Class Road Map - IT-I

- L1 (9/25): Overview, Communications, Information, Entropy
- L2 (9/30): Entropy, Mutual Information, KL-Divergence
- L3 (10/2): More KL, Jensen, more Venn, Log Sum, Data Proc. Inequality
- L4 (10/7):
- L5 (10/9):
- L6 (10/14):
- L7 (10/16):
- L8 (10/21):
- L9 (10/23):
- L10 (10/28):
- L11 (10/30):
- L12 (11/4):
- LXX (11/6): In class midterm exam
- LXX (11/11): Veterans Day holiday
- L13 (11/13):
- L14 (11/18):
- L15 (11/20):
- L16 (11/25):
- L17 (11/27):
- L18 (12/2):
- L19 (12/4):
- LXX (12/10): Final exam

Finals Week: December 9th–13th.
Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano’s inequality).
Homework

- Homework 1 on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), due Tuesday, Oct 8th, 11:55pm.
Definition 3.2.1 (Entropy)

Given a discrete random variable $X$ over a finite sized alphabet, the entropy of the random variable is:

$$H(X) \triangleq E \log \frac{1}{p(X)} = \sum_x p(x) \log \frac{1}{p(x)} = -\sum_x p(x) \log p(x)$$  \hspace{1cm} (3.1)

- Entropy is in units of “bits” since logs are base 2 (units of “nats” if base $e$ logs).
- Measures the degree of uncertainty in a distribution.
- Measures the disorder or spread of a distribution.
- Measures the “choice” that a source has in choosing symbols according to the density (higher entropy means more choice).
Joint Entropy

- Two random variables $X$ and $Y$ have joint entropy.

$$H(X, Y) = - \sum_x \sum_y p(x, y) \log p(x, y) = E \log \frac{1}{p(X, Y)} \quad (3.1)$$

- Obvious generalizations to vectors $X_{1:N} = (X_1, X_2, \ldots, X_N)$.

$$H(X_1, \ldots, X_N) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, \ldots, x_N) \log \frac{1}{p(x_1, \ldots, x_N)} \quad (3.2)$$

$$= E \log \frac{1}{p(x_1, \ldots, x_N)} \quad (3.3)$$
Why \log?: Khinchin’s uniqueness theorem

- For a distribution on \( n \) symbols with probabilities \( p = (p_1, p_2, \ldots, p_n) \), let \( H(p) = H(p_1, p_2, \ldots, p_n) \) be the entropy of that distribution.

- Consider any information measure, say \( \mathcal{H}(p) \) on \( p \), and consider the following three natural and desirable properties.
  1. \( \mathcal{H}(p) \) takes its largest value when \( p_i = 1/n \) for all \( i \).
  2. If we define the conditional information as
     \[
     \mathcal{H}(Y|X) \triangleq \sum_x p(x) \mathcal{H}(p(y_1|x), p(y_2|x), \ldots, p(y_n|x))
     \]
     \[
     = \sum_x p(x) \mathcal{H}(Y|X = x),
     \]
     then we wish to have additivity in the following way
     \[
     \mathcal{H}(X, Y) = \mathcal{H}(X) + \mathcal{H}(Y|X)
     \]
  3. For a distribution on \( n + 1 \) symbols, then if the probability of one is zero, we wish for \( \mathcal{H}(p_1, p_2, \ldots, p_n, 0) = \mathcal{H}(p_1, p_2, \ldots, p_n) \).
Why \( \log \): Khinchin’s uniqueness theorem

**Theorem 3.2.5 (Khinchin’s Theorem)**

If \( \mathcal{H}(p_1, \ldots, p_n) \) satisfies the above 3 properties for all \( n \) and for all \( p \) such that \( p_i \geq 0, \forall i \) and \( \sum_i p_i = 1 \) (i.e., all probability distributions), then

\[
\mathcal{H}(p_1, \ldots, p_n) = -\lambda \sum_i p_i \log p_i \tag{3.18}
\]

for \( \lambda \) a positive constant.

- Thus, we get entropy for some logarithmic base.
Summary so far

\[ H(X) = EI(X) = - \sum_x p(x) \log p(x) \] (3.18)

\[ H(X, Y) = - \sum_{x,y} p(x,y) \log p(x,y) \] (3.19)

\[ H(Y|X) = - \sum_{x,y} p(x,y) \log p(y|x) \] (3.20)

\[ H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \] (3.21)

and

\[ 0 \leq H(X) \leq \log n, \text{ where } n \text{ is } X\text{'s alphabet size.} \] (3.22)
A way of looking at the relationships.

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

- Given two distributions $p(x)$ and $q(x)$ over the same alphabet, i.e.,
  
  
  \[ p(x) = P_p(X = x) \quad \text{and} \quad q(x) = P_q(X = x) \]

  then the KL-divergence is defined as follows:

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

  (3.24)

- It is like an expected log-odds ratio, weighted by $p$.

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

  (so not a metric or a distance, thus a “divergence”).

- Also, limiting and continuity arguments show that $0 \log 0 = 0$ and $p \log(p/0) = \infty$. Hence, we might have $D(p||q) = \infty$. 

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

\[
D(\mu_1 || \mu_2) = \sum_{x,y} \mu_1(x, y) \log \frac{\mu_1(x, y)}{\mu_2(x, y)} = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = I(X; Y)
\]

(3.25) (3.26)

- 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 $Y$.

