Reading

- Read chapters 1, and 2 in C&T.
Announcements and Assignments

- Homework 0, was due last night. If you have not yet done it, you have a few more days.
- Late policy: 10% every 24 hour period that you are late, and no more than 3 days late accepted.
- Lowest grade out of all HW grades is not counted towards final grade (so you can skip one HW with impunity).
- Homework 1 is now posted, due next Monday at 11:45pm via our dropbox (https://catalyst.uw.edu/collectit/dropbox/karna/19164).
- Please do use the discussion forum for all questions, so that all will benefit from them being answered.

Class Road Map

- L1 (1/3): Overview, Entropy
- L2 (1/5): Props. Entropy, Mutual Information, KL-Divergence
- L3 (1/10): KL-Divergence, Jensen, properties, Data Proc. Inequality
- L5 (1/17):
- L6 (1/19):
- L7 (1/24):
- L8 (1/26):
- L9 (1/31):
- L10 (2/2):
- L11 (2/7):
- L12 (2/9):
- L13 (2/14): Midterm
- L14 (2/16):
- L15 (2/21):
- L16 (2/23):
- L175 (2/28):
- L18 (3/1):
- L19 (3/6):
- L20 (3/8):

Finals Week: March 12th–16th.
Information - we use entropy to measure it.

The communication theory model

Surprise of an event \( \{X = x\} \) is measured \( \log \frac{1}{p(x)} \), and there are reasons for using \( \log \).

Entropy is lower bound on min number of guesses (on average) to guess the value of a random variable.

Communication Theory

General model of communication

Source \rightarrow \text{encoder} \rightarrow \text{channel} \rightarrow \text{decoder} \rightarrow \text{receiver}

General model of communication expanded:

- Source
- Source Coder - compress, remove redundancy
- Channel Coder - expand, add redundancy
- Noisy Channel - distort
- Channel Decoder - compress, remove redundancy, correct errors
- Source Decoder - expand, add redundancy
- Receiver
Mutual Information

- The mutual information (MI) is the average amount of information that r.v. $X$ has about $Y$, and vice versa.

**Definition 1 (mutual information)**

\[
I(X; Y) = E_{p(x,y)} \log \frac{p(x|y)}{p(x)} = E_{p(x,y)} \log \frac{p(x|y)p(y)}{p(x)p(y)}
\]

\[
= E_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\]

\[
H(X) = EI(x) = - \sum_x p(x) \log p(x)
\]

\[
H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y)
\]

\[
H(Y|X) = - \sum_{x,y} p(x, y) \log p(y|x)
\]

\[
H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)
\]

\[
0 \leq H(X) \leq \log n, \quad \text{where } n \text{ is } X\text{'s alphabet size.}
\]
Mutual Information and Entropy - Venn Diagram

Note, these are not sets in the standard sense. Rather the area of the regions convey “degree of information” and the overlapped region correspond to the overlap in information. I.e., the intersection consists of information that is, on average, revealed by both $X$ and $Y$. 

Mutual Information and Entropy - Block Diagram

Another way of looking at the same relationships.
KL-Divergence

- Given two distributions \( p(x) \) and \( q(x) \) over the same alphabet, i.e., \( p(x) = P_p(X = x) \) and \( q(x) = P_q(X = x) \), then the KL-divergence is defined as follows:

\[
D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}
\]

It is like an expected log-odds ratio, weighted by \( p \).

- Note, KL-divergence is not symmetric in general, i.e., \( D(p||q) \neq D(q||p) \).

- Also, limiting and continuity arguments show that \( 0 \log 0 = 0 \) and \( p \log(p/0) = \infty \), so we make these assumptions. Thus, \( D(p||q) \leq \infty \).

KL-Divergence and MI

- Let \( \mu_1(x, y) = p(x, y) \) and \( \mu_2(x, y) = p(x)p(y) \) with \( p(x) = \sum_y p(x, y) \) and \( p(y) = \sum_x p(x, y) \)

then

\[
D(\mu_1||\mu_2) = \sum_{x,y} \mu_1(x,y) \log \frac{\mu_1(x,y)}{\mu_2(x,y)}
\]

\[
= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = I(X;Y)
\]

- Thus, the MI is the distance between the joint distribution on \( X \) and \( Y \) and the product of the marginal distributions respectively on \( X \) and on \( Y \).

