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)
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.
Problem sets 1 and 2

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
Universal Compression

While $K(x)$ is algorithmic, and sometimes relates to entropy, it is not a practical measure since we can’t compute it.
Universal Compression

- While $K(x)$ is algorithmic, and sometimes relates to entropy, it is not a practical measure since we can’t compute it.
- Are there any purely algorithmic forms of compression (besides $K$) that can be shown to relate to $H$ and, ideally, can be shown to compress to the entropy rate?
Universal Compression

While $K(x)$ is algorithmic, and sometimes relates to entropy, it is not a practical measure since we can't compute it.

Are there any purely algorithmic forms of compression (besides $K$) that can be shown to relate to $H$ and, ideally, can be shown to compress to the entropy rate?

Note, that if it is algorithmic, it would be useful if it doesn’t need to explicitly compute the probability distribution governing the symbols.
While $K(x)$ is algorithmic, and sometimes relates to entropy, it is not a practical measure since we can’t compute it.

Are there any purely algorithmic forms of compression (besides $K$) that can be shown to relate to $H$ and, ideally, can be shown to compress to the entropy rate?

Note, that if it is algorithmic, it would be useful if it doesn’t need to explicitly compute the probability distribution governing the symbols.

I.e., do there exist compression algorithms that do not use the probability distribution but still give the entropy rate?
Lempel Ziv Encoding

- To encode, we give the location of the prefix (which is everything but the last symbol) and then append that index with the last symbol. Use 0 as a null pointer, indicating the string didn’t occur before.
Lempel Ziv Encoding

- To encode, we give the location of the prefix (which is everything but the last symbol) and then append that index with the last symbol. Use 0 as a null pointer, indicating the string didn’t occur before.

- **Ex:** for the string “this is the thing”, we have the following parse and encoding:

| phrase | t | h | i | s | | is | | | he | | | th | | | in | | | g |
|--------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| encoding | (0,t) | (0,h) | (0,i) | (0,s) | (0,\textunderscore) | (3,s) | (5,t) | (2,e) | (7,h) | (3,n) | (0,g) |
Lempel Ziv Encoding

- To encode, we give the location of the prefix (which is everything but the last symbol) and then append that index with the last symbol. Use 0 as a null pointer, indicating the string didn’t occur before.

- Ex: for the string “this is the thing”, we have the following parse and encoding:

<table>
<thead>
<tr>
<th>phrase</th>
<th>t</th>
<th>h</th>
<th>i</th>
<th>s</th>
<th>th</th>
<th>is</th>
<th>th</th>
<th>he</th>
<th>th</th>
<th>in</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>position</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>encoding</td>
<td>(0,t)</td>
<td>(0,h)</td>
<td>(0,i)</td>
<td>(0,s)</td>
<td>(0, th)</td>
<td>(3,s)</td>
<td>(5,t)</td>
<td>(2,e)</td>
<td>(7,h)</td>
<td>(3,n)</td>
<td>(0,g)</td>
</tr>
</tbody>
</table>

- Ex: Binary string “1011010100010 . . . ”, we have the following parse

| phrase | 1 | 0 | 11 | 01 | 010 | 00 | 10 |
| position | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| encoding | (0,1) | (0,0) | (1,1) | (2,1) | (4,0) | (2,0) | (1,0) |
Lempel-Ziv

- Assume source alphabet is binary ($\mathcal{X} = \{0, 1\}$).
Lempel-Ziv

- Assume source alphabet is binary ($\mathcal{X} = \{0, 1\}$).
- Define:
Lempel-Ziv

- Assume source alphabet is binary ($\mathcal{X} = \{0, 1\}$).
- Define:

**Definition 2.1**

A parsing $S$ of a binary string $x_1, x_2, \ldots, x_n$ is a division of the string into phrases, separated by commas.
Lempel-Ziv

- Assume source alphabet is binary ($\mathcal{X} = \{0, 1\}$).
- Define:

**Definition 2.1**

A parsing $S$ of a binary string $x_1, x_2, \ldots, x_n$ is a division of the string into phrases, separated by commas.

**Definition 2.2**

A distinct parsing is a parsing where no two phrases are identical.
Lempel-Ziv

- Assume source alphabet is binary ($\mathcal{X} = \{0, 1\}$).
- Define:

**Definition 2.1**

A parsing $S$ of a binary string $x_1, x_2, \ldots, x_n$ is a division of the string into phrases, separated by commas.

**Definition 2.2**

A **distinct** parsing is a parsing where no two phrases are identical.

- Ex: 01101101. A parsing is 0,11,0,11,01 but this is not distinct.
Lempel-Ziv

- Assume source alphabet is binary ($\mathcal{X} = \{0, 1\}$).
- Define:

**Definition 2.1**

A parsing $S$ of a binary string $x_1, x_2, \ldots, x_n$ is a division of the string into phrases, separated by commas.

**Definition 2.2**

A **distinct** parsing is a parsing where no two phrases are identical.

- Ex: 01101101. A parsing is 0,11,0,11,01 but this is not distinct.
- A distinct parsing of 01101101 would be 0,1,10,11,01
Lempel-Ziv

- Assume source alphabet is binary ($\mathcal{X} = \{0, 1\}$).
- Define:

**Definition 2.1**
A parsing $S$ of a binary string $x_1, x_2, \ldots, x_n$ is a division of the string into phrases, separated by commas.

**Definition 2.2**
A **distinct** parsing is a parsing where no two phrases are identical.

- Ex: $01101101$. A parsing is $0, 11, 0, 11, 01$ but this is not distinct.
- A distinct parsing of $01101101$ would be $0, 1, 10, 11, 01$
- Of course, Lempel-Ziv produces a distinct parsing of the source sequence.
Lempel-Ziv

Let $c(n)$ be the number of phrases in a LZ parsing of the string of length $n$. Thus, $c(n)$ depends on the string $x_{1:n}$ ($c = c(n) = c(x_{1:n})$).
Lempel-Ziv

Let $c(n)$ be the number of phrases in a LZ parsing of the string of length $n$. Thus, $c(n)$ depends on the string $x_{1:n}$ ($c = c(n) = c(x_{1:n})$).

After compression, we have a sequence of $c(n)$ pairs of numbers of the form (pointer, bit) where each pointer requires $\lceil \log c(n) \rceil$ bits.
Lempel-Ziv

- Let $c(n)$ be the number of phrases in a LZ parsing of the string of length $n$. Thus, $c(n)$ depends on the string $x_{1:n}$ ($c = c(n) = c(x_{1:n})$).

- After compression, we have a sequence of $c(n)$ pairs of numbers of the form (pointer, bit) where each pointer requires $\lceil \log c(n) \rceil$ bits.

- Length of the compressed sequence is then

$$c(n)(\log c(n) + 1) \text{ bits} \quad (1)$$
Let $c(n)$ be the number of phrases in a LZ parsing of the string of length $n$. Thus, $c(n)$ depends on the string $x_{1:n}$ ($c = c(n) = c(x_{1:n})$).

After compression, we have a sequence of $c(n)$ pairs of numbers of the form (pointer, bit) where each pointer requires $\lceil \log c(n) \rceil$ bits.

Length of the compressed sequence is then

$$c(n)(\log c(n) + 1) \text{ bits}$$

Can we show that LZ compression achieves the entropy rate? How?
Lempel-Ziv

- Let $c(n)$ be the number of phrases in a LZ parsing of the string of length $n$. Thus, $c(n)$ depends on the string $x_{1:n}$ ($c = c(n) = c(x_{1:n})$).
- After compression, we have a sequence of $c(n)$ pairs of numbers of the form (pointer, bit) where each pointer requires $\lceil \log c(n) \rceil$ bits.
- Length of the compressed sequence is then

$$c(n)(\log c(n) + 1) \text{ bits}$$

(1)

- Can we show that LZ compression achieves the entropy rate? How?
- I.e., our goal is to show that:

$$\frac{c(n)(\log c(n) + 1)}{n} \to H(X)$$

(2)

for stationary ergodic sequence $x_1, x_2, \ldots, x_n$. 
Lemma 2.3

The number of phrases $c(n)$ in a distinct parsing of binary sequence $x_{1:n}$ satisfies:

$$c(n) \leq \frac{n}{(1 - \epsilon_n) \log n}$$ (3)

where $\epsilon \to 0$ and $n \to \infty$. That is

$$c(n) \leq \frac{n}{\log n} \left(1 + o(1)\right)$$ (4)
Lemma 2.4

Let $Z$ be a positive integer valued random variable with mean $\mu$. Then we can bound $H(Z)$ as:

$$H(Z) \leq (\mu + 1) \log(\mu + 1) - \mu \log \mu$$ (5)
Simple Lemma

**Lemma 2.4**

Let $Z$ be a positive integer valued random variable with mean $\mu$. Then we can bound $H(Z)$ as:

$$H(Z) \leq (\mu + 1) \log(\mu + 1) - \mu \log \mu$$

(5)
Ergodicity Intuition and Definition

- Intuition: Ergodic processes cannot be separated into different persistent modes of behavior.
Ergodicity Intuition and Definition

- **Intuition**: Ergodic processes cannot be separated into different persistent modes of behavior.
- **This yields the result that time average are the same as ensemble averages**
Ergodicity Intuition and Definition

- Intuition: Ergodic processes cannot be separated into different persistent modes of behavior.
- This yields the result that time average are the same as ensemble averages.
- Let $x = \{x_i\}$ be a sequence of source letters.
Ergodicity Intuition and Definition

- **Intuition:** Ergodic processes cannot be separated into different persistent modes of behavior.
- This yields the result that time average are the same as ensemble averages.
- Let \( x = \{x_i\} \) be a sequence of source letters.
- Let \( T^\ell x \) be the shifted sequence in time by \( \ell \) positions. I.e., if \( T^\ell x = x' \) then \( x'_i = x_{i+\ell}, \forall i. \)
Ergodicity Intuition and Definition

- Intuition: Ergodic processes cannot be separated into different persistent modes of behavior.
- This yields the result that time average are the same as ensemble averages
- Let \( x = \{ x_i \} \) be a sequence of source letters.
- Let \( T^\ell x \) be the shifted sequence in time by \( \ell \) positions. I.e., if \( T^\ell x = x' \) then \( x'_i = x_{i+\ell}, \forall i \).
- Let \( S \) be an infinite set of sequences of source letters, i.e.,
  \[ S = \{ x : x \text{ is a sequence of source letters} \} \]
Ergodicity Intuition and Definition

- Intuition: Ergodic processes cannot be separated into different persistent modes of behavior.
- This yields the result that time average are the same as ensemble averages.
- Let $x = \{x_i\}$ be a sequence of source letters.
- Let $T^\ell x$ be the shifted sequence in time by $\ell$ positions. I.e., if $T^\ell x = x'$ then $x'_i = x_{i+\ell}$, $\forall i$.
- Let $S$ be an infinite set of sequences of source letters, i.e., $S = \{x : x$ is a sequence of source letters $\}$.
- And $T^\ell S$ is the set of all sequences shifted by $\ell$ positions. I.e., if $x' = T^\ell x$, then $x' \in T^\ell S$ if $x \in S$. 


