above ideas, expressible as lemmas, first proven using set theory and second proven using information theoretic ideas.

- Let $A, B, C$ be sets.

- **lemma:**
  $$\mu((A \cap B) \setminus C) = \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C).$$

- **proof:**
  
  $$\mu((A \cap B) \setminus C) = \mu(A \setminus C) + \mu(B \setminus C) - \mu((A \cup B) \setminus C) \quad (39)$$
  $$= (\mu(A \cup C) - \mu(C)) + (\mu(B \cup C) - \mu(C))$$
  $$\quad - (\mu(A \cup B \cup C) - \mu(C)) \quad (40)$$
  $$= \mu(A \cup C) + \mu(B \cup C) - \mu(A \cup B \cup C) - \mu(C). \quad (41)$$
Examples

- As a Shannon measure (let $A, B, C$ be random variables):
Examples

- As a Shannon measure (let $A, B, C$ be random variables):
  - **Lemma:** $I(A; B|C) = H(A, C) + H(B, C) - H(A, B, C) - H(C)$
  
  $p(A \cup B|C) \quad p(A \cup B \cup C) \quad p(B \cup C) - p(A \cup B \cup C) - p(C)$
Examples

- As a Shannon measure (let $A, B, C$ be random variables):
- **Lemma**: $I(A; B|C) = H(A, C) + H(B, C) - H(A, B, C) - H(C)$
- **Proof**:

\[
I(A; B|C) = H(A|C) - H(A|B, C) \tag{42}
\]
\[
= H(A, C) - H(C) - (H(A, B, C) - H(B, C)) \tag{43}
\]
\[
= H(A, C) + H(B, C) - H(A, B, C) + H(Z) \tag{44}
\]
Examples

As a Shannon measure (let $A, B, C$ be random variables):

**lemma:** $I(A; B|C) = H(A, C') + H(B, C') - H(A, B, C) - H(C)$

**proof:**

\[
I(A; B|C) = H(A|C') - H(A|B, C') \\
= H(A, C') - H(C') - (H(A, B, C') - H(B, C')) \\
= H(A, C') + H(B, C') - H(A, B, C') + H(Z)
\]

**Key point:** this is the same as before, but in the first case we showed it using set theory.
Another key point. The information measure generates everything. That is

$$\mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n]$$

(45)

is sufficient to get all Shannon’s information theoretic quantities.
Another key point. The information measure generates everything. That is

$$\mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n]$$

(45)

is sufficient to get all Shannon’s information theoretic quantities.

Example: Start by defining $\mu(\tilde{X}_G)$ only on $\mathcal{B}$ with entropic quantities as above.
Sufficiency

- Another key point. The information measure generates everything. That is

\[ \mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n] \]  

(45)

is sufficient to get all Shannon's information theoretic quantities.

- Example: Start by defining \( \mu(\tilde{X}_G) \) only on \( B \) with entropic quantities as above.

- Then consider \( G, G', \) and \( G'' \) as sets, and form

\[
\mu((\tilde{X}_G \cap \tilde{X}_{G'}) \setminus \tilde{X}_{G''}) = \mu(\tilde{X}_G \cup \tilde{X}_{G''}) + \mu(\tilde{X}_{G'} \cup \tilde{X}_{G''}) - \mu(X_G, X_{G'}, X_{G''}) - \mu(X_{G''})
\]

(46)

\[
= I(X_G; X_{G'}|X_{G''})
\]

(47)
Sufficiency

- Another key point. The information measure generates everything. That is
  \[
  \mu(\tilde{X}_G) \triangleq H(X_G) \text{ for } G \subseteq [n]
  \]
  is sufficient to get all Shannon’s information theoretic quantities.
- Example: Start by defining \( \mu(\tilde{X}_G) \) only on \( B \) with entropic quantities as above.
- Then consider \( G, G', \) and \( G'' \) as sets, and form
  \[
  \mu((\tilde{X}_G \cap \tilde{X}_{G'}) \setminus \tilde{X}_{G''}) = \mu(\tilde{X}_G \cup \tilde{X}_{G''}) + \mu(\tilde{X}_{G'} \cup \tilde{X}_{G''}) - \mu(X_G, X_{G'}, X_{G''}) - \mu(X_{G''})
  \]
  \[
  = I(X_G; X_{G'}|X_{G''})
  \]
- That is, in most general case, we get conditional mutual information, but setting various combinations of \( G, G', \) or \( G'' \) to empty allows us to get \( I(X_G; X_{G'}) \) (if \( G'' = \emptyset \)), \( H(X_G|X_{G''}) \) (if \( G = G' \)), or \( H(X_G) \) (if \( G = G' \) and \( G'' = 0 \)).
Signed Measure

- When $n = 2$ all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
Signed Measure

- When \( n = 2 \) all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
- When \( n = 3 \), is this so?
Signed Measure

- When $n = 2$ all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
- When $n = 3$, is this so? What is:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) \quad (48)$$

and what might it mean?

$$\frac{?}{\text{?}} = I(X_1; X_2 | X_3)$$
Signed Measure

- When $n = 2$ all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
- When $n = 3$, is this so? What is:

$$
\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) \quad (48)
$$

and what might it mean?
- We have that:

$$
\mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) + \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c) = \mu(\tilde{X}_1 \cap \tilde{X}_2) \quad (49)
$$
Signed Measure

- When $n = 2$ all measures are positive given entropic instantiations (e.g., discrete entropy and MI are non-negative).
- When $n = 3$, is this so? What is:
  \[ \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) \]  (48)
and what might it mean?
- We have that:
  \[ \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) + \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c) = \mu(\tilde{X}_1 \cap \tilde{X}_2) \]  (49)
which translates as
\[ I(X_1; X_2; X_3) + I(X_1; X_2|X_3) = I(X_1; X_2) \]  (50)
Signed Measure

In fact, we have that:

\[ I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \]  \hspace{1cm} (51)
\[ = I(X_1; X_3) - I(X_1; X_3|X_2) \]  \hspace{1cm} (52)
\[ = I(X_2; X_3) - I(X_2; X_3|X_1) \]  \hspace{1cm} (53)

Why is this?
Signed Measure

In fact, we have that:

\[
I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \tag{51}
\]

\[
= I(X_1; X_3) - I(X_1; X_3|X_2) \tag{52}
\]

\[
= I(X_2; X_3) - I(X_2; X_3|X_1) \tag{53}
\]

Why is this? Due to the three ways of “de-eliminating” or “de-marginalizing” the three variables.
Signed Measure

- In fact, we have that:

\[ I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \]  \hspace{1cm} (51)

\[ = I(X_1; X_3) - I(X_1; X_3|X_2) \]  \hspace{1cm} (52)

\[ = I(X_2; X_3) - I(X_2; X_3|X_1) \]  \hspace{1cm} (53)

Why is this? Due to the three ways of “de-eliminating” or “de-marginalizing” the three variables.