- Project of marginal distributions \( p(x) = \sum_y p(x, y) \) is a projection of \( p(x, y) \) down to the independent distribution. I.e.,

\[
p(x)p(y) = \arg\max_{p'(x,y) \text{ s.t. } p'(x,y)=p'(x)p'(y)} D(p(x,y)||p'(x,y))
\]
Event Specific Conditional Mutual Information

- Information can change if we condition on a third random variable event \( \{Z = z\} \), and this is denoted \( I(X; Y | Z = z) \) where \( X, Y, Z \) are random variables.
- Joint distribution over 3 random variables \( p(x, y, z) \) is given.
- Then the event specific (where the event is \( \{Z = z\} \)) conditional mutual information is given by

\[
I(X; Y | Z = z) = \sum_{x,y} p(x, y | z) \log \frac{p(x, y | z)}{p(x | z)p(y | z)} 
\]  

(12)

- Note that this is identical to regular mutual information except in this case we are always conditioning on the event \( z \).
- I.e., relative to standard mutual information:

\[
I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} 
\]  

(13)

we use different distributions, \( p(x, y) \to p(x, y | z) \), \( p(x) \to p(x | z) \), and \( p(y) \to p(y | z) \).

Conditional Mutual Information

- Information can change on average if we condition on a third random variable, and this is denoted \( I(X; Y | Z) \) where \( X, Y, Z \) are random variables.

**Definition 3 (conditional mutual information)**

\[
I(X; Y | Z) \triangleq \sum_z p(z) I(X; Y | Z = z) 
\]  

(14)

\[
= \sum_z p(z) E_{p(x,y|z)} \log \frac{p(x,y | Z = z)}{p(x | Z = z)p(y | Z = z)} 
\]  

(15)

\[
= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y | z)}{p(x | z)p(y | z)} 
\]  

(16)

\[
= E \left[ \log \frac{1}{p(x | z)} - \log \frac{1}{p(x | y, z)} \right] 
\]  

(17)

\[
= H(X | Z) - H(X | Y, Z) 
\]  

(18)
**Chain Rule for Mutual Information**

**Proposition 4**

\[
I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y | X_1, X_2, \ldots, X_{i-1}) \tag{19}
\]

Example: 

\[
I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y | X_1) \tag{20}
\]

**Proof.**

\[
I(X_1, \ldots, X_N; Y) = H(X_1, \ldots, X_N) - H(X_1, \ldots, X_N | Y) \tag{21}
\]

\[
= \sum_i H(X_i | X_1, \ldots, X_{i-1}) - \sum_i H(X_i | X_1, \ldots, X_{i-1}, Y) \tag{22}
\]

\[
= \sum_i I(X_i; Y | X_1, \ldots, X_{i-1}) \tag{23}
\]

**Conditional Relative Entropy - KL-divergence**

**Definition 5**

\[
D(p(y|x) || q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \tag{24}
\]

- Same as standard KL-divergence but now using conditional distribution.
Proposition 6

\[ D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \]  
\[ (25) \]

Proof.

\[ D(p(x, y)||q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \]  
\[ (26) \]

\[ = \sum_{x,y} p(x, y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \]  
\[ (27) \]

\[ = \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} + \sum_{x,y} p(x, y) \log \frac{p(x)}{q(x)} \]  
\[ (28) \]

ConVex Functions

- \( f \) is said to be convex on \((a, b)\) if for all \(x_1, x_2 \in (a, b), 0 \leq \lambda \leq 1, \)
  \[ f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \]  
\[ (29) \]

- Many convex functions, \( f(x) = x^2 \), or \( f(x) = e^x \), or \( f(x) = x \log x, x \geq 0. \)

- Visualized:

  ![Convex Function Diagram]

  \( \lambda x_1 + (1 - \lambda)x_2 \)

- \( f \) is strictly convex if equality holds only at \( \lambda = 0 \) or \( \lambda = 1. \)
Theorem 7 (Jensen)

Let $f$ be a convex function and $X$ a random variable, then

$$Ef(X) = \sum_x p(x) f(x) \geq f(EX) = f\left(\sum_x x p(x)\right) \quad (30)$$

- If $f$ is strictly convex, then $\{Ef(X) = f(EX)\} \Rightarrow \{X = EX\}$ meaning $X$ is a constant random variable.