Intuition: Ergodic processes cannot be separated into different persistent modes of behavior.

This yields the result that time averages are the same as ensemble averages.

- Let $x = \{x_i\}$ be a sequence of source letters.
- Let $T^\ell x$ be the shifted sequence in time by $\ell$ positions. I.e., if $T^\ell x = x'$ then $x'_i = x_{i+\ell}$, $\forall i$.

- Let $S$ be an infinite set of sequences of source letters, i.e., $S = \{x : x$ is a sequence of source letters $\}$.
- And $T^\ell S$ is the set of all sequences shifted by $\ell$ positions. I.e., if $x' = T^\ell x$, then $x' \in T^\ell S$ if $x \in S$.

Definition: Invariant. A set $S$ is invariant if $T^\ell S = S$ for all $\ell$.  

---

Prof. Jeff Bilmes

Ergodicity Intuition and Definition

- Intuition: Ergodic processes cannot be separated into different persistent modes of behavior.
- This yields the result that time average are the same as ensemble averages.
- Let \( x = \{ x_i \} \) be a sequence of source letters.
- Let \( T^\ell x \) be the shifted sequence in time by \( \ell \) positions. I.e., if \( T^\ell x = x' \) then \( x'_i = x_{i+\ell} \), \( \forall i \).
- Let \( S \) be an infinite set of sequences of source letters, i.e., \( S = \{ x : x \text{ is a sequence of source letters} \} \).
- And \( T^\ell S \) is the set of all sequences shifted by \( \ell \) positions. I.e., if \( x' = T^\ell x \), then \( x' \in T^\ell S \) if \( x \in S \).
- Definition: Invariant. A set \( S \) is invariant if \( T^\ell S = S \) for all \( \ell \).
- Example: The set of all sequences of a discrete alphabet source is invariant.
Ergodicity Intuition and Definition

- Intuition: Ergodic processes cannot be separated into different persistent modes of behavior.
- This yields the result that time average are the same as ensemble averages.
- Let \( x = \{x_i\} \) be a sequence of source letters.
- Let \( T^\ell x \) be the shifted sequence in time by \( \ell \) positions. I.e., if \( T^\ell x = x' \) then \( x'_i = x_{i+\ell}, \forall i \).
- Let \( S \) be an infinite set of sequences of source letters, i.e., \( S = \{x : x \text{ is a sequence of source letters}\} \).
- And \( T^\ell S \) is the set of all sequences shifted by \( \ell \) positions. I.e., if \( x' = T^\ell x \), then \( x' \in T^\ell S \) if \( x \in S \).
- Definition: **Invariant**. A set \( S \) is invariant if \( T^\ell S = S \) for all \( \ell \).
- Example: The set of all sequences of a discrete alphabet source is invariant.
- Example: \( S = \{\ldots 000\ldots, \ldots 111\ldots\} \)
Ergodicity Intuition and Definition

Ex: For any sequence, the set

\[ \{ \ldots, T^{-2}x, T^{-1}x, T^0x, T^1x, T^2x, \ldots \} \]

(which is the set of all shifts of a sequence) is invariant.

\[ (6) \]
Ergodicity Intuition and Definition

- Ex: For any sequence, the set
  \[
  \{ \ldots, T^{-2}x, T^{-1}x, T^0x, T^1x, T^2x, \ldots \},
  \]
  (which is the set of all shifts of a sequence) is invariant.

- Definition: A discrete stationary source is **ergodic** if every invariant set of sequences either has probability 1 or probability 0. I.e.,
  \[
  \Pr\left\{ S : T^\ell S = S, \forall \ell \right\} = 1 \text{ or } 0 \ \forall S
  \]
Ergodicity Intuition and Definition

- Ex: For any sequence, the set

\[ \{ \ldots, T^{-2}x, T^{-1}x, T^0x, T^1x, T^2x, \ldots \} , \]  

(which is the set of all shifts of a sequence) is invariant.

- Definition: A discrete stationary source is **ergodic** if every invariant set of sequences either has probability 1 or probability 0. I.e.,

\[ \Pr\{ S : T^\ell S = S, \forall \ell \} = 1 \text{ or } 0 \quad \forall S \]  

This implies that time averages give us ensemble averages.
Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary ergodic process with probability mass function $p(x_1, x_2, \ldots, x_n)$. 
kth order Markov chain approximation

- Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary ergodic process with probability mass function $p(x_1, x_2, \ldots, x_n)$.
- For fixed integer $k$, define $k'$th order Markov approximation to $p$ as:

$$Q_k(x_{-(k-1)}, \ldots, x_0, x_1, \ldots, x_n) \triangleq p(x_{-(k-1):0}) \prod_{j=1}^{n} p(x_j|x_{j-k:j-1})$$  \hspace{1cm} (8)

where we think of $x_{-(k-1):0}$ as state “0” and $x_{j-k:j-1}$ as state “$j$”, and where $x_{i:j} = \{x_i, x_{i+1}, \ldots, x_j\}$ with $i \leq j$. 
Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary ergodic process with probability mass function $p(x_1, x_2, \ldots, x_n)$.

For fixed integer $k$, define $k'$th order Markov approximation to $p$ as:

$$Q_k(x_{-(k-1)}, \ldots, x_0, x_1, \ldots, x_n) \triangleq p(x_{-(k-1):0}) \prod_{j=1}^{n} p(x_j|x_{j-k:j-1}) \quad (8)$$

where we think of $x_{-(k-1):0}$ as state “0” and $x_{j-k:j-1}$ as state “$j$”, and where $x_{i:j} = \{x_i, x_{i+1}, \ldots, x_j\}$ with $i \leq j$.

Note also that $p(x_n|x_{n-k:n-1})$ is also stationary and ergodic since $p(x_{n-k:n})$ also is.
We then get:

\[\sum_{j=1}^{n} \log p(x_j| x_{j-k+1:k-1}) = -\frac{1}{n} \sum_{j=1}^{n} \log \frac{1}{Q_k(x_1,\ldots,x_n|x_{-k+1})} = -\frac{1}{n} \sum_{j=1}^{n} \log p(x_j| x_{j-k+1:k-1}) = H(X_j| X_{j-k+1:k-1})\] (9)

(10)

which follows since the process is stationary ergodic.

We will show that:

\[\limsup_{n \to \infty} c(n) \log c(n) / n \leq H(X_j| X_{j-k+1:k-1}) \to H(X)\]

(11)

where \(H(X)\) is the entropy rate of the stochastic process.

But first . . .
We then get:

\[- \frac{1}{n} \log Q_k(x_1, \ldots, x_n | x_{-(k-1):0})\]

(9)

\[-\mathbb{E} \log p(X_j | X_{j-k:j-1}) \leq H(X_j | X_{j-k:j-1})\]

(10)

which follows since the process is stationary ergodic.

We will show that:

\[\lim \sup_{n \to \infty} \frac{c(n) \log c(n)}{n} \leq H(X_j | X_{j-k:j-1})\]

(11)

where \(H(X)\) is the entropy rate of the stochastic process.
We then get:

\[- \frac{1}{n} \log Q_k(x_1, \ldots, x_n | x_{-(k-1):0}) = -\frac{1}{n} \sum_{j=1}^{n} \log p(x_j | x_{j-k:j-1})\]

(9)

(10)
We then get:

\[- \frac{1}{n} \log Q_k(x_1, \ldots, x_n | x_{-(k-1):0}) = - \frac{1}{n} \sum_{j=1}^{n} \log p(x_j | x_{j-k:j-1})\]

(9)

\[- \frac{1}{n} \log p(x_1, \ldots, x_n | x_{-(k-1):0}) = - \frac{1}{n} \sum_{j=1}^{n} \log p(x_j | x_{j-k:j-1})\]

(10)
We then get:

\[ -\frac{1}{n} \log Q_k(x_1, \ldots, x_n|x_{-(k-1):0}) = -\frac{1}{n} \sum_{j=1}^{n} \log p(x_j|x_{j-k:j-1}) \]  

(9)

\[ \rightarrow -E \log p(X_j|X_{j-k:j-1}) = H(X_j|X_{j-k:j-1}) \]  

(10)
\( k \)th order Markov chain approximation

- We then get:

\[
- \frac{1}{n} \log Q_k(x_1, \ldots, x_n|x_{-(k-1):0}) = - \frac{1}{n} \sum_{j=1}^{n} \log p(x_j|x_{j-k:j-1})
\]  

(9)

\[
\rightarrow -E \log p(X_j|X_{j-k:j-1}) = H(X_j|X_{j-k:j-1})
\]  

(10)

which follows since the process is stationary ergodic.
$k$th order Markov chain approximation

- We then get:

$$- \frac{1}{n} \log Q_k(x_1, \ldots, x_n|x_{-(k-1):0}) = - \frac{1}{n} \sum_{j=1}^{n} \log p(x_j|x_{j-k:j-1})$$  

(9)

$$\rightarrow -E \log p(X_j|X_{j-k:j-1}) = H(X_j|X_{j-k:j-1})$$  

(10)

which follows since the process is stationary ergodic.

- We will show that:

$$\limsup_{n \to \infty} \frac{c(n) \log c(n)}{n} \leq H(X_j|X_{j-k:j-1}) \to H(\mathcal{X})$$  

(11)

where $H(\mathcal{X})$ is the entropy rate of the stochastic process.
$k$th order Markov chain approximation

- We then get:

$$- \frac{1}{n} \log Q_k(x_1, \ldots, x_n | x_{-(k-1):0}) = - \frac{1}{n} \sum_{j=1}^{n} \log p(x_j | x_{j-k:j-1})$$

(9)

$$\rightarrow -E \log p(X_j | X_{j-k:j-1}) = H(X_j | X_{j-k:j-1})$$

(10)

which follows since the process is stationary ergodic.

- We will show that:

$$\lim_{n \to \infty} \sup_n \frac{c(n) \log c(n)}{n} \leq H(X_j | X_{j-k:j-1}) \to H(\mathcal{X})$$

(11)

where $H(\mathcal{X})$ is the entropy rate of the stochastic process. But first
... a bit more notation

- We consider $x_{1:n}$ parsed into $c$ distinct phrases, $x_i \in \mathcal{X}$, as follows
  
  $x_{1:n} = y_1 y_2 \ldots y_c$ where $y_c$ is a subsequence.
a bit more notation

- We consider $x_{1:n}$ parsed into $c$ distinct phrases, $x_i \in \mathcal{X}$, as follows
  $x_{1:n} = y_1 y_2 \ldots y_c$ where $y_c$ is a subsequence.
- Let $v_i$ be the index into $x_{1:n}$ of the start of the $i$'th phrase. I.e.,
  $y_i = x_{v_i:v_i+1-1}$ for all $i = 1, 2, \ldots, c$. 
We consider $x_{1:n}$ parsed into $c$ distinct phrases, $x_i \in \mathcal{X}$, as follows
\[ x_{1:n} = y_1 y_2 \ldots y_c \] where $y_c$ is a subsequence.