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

\[
p(x)p(y) = \arg\min_{p'(x,y)} D(p(x,y)||p'(x,y)) \quad \text{s.t. } p'(x,y) = p'(x)p'(y)
\]

(3.27)
Conditional Relative Entropy - KL-divergence

Definition 3.2.11

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

(3.37)

- Same as standard KL-divergence but now using conditional distribution.
- We take the expected value of \( \log \frac{p(y|x)}{q(y|x)} \) w.r.t. \( p(x,y) \) even though \( D(p(y|x)||q(y|x)) \) mentions only \( p(y|x) \). I.e., notation does not mention \( p(x,y) \) even though it is dependent on the full joint.
Chain Rule for KL-divergence

Proposition 3.3.1

\[
D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y \| x) \| q(y \| x)) \tag{3.1}
\]
Proposition 3.3.1

\[ D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x)) \]  \hfill (3.1)

Proof.

\[ D(p(x, y) \| q(x, y)) = \sum_{x, y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \]  \hfill (3.2)
Chain Rule for KL-divergence

Proposition 3.3.1

\[ D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x)) \]  \hspace{1cm} (3.1)\

Proof.

\[ D(p(x, y) \| q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \]  \hspace{1cm} (3.2)\

\[ = \sum_{x,y} p(x, y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \]  \hspace{1cm} (3.3)
**Proposition 3.3.1**

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

(3.1)

**Proof.**

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

(3.2)

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

(3.3)

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

(3.4)
ConVex Functions

- $f$ is said to be convex on $(a, b)$ if for all $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$, 
  \[ f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \] 
  \[(3.5)\]
ConVex Functions

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

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \quad (3.5)$$

- Many convex functions, $f(x) = x^2$, or $f(x) = e^x$, or

$f(x) = x \log x$, $x \geq 0$. 
ConVex Functions

- \( f \) is said to be convex on \((a, b)\) if for all \( x_1, x_2 \in (a, b), \ 0 \leq \lambda \leq 1, \)
  \[
  f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \quad (3.5)
  \]

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

- Visualized:

\[
\begin{align*}
\text{Visualized:} & \\
\lambda f(x_1) + (1 - \lambda)f(x_2) & \\
\lambda x_1 + (1 - \lambda)x_2
\end{align*}
\]
ConVex Functions

- $f$ is said to be convex on $(a, b)$ if for all $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$,
  \[ f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \] (3.5)

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

- Visualized:

- $f$ is strictly convex if equality holds only at $\lambda = 0$ or $\lambda = 1$. 
Theorem 3.3.2 (Jensen)

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

$$E f(X) = \sum_{x} p(x) f(x) \geq f(E X) = f\left(\sum_{x} x p(x)\right) \quad (3.6)$$
Jensen’s Inequality

**Theorem 3.3.2 (Jensen)**

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

$$Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \quad (3.6)$$

If $f$ is strictly convex, then \( \{Ef(X) = f(E(X))\} \Rightarrow \{X = EX\} \)
meaning $X$ is a constant random variable.
Lemma 3.3.3

\[ D(p\|q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  \hspace{1cm} (3.7)

Proof.

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

\[ -D(p\|q) \leq \log \left( \sum_{x \in A} p(x) q(x) / p(x) \right) \leq \log \left( \sum_{x} q(x) \right) = 0 \]  \hspace{1cm} (3.10)
KL Divergence is non-negative

Lemma 3.3.3

\[ D(p\|q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \quad (3.7) \]

Proof.

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

\[ -D(p\|q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (3.8) \]

(3.10)
KL Divergence is non-negative

Lemma 3.3.3

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \quad (3.7) \]

Proof.

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

\[ -D(p||q) = -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \quad (3.8) \]
KL Divergence is non-negative

Lemma 3.3.3

\[ D(p \| q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \] (3.7)

Proof.

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

\[
-D(p \| q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)}
\] (3.8)

\[
= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)}
\] (3.10)
KL Divergence is non-negative

Lemma 3.3.3

\[ D(p \| q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \] (3.7)

Proof.

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

\[ -D(p \| q) = - \sum_{x} p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \] (3.8)

\[ = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \] (3.9)

(3.10)
KL Divergence is non-negative

Lemma 3.3.3

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \] (3.7)

Proof.

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

\[ -D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \] (3.8)

\[ = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \] (3.9)

\[ = \log \left( \sum_{x \in A} q(x) \right) \] (3.10)
KL Divergence is non-negative

**Lemma 3.3.3**

\[ D(p\|q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  \hspace{1cm} (3.7)

**Proof.**

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

\[ -D(p\|q) = -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \]  \hspace{1cm} (3.8)

\[ = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \]  \hspace{1cm} (3.9)

\[ = \log \left( \sum_{x \in A} q(x) \right) \leq \log \left( \sum_x q(x) \right) \]  \hspace{1cm} (3.10)
KL Divergence is non-negative

**Lemma 3.3.3**

\[
D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \tag{3.7}
\]

**Proof.**

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

\[
-D(p||q) = -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \tag{3.8}
\]

\[
= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \tag{3.9}
\]

\[
= \log \left( \sum_{x \in A} q(x) \right) \leq \log \left( \sum_x q(x) \right) = \log 1 = 0 \tag{3.10}
\]
KL Divergence is non-negative

- Recall from a few slides ago: if $f$ is strictly convex, then
  \[ \{ Ef(Z) = f(E(Z)) \} \Rightarrow \{ Z = E(Z) \} \]
  meaning $Z$ is a constant random variable.
Recall from a few slides ago: if \( f \) is strictly convex, then 
\[
\{ Ef(Z) = f(E(Z)) \} \Rightarrow \{ Z = EZ \}
\]
meaning \( Z \) is a constant random variable.

- **Note that** \( \log x \) **is strictly concave.**
KL Divergence is non-negative

- Recall from a few slides ago: if $f$ is strictly convex, then $\{Ef(Z) = f(E(Z))\} \Rightarrow \{Z = EZ\}$ meaning $Z$ is a constant random variable.