Lemma 8

$$D(p\|q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \quad (31)$$

Proof.

Show that $-D(p\|q) \leq 0$. Let $A = \{x : p(x) > 0\} = \text{supp}(p)$. then

$$-D(p\|q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \quad (32)$$

$$= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \quad (33)$$

$$= \log \left( \sum_{x \in A} q(x) \right) \leq \log \left( \sum_x q(x) \right) = \log 1 = 0 \quad (34)$$
KL Divergence is non-negative

- Note that $\log x$ is strictly concave.
- Thus, equality in $\sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} = \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right)$ means $Z = EZ$ with $Z = p(X)/q(X)$, so $Z$ is a constant random variable.
- The only valid constant, with $p$ and $q$ still being probability distributions is $Z = 1$ or $p(x) = q(x)$.
- Thus, if $p(x) = q(x)$ then $D(p||q) = 0$ and vice versa.
- We’ll use this theorem to prove important properties about mutual information.

Mutual Information is non-negative

**Proposition 9**

$$I(X;Y) \geq 0 \text{ and } I(X;Y) = 0 \iff X \perp \!\!\!\!\!\!\perp Y$$ (35)

**Proof.**

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$$ (36)

and if $p(x,y) = p(x)p(y)$ we have equality, which is also condition for independence.
Mutual Information, more intuition

- So $I(X; Y)$ measures the “degree of dependence” between $X$ and $Y$.
- We have $0 \leq I(X; Y) \leq \min(H(X), H(Y))$.
- $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.
- If $X \perp \perp Y$, then $I(X; Y) = 0$ since in such case $H(X|Y) = H(X)$ and $H(Y|X) = H(Y)$.
- If $X = Y$, then $I(X; Y) = H(X) = H(Y)$ since in such case $H(Y|X) = H(X|Y) = 0$.

Conditioning can only reduce entropy

- Comparing $H(X)$ with $H(X|Y)$, knowing $Y$, on average, could tell us something about $X$ thereby reducing entropy.

**Proposition 10**

$$H(X|Y) \leq H(X) \quad \text{and} \quad H(X|Y) = H(X) \iff X \perp \perp Y$$

**Proof.**

$$0 \leq I(X; Y) = H(X) - H(X|Y)$$

- As mentioned, we could have $H(X|Y = y) > H(X)$, but (in the average case), $\sum_y p(y)H(X|Y = y) \leq H(X)$.
Independence Bounds on Entropy

- Entropy of a set of random variables is highest when the random variables are independent - the least redundancy between them

**Proposition 11**

\[
H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \quad (39)
\]

**Proof.**

\[
H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{N} H(X_i) \quad (40)
\]

Two variable instance of Proposition 11 is

\[
H(X_1, X_2) \leq H(X_1) + H(X_2) \quad (41)
\]

Note that equality in Equation 41 is achieved when all variables are mutually independent. I.e. when \(X_i \perp \perp X_j\) for all \(i, j\).
Conditioning and Mutual Information

- What about $I(X;Y)$ vs. $I(X;Y|Z)$?
- If $X \perp \perp Y|Z$ then $I(X;Y|Z) = 0$. For example, $X \perp \perp Y|Z$ whenever $X \rightarrow Z \rightarrow Y$.
- Alternatively, if $Z = Y$, then $I(X;Y|Z) = 0$.
- Thus, we can have $I(X;Y) > I(X;Y|Z)$.
- On the other hand, if $Z = X + Y$ and $X \perp \perp Y$ then $I(X;Y) = 0$ but $I(X;Y|Z) > 0$.
- Thus, no general conditioning relationship for mutual information and conditional mutual information.

Log-sum inequality

Theorem 12

Given $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$, with $a_i \geq 0$ and $b_i \geq 0$, we have

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$  \hspace{1cm} (42)

and we have equality iff $a_i/b_i = c = \text{const.}$.