Let $v_i$ be the index into $x_{1:n}$ of the start of the $i$'th phrase. I.e.,
\[ y_i = x_{v_i:v_{i+1}-1} \] for all $i = 1, 2, \ldots, c$.

Let $s_i = x_{v_i-k:v_i-1}$ which is the $k$ symbols of $x_{1:n}$ preceding $y_i$ with
\[ s_1 = x_{-(k-1):0}, \] so $s_i$ is the “state” or the prefix of the $i$'th phrase.
...a bit more notation

- We consider $x_{1:n}$ parsed into $c$ distinct phrases, $x_i \in X$, as follows $x_{1:n} = y_1 y_2 \ldots y_c$ where $y_c$ is a subsequence.
- Let $v_i$ be the index into $x_{1:n}$ of the start of the $i$’th phrase. I.e., $y_i = x_{v_i:v_{i+1}-1}$ for all $i = 1, 2, \ldots, c$.
- Let $s_i = x_{v_i-k:v_i-1}$ which is the $k$ symbols of $x_{1:n}$ preceding $y_i$ with $s_1 = x_{-(k-1):0}$, so $s_i$ is the “state” or the prefix of the $i$’th phrase.
- Let $c_{\ell s}$ be the number of phrases in $x_{1:n}$ with length $\ell$ and that have preceding state $s$, for $\ell = 1, 2, \ldots, n$ and $s \in X^k$ (length $k$ strings). I.e., $c_{\ell s}$ is the number of phrases of length $\ell$ with prefix $s$,
We consider $x_{1:n}$ parsed into $c$ distinct phrases, $x_i \in \mathcal{X}$, as follows:

$x_{1:n} = y_1 y_2 \ldots y_c$ where $y_c$ is a subsequence.

Let $v_i$ be the index into $x_{1:n}$ of the start of the $i$'th phrase. I.e.,

$y_i = x_{v_i:v_i+1-1}$ for all $i = 1, 2, \ldots, c$.

Let $s_i = x_{v_i-k:v_i-1}$ which is the $k$ symbols of $x_{1:n}$ preceding $y_i$ with

$s_1 = x_{-(k-1):0}$, so $s_i$ is the “state” or the prefix of the $i$'th phrase.

Let $c_{\ell s}$ be the number of phrases in $x_{1:n}$ with length $\ell$ and that have

preceding state $s$, for $\ell = 1, 2, \ldots, n$ and $s \in \mathcal{X}^k$ (length $k$ strings).

I.e., $c_{\ell s}$ is the number of phrases of length $\ell$ with prefix $s$, or,

$$c_{\ell s} = |\{ i \in \{1, \ldots, c\} : |y_i| = \ell, s_i = s \}| \quad (12)$$
...a bit more notation

- We consider $x_{1:n}$ parsed into $c$ distinct phrases, $x_i \in X$, as follows
  $x_{1:n} = y_1 y_2 \ldots y_c$ where $y_c$ is a subsequence.
- Let $v_i$ be the index into $x_{1:n}$ of the start of the $i$'th phrase. I.e.,
  $y_i = x_{v_i:v_{i+1}-1}$ for all $i = 1, 2, \ldots, c$.
- Let $s_i = x_{v_i-k:v_{i-1}}$ which is the $k$ symbols of $x_{1:n}$ preceding $y_i$ with
  $s_1 = x_{-(k-1):0}$, so $s_i$ is the “state” or the prefix of the $i$'th phrase.
- Let $c_{\ell,s}$ be the number of phrases in $x_{1:n}$ with length $\ell$ and that have
  preceding state $s$, for $\ell = 1, 2, \ldots, n$ and $s \in X^k$ (length $k$ strings).
  I.e., $c_{\ell,s}$ is the number of phrases of length $\ell$ with prefix $s$, or,
  \[
  c_{\ell,s} = |\{i \in \{1, \ldots, c\} : |y_i| = \ell, s_i = s\}|
  \] (12)
- Then $\sum_{\ell,s} c_{\ell,s} = c = c(n)$, where $c = c(n)$ is the total number of
  phrases in a distinct parsing of a sequence of length $n$. 
...a bit more notation and a lemma

- And \( \sum_{\ell,s} \ell c_{\ell s} = n \) which is the total string length.
... a bit more notation and a lemma

- And $\sum_{\ell,s} \ell c_{\ell s} = n$ which is the total string length.
- Key question: Can we relate (or at least bound) the probability to some deterministic aspect of the parsing?
...a bit more notation and a lemma

- And $\sum_{\ell,s} \ell c_{\ell s} = n$ which is the total string length.
- Key question: Can we relate (or at least bound) the probability to some deterministic aspect of the parsing? Yes.
And $\sum_{\ell,s} \ell c_{\ell s} = n$ which is the total string length.

Key question: Can we relate (or at least bound) the probability to some deterministic aspect of the parsing? Yes.

We have the following lemma:
...a bit more notation and a lemma

- And $\sum_{\ell,s} \ell c_{\ell s} = n$ which is the total string length.
- Key question: Can we relate (or at least bound) the probability to some deterministic aspect of the parsing? Yes.
- We have the following lemma:

**Lemma 3.1 (Ziv's inequality)**

For any distinct parsing (which includes the LZ parsing) of the string $x_1:n$, we have:

$$\log Q_k(x_1, x_2, \ldots, x_n | s_1) \leq - \sum_{\ell, s} c_{\ell s} \log c_{\ell s}$$

(13)

Note that the bound is independent of $Q$ and depends only on $c_{\ell s}$, which is the number of phrases of length $\ell$ with prefix (state) $s$. Key idea: as there is more diversity in string $x_1:n$, the max possible probability decreases. "distinct" $y_i$’s increase diversity.
... a bit more notation and a lemma

- And \( \sum_{\ell, s} \ell c_{\ell s} = n \) which is the total string length.
- Key question: Can we relate (or at least bound) the probability to some deterministic aspect of the parsing? Yes.
- We have the following lemma:

**Lemma 3.1 (Ziv's inequality)**

For any **distinct** parsing (which includes the LZ parsing) of the string \( x_1^n \), we have:

\[
\log Q_k(x_1, x_2, \ldots, x_n | s_1) \leq - \sum_{\ell, s} c_{\ell s} \log c_{\ell s}
\]  

(13)

- Note that the bound is independent of \( Q \) and depends only on \( c_{\ell s} \), which is the number of phrases of length \( \ell \) with prefix (state) \( s \).
... a bit more notation and a lemma

- And $\sum_{\ell,s} \ell c_{\ell s} = n$ which is the total string length.
- Key question: Can we relate (or at least bound) the probability to some deterministic aspect of the parsing? Yes.
- We have the following lemma:

**Lemma 3.1 (Ziv’s inequality)**

> For any **distinct** parsing (which includes the LZ parsing) of the string $x_1:n$, we have:

$$\log Q_k(x_1, x_2, \ldots x_n | s_1) \leq -\sum_{\ell,s} c_{\ell s} \log c_{\ell s} \tag{13}$$

- Note that the bound is independent of $Q$ and depends only on $c_{\ell s}$, which is the number of phrases of length $\ell$ with prefix (state) $s$.
- Key idea: as there is more diversity in string $x_1:n$, the max possible probability decreases. “distinct” $y_i$’s increase diversity.
Proof of Ziv’s inequality

First, we have:

\[ Q_k(x_1:n|s_1) = Q_k(y_1:c|s_1) = \prod_{i=1}^{c} p(y_i|s_i) \]  \hspace{1cm} (14)

where the r.h.s. follows because of the \( k' \)th order Markov assumption, that \( y_i \) depends on nothing else in the past given the immediate past \( s_i \).
Proof of Ziv’s inequality

proof of Ziv’s inequality.

First, we have:

$$Q_k(x_{1:n}|s_1) = Q_k(y_{1:c}|s_1) = \prod_{i=1}^{c} p(y_i|s_i)$$  \hspace{1cm} (14)

where the r.h.s. follows because of the $k$’th order Markov assumption, that $y_i$ depends on nothing else in the past given the immediate past $s_i$.

This gives

$$\log Q_k(x_1, x_2, \ldots, x_n|s_1) = \sum_i \log p(y_i|s_i)$$  \hspace{1cm} (15)

$$= \sum_{\ell, s} \sum_{i:|y_i|=\ell, s_i=s} \log p(y_i|s_i) = \sum_{\ell, s} c_{\ell s} \sum_{i:|y_i|=\ell, s_i=s} \frac{1}{c_{\ell s}} \log p(y_i|s_i) \ldots$$
Proof of Ziv’s inequality

But we have mixture, since

$$\sum_{i:|y_i| = \ell, s_i = s} \frac{1}{c_{\ell s}} = 1$$

(16)

since $c_{\ell s}$ is the number of length $\ell$ phrases with prefix $s$, and the sum sums over all of them.

...
Proof of Ziv’s inequality.

But we have mixture, since

$$\sum_{i: |y_i| = \ell, s_i = s} \frac{1}{c_\ell s} = 1$$

(16)

since $c_\ell s$ is the number of length $\ell$ phrases with prefix $s$, and the sum sums over all of them.

So, using Jensen’s inequality, we get that:

$$\sum_{\ell, s} c_\ell s \sum_{i: |y_i| = \ell, s_i = s} \frac{1}{c_\ell s} \log p(y_i | s_i) \leq \sum_{\ell, s} c_\ell s \log \left( \sum_{i: |y_i| = \ell, s_i = s} \frac{1}{c_\ell s} p(y_i | s_i) \right)$$

(17)
Proof of Ziv’s inequality

But the \( y_i \)'s are distinct (no double counts) which means that

\[
\sum_{i: |y_i| = \ell, s_i = s} p(y_i | s_i) \leq 1
\]  

(18)
Proof of Ziv’s inequality

But the $y_i$’s are distinct (no double counts) which means that

$$\sum_{i: |y_i| = \ell, s_i = s} p(y_i|s_i) \leq 1 \quad (18)$$

this then yields our result, that

$$\log Q_k(x_1, x_2, \ldots, x_n|s_1) \leq \sum_{\ell, s} c_{\ell s} \log \left( \sum_{i: |y_i| = \ell, s_i = s} \frac{1}{c_{\ell s}} p(y_i|s_i) \right) \quad (19)$$

$$\leq \sum_{\ell, s} c_{\ell s} \log \left( \frac{1}{c_{\ell s}} \right) \quad (20)$$
Recall limsup/liminf (from lecture 28)

- Recall,

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n>0} \left( \sup_{k>n} a_k \right) = \inf S
\]

where

\[S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{ a_n, a_{n+1}, \ldots, \} \}.\]
Recall \( \limsup \) and \( \liminf \) (from lecture 28)

- Recall,

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \tag{21}
\]

where

\[ S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{ a_n, a_{n+1}, \ldots, \} \}. \]

- For example, while \( \lim_{x \to \infty} \sin(x) \) does not exist,
  \( \limsup_{x \to \infty} \sin(x) = 1. \)
Recall limsup/liminf (from lecture 28)

• Recall,

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \tag{21}
\]

where

\[ S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{ a_n, a_{n+1}, \ldots, \} \}. \]

• For example, while \( \lim_{x \to \infty} \sin(x) \) does not exist, \( \limsup_{x \to \infty} \sin(x) = 1 \).