- So we have a “new” information theoretic quantity \( I(X_1; X_2; X_3) \) for discrete random variables. Must it be positive?
In fact, we have that:

$$I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3)$$  \hspace{1cm} (51)

$$= I(X_1; X_3) - I(X_1; X_3|X_2)$$  \hspace{1cm} (52)

$$= I(X_2; X_3) - I(X_2; X_3|X_1)$$  \hspace{1cm} (53)

Why is this? Due to the three ways of “de-eliminating” or “de-marginalizing” the three variables.

So we have a “new” information theoretic quantity $I(X_1; X_2; X_3)$ for discrete random variables. Must it be positive?

Consider uniform binary $X_1 \perp \perp X_2$ and $X_3 = X_1 \oplus X_2$ (xor). Then $X_i \perp \perp X_j$ for $i \neq j$ and $I(X_i; X_j) = 0$ and $H(X_i|X_j) = 1$ for $i \neq j$. 
Signed Measure

- In fact, we have that:

\[ I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \]  
\[ = I(X_1; X_3) - I(X_1; X_3|X_2) \]  
\[ = I(X_2; X_3) - I(X_2; X_3|X_1) \]

(51)  
(52)  
(53)

Why is this? Due to the three ways of “de-eliminating” or “de-marginalizing” the three variables.

- So we have a “new” information theoretic quantity \( I(X_1; X_2; X_3) \) for discrete random variables. Must it be positive?

- Consider uniform binary \( X_1 \perp \perp X_2 \) and \( X_3 = X_1 \oplus X_2 \) (xor). Then \( X_i \perp \perp X_j \) for \( i \neq j \) and \( I(X_i; X_j) = 0 \) and \( H(X_i|X_j) = 1 \) for \( i \neq j \).

- Furthermore, \( H(X_i|X_j, X_k) = 0 \) for \( i \neq j, j \neq k, k \neq i \), and \( I(X_1; X_2|X_3) = H(X_1|X_3) - H(X_1|X_3, X_2) = 1 - 0 = 1 \).
Signed Measure

In fact, we have that:

\[ I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) \]  \hspace{1cm} (51)

\[ = I(X_1; X_3) - I(X_1; X_3|X_2) \]  \hspace{1cm} (52)

\[ = I(X_2; X_3) - I(X_2; X_3|X_1) \]  \hspace{1cm} (53)

Why is this? Due to the three ways of “de-eliminating” or “de-marginalizing” the three variables.

So we have a “new” information theoretic quantity \( I(X_1; X_2; X_3) \) for discrete random variables. Must it be positive?

Consider uniform binary \( X_1 \perp \perp X_2 \) and \( X_3 = X_1 \oplus X_2 \) (xor). Then \( X_i \perp \perp X_j \) for \( i \neq j \) and \( I(X_i; X_j) = 0 \) and \( H(X_i|X_j) = 1 \) for \( i \neq j \).

Furthermore, \( H(X_i|X_j, X_k) = 0 \) for \( i \neq j, j \neq k, k \neq i \), and \( I(X_1; X_2|X_3) = H(X_1|X_3) - H(X_1|X_3, X_2) = 1 - 0 = 1. \)

Therefore, \( I(X_1; X_2; X_3) = 0 - 1 = -1! \)
More on $I(X_1; X_2; X_3)$

This above a simple consequence of that conditioning can either increase or decrease mutual information.
More on $I(X_1; X_2; X_3)$

- This above a simple consequence of that conditioning can either increase or decrease mutual information.

- Example: Increase, example from before $X_3 = X_1 \oplus X_2$ with $0 = I(X_1; X_2) < I(X_1; X_2|X_3) = 1$
More on $I(X_1; X_2; X_3)$

- This above a simple consequence of that conditioning can either increase or decrease mutual information.

- Example: Increase, example from before $X_3 = X_1 \oplus X_2$ with $0 = I(X_1; X_2) < I(X_1; X_2|X_3) = 1$

- Example: Decrease, Markov chain $X_1 \rightarrow X_2 \rightarrow X_3$ so $I(X_1; X_3) \geq I(X_1; X_3|X_2) = 0$. 
More on $I(X_1; X_2; X_3)$

- This above a simple consequence of that conditioning can either increase or decrease mutual information.

- Example: Increase, example from before $X_3 = X_1 \oplus X_2$ with $0 = I(X_1; X_2) < I(X_1; X_2|X_3) = 1$

- Example: Decrease, Markov chain $X_1 \rightarrow X_2 \rightarrow X_3$ so $I(X_1; X_3) \geq I(X_1; X_3|X_2) = 0$.

- Thus $I(X_1; X_2; X_3)$ might be either positive or negative.
More on $I(X_1; X_2; X_3)$

- We’ve seen $I(X_1; X_2; X_3)$ before in Venn diagrams.

![Venn Diagram illustrating information measures]

- $I(X_1; X_2; X_3)$ can be viewed as a form of information “among” three random variables, or the information “common to” the three r.v.s.
More on $I(X_1; X_2; X_3)$

- We’ve seen $I(X_1; X_2; X_3)$ before in Venn diagrams

- We can thus view $I(X_1; X_2; X_3)$ as a form of information “among” three random variables, or the information “common to” the three r.v.s.
More on $I(X_1; X_2; X_3)$

In fact,

$$I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3)$$  \hspace{1cm} (54)

$$= I(X_1; X_2) - [I(X_1; X_2, X_3) - I(X_1; X_3)]$$  \hspace{1cm} (55)

$$= [I(X_1; X_2) + I(X_1; X_3)] - I(X_1; X_2, X_3)$$  \hspace{1cm} (56)
More on $I(X_1; X_2; X_3)$

- In fact,

$$I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2 | X_3)$$

$$= I(X_1; X_2) - [I(X_1; X_2, X_3) - I(X_1; X_3)]$$

$$= [I(X_1; X_2) + I(X_1; X_3)] - I(X_1; X_2, X_3)$$

- So $I(X_1; X_2; X_3)$ can be seen as the difference between the joint information $X_2, X_3$ has about $X_1$, and the sum of the individual amounts of information $X_2$ and $X_3$ have about $X_1$. 

\[\hat{\text{Comment}}\]
More on $I(X_1; X_2; X_3)$

- In fact,

$$I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3)$$

$$= I(X_1; X_2) - [I(X_1; X_2, X_3) - I(X_1; X_3)]$$

$$= [I(X_1; X_2) + I(X_1; X_3)] - I(X_1; X_2, X_3)$$

So $I(X_1; X_2; X_3)$ can be seen as the difference between the joint information $X_2, X_3$ has about $X_1$, and the sum of the individual amounts of information $X_2$ and $X_3$ have about $X_1$.