- Recall, by limiting arguments, we have $0 \log 0 = 0$, $a \log a/0 = \infty$ for $a > 0$, and $0 \log 0/0 = 0$.
- This inequality is used for showing a number of important properties.
Log-sum inequality

- Consider \( f(t) = t \log t = t(\ln t)(\log e) \) which is strictly convex.

![Graph of f(t) = t log t]

- Why is this convex? \( f''(t) = 1/t \log e > 0 \) for all \( t > 0 \)

Proof of log-sum inequality.

- Since \( f \) is convex, Jensen’s inequality says that:

\[
\sum_i \alpha_i f(t_i) \geq f\left( \sum_i \alpha_i t_i \right) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum \alpha_i = 1. \quad (43)
\]

- Set \( \alpha_i = b_i / \sum_{j=1}^{n} b_j \) and \( t_i = a_i / b_i \) in the following:

\[
\sum_{i=1}^{n} \frac{a_i}{\sum_{j} b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{b_i}{\sum_{j} b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \alpha_i f(t_i) \quad (44)
\]

\[
\geq f\left( \sum_i \alpha_i t_i \right) = \left( \sum_i \alpha_i t_i \right) \log \left( \sum_j \alpha_i t_i \right) \quad (45)
\]

\[
= \left( \sum_i \frac{a_i}{\sum_{j} b_j} \right) \log \left( \frac{\sum_i a_i}{\sum_{j} b_j} \right) \quad (46)
\]
Convexity of $D(p\|q)$ in the pair

- $D(p\|q)$ is convex in the pair, meaning
- Let $(p_1, q_1)$ and $(p_2, q_2)$ be two probability mass pairs (i.e., each of $p_i$ and $q_i$ is a complete distribution).
- Then $(p, q) = \lambda(p_1, q_1) + (1 - \lambda)(p_2, q_2)$ is a mixture of pairs.
- Convex in the pair means that

$$D(\lambda p_1 + (1 - \lambda)p_2 \| \lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1 \| q_1) + (1 - \lambda)D(p_2 \| q_2)$$

- Proof: Use log-sum inequality with, i.e., we have

$$\left(\lambda p_1 + (1 - \lambda)p_2\right)(x) \frac{(\lambda p_1 + (1 - \lambda)p_2)(x) \log}{(\lambda q_1 + (1 - \lambda)q_2)(x)} \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \log \frac{1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)}$$

$$= \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)}$$

- And then sum over $x$.

Note that we can set $q_1 = q_2$ to get convexity just in $p$.

This is the basis for the alternating minimization procedure, which is a special case of the EM algorithm, the computation of the rate-distortion function, and the computation of the general-case channel capacity function (we’ll go over this more next quarter).

With this result, we can formalize many of the things we saw empirically or intuitively.
Entropy is concave in $p$

- We saw this before, mixing distributions can only increase entropy relative to the same mixture of the entropies.

**Proof.**

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{50}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \tag{51}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i - p_i \log 1/|\mathcal{X}| \right) \tag{52}
\]

\[
= \log |\mathcal{X}| - D(p||u) \tag{53}
\]

where $u$ is the uniform distribution. So $H(p)$ is a constant minus something convex in $p$.

**Consequences for MI**

- Let $(X, Y)$ be a joint r.v. space, so $p(x, y) = p(y|x)p(x)$.
- Then $I(X; Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$.
- That is, with $I_p(x)(X; Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)}$,

\[
I_{\lambda p_1(x)+(1-\lambda)p_2(x)}(X; Y) \geq \lambda I_{p_1(x)}(X; Y) + (1 - \lambda) I_{p_2(x)}(X; Y)
\]

- Also, $I(X; Y)$ is a convex function of $p(y|x)$ for fixed $p(x)$.
- That is, with $I_p(y|x)(X; Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)}$,

\[
I_{\lambda p_1(y|x)+(1-\lambda)p_2(y|x)}(X; Y) \leq \lambda I_{p_1(y|x)}(X; Y) + (1 - \lambda) I_{p_2(y|x)}(X; Y)
\]

- This will be quite important for channel capacity, and various other optimizations involving mutual information and distributions.
MI and communications and convexity

- Consider the problem of sending information from a sender $X$ to receiver $Y$ via a noisy process $p(y|x)$. i.e., for every $x$, we have a distribution over possible $y$ received.