- 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 $Z = q(X)/p(X)$, so $Z$ is a constant random variable.
Recall from a few slides ago: if $f$ is strictly convex, then 
\[ \{ E f(Z) = f(E(Z)) \} \Rightarrow \{ Z = EZ \} \] meaning $Z$ is a constant random variable.

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 $Z = q(X)/p(X)$, so $Z$ is a constant random variable.

The only valid constant, with $p$ and $q$ still being probability distributions is $Z = 1$ w.p.1. meaning $p(x) = q(x)$. 

KL Divergence is non-negative 

KL Divergence & Cond. MI  
Review and Venn 
log sum  
Data Proc. Inequality 

Prof. Jeff Bilmes  
EE514a/Fall 2019/Info. Theory I – Lecture 3 - Oct 2th, 2019  
L3 F18/47 (pg.35/170)
Recall from a few slides ago: if $f$ is strictly convex, then 
\[ \{ Ef(Z) = f(E(Z)) \} \Rightarrow \{ Z = EZ \} \] meaning $Z$ is a constant random variable.

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 $Z = q(X)/p(X)$, so $Z$ is a constant random variable.

The only valid constant, with $p$ and $q$ still being probability distributions is $Z = 1$ w.p.1. meaning $p(x) = q(x)$.

Thus, if $p(x) = q(x)$ then $D(p\|q) = 0$ and vice versa.
Recall from a few slides ago: if $f$ is strictly convex, then 
\[ \{ Ef(Z) = f(E(Z)) \} \Rightarrow \{ Z = EZ \} \] meaning $Z$ is a constant random variable.

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 $Z = q(X)/p(X)$, so $Z$ is a constant random variable.

The only valid constant, with $p$ and $q$ still being probability distributions is $Z = 1$ w.p.1. meaning $p(x) = q(x)$.

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.
Proposition 3.3.4

\[ I(X; Y) \geq 0 \text{ and } I(X; Y) = 0 \iff X \perp \!\!\!\!\!\!\!\!\!\perp Y \] (3.11)
Proposition 3.3.4

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

(3.11)

Proof.

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

(3.12)

and if \( p(x, y) = p(x)p(y) \) we have equality, which is also condition for independence.
So \(I(X;Y)\) measures the “degree of dependence” between \(X\) and \(Y\).
So $I(X; Y)$ measures the “degree of dependence” between $X$ and $Y$.

We have $0 \leq I(X; Y) \leq \min(H(X), H(Y))$. 

Mutual Information, more intuition
Mutual Information, more intuition

- So $I(X;Y)$ measures the “degree of dependence” between $X$ and $Y$.
- We have $0 \leq I(X;Y) \leq \min(H(X), H(Y))$.
- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. 
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 $X \perp Y$, then $I(X;Y) = 0$ since in such case $H(X|Y) = H(X)$ and $H(Y|X) = H(Y)$. 
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 never increase entropy

- Comparing $H(X)$ with $H(X|Y)$: knowing $Y$, on average, could (1) tell us something about $X$ thereby reducing entropy, or could (2) tell us nothing about $X$ leaving its entropy unchanged.

**Proposition 3.3.5**

$$H(X|Y) \leq H(X) \text{ and } H(X|Y) = H(X) \text{ iff } X \perp \!\!\!\!\!\perp Y$$ (3.13)
Conditioning can never increase entropy

- Comparing $H(X)$ with $H(X|Y)$: knowing $Y$, on average, could (1) tell us something about $X$ thereby reducing entropy, or could (2) tell us nothing about $X$ leaving its entropy unchanged.

**Proposition 3.3.5**

\[ H(X|Y) \leq H(X) \text{ and } H(X|Y) = H(X) \iff X \perp \!\!\!\!\!\!\perp Y \]  

(3.13)
Conditioning can never increase entropy

- Comparing $H(X)$ with $H(X|Y)$: knowing $Y$, on average, could (1) tell us something about $X$ thereby reducing entropy, or could (2) tell us nothing about $X$ leaving its entropy unchanged.

Proposition 3.3.5

$$H(X|Y) \leq H(X) \text{ and } H(X|Y) = H(X) \text{ iff } X \perp \perp Y$$  \hspace{1cm} (3.13)

Proof.

$$0 \leq I(X;Y) = H(X) - H(X|Y)$$  \hspace{1cm} (3.14)

- As mentioned, we could have $H(X|Y = y) > H(X)$, but (in the average case), $\sum_y p(y) H(X|Y = y) \leq H(X)$
Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions).
Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions)

Let $T$ be an independent random shuffle operation (permutation), i.e., $T \perp \! \! \! \perp X$. 

Follows since $H(TX) \geq H(TX | T) = H(T^{-1}TX | T) = H(X | T) = H(X)$ (conditioning reduces entropy, as we just learned).

Consider an example: $X$ is low entropy, one set of card positions very likely, the rest very unlikely. How could $T$ increase entropy?

Question: What if $T$ is deterministic? Would this contradict the above result (yes or now)? Why?
Shuffles

- Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions)
- Let $T$ be an independent random shuffle operation (permutation), i.e., $T \perp X$.
- Then $H(TX) \geq H(X)$. 
Shuffles

- Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions)
- Let $T$ be an independent random shuffle operation (permutation), i.e., $T \perp X$.
- Then $H(TX) \geq H(X)$.
- Follows since

$$H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$$

(conditioning reduces entropy, as we just learned).

Consider an example: $X$ is low entropy, one set of card positions very likely, the rest very unlikely. How could $T$ increase entropy?

Question: What if $T$ is deterministic? Would this contradict the above result (yes or now)? Why?
Shuffles

- Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions).
- Let $T$ be an independent random shuffle operation (permutation), i.e., $T \perp \!\!\!\!\perp X$.
- Then $H(TX) \geq H(X)$.
- Follows since $H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$ (conditioning reduces entropy, as we just learned).
- Consider an example: $X$ is low entropy, one set of card positions very likely, the rest very unlikely. How could $T$ increase entropy?
Shuffles

- Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions).