- Of course, this is symmetric so this could be regarding the information between any two $X_i, X_j$ and a third $X_k$ for $i \neq j$, $j \neq k$, and $k \neq i$. 
More on $I(X_1; X_2; X_3)$

- As an alternative, consider:

$$I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - I(X_1; X_2; X_3) \quad (57)$$

$$= \sum_{x_1, x_2, x_3} p(x_1, x_2, x_3) \log \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)} \quad (58)$$

$$\mathbb{E} \log \frac{\rho(x_1, x_2)}{\rho(x_1)\rho(x_2)}$$
More on $I(X_1; X_2; X_3)$

- As an alternative, consider:

$$I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2, X_3) \quad (57)$$

$$= \sum_{x_1, x_2, x_3} p(x_1, x_2, x_3) \log \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)} \quad (58)$$

- So $I'(X_1; X_2; X_3)$ is the difference in coding length between coding independently and coding jointly (which is 0 if all variables are independent, and large if the variables are highly correlated).
More on $I(X_1; X_2; X_3)$

- As an alternative, consider:

$$I'(X_1; X_2; X_3) \triangleq H(X_1) + H(X_2) + H(X_3) - I(X_1; X_2; X_3) \quad (57)$$

$$= \sum_{x_1, x_2, x_3} p(x_1, x_2, x_3) \log \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)} \quad (58)$$

- So $I'(X_1; X_2; X_3)$ is the difference in coding length between coding independently and coding jointly (which is 0 if all variables are independent, and large if the variables are highly correlated).

- Visualized, we have:
Applications of $I(X_1; X_2; X_3)$

- **EAR measure** in pattern classification:
  
  $$-I(X_i; X_j; C) = I(X_i; X_j|C) - I(X_i; X_j)$$
  
  where $C$ is a class variable and $X_i$ are features. If EAR is negative, then $X_i, X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

Synergy in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy $(R_1, R_2)$” is defined as:

$$\text{synergy}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2) = I(R_1; R_2|S) - I(R_1; R_2)$$

If synergy $< 0$, there exists “redundancy” in the neural code.
If synergy $> 0$, there the cells are “synergistic”, they encode more information in their neural response about a stimuli jointly than the sum of what they cover separately.
Applications of $I(X_1; X_2; X_3)$

- **EAR measure** in pattern classification:
  
  
  $$-I(X_i; X_j; C) = I(X_i; X_j|C) - I(X_i; X_j)$$

  where $C$ is a class variable and $X_i$ are features. If EAR is negative, then $X_i, X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

- **Synergy** in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy($R_1, R_2$)” is defined as:

  $$\text{synergy}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2) \quad (59)$$

  $$\quad = I(R_1; R_2|S) - I(R_1; R_2) \quad (60)$$

  $$\quad = -I(R_1; R_2; S) \quad (61)$$
Applications of $I(X_1; X_2; X_3)$

- **EAR measure** in pattern classification:
  \[-I(X_i; X_j; C) = I(X_i; X_j|C) - I(X_i; X_j)\]
  where $C$ is a class variable and $X_i$ are features. If EAR is negative, then $X_i, X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

- **Synergy** in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy$(R_1, R_2)$” is defined as:
  \[
  \text{synergy}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2) \quad (59)
  \]
  \[
  = I(R_1; R_2|S) - I(R_1; R_2) \quad (60)
  \]
  \[
  = -I(R_1; R_2; S) \quad (61)
  \]
  If synergy < 0, there exists “redundancy” in the neural code.
Applications of $I(X_1; X_2; X_3)$

- **EAR measure** in pattern classification:
  
  $-I(X_i; X_j; C) = I(X_i; X_j|C) - I(X_i; X_j)$ where $C$ is a class variable and $X_i$ are features. If EAR is negative, then $X_i, X_j$ are more dependent conditioned on $C$ than otherwise (so is indicative of a “good” direct interaction to model for classification).

- **Synergy** in a nerve cell (neural code of sensory stimuli). Here, $S =$ stimuli, $R_1 =$ response of one neuron, $R_2 =$ response of another neuron. Then “synergy($R_1, R_2$)” is defined as:

  $$\text{synergy}(R_1, R_2) = I(S; R_1, R_2) - I(S; R_1) - I(S; R_2) \quad (59)$$

  $$= I(R_1; R_2|S) - I(R_1; R_2) \quad (60)$$

  $$= -I(R_1; R_2; S) \quad (61)$$

  - If synergy $< 0$, there exists “redundancy” in the neural code.
  - If synergy $> 0$, there the cells are “synergistic”, they encode more information in their neural response about a stimuli jointly than the sum of what they cover separately.
Generalizations $I(X_1; X_2; X_3)$

- For $n > 3$ random variables, and multiple groups $\{G_i\}_i$, we can define:

$$
\mu((\hat{X}_{G_1} \cap \hat{X}_{G_2} \cap \cdots \cap \hat{X}_{G_m}) \setminus \hat{X}_F) \triangleq I(X_{G_1}; X_{G_2}; \ldots; X_{G_m} | X_F)
$$

(62)

as the mutual information “between” or “among” the groups $\{X_{G_i}\}_i$ conditioned on $X_F$. 
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

\[ X \rightarrow Y_1 \rightarrow Z_1 \]
\[ X \rightarrow Y_2 \rightarrow Z_2 \]

vs.

\[ X \rightarrow Y_1 \]
\[ Y_1 \rightarrow W_1 \]
\[ Y_1 \rightarrow W_2 \]

I.e., you have either $(Z_1, Z_2)$ or $(W_1, W_2)$ but not both, where $I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X)$, so each individual variable is equally useful to learn about $X$.

Example: communications: might get "stuck" with noise given in $Y$.

Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

$$X \xrightarrow{Y_1} Z_1 \quad \text{vs.} \quad X \xrightarrow{Y_1} W_1$$

$$X \xrightarrow{Y_2} Z_2 \quad \text{vs.} \quad X \xrightarrow{Y_1} W_2$$

- I.e., you have either $(Z_1, Z_2)$ or $(W_1, W_2)$ but not both, where $I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X)$, so each individual variable is equally useful to learn about $X$. 

Example: communications: might get "stuck" with noise given in $Y$.

Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

![Diagram showing two sets of signals Y1, Y2, Z1, Z2 vs. W1, W2]

- I.e., you have either $(Z_1, Z_2)$ or $(W_1, W_2)$ but not both, where $I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X)$, so each individual variable is equally useful to learn about $X$.
- Example: communications: might get “stuck” with noise given in $Y$. 

---

Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Example: Two views of two views of a source

- Can correlated noisy signals ever reveal more about a source than independent (noisy) looks at a source?
- We wish to learn about $X$ and we have two ways we might do it:

![Diagram](image)

I.e., you have either $(Z_1, Z_2)$ or $(W_1, W_2)$ but not both, where

$$I(Z_1; X) = I(Z_2; X) = I(W_1; X) = I(W_2; X),$$

so each individual variable is equally useful to learn about $X$.