• Also, \( \limsup_{x \to \infty} (\sin(x) - \sin^2(x)) = \)
recall limsup/liminf (from lecture 28)

- Recall,

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n>0} \left( \sup_{k>n} a_k \right) = \inf S \tag{21}
\]

where

\[S = \{a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{a_n, a_{n+1}, \ldots, \}\}\].

- For example, while \(\lim_{x \to \infty} \sin(x)\) does not exist, \(\limsup_{x \to \infty} \sin(x) = 1\).

- Also, \(\limsup_{x \to \infty} (\sin(x) - \sin^2(x)) = \)
recall limsup/liminf (from lecture 28)

- Recall,

\[ \limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \quad (21) \]

where

\[ S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{a_n, a_{n+1}, \ldots, \} \} \].

- For example, while \( \lim_{x \to \infty} \sin(x) \) does not exist, \( \limsup_{x \to \infty} \sin(x) = 1 \).

- Also, \( \limsup_{x \to \infty} (\sin(x) - \sin^2(x)) = 1/4 \).

- Thus, \( \limsup \) allows for oscillation in the sequences and in some sense \( \limsup \) asks for infimum convergence in the local maxima (or perhaps better, “reverse-time cumulative” local maxima).
recall limsup/liminf (from lecture 28)

Recall,

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S
\]

(21)

where

\[ S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{ a_n, a_{n+1}, \ldots, \} \} \]

For example, while \( \lim_{x \to \infty} \sin(x) \) does not exist,
\( \limsup_{x \to \infty} \sin(x) = 1 \).

Also, \( \limsup_{x \to \infty} (\sin(x) - \sin^2(x)) = 1/4 \).

Thus, \( \limsup \) allows for oscillation in the sequences and in some sense \( \limsup \) asks for infimum convergence in the local maxima (or perhaps better, “reverse-time cumulative” local maxima).

Also,

\[
\liminf_{n \to \infty} a_n \triangleq \sup_{n > 0} \left( \inf_{k > n} a_k \right)
\]

(22)

so \( \liminf \) asks for supremum convergence in the local minima.
Main Theorem

Theorem 3.2

Let $X_{1:n}$ be a stationary ergodic process with entropy rate $H(\mathcal{X})$, and $c(n)$ be the number of phrases in a distinct parsing of a sample of length $n$ from this process. Then

$$\limsup_{n \to \infty} \frac{c(n) \log c(n)}{n} \leq H(\mathcal{X}) \quad \text{w.p.1} \quad (23)$$

Proof.

- We write $c$ for $c(n)$. ...
Main Theorem

Theorem 3.2

Let $X_{1:n}$ be a stationary ergodic process with entropy rate $H(\mathcal{X})$, and $c(n)$ be the number of phrases in a distinct parsing of a sample of length $n$ from this process. Then

$$\limsup_{n \to \infty} \frac{c(n) \log c(n)}{n} \leq H(\mathcal{X}) \quad \text{w.p.1} \quad (23)$$

Proof.

- We write $c$ for $c(n)$. By Ziv’s inequality, and since $\sum_{\ell,s} c_{\ell s} = c$:

$$\log Q_k(x_1, x_2, \ldots, x_n | s_1) \leq - \sum_{\ell,s} c_{\ell s} \log \frac{c_{\ell s} c}{c} \quad (24)$$

$$= -c \log c - c \sum_{\ell,s} \frac{c_{\ell s} c}{c} \log \frac{c_{\ell s} c}{c} \quad (25)$$
... proof continued.

- Lets write \( \pi_{\ell s} = \frac{c_{\ell s}}{c} \), which can be treated as a probability since \( \pi_{\ell s} \geq 0 \) and \( \sum_{\ell s} \pi_{\ell s} = 1 \).
... proof continued.

- Let's write \( \pi_{\ell s} = \frac{c_{\ell s}}{c} \), which can be treated as a probability since \( \pi_{\ell s} \geq 0 \) and \( \sum_{\ell s} \pi_{\ell s} = 1 \).

- Then since \( \sum_{\ell s} \ell c_{\ell s} = n \), we have

\[
\sum_{\ell s} \ell \pi_{\ell s} = \frac{n}{c} \tag{26}
\]
Main Theorem

... proof continued.

- Let's write $\pi_{\ell s} = c_{\ell s}/c$, which can be treated as a probability since $\pi_{\ell s} \geq 0$ and $\sum_{\ell s} \pi_{\ell s} = 1$.
- Then since $\sum_{\ell s} \ell c_{\ell s} = n$, we have
  \[ \sum_{\ell s} \ell \pi_{\ell s} = \frac{n}{c} \]  \hfill (26)

- Define new random variables $U, V$ s.t.,
  \[ p(U = \ell, V = s) = \pi_{\ell s} \]  \hfill (27)
Main Theorem

... proof continued.

- Lets write $\pi_{\ell s} = c_{\ell s} / c$, which can be treated as a probability since $\pi_{\ell s} \geq 0$ and $\sum_{\ell s} \pi_{\ell s} = 1$.
- Then since $\sum_{\ell s} \ell c_{\ell s} = n$, we have

$$\sum_{\ell s} \ell \pi_{\ell s} = n / c \quad (26)$$

- Define new random variables $U, V$ s.t.,

$$p(U = \ell, V = s) = \pi_{\ell s} \quad (27)$$

- So that

$$EU = \sum_{\ell} \ell \pi_\ell = \sum_{\ell} \ell \sum_{s} \pi_{\ell s} = n / c \quad (28)$$
Main Theorem

... proof continued.

This immediately gives us:

\[ \log Q_k(x_1:n | s_1) \leq \sum_{\ell s} c_{\ell s} \log 1/c_{\ell s} \] (29)

\[ = cH(U, V) - c \log c \] (30)
Main Theorem

... proof continued.

- This immediately gives us:

\[
\log Q_k(x_1:n|s_1) \leq \sum_{\ell s} c_{\ell s} \log \frac{1}{c_{\ell s}} \quad (29)
\]

\[
= c H(U, V) - c \log c \quad (30)
\]

- Or

\[
- \frac{1}{n} \log Q_k(x_1:n|s_1) \geq \frac{c}{n} \log c - \frac{c}{n} H(U, V) \quad (31)
\]
Main Theorem

... proof continued.

- This immediately gives us:

\[
\log Q_k(x_{1:n} | s_1) \leq \sum_{\ell_s} c_{\ell_s} \log \frac{1}{c_{\ell_s}}
\] (29)

\[
= cH(U, V) - c \log c
\] (30)

- Or

\[
-\frac{1}{n} \log Q_k(x_{1:n} | s_1) \geq \frac{c}{n} \log c - \frac{c}{n} H(U, V)
\] (31)

\[
\Rightarrow \text{entropy rate as } k \to \infty
\]

and \( n \to \infty \)
Main Theorem

...proof continued.

- This immediately gives us:

\[
\log Q_k(x_{1:n} | s_1) \leq \sum \ell_s c_{\ell_s} \log 1/c_{\ell_s} \tag{29}
\]

\[
= cH(U, V) - c \log c \tag{30}
\]

- Or

\[
-\frac{1}{n} \log Q_k(x_{1:n} | s_1) \geq \left( \frac{c}{n} \log c \right) - \frac{c}{n} H(U, V) \tag{31}
\]

\[
\rightarrow \text{entropy rate as } k \rightarrow \infty \text{ and } n \rightarrow \infty
\]

\[
c = c(n), \text{ so this is what we wish to show converges to entropy of } X
\]
Main Theorem

... proof continued.

- This immediately gives us:

\[
\log Q_k(x_1:n|s_1) \leq \sum_{\ell s} c_{\ell s} \log 1/c_{\ell s} \tag{29}
\]

\[= cH(U,V) - c \log c \tag{30}\]

- Or

\[c = c(n), \text{ so this is what we wish to show converges to entropy of } X\]

\[- \frac{1}{n} \log Q_k(x_1:n|s_1) \geq \left\{ \frac{c}{n} \log c \right\} - \left\{ \frac{c}{n} H(U,V) \right\} \tag{31}\]

\[\rightarrow \text{ entropy rate as } k \rightarrow \infty \]

\[\text{and } n \rightarrow \infty \]

Ideally, this will \(\rightarrow 0\) as \(n \rightarrow \infty\)
Main Theorem

... proof continued.

- Now, we know we have $H(U, V) \leq H(U) + H(V)$. 

...
Now, we know we have $H(U, V) \leq H(U) + H(V)$.

Also, we have $H(V) \leq \log |\{0, 1\}|^k = k$, we can think of $V$ as a state (binary string of length $k$) variable.
...proof continued.

- And also by Lemma 2.4,
Main Theorem

... proof continued.

- And also by Lemma 2.4,

\[
H(U)
\]
Main Theorem

\[ H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU \]  

(32)

... proof continued.

And also by Lemma 2.4,

\[ H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU \]  

(33)
Main Theorem

... proof continued.

- And also by Lemma 2.4,

\[
H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU
\]  
(32)

\[
= (\frac{n}{c} + 1) \log(\frac{n}{c} + 1) - \frac{n}{c} \log \frac{n}{c}
\]  
(33)

\[
= (n + 1) \log(cn + 1) + \log(n/c) + \log(cn) - \log(cn + 1)
\]  
(34)
Main Theorem

... proof continued.

And also by Lemma 2.4,

\[
H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU
\]

\[
= \left(\frac{n}{c} + 1\right) \log\left(\frac{n}{c} + 1\right) - \frac{n}{c} \log \frac{n}{c}
\]

\[
= \frac{n}{c} \log\left(\frac{n}{c} + 1\right) + \log\left(\frac{n}{c} + 1\right) - \frac{n}{c} \log \frac{n}{c}
\]
Main Theorem

... proof continued.

- And also by Lemma 2.4,

$$H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU$$

$$= \left(\frac{n}{c} + 1\right) \log\left(\frac{n}{c} + 1\right) - \frac{n}{c} \log\frac{n}{c}$$

$$= \frac{n}{c} \log\left(\frac{n}{c} + 1\right) + \log\left(\frac{n}{c} + 1\right) - \frac{n}{c} \log\frac{n}{c}$$

$$= \frac{n}{c} \log\frac{c}{n}\left(\frac{n}{c} + 1\right) + \log\left(\frac{n}{c} + 1\right)$$

$$= \left(\frac{n}{c} + 1\right) \log\left(\frac{c}{n}\right) + \log\left(\frac{n}{c} + 1\right)$$

(32)

(33)

(34)

(35)
Main Theorem

... proof continued.