- Let $T$ be an independent random shuffle operation (permutation), i.e., $T \perp \perp X$.

- Then $H(TX) \geq H(X)$.

- Follows since

  $H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$

  (conditioning reduces entropy, as we just learned).

- Consider an example: $X$ is low entropy, one set of card positions very likely, the rest very unlikely. How could $T$ increase entropy?

- Question: What if $T$ is deterministic? Would this contradict the above result (yes or now)? Why?
Additive 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 3.3.6

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

(3.15)
Additive 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 3.3.6**

\[ H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \]  \hspace{1cm} (3.15)

**Proof.**

\[ H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{N} H(X_i) \]  \hspace{1cm} (3.16)
Additive 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 3.3.6

\[ H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \]  \hspace{1cm} (3.15)

Proof.

\[ H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, \ldots X_{i-1}) \leq \sum_{i=1}^{N} H(X_i) \]  \hspace{1cm} (3.16)
Independence Bounds on Entropy

Two variable instance of Proposition 3.3.6 is

\[ H(X_1, X_2) \leq H(X_1) + H(X_2) \]  (3.17)
Independence Bounds on Entropy

- Two variable instance of Proposition 3.3.6 is

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

(3.17)

- Note that equality the Equation is achieved when all variables are mutually independent. I.e. when \( X_i \perp \perp X_j \) for all \( i, j \).
What about $I(X; Y)$ vs. $I(X; Y|Z)$?
Conditioning and Mutual Information

- What about $I(X; Y)$ vs. $I(X; Y|Z)$?
- If $X \perp Y|Z$ then $I(X; Y|Z) = 0$. For example, $X \perp Y|Z$ 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 $X \perp Y$ then $I(X; Y) = 0$ but $I(X; Y|Z) > 0$.

Thus, no general conditioning relationship for mutual information and conditional mutual information.
What about $I(X; Y)$ vs. $I(X; Y|Z)$?

If $X \perp Y|Z$ then $I(X; Y|Z) = 0$. For example, $X \perp Y|Z$ whenever $X \rightarrow Z \rightarrow Y$.

Alternatively, if $Z = Y$, then $I(X; Y|Z) = 0$. 
What about $I(X; Y)$ vs. $I(X; Y|Z)$?

If $X \perp Y|Z$ then $I(X; Y|Z) = 0$. For example, $X \perp Y|Z$ whenever $X \rightarrow Z \rightarrow Y$.

Alternatively, if $Z = Y$, then $I(X; Y|Z) = 0$.

Thus, we can have $I(X; Y) > I(X; Y|Z)$. 
Conditioning and Mutual Information

What about $I(X; Y)$ vs. $I(X; Y|Z)$?

If $X \perp Y|Z$ then $I(X; Y|Z) = 0$. For example, $X \perp Y|Z$ 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 $X \perp Y$ then $I(X; Y) = 0$ but $I(X; Y|Z) > 0$. 
What about $I(X; Y)$ vs. $I(X; Y|Z)$?

If $X \perp Y|Z$ then $I(X; Y|Z) = 0$. For example, $X \perp Y|Z$ 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 $X \perp Y$ then $I(X; Y) = 0$ but $I(X; Y|Z) > 0$.

Thus, no general conditioning relationship for mutual information and conditional mutual information.
$$H(X) = EI(x) = - \sum_x p(x) \log p(x)$$ \hspace{1cm} (3.18)

$$H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y)$$ \hspace{1cm} (3.19)

$$H(Y|X) = - \sum_{x,y} p(x, y) \log p(y|x)$$ \hspace{1cm} (3.20)

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$ \hspace{1cm} (3.21)

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$ \hspace{1cm} (3.22)

$$0 \leq H(X) \leq \log n, \quad \text{where } n \text{ is } X\text{'s alphabet size.}$$ \hspace{1cm} (3.23)
Review and Summary

- KL-D: $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$
KL-Divergence & Cond. MI

Review and Summary

KL-D: \( D(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)

MI: \( I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x, y) \| p(x)p(y)) \)
Review and Summary

- **KL-D**: \( D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)
- **MI**: \( I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x, y)||p(x)p(y)) \)
- **CMI**: \( I(X; Y|Z) = \sum_{x,y,z} p(x, y, z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = H(X|Z) - H(X|Y, Z) \)
KL-Divergence & Cond. MI Review and Venn log sum Data Proc. Inequality Review and Summary

KL-D: $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

MI: $I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x, y)||p(x)p(y))$

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

Chain Rule MI: $I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1})$
KL-Divergence & Cond. MI

Review and Summary

**KL-D:**
\[ D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \]

**MI:**
\[ I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \]

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

**Chain Rule MI:**
\[ I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \]

**Cond Rel Ent:**
\[ D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \]
Review and Summary

- **KL-D**: \( D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)
- **MI**: \( I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x, y)||p(x)p(y)) \)
- **CMI**: 
  \[ I(X; Y|Z) = \sum_{x,y,z} p(x, y, z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = H(X|Z) - H(X|Y, Z) \]
- **Chain Rule MI**: \( I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \)
- **Cond Rel Ent**: 
  \[ D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \]
- **Chain Rule KL**: 
  \[ D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \]
**KL-Divergence & Cond. MI**

**Review and Venn**

**log sum**

**Data Proc. Inequality**

---

**KL-D**: \( D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)

**MI**: \( I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \)

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

**Chain Rule MI**: \( I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \)

**Cond Rel Ent**: \( D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \)

**Chain Rule KL**: \( D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \)

**Jensen**: \( f \) convex \( \Rightarrow \) \( Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \)

---

**Prof. Jeff Bilmes**
Review and Summary

KL-D: \( D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)