- Example: communications: might get “stuck” with noise given in $Y$.
- Example: population theory. When testing about the instance, say, of a disease in a population, we can either test one individual twice (each might have errors) or two individuals once. Which is better?
Example: Two views of two views of a source

Since each variable is individually equally informative, correlated views of a source being better entirely depends on if
\[ I(X; Z_2 | Z_1) < I(X; W_2 | W_1) \] or not.

\[ \text{Then} \quad I(X; Z_1) = I(X; W) \]

\[ I(X; Z_1, Z_2) \text{ vs. } I(X; W_1, W_2) \]
Example: Two views of two views of a source

- Since each variable is individually equally informative, correlated views of a source being better entirely depends on if
  \[ I(X; Z_2|Z_1) < I(X; W_2|W_1) \]
  or not.
- But \( Z_1 \perp \perp Z_2 | X \) so \( Z_1 \rightarrow X \rightarrow Z_2 \) meaning

\[
I(X; Z_2|Z_1) \leq I(X; Z_2) = I(X; W_2)
\]  \quad (63)
Example: Two views of two views of a source

- Since each variable is individually equally informative, correlated views of a source being better entirely depends on if
  \[ I(X; Z_2|Z_1) < I(X; W_2|W_1) \]
  or not.

- But \( Z_1 \perp Z_2 | X \) so \( Z_1 \rightarrow X \rightarrow Z_2 \) meaning

  \[ I(X; Z_2|Z_1) \leq I(X; Z_2) = I(X; W_2) \]  \hspace{1cm} (63)

- Thus, a sufficient condition for this to be true is:

  \[ I(X; W) < I(X; W_2|W_1) \]  \hspace{1cm} (64)

  or

  \[ I(X; W_2) - I(X; W_2|W_1) = I(X; W_2; W_1) < 0 \]  \hspace{1cm} (65)
Example: Two views of two views of a source

Since each variable is individually equally informative, correlated views of a source being better entirely depends on if 
\[ I(X; Z_2|Z_1) < I(X; W_2|W_1) \] or not.

But \( Z_1 \perp \perp Z_2 | X \) so \( Z_1 \rightarrow X \rightarrow Z_2 \) meaning

\[ I(X; Z_2|Z_1) \leq I(X; Z_2) = I(X; W_2) \] (63)

Thus, a sufficient condition for this to be true is:

\[ I(X; W) < I(X; W_2|W_1) \] (64)

or

\[ I(X; W_2) - I(X; W_2|W_1) = I(X; W_2; W_1) < 0 \] (65)

Exercise: find an example of this where \( X \rightarrow Y \) are Z-channels, and \( Y \rightarrow Z \) and \( Y \rightarrow W \) are BSCs.
How far can pictures go? We saw earlier that for \( n = 3 \), we can show all quantities. But if \( I(X_1; X_2; X_3) < 0 \) then we need a way of showing “negative area”, and one way would be to do something like:
How far can pictures go? We saw earlier that for $n = 3$, we can show all quantities. But if $I(X_1; X_2; X_3) < 0$ then we need a way of showing “negative area”, and one way would be to do something like:

For $n > 3$, it is not possible to perfectly display such a diagram in 2D since for $n$ random variables we need $n - 1$ dimensions to be displayed perfectly.
The Zen and Art of Venn

- How far can pictures go? We saw earlier that for $n = 3$, we can show all quantities. But if $I(X_1; X_2; X_3) < 0$ then we need a way of showing “negative area”, and one way would be to do something like:

- For $n > 3$, it is not possible to perfectly display such a diagram in 2D since for $n$ random variables we need $n - 1$ dimensions to be displayed perfectly.

- “perfectly” means that atoms are adjacent if, say, only one variable is complemented between the atoms. For example, $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \cap \tilde{X}_4^c$ should be adjacent to $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c \cap \tilde{X}_4^c$
One way to try to display this, almost perfectly, is as follows:
The Zen and Art of Venn

- One way to try to display this, almost perfectly, is as follows:

- Here we see that the set $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_4^c$ is partitioned into $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3 \cap \tilde{X}_4^c$ and $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c \cap \tilde{X}_4^c$ which are not adjacent.
Actually, the measures can take any non-negative value on the atoms due to the flexibility of entropy for independent random variables. That is, we have
Flexibility

- Actually, the measures can take any non-negative value on the atoms due to the flexibility of entropy for independent random variables. That is, we have

**Theorem 3.2**

If there is no constraint on random variables $X_1, X_2, \ldots, X_n$, then $\mu^*$ can take on any set of nonnegative values on the non-empty atoms of $\mathcal{F}_n$ (namely $A$).
Flexibility

- Actually, the measures can take any non-negative value on the atoms due to the flexibility of entropy for independent random variables. That is, we have

Theorem 3.2

*If there is no constraint on random variables \(X_1, X_2, \ldots, X_n\), then \(\mu^*\) can take on any set of nonnegative values on the non-empty atoms of \(\mathcal{F}_n\) (namely \(A\)).*

Proof sketch.

We associate the atoms with a set of mutually independent random variables. Any random variable \(X_i\) than has entropy equal to the sum of the atoms contained in \(X_i\)’s set (i.e., \(\tilde{X}_i\)) but these individual entropies are unrestricted.
Markov Chain Restrictions

- What if there are restrictions? I.e., what if we have a Markov chain $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$. 

Ex. For $n = 3$, we have $X_1 \rightarrow X_2 \rightarrow X_3$ which means that $I(X_1; X_3 | X_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) = 0$ (66)
Markov Chain Restrictions

- What if there are restrictions? I.e., what if we have a Markov chain $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$.

- Ex. For $n = 3$, we have $X_1 \rightarrow X_2 \rightarrow X_3$ which means that

$$I(X_1; X_3|X_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3) = 0 \quad (66)$$
Markov Chain Restrictions

- What if there are restrictions? I.e., what if we have a Markov chain $X_1 \to X_2 \to \cdots \to X_n$.
- Ex. For $n = 3$, we have $X_1 \to X_2 \to X_3$ which means that
  \[
  I(X_1; X_3|X_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3) = 0 \tag{66}
  \]
- One way of plotting a Venn-like diagram is as follows:
Markov Chain Restrictions

In this case, what happens to $I(X_1; X_2; X_3)$?

$$I(X_1; X_2; X_3) = \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) = \mu(\tilde{X}_1 \cap \tilde{X}_3) = I(X_1; X_3) \geq 0$$

(67)
In this case, what happens to $I(X_1; X_2; X_3)$?

\[
I(X_1; X_2; X_3) = \mu(\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3) = \mu(\tilde{X}_1 \cap \tilde{X}_3) = I(X_1; X_3) \geq 0
\]  

So with a Markov chain, we have $I(X_1; X_2; X_3) \geq 0$. Another Venn-like figure would show this:

but it does not nicely generalize to $n > 3$. 
Markov Chain Restrictions

- With larger $n$, we can deduce that certain sets are empty.
Markov Chain Restrictions

- With larger $n$, we can deduce that certain sets are empty.
- For example with $n = 4$ and $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, we have:

\[
\mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4^c \cap \tilde{X}_2^c) = \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_2^c)
\]
Markov Chain Restrictions