- And also by Lemma 2.4,

\[
H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU \\
= \left( \frac{n}{c} + 1 \right) \log \left( \frac{n}{c} + 1 \right) - \frac{n}{c} \log \frac{n}{c} \\
= \frac{n}{c} \log \left( \frac{n}{c} + 1 \right) + \log \left( \frac{n}{c} + 1 \right) - \frac{n}{c} \log \frac{n}{c} \\
= \frac{n}{c} \log \left( \frac{n}{c} + 1 \right) + \log \left( \frac{n}{c} + 1 \right) \\
= \frac{n}{c} \log \left( \frac{n}{c} + 1 \right) + \log \left( \frac{n}{c} + 1 \right) - \log \left( \frac{c}{n} + 1 \right)
\]
... proof continued.

And also by Lemma 2.4,

\[
H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU
\]  
\[
= \left(\frac{n}{c} + 1\right) \log \left(\frac{n}{c} + 1\right) - \frac{n}{c} \log \frac{n}{c}
\]  
\[
= \frac{n}{c} \log \left(\frac{n}{c} + 1\right) + \log \left(\frac{n}{c} + 1\right) - \frac{n}{c} \log \frac{n}{c}
\]  
\[
= \frac{n}{c} \log \frac{n}{c} + \frac{n}{c} \log (n + 1) + \log \left(\frac{n}{c} + 1\right) - \log \frac{n}{c} + 1
\]  
\[
= \left(\frac{n}{c} + 1\right) \log \left(\frac{n}{c} + 1\right) + \log \frac{n+c}{n+c+n}
\]  
\[
= (\frac{n}{c} + 1) \log (\frac{n}{c} + 1) + \log \frac{n+c}{c+n} + \log \frac{n+c}{n+c+n}
\]
Main Theorem

... proof continued.

- And also by Lemma 2.4,

\[
H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU
\]  

(32)

\[
= \left(\frac{n}{c} + 1\right) \log\left(\frac{n}{c} + 1\right) - \frac{n}{c} \log \frac{n}{c}
\]  

(33)

\[
= \frac{n}{c} \log\left(\frac{n}{c} + 1\right) + \log\left(\frac{n}{c} + 1\right) - \frac{n}{c} \log \frac{n}{c}
\]  

(34)

\[
= \frac{n}{c} \log \left(\frac{n}{c} + 1\right) + \log\left(\frac{n}{c} + 1\right)
\]  

(35)

\[
= \frac{n}{c} \log\left(\frac{n}{c} + 1\right) + \log\left(\frac{n}{c} + 1\right) + \log\left(\frac{c}{n} + 1\right) - \log\left(\frac{c}{n} + 1\right)
\]

(36)

\[
= \left(\frac{n}{c} + 1\right) \log\left(\frac{c}{n} + 1\right) + \log\left(\frac{c}{c+n} + \frac{n+c}{n}\right)
\]

(37)

...
Main Theorem

... proof continued.

1. Thus, we have

\[
\begin{align*}
\mathcal{C}_n & : \mathbb{H}(U,V) \\
& \leq \mathcal{C}_n \mathbb{H}(V) + \mathcal{C}_n \mathbb{H}(U) \\
& \leq \mathcal{C}_n k + \mathcal{C}_n \log n + (\mathcal{C}_n + 1) \log (\mathcal{C}_n n) \\
& = \mathcal{C}_n k + \mathcal{C}_n \log n + \mathcal{C}_n \log (\mathcal{C}_n n + 1) \\
& \to 0 \quad \text{as} \quad n \to \infty
\end{align*}
\]

(42)
... proof continued.

Thus, we have

\[
\frac{c}{n} H(U, V)
\] (42)
Thus, we have

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} H(V) + \frac{c}{n} H(U)
\]  

(38)
Main Theorem

... proof continued.

Thus, we have

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} H(V) + \frac{c}{n} H(U)
\]

\[
\leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} \left(\frac{n}{c} + 1\right) \log \left(\frac{c}{n} + 1\right)
\]
Thus, we have

\[ \frac{c}{n} H(U, V) \leq \frac{c}{n} H(V) + \frac{c}{n} H(U) \]  

(38)

\[ \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} \left( \frac{n}{c} + 1 \right) \log \left( \frac{c}{n} + 1 \right) \]  

(39)

\[ = \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \left( \frac{n}{c} + 1 \right) \log \left( \frac{c}{n} + 1 \right) \]  

(40)

(42)
... proof continued.

Thus, we have

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} H(V) + \frac{c}{n} H(U) \tag{38}
\]

\[
\leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} (\frac{n}{c} + 1) \log(\frac{c}{n} + 1) \tag{39}
\]

\[
= \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + (\frac{n}{c} + 1) \log(\frac{c}{n} + 1) \tag{40}
\]

\[
= \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + (\frac{n}{c} + 1) \log(\frac{c + n}{n}) \tag{41}
\]

\[
\rightarrow 0 \text{ as } n \to \infty \tag{42}
\]
Thus, we have

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} H(V) + \frac{c}{n} H(U) \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} (\frac{n}{c} + 1) \log \frac{c}{n} + \frac{c}{n} \log(1 + \frac{c}{n}) + \log(1 + \frac{c}{n})
\]
... proof continued.

Thus, we have

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} H(V) + \frac{c}{n} H(U) \\
\leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} (\frac{n}{c} + 1) \log (\frac{n}{c} + 1) \\
= \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} (\frac{n}{c} + 1) \log (\frac{c + n}{n}) \\
= \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} \log(1 + \frac{c}{n}) + \log(1 + \frac{c}{n}) \\
\rightarrow 0 \text{ as } n \rightarrow \infty
\]
Thus, we have

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} H(V) + \frac{c}{n} H(U)
\]

\[
\leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} \left( \frac{n}{c} + 1 \right) \log \left( \frac{c}{n} + 1 \right)
\]

\[
= \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \left( \frac{n}{c} + 1 \right) \log \left( \frac{c}{n} + 1 \right)
\]

\[
= \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} \log \left( 1 + \frac{c}{n} \right) + \log \left( 1 + \frac{c}{n} \right)
\]

\[
\to 0 \text{ as } n \to \infty
\]

\[
\to 0 \text{ as } n \to \infty
\]
Thus, we have that

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + o(1) \quad (43)
\]

\[\rightarrow 0 \text{ as } n \rightarrow \infty \quad \text{we'll look at this} \]

\[\rightarrow 0 \]
Main Theorem

... proof continued.

• Thus, we have that

\[ \frac{c}{n} H(U, V) \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + o(1) \]  \tag{43}

\[ \rightarrow 0 \text{ as } n \rightarrow \infty \]

we’ll look at this

• Now, by Lemma 2.3 we have

\[ c(n) \leq \frac{n}{(1-\epsilon_n) \log n} = \frac{n}{\log n} (1 + o(1)) < n/c \text{ for big enough } n \]
Main Theorem

... proof continued.

Thus, we have that

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} k \rightarrow 0 \text{ as } n \rightarrow \infty \quad \text{we'll look at this}
\]

\[
\frac{c}{n} \log \frac{n}{c} \rightarrow 0
\]

Now, by Lemma 2.3 we have

\[
c(n) \leq \frac{n}{(1-\epsilon_n) \log n} = \frac{n}{\log n} (1 + o(1)) < \frac{n}{c} \text{ for big enough } n
\]

Then since \( c/n \log(n/c) \) is monotone up to its peak at \( n/c = e \).
Main Theorem

... proof continued.

- Thus, we have that
  \[
  \frac{c}{n} H(U, V) \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + o(1)
  \]
  \[
  \rightarrow 0 \text{ as } n \rightarrow \infty
  \]
  we'll look at this

- Now, by Lemma 2.3 we have
  \[
  c(n) \leq \frac{n}{(1-\epsilon_n) \log n} = \frac{n}{\log n} (1 + o(1)) < \frac{n}{c}
  \]
  for big enough \( n \)

- Then since \( c/n \log(n/c) \) is monotone up to its peak at \( n/c = e \),
  \[
  \frac{c}{n} \log \frac{n}{c}
  \]

(46)
Main Theorem

... proof continued.

- Thus, we have that

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + o(1) \quad (43)
\]

→0 as \( n \to \infty \)

we'll look at this →0

- Now, by Lemma 2.3 we have

\[
c(n) \leq \frac{n}{(1 - \epsilon_n) \log n} = \frac{n}{\log n} (1 + o(1)) < \frac{n}{c} \text{ for big enough } n
\]

- Then since \( \frac{c}{n} \log(\frac{n}{c}) \) is monotone up to its peak at \( \frac{n}{c} = e \),

\[
\frac{c}{n} \log \frac{n}{c} \leq \frac{n}{\log n} (1 + o(1)) \log \frac{n}{\log n} (1 + o(1)) \quad (44)
\]

(46)
Main Theorem

... proof continued.

- Thus, we have that

\[
\frac{c}{n} H(U, V) \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + o(1) \tag{43}
\]

\[
\rightarrow 0 \text{ as } n \rightarrow \infty \quad \text{we'll look at this}\]

- Now, by Lemma 2.3 we have

\[
c(n) \leq \frac{n}{(1-\epsilon_n) \log n} = \frac{n}{\log n} (1 + o(1)) < \frac{n}{c} \quad \text{for big enough } n
\]

- Then since \(c/n \log(n/c)\) is monotone up to its peak at \(n/c = e\),

\[
\frac{c}{n} \log \frac{n}{c} \leq \frac{n}{\log n} (1 + o(1)) \log \frac{n}{\log n} \tag{44}
\]

\[
= \log[\log n/(1 + o(1))] \frac{1 + o(1)}{\log n} \tag{45}
\]

\[
\rightarrow 0 \text{ as } n \rightarrow \infty \tag{46}
\]
Main Theorem

... proof continued.

- Thus, we have that
  \[
  \frac{c}{n} H(U, V) \leq \frac{c}{n} k + \frac{c}{n} \log \frac{n}{c} + o(1) \rightarrow 0 \text{ as } n \rightarrow \infty
  \]  
  we'll look at this

- Now, by Lemma 2.3 we have
  \[
  c(n) \leq \frac{n}{(1-\epsilon_n) \log n} = \frac{n}{\log n} (1 + o(1)) < \frac{n}{c} \text{ for big enough } n
  \]

- Then since \( c/n \log(n/c) \) is monotone up to its peak at \( n/c = e \),
  \[
  \frac{c}{n} \log \frac{n}{c} \leq \frac{n}{\log n} (1 + o(1)) \log \frac{n}{\log n} (1 + o(1))
  \]
  \[
  = \log[\log n/(1 + o(1))] \frac{1 + o(1)}{\log n}
  \]
  \[
  \leq O\left(\frac{\log \log n}{\log n}\right) \rightarrow 0 \text{ as } n \rightarrow \infty
  \]
Main Theorem

... proof continued.