MI: \( I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \)

CMI:
\( I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = H(X|Z) - H(X|Y, Z) \)

Chain Rule MI: \( I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \)

Cond Rel Ent: \( D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \)

Chain Rule KL: \( D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \)

Jensen: \( f \) convex \( \Rightarrow Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \)

KL non-negative: \( D(p||q) \geq 0, \ D(p||q) = 0 \iff p = q. \)
Review and Summary

**KL-D:** \( D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)

**MI:** \( I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x, y)\|p(x)p(y)) \)

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

**Chain Rule MI:** \( I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \)

**Cond Rel Ent:** \( D(p(y|x)\|q(y|x)) \triangleq \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \)

**Chain Rule KL:** \( D(p(x, y)\|q(x, y)) = D(p(x)\|q(x)) + D(p(y|x)\|q(y|x)) \)

**Jensen:** \( f \) convex \( \Rightarrow \) \( Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \)

**KL non-negative:** \( D(p\|q) \geq 0, \ D(p\|q) = 0 \Leftrightarrow p = q. \)

**MI non-negative:** \( I(X; Y) \geq 0, \ I(X; Y) = 0 \Leftrightarrow X \perp \perp Y. \)
Review and Summary

KL-D: \( D(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)

MI: \( I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = D(p(x, y) \| p(x)p(y)) \)

CMI:
\[
I(X; Y \mid Z) = \sum_{x,y,z} p(x, y, z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} = H(X \mid Z) - H(X \mid Y, Z)
\]

Chain Rule MI: \( I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y \mid X_1, X_2, \ldots, X_{i-1}) \)

Cond Rel Ent: \( D(p(y \mid x) \| q(y \mid x)) \triangleq \sum_{x,y} p(x, y) \log \frac{p(y \mid x)}{q(y \mid x)} \)

Chain Rule KL: \( D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y \mid x) \| q(y \mid x)) \)

Jensen: \( f \) convex \( \Rightarrow \) \( E f(X) = \sum_x p(x) f(x) \geq f(EX) = f(\sum_x x p(x)) \)

KL non-negative: \( D(p \| q) \geq 0, \ D(p \| q) = 0 \iff p = q. \)

MI non-negative: \( I(X; Y) \geq 0, \ I(X; Y) = 0 \iff X \perp \perp Y. \)

Conditioning reduces entropy: \( H(X) \geq H(X \mid Y), \)
\[
H(X) = H(X \mid Y) \iff X \perp \perp Y.
\]
KL-Divergence & Cond. MI

Review and Venn

KL-D:
\[ D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \]

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

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

Chain Rule MI:
\[ I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \]

Cond Rel Ent:
\[ D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \]

Chain Rule KL:
\[ D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \]

Jensen:
If \( f \) convex \( \Rightarrow \) \[ Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \]

KL non-negative:
\[ D(p||q) \geq 0, \quad D(p||q) = 0 \iff p = q. \]

MI non-negative:
\[ I(X; Y) \geq 0, \quad I(X; Y) = 0 \iff X \perp \! \! \! \perp Y. \]

Conditioning reduces entropy:
\[ H(X) \geq H(X|Y), \quad H(X) = H(X|Y) \iff X \perp \! \! \! \perp Y. \]

Indep. bound on \( H \):
\[ H(X_1, \ldots, X_N) \leq \sum_i H(X_i) \], equality iff all independent.
Given random variable $X$, the entropy of (uncertainty in, average surprise in, information contained within, etc.) a random variable can be displayed using a 2D area, as given above.

The area of the set conveys the “degree of information”
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 $Y$. 

$H(X, Y)$

$H(X)$

$H(Y)$

$I(X; Y)$

$H(X | Y)$

$H(Y | X)$
Another way of looking at the same relationships.

\[ H(X, Y) \]

\[ H(X) \]

\[ H(Y) \]

\[ H(X|Y) \]

\[ I(X; Y) \]

\[ H(Y|X) \]
Given three random variables $X_1, X_2, X_3$ related by $p(x_1, x_2, x_3)$, the following Venn diagram characterizes the relationships.
Given three random variables $X_1, X_2, X_3$ related by $p(x_1, x_2, x_3)$, the following Venn diagram characterizes the relationships.
Note in the diagram, $I(X_1; X_2) = I(X_1; X_2|X_3) + I(X_1; X_2; X_3)$
Entropy, MI, CMI, 3 RVs, in a Venn Diagram

- Note in the diagram, \( I(X_1; X_2) = I(X_1; X_2|X_3) + I(X_1; X_2; X_3) \)

- We saw that \( I(X_1; X_2) \geq I(X_1; X_2|X_3) \), but that neither is ever negative.
Note in the diagram, \( I(X_1; X_2) = I(X_1; X_2 | X_3) + I(X_1; X_2; X_3) \)

We saw that \( I(X_1; X_2) \geq I(X_1; X_2 | X_3) \), but that neither is ever negative.

thus, \( I(X_1; X_2; X_3) \geq I(X_1; X_2) - I(X_1; X_2 | X_3) \) can be negative.
Entrophy, MI, CMI, 3 RVs, in a Venn Diagram

- Note in the diagram, \( I(X_1; X_2) = I(X_1; X_2|X_3) + I(X_1; X_2; X_3) \)

- We saw that \( I(X_1; X_2) \geq I(X_1; X_2|X_3) \), but that neither is ever negative.

- thus, \( I(X_1; X_2; X_3) \triangleq I(X_1; X_2) - I(X_1; X_2|X_3) \) can be negative.