- With larger $n$, we can deduce that certain sets are empty.
- For example with $n = 4$ and $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, we have:

$$\mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4^c \cap \tilde{X}_2^c) = \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_2^c)$$

which means that

$$I(X_1; X_3; X_4|X_2) + I(X_1; X_3|X_2, X_4) = I(X_1; X_3|X_2) = 0 \quad (68)$$
With larger $n$, we can deduce that certain sets are empty. For example with $n = 4$ and $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, we have:

$$\mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_4^c \cap \tilde{X}_2^c) = \mu(\tilde{X}_1 \cap \tilde{X}_3 \cap \tilde{X}_2^c)$$

which means that

$$I(X_1; X_3; X_4|X_2) + I(X_1; X_3|X_2, X_4) = I(X_1; X_3|X_2) = 0 \quad (68)$$

We can see this from the diagram:
In fact, 5 atoms have empty measures.

\[ A = H(X_1|X_2, X_3, X_4), \quad B = I(X_1; X_2|X_3, X_4), \]
\[ C = I(X_1; X_3|X_4), \ldots, \quad \text{and} \quad G = I(X_2; X_4|X_1). \]
Markov Chain Restrictions: Empty measures

- In fact, 5 atoms have empty measures.

Then

$$A = H(X_1|X_2, X_3, X_4), \quad B = I(X_1; X_2|X_3, X_4), \quad C = I(X_1; X_3|X_4), \ldots, \quad \text{and} \quad G = I(X_2; X_4|X_1).$$

- For example: $I(X_1; X_3|X_2) = 0$. What else?
Markov Chain Restrictions: Empty measures

- In fact, 5 atoms have empty measures.

Then

\[ A = H(X_1|X_2, X_3, X_4), \quad B = I(X_1; X_2|X_3, X_4), \]
\[ C = I(X_1; X_3|X_4), \ldots, \quad \text{and} \quad G = I(X_2; X_4|X_1). \]

- For example: \( I(X_1; X_3|X_2) = 0 \). What else? \( I(X_1; X_4|X_2) = 0, \)
\( I(X_1; X_4|X_3) = 0, \quad I(X_2; X_4|X_3) = 0, \) and \( I(X_1, X_2; X_4|X_3) = 0. \)
Markov Chain Restrictions: Empty measures

- In fact, 5 atoms have empty measures.

\[ A = H(X_1 | X_2, X_3, X_4), \quad B = I(X_1; X_2 | X_3, X_4), \]
\[ C = I(X_1; X_3 | X_4), \quad \ldots \quad \text{and} \quad G = I(X_2; X_4 | X_1). \]

- For example: \( I(X_1; X_3 | X_2) = 0 \). What else? \( I(X_1; X_4 | X_2) = 0 \),
\( I(X_1; X_4 | X_3) = 0 \), \( I(X_2; X_4 | X_3) = 0 \), and \( I(X_1, X_2; X_4 | X_3) = 0 \).

- The 10 non-empty atoms are then: \( H(X_1 | X_2, X_3, X_4), \)
\( I(X_1; X_2 | X_3, X_4), \quad I(X_1; X_3 | X_4), \quad I(X_1; X_4), \quad H(X_2 | X_1, X_2, X_4), \)
\( I(X_2; X_3 | X_1, X_4), \quad I(X_2; X_4 | X_1), \quad H(X_3 | X_1, X_2, X_4), \)
\( I(X_3; X_4 | X_1, X_2), \quad \text{and} \quad H(X_4 | X_1, X_2, X_3). \)
Markov chains

Theorem 3.3

Given $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$, then $\mu()$ is always non-negative.

- It is also possible, using $\mu$ to prove the concavity of entropy, convexity of MI in $p(y|x)$ for fixed $p(x)$ and the concavity of MI in $p(x)$ for fixed $p(y|x)$. 

Aside: perhaps an easy way to remember this: 1) remember that $H(X)$ is concave in $p(x)$ via the plot; 2) since $I(X;Y) = H(X) - H(X|Y) = H(X) - \sum_{x,y} p(x) p(y|x) \log p(y|x)$ then fixing $p(x)$, we have a "convex-looking" function of $p(y|x)$ (not really though since $-p \log p$ is concave in $p$), and fixing $p(y|x)$ we have a concave-looking function of $p(x)$. 

Prof. Jeff Bilmes
Markov chains

**Theorem 3.3**

*Given* $X_1 \to X_2 \to \cdots \to X_n$, *then* $\mu()$ *is always non-negative.*

- It is also possible, using $\mu$ to prove the concavity of entropy, convexity of MI in $p(y|x)$ for fixed $p(x)$ and the concavity of MI in $p(x)$ for fixed $p(y|x)$.

- Aside: perhaps an easy way to remember this: 1) remember that $H(X)$ is concave in $p(x)$ via the plot; 2) since

$$I(X;Y) = H(X) - H(X|Y) = H(X) - \sum_{x,y} p(x)p(y|x) \log p(y|x)$$

then fixing $p(x)$, we have a “convex-looking” function of $p(y|x)$ (not really though since $-p \log p$ is concave in $p$), and fixing $p(y|x)$ we have a concave-looking function of $p(x)$. 
Imperfect Secrecy Theorem

- Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

\[
\begin{align*}
X &\quad \rightarrow \quad \text{encryption} \\
Z &\quad \downarrow \\
Y &\quad \leftarrow \\
\end{align*}
\]
Imperfect Secrecy Theorem

- Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

- Goal: We need to make $I(X;Y)$ small (ideally zero).
Imperfect Secrecy Theorem

- Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

- Goal: We need to make $I(X;Y)$ small (ideally zero).

- Also, need to have $H(X|Y, Z) = 0$ to be able to decrypt without error (i.e., $X \rightarrow Y \leftarrow Z$).
Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:

Goal: We need to make $I(X; Y)$ small (ideally zero).

Also, need to have $H(X|Y, Z) = 0$ to be able to decrypt without error (i.e., $X \rightarrow Y \leftarrow Z$).

Under this constraint, however, we have lower bound:

$$I(X; Y) \geq H(X) - H(Z)$$

thus, to make $I(X; Y)$ small, need to make $H(Z)$ big (long random passwords, best is to have password as long or as random as the cleartext!).
Imperfect Secrecy Theorem

- Suppose that $X$ is plain text, $Y$ is cipher text, and $Z$ is a key (password) in the following model:
- Goal: We need to make $I(X; Y)$ small (ideally zero).
- Also, need to have $H(X|Y, Z) = 0$ to be able to decrypt without error (i.e., $X \rightarrow Y \leftarrow Z$).
- Under this constraint, however, we have lower bound:

$$I(X; Y) \geq H(X) - H(Z) \quad (69)$$

thus, to make $I(X; Y)$ small, need to make $H(Z)$ bit (long random passwords, best is to have password as long or as random as the cleartext!).