Thus, \( \frac{c}{n} \mathcal{H}(U, V) \to 0 \) as \( n \to \infty \).
Main Theorem

... proof continued.

- Thus, $\frac{c}{n} H(U, V) \to 0$ as $n \to \infty$.
- Therefore,

$$\frac{c(n) \log c(n)}{n} \leq -\frac{1}{n} \log Q_k(x_{1:n}|s_1) + \epsilon_k(n)$$  \hspace{1cm} (47)

where $\epsilon_k(n) \to 0$ as $n \to \infty$
... proof continued.

- Thus, \( \frac{c}{n} H(U, V) \to 0 \) as \( n \to \infty \).
- Therefore,

\[
\frac{c(n) \log c(n)}{n} \leq -\frac{1}{n} \log Q_k(x_1:n|s_1) + \epsilon_k(n) \quad (47)
\]

where \( \epsilon_k(n) \to 0 \) as \( n \to \infty \)

- Therefore,

\[
\limsup_{n \to \infty} \frac{c(n) \log c(n)}{n} \leq \lim_{n \to \infty} -\frac{1}{n} Q_k(X_1:n|X_{-(k-1):0}) \\
= H(X_0|X_{-1}, X_0, \ldots, X_k) \quad // \text{for stationary ergodic source} \\
\to H(\mathcal{X}) \text{ as } k \to \infty \quad (48)
\]
Theorem 3.3

Let $X_i$ be an infinite length stationary ergodic stochastic process. Let $\ell(x_1:n)$ be the LZ codeword length for $n$ symbols. Then

$$\limsup_{n \to \infty} \frac{1}{n} \ell(x_1:n) \leq H(X)$$

(50)

Proof.

- We know that $\ell(x_1:n) = c(n)(\log(c(n)) + 1)$, where $c(n)$ is the number of phrases in the LZ parse (so they are distinct).
Theorem 3.3

Let $X_i$ be an infinite length stationary ergodic stochastic process. Let $\ell(x_1:n)$ be the LZ codeword length for $n$ symbols. Then

$$\limsup_{n \to \infty} \frac{1}{n} \ell(x_1:n) \leq H(X) \quad (50)$$

Proof.

- We know that $\ell(x_1:n) = c(n)(\log(c(n)) + 1)$, where $c(n)$ is the number of phrases in the LZ parse (so they are distinct).

- But from Lemma 2.3, we have

$$\limsup_{n \to \infty} \frac{c(n)}{n} = \limsup_{n \to \infty} \frac{1 + o(1)}{\log n} = 0 \quad (51)$$
Related to lengths

**Theorem 3.3**

Let $X_i$ be an infinite length stationary ergodic stochastic process. Let $\ell(x_1:n)$ be the LZ codeword length for $n$ symbols. Then

$$\limsup_{n \to \infty} \frac{1}{n} \ell(x_1:n) \leq H(X)$$

(50)

**Proof.**

- We know that $\ell(x_1:n) = c(n)(\log(c(n)) + 1)$, where $c(n)$ is the number of phrases in the LZ parse (so they are distinct).
- But from Lemma 2.3, we have

$$\limsup_{n \to \infty} \frac{c(n)}{n} = \limsup_{n \to \infty} \frac{1 + o(1)}{\log n} = 0$$

(51)

Therefore,
Related to lengths

... proof continued.

Therefore,

\[
\limsup_{n \to \infty} \frac{\ell(x_1^n)}{n} = \limsup_{n \to \infty} \left( \frac{c(n) \log c(n)}{n} \right)
\]

\[
\leq H(X)
\]

(53)

(54)
Related to lengths

... proof continued.

Therefore,

\[
\limsup_{n \to \infty} \frac{\ell(x_1:n)}{n} = \limsup_{n \to \infty} \left( \frac{c(n) \log c(n)}{n} + \frac{c(n)}{n} \right) \rightarrow H(X)
\]

\leq H(X)

(53)

(54)

In other words, a purely algorithmic procedure (LZ), when faced with a (stationary ergodic) stochastic process governed by some distribution, but without knowing anything about the distribution and by only following the algorithm, will in the limit converge to the entropy rate of the stochastic process.
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.).

The Venn diagram illustrates the following:

- $H(X)$ represents the information content of variable $X$
- $H(Y)$ represents the information content of variable $Y$
- $H(X,Y)$ represents the joint information content of variables $X$ and $Y$
- $H(X|Y)$ represents the conditional information content of $X$ given $Y$
- $H(Y|X)$ represents the conditional information content of $Y$ given $X$
- $I(X;Y)$ represents the mutual information between $X$ and $Y$

But is within these sets? So far, we’ve only said it is “information”. We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:

1. Gain intuition
2. Help prove theorems
3. Lead to new (useful) information theoretic inequalities that are “non-Shannon” (i.e., not previously known).
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.).

But is within these sets? So far, we’ve only said it is “information”
We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

But is within these sets? So far, we’ve only said it is “information”

We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:
Information and Venn Diagrams

- We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

\[ H(X), H(Y), H(X|Y), I(X;Y), H(Y|X) \]

But is within these sets? So far, we’ve only said it is “information”

- We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:
  1. gain intuition
Information and Venn Diagrams

- We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

- But is within these sets? So far, we’ve only said it is “information”

- We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:
  1. gain intuition
  2. help prove theorems
Information and Venn Diagrams

- We’ve previously seen that Venn diagrams are a useful way to visualize the relationship between information measures (entropy, etc.)

\[
\begin{align*}
H(X) & \quad I(X;Y) \quad H(Y) \\
H(X) & \quad H(X|Y) \quad H(Y|X) \\
H(X,Y) & \\
H(X) & \quad H(Y)
\end{align*}
\]

- But is within these sets? So far, we’ve only said it is “information”

- We want now to show that set theory and the relation between set theory and information theory can be made more precise in order to:
  1. gain intuition
  2. help prove theorems
  3. lead to new (useful) information theoretic inequalities that are “non-Shannon” (i.e., not previously known).
Definitions: field, atom

- We have a set of random variables $X_1, X_2, \ldots, X_n$. 

  - An atom of $F_n$ are sets of the form $\bigcap_{i=1}^{n} Y_i$ where $Y_i = \{ \tilde{X}_i \text{ or } \tilde{X}_i^c \}$ (55)
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 $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$. 
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 ($\tilde{X}^c$), set subtractions/difference ($\setminus$) on combinations of $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$. 

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

Ex: $n = 3$, 8 atoms.
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 ($\cup$), intersections ($\cap$), complementation ($\tilde{X}^c$), 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

$$\text{an atom} = \bigcap_{i=1}^{n} Y_i \text{ where } Y_i = \begin{cases} \tilde{X}_i, & \text{or} \\ \tilde{X}_i^c \end{cases}$$  \hspace{1cm} (55)
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

\[
\text{an atom } = \bigcap_{i=1}^{n} Y_i \text{ where } Y_i = \begin{cases} 
\tilde{X}_i \\
\tilde{X}_i^c 
\end{cases} \text{ or (55)}
\]

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

Ex: $n = 3$, 8 atoms
Definitions: field, atom

- In general, there are $|A| = 2^n$ atoms where $A$ is the set of atoms.
Definitions: field, atom

- In general, there are $|A| = 2^n$ atoms where $A$ is the set of atoms.
- The atoms are disjoint: Why?
Definitions: field, atom

- In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.
- The atoms are disjoint: Why? For any two distinct atoms, there is at least one factor which is a complement.
Definitions: field, atom

- In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.
- The atoms are disjoint: Why? For any two distinct atoms, there is at least one factor which is a complement.
- There are $2^{2^n}$ elements in the field. Why?
Definitions: field, atom

- In general, there are $|\mathcal{A}| = 2^n$ atoms where $\mathcal{A}$ is the set of atoms.
- The atoms are disjoint: Why? For any two distinct atoms, there is at least one factor which is a complement.
- There are $2^{2^n}$ elements in the field. Why? union of disjoint atoms and each atom may be chosen or not chosen.
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

\[
\mu(A \cup B) = \mu(A) + \mu(B)
\] (56)
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

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

(56)

- 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 $A - B \equiv A \cap B^c$. 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.
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

\[
\mu(A \cup B) = \mu(A) + \mu(B)
\]  

(56)

- 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 \( A \setminus B \equiv A \cap B^c \).
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

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

(56)

- 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.
Definitions: signed measure

- Example: Consider two sets $\tilde{X}_1, \tilde{X}_2$

![Venn Diagram](image_url)
Definitions: signed measure

- Example: Consider two sets $\tilde{X}_1, \tilde{X}_2$

Here, the signed measure $\mu$ on $\mathcal{F}_n$ is determined by the four values:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2), \mu(\tilde{X}_1^c \cap \tilde{X}_2), \mu(\tilde{X}_1 \cap \tilde{X}_2^c), \mu(\tilde{X}_1^c \cap \tilde{X}_2^c),$$

(57)
Definitions: signed measure

- Example: Consider two sets $\tilde{X}_1, \tilde{X}_2$

Here, the signed measure $\mu$ on $\mathcal{F}_n$ is determined by the four values:

$$
\mu(\tilde{X}_1 \cap \tilde{X}_2), \mu(\tilde{X}_1^c \cap \tilde{X}_2), \mu(\tilde{X}_1 \cap \tilde{X}_2^c), \mu(\tilde{X}_1^c \cap \tilde{X}_2^c),
$$

(57)

- For example, we have

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

(58)
Random variables

- We said earlier that $\tilde{X}_i$ was associated with a random variable $X_i$. 
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 $X_i$. 
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 $X \in \mathcal{F}_n$. 

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.
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: $H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2)$ $I(X_1; X_2)$. 

Prof. Jeff Bilmes
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 $\mu$. 

Prof. Jeff Bilmes
Signed Measures and Shannon Measures

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

\[
\begin{align*}
\mu^*(\tilde{X}_1 \cap \tilde{X}_2) &= I(X_1; X_2) \quad (59) \\
\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) &= \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \quad (60) \\
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) &= \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \quad (61) \\
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) &= \mu^*(\emptyset) = 0 \quad (62)
\end{align*}
\]
Signed Measures and Shannon Measures

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

$$
\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2) \quad (59)
$$

$$
\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \quad (60)
$$

$$
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \quad (61)
$$

$$
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0 \quad (62)
$$

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

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

\[
\begin{align*}
\mu^*(\tilde{X}_1 \cap \tilde{X}_2) & = I(X_1; X_2) \\
\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) & = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \\
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) & = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \\
\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) & = \mu^*(\emptyset) = 0
\end{align*}
\]

- 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?
Signed Measures and Shannon Measures

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

\[ \mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2) \quad (59) \]
\[ \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \quad (60) \]
\[ \mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \quad (61) \]
\[ \mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0 \quad (62) \]

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?