- \( I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2|X_3) = I(X_2; X_3) - I(X_2; X_2|X_1) = I(X_3; X_1) - I(X_3; X_1|X_2) \)
Prof. Jeff Bilmes  
EE514a/Fall 2019/Info. Theory I – Lecture 3 - Oct 2th, 2019  
L3 F32/47(pg.86/170)

Entrophy, MI, CMI, 3 RVs, in a Venn Diagram

- Note in the diagram, $I(X_1; X_2) = I(X_1; X_2 | X_3) + I(X_1; X_2; X_3)$
- We saw that $I(X_1; X_2) \geq I(X_1; X_2 | X_3)$, but that neither is ever negative.
- thus, $I(X_1; X_2; X_3) \triangleq I(X_1; X_2) - I(X_1; X_2 | X_3)$ can be negative.

- $I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2 | X_3) = I(X_2; X_3) - I(X_2; X_2 | X_1) = I(X_3; X_1) - I(X_3; X_1 | X_2)$
- Also, $I(X_1; X_2; X_3) = H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) - H(X_3, X_1) + H(X_1, X_2, X_3)$
Entropy, MI, CMI, 3 RVs, in a Venn Diagram

- Note in the diagram, \( I(X_1; X_2) = I(X_1; X_2|X_3) + I(X_1; X_2; X_3) \)
- We saw that \( I(X_1; X_2) \gtrless I(X_1; X_2|X_3) \), but that neither is ever negative.
- thus, \( I(X_1; X_2; X_3) \triangleq I(X_1; X_2) - I(X_1; X_2|X_3) \) can be negative.

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

Also, \( I(X_1; X_2; X_3) = H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) - H(X_3, X_1) + H(X_1, X_2, X_3) \)

\(-I(X_1; X_2; X_3)\) called the EAR (explaining away residual) measure in pattern recognition, and "synergy" in neuroscience.
Note in the diagram, $I(X_1; X_2) = I(X_1; X_2 | X_3) + I(X_1; X_2; X_3)$

We saw that $I(X_1; X_2) \geq I(X_1; X_2 | X_3)$, but that neither is ever negative.

Thus, $I(X_1; X_2; X_3) \triangleq I(X_1; X_2) - I(X_1; X_2 | X_3)$ can be negative.

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

Also, $I(X_1; X_2; X_3) = H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) - H(X_3, X_1) + H(X_1, X_2, X_3)$

$-I(X_1; X_2; X_3)$ called the EAR (explaining away residual) measure in pattern recognition, and “synergy” in neuroscience. Also, $I(X_1; X_2; X_3) = I(X_1; X_2) + I(X_3; X_2) - I(X_1, X_3; X_2)$
Theorem 3.5.1

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

\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}
\]

(3.24)

and we have equality iff \(a_i/b_i = c = \text{const.}\)
**Log-sum inequality**

**Theorem 3.5.1**

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

\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}
\]

(3.24)

*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 \(0 \log 0/0 = 0\).
Theorem 3.5.1

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

\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{3.24}
\]

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

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

Why is this convex?
Consider \( f(t) = t \log t = t(\ln t)(\log e) \) which is strictly convex.

\[ f(t) = t \log t \]

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

Proof of log-sum inequality.

- Since $f$ is convex, Jensen’s inequality says that:

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (3.25)$$
Proof of log-sum inequality.

Since $f$ is convex, Jensen’s inequality says that:

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (3.25)$$

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

$$\sum_{i=1}^{n} \frac{a_i}{\sum_{j} b_j} \log \frac{a_i}{b_i} = $$
Proof of log-sum inequality.

- Since $f$ is convex, Jensen’s inequality says that:

$$\sum_{i} \alpha_i f(t_i) \geq f\left(\sum_{i} \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_{i} \alpha_i = 1. \quad (3.25)$$

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

$$\sum_{i=1}^{n} \frac{a_i}{\sum_{j} b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{b_i}{\sum_{j} b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \quad (3.28)$$
Proof of log-sum inequality.

Since $f$ is convex, Jensen’s inequality says that:

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_i \alpha_i = 1. \quad (3.25)$$

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

$$\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^n \alpha_i f(t_i) \quad (3.26)$$
Proof of log-sum inequality.

Since $f$ is convex, Jensen’s inequality says that:

$$
\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right)
$$
with $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. (3.25)

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

$$
\sum_{i=1}^{n} \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \alpha_i f(t_i)
$$

$$
\geq f\left(\sum_i \alpha_i t_i\right)
$$

(3.26)
Proof of log-sum inequality.

Since $f$ is convex, Jensen’s inequality says that:

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (3.25)$$

Set $\alpha_i = \frac{b_i}{\sum_{j=1}^n b_j}$ and $t_i = \frac{a_i}{b_i}$ in the following:

$$\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^n \alpha_i f(t_i) \quad (3.26)$$

$$\geq f\left(\sum_i \alpha_i t_i\right) = \left(\sum_i \alpha_i t_i\right) \log \left(\sum_j \alpha_j t_j\right) \quad (3.27)$$
Proof of log-sum inequality.