  $\begin{array}{c}
  X \\
  \downarrow \\
  p(y|x) \\
  \uparrow \\
  Y
  \end{array}$

- The rate of information transmitted from $X$ to $Y$, per channel use, in units of bits, is $I(X;Y)$.

- “Mixing up” $p(x)$ can only increase information transmission for a fixed channel, relative to the original mixture of rates.

- “Mixing up” $p(y|x)$ for the noisy channel for a fixed source can only reduce the rate of transmission, relative to original mixture of rates.

- We will make this precise when we study Shannon’s channel coding theorem and his proof.

Data Processing Inequality

- Question: Given an information source, can additional processing gain more amount of information about that source?

- Lets view this as a picture:

  ![Diagram of data processing inequality]

- Question: Is it possible to obtain more information about a source given additional processing? Before you answer this, consider the following scenario:
**Data Processing Inequality**

- Image denoising, important problem in computer vision, big commercial market.
- High ISO images are noisy, but they are the only way to take pictures in low light with narrow aperture (meaning wide depth-of-field).
- Goal of image denoising is to remove the noise from the image and recover the original image.
- Example in our current context:

  ![Diagram of image denoising process](image)

  - Information Source (state of nature)
  - True Amount Of Information In Source
  - Imperfect Observation Process
  - Further Processing
  - Refined Observed Process & Information

- Question: Is it possible to obtain more information about a source given additional processing? Unfortunately, no!

**Markov Chain**

**Definition 13**

Random variables $X$, $Y$, and $Z$ form a Markov chain if $Z \perp X \mid Y$. I.e.,

$$p(z, x \mid y) = p(z \mid y)p(x \mid y) \quad \forall x, y, z$$  \hspace{1cm} (54)

- This means that
  $$p(x, y, z) = p(z \mid x, y)p(y \mid x)p(x) = p(z \mid y)p(y \mid x)p(x)$$
- Graphs (i.e., Bayesian networks) that can describe this.

  ![Graphs of Markov chains](graphs)

- Ex: If $Z = f(Y)$, then $X \rightarrow Y \rightarrow Z$ is true (i.e., $X, Y, Z$ form a Markov chain). $f(\cdot)$ can either random or deterministic. Key is that $X$ is irrelevant to determine $Z$ given $Y$.
- Notationally, when we state “$X \rightarrow Y \rightarrow Z$”, this means that we assert that $X, Y, Z$ form a Markov chain.
Theorem 14 (Data Processing Inequality)

If \( X \rightarrow Y \rightarrow Z \) then

\[
I(X; Y) \geq I(X; Z)
\]  \hfill (55)

- So in the Markov chain, the “arrows” correspond to processing and the random variables correspond to data.
- The processing can be either random or deterministic.
- The data processing inequality says that as we perform further processing of a data source, we move away from it, in a Markov chain, and we can (only at best) lose information about the original source, as measured by mutual information.

proof of data processing inequality.

By the chain rule of mutual information:

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

\[
= I(X; Z) + I(X; Y|Z)
\]  \hfill (56)

Now, \( X \perp Z|Y \) iff \( I(X; Z|Y) = 0 \). Also, \( I(X; Y|Z) \geq 0 \). So

\[
I(X; Y) \geq I(X; Z).
\]

- Similarly \( I(Y; Z) \geq I(X; Z) \).
- Chain distance decreases information. How fast?
- Corollary: If \( Z = f(Y) \), then \( X \rightarrow Y \rightarrow Z \), or \( X \rightarrow Y \rightarrow f(Y) \). Thus, \( I(X; Y) \geq I(X; f(Y)) \).
Processing can only lose information about $X$. When $X$ is source and $Y$ is receiver, no processing will increase information about $X$.

Consider pattern recognition: $X$ is an object, $Y$ is a list of features about the object, and $f(Y)$ is further processing.

Then any further processing can only reduce information about original object.

Question: What then is image de-noising? Why do we see result as higher fidelity but it has less information about original image? Other examples: audio restoration.

Another corollary: If $X \rightarrow Y \rightarrow Z$, then $I(X; Y|Z) \leq I(X; Y)$. I.e., $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) \leq H(X) - H(X|Y)$.