(64)
Signed Measures and Shannon Measures

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

$\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2)$ \hspace{1cm} (59)

$\mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1 | X_2)$ \hspace{1cm} (60)

$\mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2 | X_1)$ \hspace{1cm} (61)

$\mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0$ \hspace{1cm} (62)

- 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?

$\mu^*(\tilde{X}_1)$
We can make the following associations/definitions with signed measure $\mu^*$:

\[ \mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2) \] (59)
\[ \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \] (60)
\[ \mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \] (61)
\[ \mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0 \] (62)

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?

\[ \mu^*(\tilde{X}_1) = \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \] (63)
\[ = I(X_1; X_2) + H(X_1|X_2) = H(X_1) \] (64)
Signed Measures and Shannon Measures

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

  \[
  \mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2) \quad (59)
  \]
  \[
  \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) = \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2) \quad (60)
  \]
  \[
  \mu^*(\tilde{X}_1^c \cap \tilde{X}_2) = \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1) \quad (61)
  \]
  \[
  \mu^*(\tilde{X}_1^c \cap \tilde{X}_2^c) = \mu^*(\emptyset) = 0 \quad (62)
  \]

- 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?

  \[
  \mu^*(\tilde{X}_1) = \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c))
  \]
  \[
  = I(X_1; X_2) + H(X_1|X_2) = H(X_1) \quad (64)
  \]
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2) \quad (65)$$
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2)$$  \hspace{1cm} (65)

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

$$= H(X_1, X_2)$$  \hspace{1cm} (68)
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2)$$  \hspace{1cm} (65)

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$$

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2)$$  \hspace{1cm} (68)
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2) \tag{65}$$

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu \left( \bigcup_{\text{all atoms } A} A \right) \tag{66}$$
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?
  \[
  \mu^*(\tilde{X}_2) = H(X_2)
  \]  
  (65)

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?
  \[
  \mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu\left( \bigcup_{\text{all atoms } A} A \right)
  = I(X_1; X_2) + H(X_1|X_2) + H(X_2|X_1) + 0
  \]  
  (66)
Signed Measures and Shannon Measures

- What would \( \mu^*(\tilde{X}_2) \) be?

\[
\mu^*(\tilde{X}_2) = H(X_2)
\]  
\( (65) \)

- What about \( \mu^*(\tilde{X}_1 \cup \tilde{X}_2) \)?

\[
\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu \left( \bigcup_{\text{all atoms } A} A \right)
\]
\[
= I(X_1; X_2) + H(X_1 | X_2) + H(X_2 | X_1) + 0
\]  
\( (67) \)

\[
= H(X_1, X_2)
\]  
\( (68) \)
Signed Measures and Shannon Measures

- What would $\mu^*(\tilde{X}_2)$ be?

$$\mu^*(\tilde{X}_2) = H(X_2) \quad (65)$$

- What about $\mu^*(\tilde{X}_1 \cup \tilde{X}_2)$?

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = \mu \left( \bigcup_{\text{all atoms } A} A \right) \quad (66)$$

$$= I(X_1; X_2) + H(X_1|X_2) + H(X_2|X_1) + 0 \quad (67)$$

$$= H(X_1, X_2) \quad (68)$$

- So, we have defined $\mu^*$ only on the atoms, and from this we have, using the signed measure property and set theory, fully recovered all the rest of the Shannon information values.
Unions of sets

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

\begin{align*}
\mu^*(\emptyset) &= 0 \\
\mu^*(\tilde{X}_1) &= H(X_1) \\
\mu^*(\tilde{X}_2) &= H(X_2) \\
\mu^*(\tilde{X}_1 \cup \tilde{X}_2) &= H(X_1, X_2)
\end{align*}

(69) (70) (71) (72)
What if we define $\mu^*$ only on the unions of sets. I.e., we make the following definitions:

\begin{align*}
\mu^*(\emptyset) &= 0 \quad (69) \\
\mu^*(\tilde{X}_1) &= H(X_1) \quad (70) \\
\mu^*(\tilde{X}_2) &= H(X_2) \quad (71) \\
\mu^*(\tilde{X}_1 \cup \tilde{X}_2) &= H(X_1, X_2) \quad (72)
\end{align*}

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

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

\[ \mu^*(\emptyset) = 0 \]  
\[ \mu^*(\tilde{X}_1) = H(X_1) \]  
\[ \mu^*(\tilde{X}_2) = H(X_2) \]  
\[ \mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2) \]

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:

\[ \mu(\tilde{X}_1 \cap \tilde{X}_2) = \mu(\tilde{X}_1) + \mu(\tilde{X}_2) - \mu(\tilde{X}_1 \cup \tilde{X}_2) \]  
\[ = H(X_1) + H(X_2) - H(X_1, X_2) = I(X_1; X_2) \]
Recovering Shannon

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

\[ H/I \leftrightarrow \mu^* \]  \hspace{1cm} (75)

\[ , \leftrightarrow \cup \]  \hspace{1cm} (76)

\[ ; \leftrightarrow \cap \]  \hspace{1cm} (77)

\[ | \leftrightarrow \setminus \]  \hspace{1cm} //set minus  \hspace{1cm} (78)

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.
Recuperating Shannon

- So we have recovered Shannon’s information measures with the following correspondence:
  
  \[
  H/I \longleftrightarrow \mu^* 
  \]  
  \[
  , \longleftrightarrow \cup 
  \]  
  \[
  ; \longleftrightarrow \cap 
  \]  
  \[
  \setminus \longleftrightarrow \setminus \quad \text{set minus} 
  \]  

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

\[ H/I \leftrightarrow \mu^* \]  \hspace{1cm} (75)

\[ , \leftrightarrow \cup \]  \hspace{1cm} (76)

\[ ; \leftrightarrow \cap \]  \hspace{1cm} (77)

\[ | \leftrightarrow \setminus \hspace{0.5cm} //set \ minus \]  \hspace{1cm} (78)

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.
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[
\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)
\]  

which follows since

\[
\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) 
\]
Recovering Shannon

For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[
\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)
\]  \hspace{1cm} (79)

which follows since

\[
\mu(A \cup B)
\]

\hspace{1cm} (82)
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  

(79)

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) \]  

(82)
Recovering Shannon

- For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \]  

(80)

(82)
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[ \mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \]  

which follows since

\[ \mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \]  
\[ = \mu(A) + \mu(B \setminus A) + \mu(A \cap B) - \mu(A \cap B) \]
Recovering Shannon

For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[
\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)
\]  

(79)

which follows since

\[
\mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A)
\]  

(80)

\[
= \mu(A) + \mu(B \setminus A) + \mu(A \cap B) - \mu(A \cap B)
\]  

(81)

\[
= \mu(A) + \mu(B) - \mu(A \cap B)
\]  

(82)
For example, a simple example of a well-known property in set theory: The inclusion-exclusion formula for measures, is as follows:

\[
\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B) \tag{79}
\]

which follows since

\[
\mu(A \cup B) = \mu(A \cup (B \setminus A)) = \mu(A) + \mu(B \setminus A) \tag{80}
\]
\[
= \mu(A) + \mu(B \setminus A) + \mu(A \cap B) - \mu(A \cap B) \tag{81}
\]
\[
= \mu(A) + \mu(B) - \mu(A \cap B) \tag{82}
\]

Equating our measures, we see that the entropy/mutual information formula is just inclusion-exclusion:

\[
H(X_1, X_2) = H(X_1) + H(X_2) - I(X_1; X_2) \tag{83}
\]
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.
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:

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

(84)

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):

\[ B \triangleq \{ \tilde{X}_G : \emptyset \subset G \subseteq [n] \} \]  

(85)
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:
  \[
  A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset
  \]

- Non-empty atoms are $A \triangleq \{\text{all atoms}\} \setminus \{A_0\}$ which are those that are not assuredly empty, so $|A| = 2^n - 1$. 

\[(84)\]
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:

$$A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset \quad (84)$$

- 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?
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:

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

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 $\Omega$. 
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:

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

(84)

- 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 $G$. 
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:
  \[
  A_0 = \bigcap_{i \in [n]} \tilde{X}_i^c = \left( \bigcup_i \tilde{X}_i \right)^c = \emptyset
  \]

- 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 \( G \subseteq [n] \), \( \tilde{X}_G = \bigcup_{i \in G} \tilde{X}_i \).
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:

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

(84)

- 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):

$$\mathcal{B} \triangleq \left\{ \tilde{X}_G : \emptyset \subset G \subseteq [n] \right\}$$

(85)
What needs to be specified

**Theorem 4.1**

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 4.1

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 4.1**

*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 $\mu$. 
Inclusion/Exclusion

- Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure)
### Inclusion/Exclusion

- Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure).
- First note, from the binomial expansion:

\[
0 = 1 - 1 = (1 - 1)^n \\
= \sum_{\ell=0}^{n} \binom{n}{\ell} (-1)^\ell (1)^{n-\ell} \\
= 1 - \binom{n}{1} + \binom{n}{2} - \ldots + (-1)^n \binom{n}{n}
\]
Inclusion/Exclusion

- Consider a simple signed measure $\mu(A) = |A|$, the cardinality (or size or counting) measure (but the same idea works for any signed measure).

- First note, from the binomial expansion:

\[
0 = 1 - 1 = (1 - 1)^n \\
= \sum_{\ell=0}^{n} \binom{n}{\ell} (-1)^{\ell} (1)^{n-\ell} \\
= 1 - \binom{n}{1} + \binom{n}{2} - \ldots + (-1)^{n} \binom{n}{n}
\]

- We have $n$ sets $A_i$, for $i = 1\ldots n$ such that $A_i \subseteq \Omega$. 
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.

\[
| \bigcap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \ldots + (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n| \quad (89)
\]

\[
+ \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots \quad (90)
\]

\[
+ (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n| \quad (91)
\]
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.

\[
| \cap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots + (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n| \tag{89}
\]

\[
\sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots + (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n| \tag{90}
\]

\[
+ (-1)^{n-1} |A_1 \cup A_2 \cup \ldots \cup A_n| \tag{91}
\]

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

\[
| \bigcap_{i=1}^{n} A_i | = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cup A_j| 
+ \sum_{1 \leq i < j < k \leq n} |A_i \cup A_j \cup A_k| + \ldots 
+ (-1)^{n-1}|A_1 \cup A_2 \cup \ldots \cup A_n|
\] (89)

(90)

(91)

- 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.
Consider an \( x \in \Omega \) where \( x \in A_i \) for all \( i = 1 \ldots n \).
Consider an \( x \in \Omega \) where \( x \in A_i \) for all \( i = 1 \ldots n \).

Then the l.h.s. of Equation 91 contributes only 1 (unity) for this \( x \).
Consider an \( x \in \Omega \) where \( x \in A_i \) for all \( i = 1 \ldots n \).

Then the l.h.s. of Equation 91 contributes only 1 (unity) for this \( x \).