- Since $f$ is convex, Jensen’s inequality says that:

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (3.25)$$

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

$$\sum_{i=1}^{n} \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \alpha_i f(t_i) \quad (3.26)$$

$$\geq f\left(\sum_i \alpha_i t_i\right) = \left(\sum_i \alpha_i t_i\right) \log \left(\sum_j \alpha_i t_i\right) \quad (3.27)$$

$$= \left(\sum_i \frac{a_i}{\sum_j b_j}\right) \log \left(\frac{\sum_i a_i}{\sum_j b_j}\right) \quad (3.28)$$
Convexity of $D(p \| q)$ in the pair

- $D(p \| q)$ in convex in the pair, meaning

\[
D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1 \| q_1) + (1-\lambda) D(p_2 \| q_2)
\]
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).
Convexity of $D(p \parallel q)$ in the pair

- $D(p \parallel q)$ in 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.
Convexity of $D(p||q)$ in the pair

- $D(p||q)$ in 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

$$D(\lambda p_1 + (1 - \lambda)p_2||\lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1||q_1) + (1 - \lambda)D(p_2||q_2)$$
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, i.e., we have

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

(3.29)
Convexity of $D(p||q)$ in the pair

- $D(p||q)$ in 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, i.e., we have

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

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

(3.29)
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, i.e., we have

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

$$= \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)}$$
Convexity of $D(p||q)$ in the pair

- $D(p||q)$ in 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, i.e., we have

$$\begin{align*}
(\lambda p_1 + (1 - \lambda)p_2)(x) \log \frac{(\lambda p_1 + (1 - \lambda)p_2)(x)}{(\lambda q_1 + (1 - \lambda)q_2)(x)} & \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)} \\
& = \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)}
\end{align*}$$

(3.29) (3.30) (3.31)

- And then sum over $x$. 
Convexity of $D(p\|q)$ in the pair

- Note that we can set $q_1 = q_2$ to get convexity just in $p$. 
Convexity of $D(p||q)$ in the pair

- 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).
Convexity of $D(p||q)$ in the pair

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.**

\[
H(p)
\]
Entropy is concave in $p$

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

**Proof.**

\[
H(p) = - \sum_i p_i \log p_i
\]
Entropy is concave in $p$

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

Proof.

$$H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \quad (3.32)$$
Entropy is concave in $p$

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

**Proof.**

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{3.32}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \tag{3.33}
\]
Entropy is concave in $p$

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

**Proof.**

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \quad (3.32)
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \quad (3.33)
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i - p_i \log \frac{1}{|\mathcal{X}|} \right) \quad (3.34)
\]
Entropy is concave in $p$

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

Proof.

$$H(p) = -\sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \quad (3.32)$$

$$= \log |\mathcal{X}| - \left(\sum_i p_i \log p_i + p_i \log |\mathcal{X}|\right) \quad (3.33)$$

$$= \log |\mathcal{X}| - \left(\sum_i p_i \log p_i - p_i \log 1/|\mathcal{X}|\right) \quad (3.34)$$

$$= \log |\mathcal{X}| - D(p||u) \quad (3.35)$$
Entropy is concave in \( p \)

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

**Proof.**

\[
H(p) = -\sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{3.32}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \tag{3.33}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i - p_i \log 1/|\mathcal{X}| \right) \tag{3.34}
\]

\[
= \log |\mathcal{X}| - D(p||u) \tag{3.35}
\]

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 $p(x, y) = p(y|x)p(x)$. 
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 \(p(y|x)\).
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 \(I_{p(x)}(X; Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)}\),

\[
I_{\lambda p_1(x) + (1-\lambda)p_2(x)}(X; Y) \geq \lambda I_{p_1(x)}(X; Y) + (1 - \lambda) I_{p_2(x)}(X; Y)
\]
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 \(p(x)\).
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 \(I_{p(y|x)}(X; Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)}\),

\[
I_{\lambda p_1(y|x) + (1-\lambda)p_2(y|x)}(X; Y) \leq \lambda I_{p_1(y|x)}(X; Y) + (1 - \lambda) I_{p_2(y|x)}(X; Y)
\]
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 
\[
I_{p(x)}(X; Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)},
\]

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

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.
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.

\[ X \rightarrow p(y|x) \rightarrow Y \]
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.

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

  ![Diagram](image)

- 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.
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.

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

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.
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.

  ![Diagram of a noisy channel](image)

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

Let's view this as a picture:
Question: Given an information source, can additional processing gain more amount of information about that source?

Lets view this as a picture:

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.
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).
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.
**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 showing the process of data processing from source to refined observation.](image)

**Question:** Is it possible to obtain more information about a source given additional processing?

**Answer:** Unfortunately, no!
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:

**Question:** Is it possible to obtain more information about a source given additional processing?
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:

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

Definition 3.6.1

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

$$p(z, x | y) = p(z | y)p(x | y) \quad \forall x, y, z$$  \hspace{1cm} (3.36)

- This means that
  $$p(x, y, z) = p(z | x, y)p(y | x)p(x) = p(z | y)p(y | x)p(x)$$
Markov Chain

Definition 3.6.1
Random variables $X$, $Y$, and $Z$ form a Markov chain if $Z \perp \perp X \mid Y$. I.e.,

$$p(z, x \mid y) = p(z \mid y)p(x \mid y) \quad \forall x, y, z$$

(3.36)

- 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.
Random variables \( X, Y, \) and \( Z \) form a **Markov chain** if \( Z \perp X \mid Y \). I.e.,

\[
p(z, x \mid y) = p(z \mid y)p(x \mid y) \quad \forall x, y, z
\]  
(3.36)

- 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.

- Ex: If \( Z = f(Y) \), then \( X \rightarrow Y \rightarrow Z \) is true (i.e., \( X, Y, Z \) form a Markov chain).
Markov Chain

**Definition 3.6.1**
Random variables $X$, $Y$, and $Z$ form a **Markov chain** if $Z \perp\!\!\!\!\!\!\perp X \mid Y$. I.e.,

$$p(z, x \mid y) = p(z \mid y)p(x \mid y) \quad \forall x, y, z \quad (3.36)$$

- 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.

- 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 be random (independent of $X$) or deterministic. Key is that $X$ is irrelevant to determine $Z$ given $Y$. 

---

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 3 - Oct 2th, 2019
Markov Chain

Definition 3.6.1

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

$$p(z, x|y) = p(z|y)p(x|y) \quad \forall x, y, z$$  \hspace{1cm} (3.36)

- This means that
  $$p(x, y, z) = p(z|x, y)p(y|x)p(x) = p(z|y)p(y|x)p(x)$$

- Graphs (i.e., Bayesian networks) that can describe this.