Intuition: Knowing $Z$ reduces amount learnt between $X$ and $Y$.

If $X \rightarrow Y \rightarrow Z$, then $I(X; Y|Z) \leq I(X; Y)$, as we just saw.

Recall, if $X \rightarrow Z \leftarrow Y$, then $I(X; Y|Z) \geq I(X; Y)$.

E.g., $X \perp \perp Y$ and $Z = X + Y$, the example we saw earlier.

So, the relationship between $I(X; Y|Z)$ and $I(X; Y)$ depends on the underlying “causal” relationship between the variables.
Entropy Always Increases

- We have probably heard that entropy, in the universe, always increases. How does this relate to our entropy?
- First law of thermodynamics: when all types of energy transfer, including work & heat, are taken into account, energy in an isolated system remains constant (conservation of energy).
- Second law of thermodynamics: The total entropy of an isolated system cannot decrease. It may (and generally does) increase.
- Entropy, here, is availability of energy in a closed system (it is not related to the total amount of energy in that closed system relative to other systems).

Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space $T$
- Uneven distribution, smaller $H$

- more entropy
- same overall energy (1st law)
- less usable energy
- Uniform temperature
- Temp. distribution over space $T$
Entropy and thermodynamics

- Claim: When there exists a transaction, the entropy always increases.
- Consequence: someday universe will be a gray blur of uniformly distributed matter.
- Model closed system as time-homogeneous 1st-order Markov chain, $p(x_{n+1}|x_n)$
- $X_n$ is a random variable at time $n$
- We have $\cdots \rightarrow x_{n-1} \rightarrow x_n \rightarrow x_{n+1} \rightarrow \cdots$
- The question we wish to study is: What happens to $H(X_n)$ as $n$ increases.
- Given two starting distributions $p_n(x_n)$ and $q_n(x_n)$, then $D(p_n(x_n)||q_n(x_n))$ will decrease with $n$. Why?
- We’ll write $p(x_n)$ for $p_n(x_n)$, also $q(x_n)$ for $q_n(x_n)$.
- Joint of two successive variables, via the homogeneous dynamics, is:

$$p(x_n, x_{n+1}) = p(x_{n+1}|x_n)p_n(x_n) \quad (58)$$
$$q(x_n, x_{n+1}) = p(x_{n+1}|x_n)q_n(x_n) \quad (59)$$

Then we have:

$$D(p(x_n, x_{n+1})||q(x_n, x_{n+1}))$$
$$= D(p(x_n)||q(x_n)) + D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n))$$
$$= 0 \quad (60)$$
$$= D(p(x_{n+1})||q(x_{n+1})) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1})) \geq 0 \quad (62)$$

This means that

$$D(p(x_n)||q(x_n)) \geq D(p(x_{n+1})||q(x_{n+1})) \quad (63)$$
Entropy and thermodynamics

- This means that

\[ D(p(x_n)||q(x_n)) \geq D(p(x_{n+1})||q(x_{n+1})) \]  

(64)

- Consider \( q(x) \) to be a stationary distribution of the Markov process, i.e., \( q(x) = \sum_y p(x|y)q(y) \) so that \( q(X_n = x_n) = q(X_{n+1} = x_n) \) for all \( n \).

- Then \( p(x_n) \) approaches this stationary distribution in the sense \( D(p(x_n)||q) \geq D(p(x_{n+1})||q) \). We cannot move farther away from any stationary distribution.

Consider the stationary distribution \( q(x) \) is the uniform distribution \( u(x) \).

Then

\[ D(p(x_n)||u) = -H(X_n) + \log n \geq -H(X_{n+1}) + \log n \]  

(65)

which means that \( H(X_{n+1}) \geq H(X_n) \) and indeed entropy can never decrease (can only increase) as expected in statistical physics.

- On the other hand, suppose the stationary distribution \( q \) is a non-uniform distribution, and in fact, \( H(q) \) could be very small.

- Then, entropy of the Markov chain could decrease over time.

- Thus, the increase/decrease of entropy over time is entirely dependent on the transition distribution, \( p(x_{n+1}|x_n) \).