- We can use information measures to show this.
Imperfect Secrecy Theorem

- Consider the next info diagram
Imperfect Secrecy Theorem

Consider the next info diagram

\[ \begin{align*}
I(X; Y | Z) &= a \geq 0 \\
I(Y; Z | X) &= b \geq 0 \\
H(Z | X, Y) &= c \geq 0 \\
I(X; Y; Z) &= d
\end{align*} \]
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

(71)
Imperfect Secrecy Theorem

Consider the next info diagram

\[ \Rightarrow \]

Then we have

\[ 0 \]

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

(71)
Imperfect Secrecy Theorem

- Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) \]

(71)
Imperfect Secrecy Theorem

- Consider the next info diagram

\[
\begin{align*}
I(X; Y|Z) &= a \geq 0 \\
I(Y; Z|X) &= b \geq 0 \\
H(Z|X, Y) &= c \geq 0 \\
I(X; Y; Z) &= d
\end{align*}
\]

Then we have

\[
0 \leq I(Y; Z) = \mu(Y \cap Z)
\]

(71)
Imperfect Secrecy Theorem

Consider the next info diagram

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]  

(70)

\[ I(X; Y | Z) = a \geq 0 \]

\[ I(Y; Z | X) = b \geq 0 \]

\[ H(Z | X, Y) = c \geq 0 \]

\[ I(X; Y; Z) = d \]

(71)
Imperfect Secrecy Theorem

- Consider the next info diagram

\[
\begin{align*}
I(X; Y|Z) &= a \geq 0 \\
I(Y; Z|X) &= b \geq 0 \\
H(Z|X, Y) &= c \geq 0 \\
I(X; Y; Z) &= d
\end{align*}
\]

- Then we have

\[
0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \tag{70}
\]

\[
= I(X; Y; Z) + I(Y; Z|X) \tag{71}
\]
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X, Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]  \hspace{1cm} (70)
\[ = I(X; Y; Z) + I(Y; Z|X) = b + d \]  \hspace{1cm} (71)
Imperfect Secrecy Theorem

- Consider the next info diagram

- Then we have

\[
0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c)
\]

\[
= I(X; Y; Z) + I(Y; Z|X) = b + d
\]

- Thus, \( d \geq -b \), and

\[
I(X; Y) = a \geq a - b - c = H(X) - H(Z)
\]
Imperfect Secrecy Theorem

Consider the next info diagram

\[ I(X; Y|Z) = a \geq 0 \]
\[ I(Y; Z|X) = b \geq 0 \]
\[ H(Z|X,Y) = c \geq 0 \]
\[ I(X; Y; Z) = d \]

Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \quad (70) \]
\[ = I(X; Y; Z) + I(Y; Z|X) = b + d \quad (71) \]

Thus, \( d \geq -b \), and

\[ H(X) - H(Z) \quad (72) \]
Imperfect Secrecy Theorem

- Consider the next info diagram

- Then we have

\[ 0 \leq I(Y;Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]
\[ = I(X;Y;Z) + I(Y;Z|X) = b + d \] (70)

- Thus, \( d \geq -b \), and

\[ H(X) - H(Z) = (a + d) - (d + b + c) \] (72)
Imperfect Secrecy Theorem

1. Consider the next info diagram

   ![Info Diagram]

2. Then we have

   \[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \]  \[= I(X; Y; Z) + I(Y; Z|X) = b + d \]  \[\Rightarrow I(X; Y|Z) = a \geq 0 \]

   \[ I(Y; Z|X) = b \geq 0 \]

   \[ H(Z|X, Y) = c \geq 0 \]

   \[ I(X; Y; Z) = d \]

3. Thus, \( d \geq -b \), and

   \[ H(X) - H(Z) = (a + d) - (d + b + c) = a - b - c \]
Imperfect Secrecy Theorem

- Consider the next info diagram

- Then we have

\[ 0 \leq I(Y; Z) = \mu(Y \cap Z) = \mu(Y \cap Z \cap X) + \mu(Y \cap Z \cap X^c) \] (70)

\[ = I(X; Y; Z) + I(Y; Z|X) = b + d \] (71)

- Thus, \( d \geq -b \), and

\[ H(X) - H(Z) = (a + d) - (d + b + c) = a - b - c \] (72)

- so \( I(X; Y) = a + d \geq a - b \geq a - b - c = H(X) - H(Z) \)
Imperfect Secrecy Theorem

- Note that a number of (nice) assumptions were not used in this derivation.
Imperfect Secrecy Theorem

- Note that a number of (nice) assumptions were not used in this derivation.

1. Did not use $H(Y|X,Z) = 0$, so we could have random cipher text as a function of the text and password.
Imperfect Secrecy Theorem

- Note that a number of (nice) assumptions were **not** used in this derivation.

1. Did not use $H(Y|X, Z) = 0$, so we could have random cipher text as a function of the text and password.

2. $I(X; Z) = 0$, so don’t require that the text and password to be independent (so it can be made to be easy to remember the password even if it is long).
Imperfect Secrecy Theorem

- Note that a number of (nice) assumptions were not used in this derivation.

1. Did not use $H(Y|X, Z) = 0$, so we could have random cipher text as a function of the text and password.

2. $I(X; Z) = 0$, so don’t require that the text and password to be independent (so it can be made to be easy to remember the password even if it is long).

- But then, this is only a lower bound, it shows a possible worst case.
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ gives:

Considering the above figure, we have:

$$\mu(\tilde{X}_1 \cap \tilde{X}_4) = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3)$$ (73)

which is data processing inequality.

We also have

$$\mu(\tilde{X}_2 \cap \tilde{X}_3) = I(X_2; X_3) \geq I(X_2; X_4) = \mu(\tilde{X}_2 \cap \tilde{X}_4)$$ (74)

And all of the others...
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ gives:

$$X_1 X_2 X_3 X_4$$

Considering the above figure, we have:

$$\mu(\tilde{X}_1 \cap \tilde{X}_4) = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3) \quad (73)$$

which is data processing inequality.
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ gives:

\[
\begin{align*}
\mu(\tilde{X}_1 \cap \tilde{X}_4) & = I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3) \\
& \text{(73)}
\end{align*}
\]

which is data processing inequality.

- Considering the above figure, we have:

\[
\begin{align*}
\mu(\tilde{X}_2 \cap \tilde{X}_3) & = I(X_2; X_3) \geq I(X_2; X_4) = \mu(\tilde{X}_2 \cap \tilde{X}_4) \\
& \text{(74)}
\end{align*}
\]
The Venn and Art of the Data Processing Inequality

- Markov chain $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ gives:

\[
\begin{align*}
\mu(\tilde{X}_1 \cap \tilde{X}_4) &= I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3) \\
\mu(\tilde{X}_2 \cap \tilde{X}_3) &= I(X_2; X_3) \geq I(X_2; X_4) = \mu(\tilde{X}_2 \cap \tilde{X}_4)
\end{align*}
\] (73)

which is data processing inequality.