- 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 be random (independent of $X$) 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 3.6.2 (Data Processing Inequality)

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

\[
I(X; Y) \geq I(X; Z)
\]  \hspace{1cm} (3.37)

- So in the Markov chain, the “arrows” correspond to processing and the random variables correspond to data.
Theorem 3.6.2 (Data Processing Inequality)

If $X \rightarrow Y \rightarrow Z$ then

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

(3.37)

- 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.
Theorem 3.6.2 (Data Processing Inequality)

If $X \rightarrow Y \rightarrow Z$ then

$$I(X; Y) \geq I(X; Z)$$ (3.37)

- 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) lose information about the original source, as measured by mutual information.
Data Processing Inequality

By the chain rule of mutual information:

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

Example: what if \( X = Y \)?

Going the other direction, we similarly get

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

Corollary: If \( Z = f(Y) \), then \( X \rightarrow Y \rightarrow Z \), or \( X \rightarrow Y \rightarrow f(Y) \).

Thus, \[ I(X; Y) \geq I(X; f(Y)) \].
Data Processing Inequality

By the chain rule of mutual information:

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

(3.38)
(3.39)

Now, \( X \perp Z|Y \) iff \( I(X; Z|Y) = 0 \).
Data Processing Inequality

proof of data processing inequality.

By the chain rule of mutual information:

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

(3.38)  
(3.39)

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

proof of data processing inequality.

By the chain rule of mutual information:

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

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

By the chain rule of mutual information:

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \]
\[ = I(X; Z) + I(X; Y|Z) \]  \hspace{1cm} (3.38) \hspace{1cm} (3.39)

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

Example: what if \( X = Y \)?
Data Processing Inequality

**proof of data processing inequality.**

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \]  \hspace{1cm} (3.38)

\[ = I(X; Z) + I(X; Y|Z) \]  \hspace{1cm} (3.39)

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

- Example: what if \( X = Y \)?
- What if \( H(Y|Z) = 0 \)?
Data Processing Inequality

proof of data processing inequality.

By the chain rule of mutual information:

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \]
\[ = I(X; Z) + I(X; Y|Z) \]
\[ (3.38) \]
\[ (3.39) \]

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

- Example: what if \( X = Y \)? What if \( H(Y|Z) = 0 \)?
- Going the other direction, we similarly get \( I(Y; Z) \geq I(X; Z) \).
Proof of Data Processing Inequality

By the chain rule of mutual information:

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \]  \hspace{1cm} (3.38)
\[ = I(X; Z) + I(X; Y|Z) \]  \hspace{1cm} (3.39)

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

- Example: what if \( X = Y \)? What if \( H(Y|Z) = 0 \)?
- Going the other direction, we similarly get \( I(Y; Z) \geq I(X; Z) \).
- Chain distance decreases information.
By the chain rule of mutual information:

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

(3.38) \hspace{1cm} (3.39)

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

- Example: what if \( X = Y \)? What if \( H(Y|Z) = 0 \)?
- Going the other direction, we similarly get \( I(Y; Z) \geq I(X; Z) \).
- Chain distance decreases information. How fast?
Data Processing Inequality

proof of data processing inequality.

By the chain rule of mutual information:

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

(3.38)
(3.39)

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

- Example: what if \( X = Y \)? What if \( H(Y|Z) = 0 \)?
- Going the other direction, we similarly get \( I(Y; Z) \geq I(X; Z) \).
- Chain distance decreases information. How fast?
- Corollary: If \( Z = f(Y) \), then \( X \rightarrow Y \rightarrow Z \), or \( X \rightarrow Y \rightarrow f(Y) \).
By the chain rule of mutual information:

\[
I(X; Y, Z) = I(X; Y) + I(X; Z | Y)
\]  \hspace{1cm} (3.38)
\[
= I(X; Z) + I(X; Y | Z)
\]  \hspace{1cm} (3.39)

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

- Example: what if \(X = Y\)? What if \(H(Y | Z) = 0\)?
- Going the other direction, we similarly get \(I(Y; Z) \geq I(X; Z)\).
- 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, \(I(X; Y) \geq I(X; f(Y))\).
Data Processing Inequality

- Processing can only lose information about $X$. When $X$ is source and $Y$ is receiver, no processing will increase information about $X$.
Data Processing Inequality

- 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.
Data Processing Inequality

- 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.
- Any further processing can only reduce information about original object.
Data Processing Inequality

- 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.
- Any further processing can only reduce information about original object. Deep neural networks: more layers, no more information.
Data Processing Inequality

- 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.
- Any further processing can only reduce information about original object. Deep neural networks: more layers, no more information.
- 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.
Data Processing Inequality

- 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.

- Any further processing can only reduce information about original object. Deep neural networks: more layers, no more information.

- 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 $I(X; Y|Z) \leq I(X; Y)$. 

\begin{align*}
\text{Prof. Jeff Bilmes} & \\
\text{EE514a/Fall 2019/Info. Theory I – Lecture 3 - Oct 2th, 2019} & \\
\text{L3 F46/47 (pg.164/170)} &
\end{align*}
Data Processing Inequality

- 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.
- Any further processing can only reduce information about original object. Deep neural networks: more layers, no more information.
- 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 $Y$. 

X Y Z
Raining
Outside
Ground is
Wet
Worms are
on sidewalk
Data Processing Inequality

- 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.
- Any further processing can only reduce information about original object. Deep neural networks: more layers, no more information.
- 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.
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 $I(X; Y|Z) \geq I(X; Y)$. 

- Data Processing Inequality
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.
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.