For the r.h.s. of Equation 91, we want to look what the contribution for this particular \( x \) will be, and we do this on the next slide.
For the r.h.s. of Equation 91, we want to look what the contribution for this particular $x$ will be, and it is:

$$n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$

(92)
For the r.h.s. of Equation 91, we want to look what the contribution for this particular $x$ will be, and it is:

$$n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$  \hspace{1cm} (92)

$$= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$  \hspace{1cm} (93)
For the r.h.s. of Equation 91, we want to look what the contribution for this particular $x$ will be, and it is:

$$n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1}\binom{n}{n}$$  \hspace{1cm} (92)

$$= (-1)^0\binom{n}{1} + (-1)^1\binom{n}{2} + (-1)^2\binom{n}{3} + (-1)^3\binom{n}{4} + \ldots + (-1)^{n-1}\binom{n}{n}$$  \hspace{1cm} (93)

$$= (-1)\left[ (-1)^1\binom{n}{1} + (-1)^2\binom{n}{2} + (-1)^3\binom{n}{3} + (-1)^4\binom{n}{4} + \ldots + (-1)^n\binom{n}{n} \right]$$  \hspace{1cm} (94)
Inclusion/Exclusion

For the r.h.s. of Equation 91, we want to look what the contribution for this particular $x$ will be, and it is:

$$n-\binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$  \hspace{1cm} (92)

$$= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}$$  \hspace{1cm} (93)

$$= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} \right]$$  \hspace{1cm} (94)

$$= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} + (1 - 1) \right]$$  \hspace{1cm} (95)

(97)
For the r.h.s. of Equation 91, we want to look what the contribution for this particular $x$ will be, and it is:

\[
\begin{align*}
&n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n} \\
&= (-1)^{0} \binom{n}{1} + (-1)^{1} \binom{n}{2} + (-1)^{2} \binom{n}{3} + (-1)^{3} \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n} \\
&= (-1) \left[ (-1)^{1} \binom{n}{1} + (-1)^{2} \binom{n}{2} + (-1)^{3} \binom{n}{3} + (-1)^{4} \binom{n}{4} + \ldots + (-1)^{n} \binom{n}{n} \right] \\
&= (-1) \left[ (-1)^{0} \binom{n}{0} + (-1)^{1} \binom{n}{1} + (-1)^{2} \binom{n}{2} + (-1)^{3} \binom{n}{3} + (-1)^{4} \binom{n}{4} + \ldots + (-1)^{n} \binom{n}{n} - 1 \right]
\end{align*}
\]
For the r.h.s. of Equation 91, we want to look what the contribution for this particular $x$ will be, and it is:

\[
n - \binom{n}{2} + \binom{n}{3} - \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n} = (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

(92)

\[
= (-1)^0 \binom{n}{1} + (-1)^1 \binom{n}{2} + (-1)^2 \binom{n}{3} + (-1)^3 \binom{n}{4} + \ldots + (-1)^{n-1} \binom{n}{n}
\]

(93)

\[
= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} \right]
\]

(94)

\[
= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} + (1 - 1) \right]
\]

(95)

\[
= (-1) \left[ (-1)^0 \binom{n}{0} + (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + (-1)^3 \binom{n}{3} + (-1)^4 \binom{n}{4} + \ldots + (-1)^n \binom{n}{n} - 1 \right]
\]

(96)

\[
= (-1) \left( (1 - 1)^n - 1 \right) = 1
\]

(97)
Next, suppose that $x \in A_i$ for $i \in S$, where $|S| = k$. In other words, $x$ is in only exactly $k$ of the sets $A_k$ rather than all of them, where $0 \leq k < n$. 
Next, suppose that $x \in A_i$ for $i \in S$, where $|S| = k$. In other words, $x$ is in only exactly $k$ of the sets $A_k$ rather than all of them, where $0 \leq k < n$.

Then the l.h.s. of Equation 91 now contributes only 0 (zero) for this $x$. 
Next, suppose that $x \in A_i$ for $i \in S$, where $|S| = k$. In other words, $x$ is in only exactly $k$ of the sets $A_k$ rather than all of them, where $0 \leq k < n$.

Then the l.h.s. of Equation 91 now contributes only 0 (zero) for this $x$.

For the r.h.s. of Equation 91, we have the contribution for this particular $x$ be ...
For the r.h.s. of Equation 91, we have the contribution for this particular $x$ be:

$$k - \left[ \binom{n}{2} - \binom{n-k}{2} \right] + \left[ \binom{n}{3} - \binom{n-k}{3} \right] \ldots + (-1)^{n-k-1} \left[ \binom{n}{n-k} - \binom{n-k}{n-k} \right]$$

$$+ (-1)^{n-k} \left[ \binom{n}{n-k+1} \right] \ldots + (-1)^{n-1} \left[ \binom{n}{n} \right]$$

$$= (-1)^0 \left[ \binom{n}{1} - \binom{n-k}{1} \right] - \left[ \binom{n}{2} - \binom{n-k}{2} \right] + \left[ \binom{n}{3} - \binom{n-k}{3} \right] \ldots + (-1)^{n-k-1} \left[ \binom{n}{n-k} - \binom{n-k}{n-k} \right]$$

$$+ (-1)^{n-k} \left[ \binom{n}{n-k+1} \right] \ldots + (-1)^{n-1} \left[ \binom{n}{n} \right]$$

$$= (-1) \left[ (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + \ldots + (-1)^n \binom{n}{n} + 1 - 1 \right]$$

$$- \left[ (-1)^0 \binom{n-k}{1} + (-1)^1 \binom{n-k}{2} + (-1)^2 \binom{n-k}{3} + \ldots + (-1)^{n-k-1} \binom{n-k}{n-k} \right]$$

$$= (-1) \left[ (-1)^0 \binom{n}{0} (-1)^1 \binom{n}{1} + (-1)^2 \binom{n}{2} + \ldots + (-1)^n \binom{n}{n} - 1 \right]$$

$$+ \left[ (-1)^0 \binom{n-k}{0} + (-1)^1 \binom{n-k}{1} + (-1)^2 \binom{n-k}{2} + (-1)^3 \binom{n-k}{3} + \ldots + (-1)^{n-k} \binom{n-k}{n-k} - 1 \right]$$

$$= (-1) [(1 - 1)^{n - 1}] + [(1 - 1)^{n-k - 1} - 1] = 1 - 1 = 0$$
The same exact argument can be used to show exclusion/exclusion formula for $\mu$, i.e.,

$$
\mu(\bigcap_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cup A_j) \quad (103)
$$

$$
+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cup A_j \cup A_k) + \ldots \quad (104)
$$

$$
+ (-1)^{n-1} \mu(A_1 \cup A_2 \cup \ldots \cup A_n) \quad (105)
$$
The same exact argument can be used to show exclusion/exclusion formula for $\mu$, i.e.,

$$\mu(\bigcap_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cup A_j)$$ (103)

$$+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cup A_j \cup A_k) + \ldots$$ (104)

$$+ (-1)^{n-1} \mu(A_1 \cup A_2 \cup \ldots \cup A_n)$$ (105)

A “dual” form has the form:

$$\mu(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i) - \sum_{1 \leq i < j \leq n} \mu(A_i \cap A_j)$$ (106)

$$+ \sum_{1 \leq i < j < k \leq n} \mu(A_i \cap A_j \cap A_k) + \ldots$$ (107)

$$+ (-1)^{n-1} \mu(A_1 \cap A_2 \cap \ldots \cap A_n)$$ (108)
Another (easier?, shorter) way of writing these is as:

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

(109)

and

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

(110)
Proof of Theorem 4.1

Note that $|\mathcal{A}| = |\mathcal{B}| = 2^n - 1 \triangleq k$
Proof of Theorem 4.1

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

Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in A$. $\text{length}(\vec{a}) = k$
Proof of Theorem 4.1

Proof of Theorem 4.1.

- 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 $B \in \mathcal{B}$. $\text{length}(\vec{b}) = k$
Proof of Theorem 4.1

Proof of Theorem 4.1.

- Note that $|\mathcal{A}| = |\mathcal{B}| = 2^n - 1 \triangleq k$
- Define $\vec{a} = [\ldots \mu(A) \ldots]^\top$ for all $A \in \mathcal{A}$. length($\vec{a}$) = $k$
- Define $\vec{b} = [\ldots \mu(B) \ldots]^\top$ for all $B \in \mathcal{B}$. 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 $B$. 

...
Proof of Theorem 4.1

Proof of Theorem 4.1.

- 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 4.1.

- 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 4.1

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

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

(111)
Here, the inclusion/exclusion principle takes the form conditioned on (or excluding) $B$:

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

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 4.1

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

\[
A = \bigcap_{i=1}^{n} Y_i = \left( \bigcap_{j:Y_j = \tilde{X}_j} \tilde{X}_j \right) \bigcap \left( \bigcap_{j:Y_j = \tilde{X}_j^c} \tilde{X}_j^c \right) 
\]

(112)

\[
= \left( \bigcap_{j:Y_j = \tilde{X}_j} \tilde{X}_j \right) \bigcap \left( \bigcup_{j:Y_j = \tilde{X}_j} \tilde{X}_j \right) 
\]

(113)

\[
= \left( \bigcap_{j:Y_j = \tilde{X}_j} \tilde{X}_j \right) \setminus \left( \bigcup_{j:Y_j = \tilde{X}_j^c} \tilde{X}_j \right) 
\]

(114)

(115)

...
Proof of Theorem 4.1

Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(111)) may take the form:

$$\mu(A_i \cup A_j \cup \cdots \cup A_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_{\ell})$$  \hspace{1cm} (116)

$$= \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_{\ell}) - \mu(\bigcup_{\ell} \tilde{X}_{\ell})$$  \hspace{1cm} (117)

which is true since \(\mu(A \setminus B) = \mu(A \cup B) - \mu(B)\).
Proof of Theorem 4.1

proof of Theorem 4.1.

Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn. (111)) may take the form:

\[
\mu(A_i \cup A_j \cup \cdots \cup A_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup_{\ell} \tilde{X}_\ell) \quad (116)
\]

\[
= \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup_{\ell} \tilde{X}_\ell) - \mu(\bigcup_{\ell} \tilde{X}_\ell) \quad (117)
\]

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 \(\mathcal{B}\)!
Also, each of the terms of the r.h.s. of the inclusion/exclusion formula (Eqn.(111)) may take the form:

\[ \mu(A_i \cup A_j \cup \cdots \cup A_k \setminus B) = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \setminus \bigcup \tilde{X}_\ell) \quad (116) \]

\[ = \mu(\tilde{X}_i \cup \tilde{X}_j \cup \cdots \cup \tilde{X}_k \cup \bigcup \tilde{X}_\ell) - \mu(\bigcup \tilde{X}_\ell) \quad (117) \]

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 \( \vec{b} = C_n \vec{a} \)).

\[ \ldots \]
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 4.1

proof of Theorem 4.1.

- 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 $\mathcal{F}_n$. 
Proof of Theorem 4.1

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.