- Considering the above figure, we have:

- We also have

\[
\begin{align*}
\mu(\tilde{X}_1 \cap \tilde{X}_4) &= I(X_1; X_4) \leq I(X_2; X_3) = \mu(\tilde{X}_2 \cap \tilde{X}_3) \\
\mu(\tilde{X}_2 \cap \tilde{X}_3) &= I(X_2; X_3) \geq I(X_2; X_4) = \mu(\tilde{X}_2 \cap \tilde{X}_4)
\end{align*}
\] (74)

- And all of the others . . .
Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).
Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

Note that for conditional mutual information we have the following:

\[
\begin{align*}
0 & \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C) \\
& = H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C) \\
& = H(X_{\alpha}) + H(X_{\beta}) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta}) \\
& = H(X_{\alpha}) + H(X_{\beta}) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta})
\end{align*}
\]

Thus, defining a function of the form \(f(A) = H(X_A)\) this proves that for any \(A, B \subseteq [n]\)

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B)
\]

meaning that the entropy function is submodular in the index set of random variables.
Again, let \([n]\) be an index set of random variables and let 
\(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and 
\(\beta \triangleq B \cup C\).

Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C) =
H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)
\]

\[
= H(X_\alpha) + H(X_\beta) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta})
\]

(78)

Thus, defining a function of the form 
\(f(A) = H(X_A)\) this proves that for any 
\(A, B \subseteq [n]\)

\[f(A) + f(B) \geq f(A \cup B) + f(A \cap B)\]

meaning that the entropy function is submodular in the index set of 
random variables.
Inequalities

- Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C)
\]

(78)
Inequalities

- Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).
- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B|X_C) = H(X_A|X_C) - H(X_A|X_B, X_C)
\]  
\text{(75)}

\[
= H(X_A \cup C) + H(X_B \cup C) - H(X_B \cup C \cap (A \cup B)) - H(X_B \cup C \cup (A \cup B))
\]  
\text{(78)}
Inequalities

- Again, let $[n]$ be an index set of random variables and let $A, B, C \subseteq [n]$ and define $\alpha \triangleq A \cup C$ and $\beta \triangleq B \cup C$.

- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B|X_C) = H(X_A|X_C) - H(X_A|X_B, X_C) = H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)
\]  

(75)  
\[
= H(X_\alpha, X_\beta) - H(X_\alpha \cap X_\beta) - H(X_\alpha \cup X_\beta)  
\]  

(78)
Inequalities

- Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C)
\]

\[
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_C) - H(X_{A \cup B \cup C})
\]

Thus, defining a function of the form \(f(A) = H(X_A)\) this proves that for any \(A, B \subseteq [n]\)

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B)
\]

meaning that the entropy function is submodular in the index set of random variables.
Inequalities

- Again, let $[n]$ be an index set of random variables and let $A, B, C \subseteq [n]$ and define $\alpha \triangleq A \cup C$ and $\beta \triangleq B \cup C$.
- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C) \tag{75}
\]

\[
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C) \tag{76}
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_C) - H(X_{A \cup B \cup C}) \tag{77}
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_{(A \cup C) \cap (B \cup C)}) - H(X_{(A \cup B) \cup (B \cup C)}) \tag{78}
\]
Inequalities

- Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B|X_C) = H(X_A|X_C) - H(X_A|X_B, X_C) \\
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C) \\
= H(X_{A\cup C}) + H(X_{B\cup C}) - H(X_C) - H(X_{A\cup B\cup C}) \\
= H(X_{A\cup C}) + H(X_{B\cup C}) - H(X_{(A\cup B)\cap(B\cup C)}) - H(X_{(A\cup B)\cup(B\cup C)}) \\
= H(X_\alpha) + H(X_\beta) - H(X_{\alpha\cap\beta}) - H(X_{\alpha\cup\beta})
\]
Inequalities

- Again, let \([n]\) be an index set of random variables and let \(A, B, C \subseteq [n]\) and define \(\alpha \triangleq A \cup C\) and \(\beta \triangleq B \cup C\).

- Note that for conditional mutual information we have the following:

\[
0 \leq I(X_A; X_B | X_C) = H(X_A | X_C) - H(X_A | X_B, X_C)
\]

\[
= H(X_A, X_C) - H(X_C) - H(X_A, X_B, X_C) + H(X_B, X_C)
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_C) - H(X_{A \cup B \cup C})
\]

\[
= H(X_{A \cup C}) + H(X_{B \cup C}) - H(X_{(A \cup B) \cap (B \cup C)}) - H(X_{(A \cup B) \cup (B \cup C)})
\]

\[
= H(X_\alpha) + H(X_\beta) - H(X_{\alpha \cap \beta}) - H(X_{\alpha \cup \beta})
\]

(75) (76) (77) (78)

- Thus, defining a function of the form \(f(A) = H(X_A)\) this proves that for any \(A, B \subseteq [n]\)

\[
f(A) + f(B) \geq f(A \cup B) + f(A \cap B)
\]

(79)

meaning that the entropy function is submodular in the index set of random variables.
More inequalities

- We also know other properties of the entropy, namely

  \[ f(\emptyset) = H(X_\emptyset) = 0 \]  

  and also that for any \( A \subseteq B \),

  \[ f(A) = H(X_A) \leq H(X_B) = f(B) \]

  meaning that entropy is a monotone non-decreasing function of the set of random variables (adding more random variables to discrete entropy function can only increase the uncertainty/information).
More inequalities

This begs the question: Consider all functions \( f : 2^{[n]} \rightarrow \mathbb{R}_+ \) that satisfy the following three properties:

\[
\begin{align*}
    f(\emptyset) & = 0 \quad (82) \\
    f(A) & \leq f(B) \text{ whenever } A \subseteq B \subseteq [n] \quad (83) \\
    f(A) + f(B) & \geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n] \quad (84)
\end{align*}
\]

and call this set of functions \( \Gamma_n \) (note that this set is called the cone of polymatroidal functions, the reason being that they are closed under conic combinations).
More inequalities

- This begs the question: Consider all functions $f : 2^{[n]} \to \mathbb{R}_+$ that satisfy the following three properties:

$$f(\emptyset) = 0$$  \hspace{1cm} (82)

$$f(A) \leq f(B) \text{ whenever } A \subseteq B \subseteq [n]$$  \hspace{1cm} (83)

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n]$$  \hspace{1cm} (84)

and call this set of functions $\Gamma_n$ (note that this set is called the cone of polymatroidal functions, the reason being that they are closed under conic combinations).

- Now, consider the set of all discrete entropy functions on $n$ random variables, call it $\Gamma^H_n$ (also close this set as well).
This begs the question: Consider all functions $f : 2^{[n]} \rightarrow \mathbb{R}_+$ that satisfy the following three properties:

\begin{align*}
    f(\emptyset) &= 0 & \text{(82)} \\
    f(A) &\leq f(B) \quad \text{whenever} \quad A \subseteq B \subseteq [n] & \text{(83)} \\
    f(A) + f(B) &\geq f(A \cup B) + f(A \cap B) \quad \text{for all} \quad A, B \subseteq [n] & \text{(84)}
\end{align*}

and call this set of functions $\Gamma_n$ (note that this set is called the cone of polymatroidal functions, the reason being that they are closed under conic combinations).

Now, consider the set of all discrete entropy functions on $n$ random variables, call it $\Gamma^H_n$ (also close this set as well).

Is there a relationship between $\Gamma^H_n$ and $\Gamma_n$?
This begs the question: Consider all functions $f : 2^{[n]} \rightarrow \mathbb{R}_+$ that satisfy the following three properties:

$$f(\emptyset) = 0$$  \hspace{1cm} (82)

$$f(A) \leq f(B) \text{ whenever } A \subseteq B \subseteq [n]$$  \hspace{1cm} (83)

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \text{ for all } A, B \subseteq [n]$$  \hspace{1cm} (84)

and call this set of functions $\Gamma_n$ (note that this set is called the cone of polymatroidal functions, the reason being that they are closed under conic combinations).

Now, consider the set of all discrete entropy functions on $n$ random variables, call it $\Gamma^H_n$ (also close this set as well).

Is there a relationship between $\Gamma^H_n$ and $\Gamma_n$?

Clearly, $\Gamma^H_n \subseteq \Gamma_n$ since entropy satisfies the three properties.
The following can be shown:
Equivalence?

The following can be shown:

**Theorem 3.4**

For \( n = 2 \) or \( n = 3 \), we have that \( \Gamma_n^{H_n} = \Gamma_n \).
The following can be shown:

**Theorem 3.4**

For $n = 2$ or $n = 3$, we have that $\Gamma_n^H = \Gamma_n$.

So it may be the case that this would generalize for $n \geq 4$, right?
The following can be shown:

**Theorem 3.4**

*For \( n = 2 \) or \( n = 3 \), we have that \( \Gamma_n^H = \Gamma_n \).*

- So it may be the case that this would generalize for \( n \geq 4 \), right?
- Would be a nice result as it would mean that entropy functions are as general as the set of polymatroidal functions (examples include, say, rank of a matrix, matroid rank, grouped matroid rank, weighted matroid rank, and many others).
Via information measures ideas presented earlier, we can get:

**Theorem 3.5**

*For any four discrete random variables \( X, Y, Z, U \), define*

\[
\Delta(Z, U|X, Y) = I(Z; U) - I(Z; U, X) - I(Z; U, Y) \quad (85)
\]

*then the following inequality holds:*

\[
\Delta(Z, U|X, Y) \leq \frac{1}{2} [I(X; Y) + I(X; Z, U) + I(Z; U, X) - I(Z; U|Y)] \quad (86)
\]

*and while the l.h.s. is symmetric in \( X \) and \( Y \) but the r.h.s. is not, we also have:*

\[
\Delta(Z, U|X, Y) \leq \frac{1}{2} [I(X; Y) + I(Y; Z, U) - I(Z; U, X) + I(Z; U|Y)] \quad (87)
\]
Consequences of this Inequality

Define $\Delta(i, j|k, \ell) \triangleq I(\{i\}; \{j\}) - I(\{i\}; \{j\}, \{k\}) - I(\{i\}; \{j\}, \{\ell\})$ where whenever we mention $\{i\}$ we really mean the r.v. $X_{\{i\}}$. 
Consequences of this Inequality

- Define $\Delta(i, j | k, \ell) \triangleq I(\{i\}; \{j\}) - I(\{i\}; \{j\}, \{k\}) - I(\{i\}; \{j\}, \{\ell\})$ where whenever we mention $\{i\}$ we really mean the r.v. $X_{\{i\}}$.

- Define $\tilde{\Gamma}_4$ as follows:

$$\tilde{\Gamma}_4 = \left\{ F \in \Gamma_4 : \text{for all permutations } \sigma \text{ of } \{1, 2, 3, 4\} , \right.$$  

$$\Delta(\sigma(1), \sigma(2) | \sigma(3), \sigma(4)) \leq \frac{1}{2} \left[ I(\sigma(3); \sigma(4)) + I(\sigma(1); \sigma(2), \sigma(3)) - I(\sigma(1); \sigma(2), \sigma(4)) + I(\sigma(3); \sigma(1), \sigma(2)) \right] \right\}$$
Consequences of this Inequality

- Define $\Delta(i, j | k, \ell) \triangleq I(\{i\}; \{j\}) - I(\{i\}; \{j\}, \{k\}) - I(\{i\}; \{j\}, \{\ell\})$ where whenever we mention $\{i\}$ we really mean the r.v. $X_{\{i\}}$.

- Define $\tilde{\Gamma}_4$ as follows:

$$
\tilde{\Gamma}_4 = \left\{ F \in \Gamma_4 : \text{ for all permutations } \sigma \text{ of } \{1, 2, 3, 4\}, \right. \\
\Delta(\sigma(1), \sigma(2) | \sigma(3), \sigma(4)) \\
\leq \frac{1}{2} \left[ I(\sigma(3); \sigma(4)) + I(\sigma(1); \sigma(2), \sigma(3)) \\
- I(\sigma(1); \sigma(2), \sigma(4)) \\
+ I(\sigma(3); \sigma(1), \sigma(2)) \right] \left. \right\}
$$

- So, $\tilde{F}_4$ are the polymatroidal functions that also satisfy the additional information theoretic inequality and so are a subclass that includes the entropy functions.
Consequences of this Inequality

Theorem 3.6

\[ \Gamma_n^H \neq \Gamma_n \]  \hfill (88)

Proof.
Consequences of this Inequality

Theorem 3.6

\[ \Gamma_n^H \neq \Gamma_n \]  \hspace{1cm} (88)

Proof.

- Define a function \( F \) as follows:

\[
\begin{align*}
F(\emptyset) &= 0 \\
F(X) &= F(Y) = F(Z) = F(U) = 2a \geq 0 \\
F(X, Y) &= 4a \\
F(X, U) &= F(X, Z) = F(Y, U) = F(Y, Z) = F(Z, U) = 3a \\
F(X, Y, Z) &= F(X, Y, U) = F(X, Z, U) \\
&= F(Y, Z, U) = F(X, Y, Z, U) = 4a
\end{align*}
\]
Consequences of this Inequality

Proof.

- This function is polymatroidal (i.e., $F \in \Gamma_4$) but it is not in the class that contains the entropy function (i.e., $F \notin \tilde{\Gamma}_4$) (exercise: verify this)
Consequences of this Inequality

Proof.

- This function is polymatroidal (i.e., $F \in \Gamma_4$) but it is not in the class that contains the entropy function (i.e., $F \notin \tilde{\Gamma}_4$) (exercise: verify this)

- Therefore, the polymatroidal class is strictly greater than the class of entropy functions.