EE514a – Information Theory I
Fall Quarter 2013

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Fall Quarter, 2013
http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/

Lecture 3 - Oct 3rd, 2013
Class Road Map - IT-I

- L1 (9/26): Overview, Communications, Information, Entropy
- L2 (10/1): Props. Entropy, Mutual Information,
- L3 (10/3): KL-Divergence, Jensen, properties, Data Proc. Inequality
- L4
- L5
- L6
- L7
- L8
- L9
- L10
- L11
- L12
- L13
- L14
- L15
- L16
- L17
- L18
- L19

Finals Week: December 12th–16th.
Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano's inequality).
Homework

- Homework 0 was due Tuesday (October 1st) 11:45pm electronically.
- It is not graded but it counts credit/no-credit.
- Homework 1 is posted on our assignment dropbox (https://canvas.uw.edu/courses/847774/assignments), due next Tuesday night.
- How do you like canvas?
Entropy

Definition 3.2.1 (Entropy)

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

$$H(X) \triangleq E \log \frac{1}{p(X)} = \sum_x p(x) \log \frac{1}{p(x)} = -\sum_x p(x) \log p(x) \quad (3.2)$$

- 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).
Entropy Of Distributions

Low Entropy

High Entropy

In Between
Binary Entropy

- Binary alphabet, \( X \in \{0, 1\} \) say.
- \( p(X = 1) = p = 1 - p(X = 0) \).
- \( H(X) = -p \log p - (1 - p) \log (1 - p) = H(p) \).
- As a function of \( p \), we get:

![Graph showing binary entropy](image)

- Note, greatest uncertainty (value 1) when \( p = 0.5 \) and least uncertainty (value 0) when \( p = 0 \) or \( p = 1 \).
- Note also: concave in \( p \).
Why log?

• 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

$$\mathcal{H}(Y|X) \triangleq \sum_x p(x)\mathcal{H}(Y|X = x) \quad (3.22)$$

then we wish to have additivity

$$\mathcal{H}(X,Y) = \mathcal{H}(X) + \mathcal{H}(Y|X) \quad (3.23)$$

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?

Theorem 3.2.5 (Khinchin’s Theorem)

If \( H(p_1, \ldots, p_n) \) 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

\[
H(p_1, \ldots, p_n) = -\lambda \sum_i p_i \log p_i
\]

(3.22)

for \( \lambda \) a positive constant.

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

\[ H(X) = EI(x) = - \sum_x p(x) \log p(x) \quad (3.22) \]

\[ H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y) \quad (3.23) \]

\[ H(Y|X) = - \sum_{x,y} p(x, y) \log p(y|x) \quad (3.24) \]

\[ H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \quad (3.25) \]

\[ 0 \leq H(X) \leq \log n, \quad \text{where } n \text{ is } X\text{'s alphabet size.} \quad (3.26) \]
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.

This is **not** set in the standard sense. Rather the area of the regions convey “degree of information.”
Mutual Information and Entropy - Venn Diagram

A way of looking at the relationships.

- $H(X)$
- $H(Y)$
- $H(X|Y)$
- $H(Y|X)$
- $I(X;Y)$
- $H(X,Y)$
KL-Divergence Intuition

- Another fundamental relationship, here between two probability distributions, say $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ over the same alphabet size.
KL-Divergence Intuition

- Another fundamental relationship, here between two probability distributions, say $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ over the same alphabet size.
- Has important relationships to entropy and mutual information.
KL-Divergence Intuition

- Another fundamental relationship, here between two probability distributions, say \( p = (p_1, \ldots, p_n) \) and \( q = (q_1, \ldots, q_n) \) over the same alphabet size.

- Has important relationships to entropy and mutual information

- How do we measure the “distance” between two probability distributions \( p \) and \( q \) in a way that is useful?
KL-Divergence Intuition

- Another fundamental relationship, here between two probability distributions, say \( p = (p_1, \ldots, p_n) \) and \( q = (q_1, \ldots, q_n) \) over the same alphabet size.
- Has important relationships to entropy and mutual information
- How do we measure the “distance” between two probability distributions \( p \) and \( q \) in a way that is useful?
- One (of many) ways \( D(p, q) = \sum_{i=1}^{n} (p_i - q_i)^2 \)
KL-Divergence Intuition

- Another fundamental relationship, here between two probability distributions, say \( p = (p_1, \ldots, p_n) \) and \( q = (q_1, \ldots, q_n) \) over the same alphabet size.
- Has important relationships to entropy and mutual information.
- How do we measure the “distance” between two probability distributions \( p \) and \( q \) in a way that is useful?
- One (of many) ways \( D(p, q) = \sum_{i=1}^{n} (p_i - q_i)^2 \)
- But we’d like a form of “information distance”, i.e., if we think the distribution is \( q \) but it is really \( p \), what cost do we incur for this error.
Another fundamental relationship, here between two probability distributions, say $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$ over the same alphabet size.

Has important relationships to entropy and mutual information

How do we measure the “distance” between two probability distributions $p$ and $q$ in a way that is useful?

One (of many) ways $D(p, q) = \sum_{i=1}^{n}(p_i - q_i)^2$

But we’d like a form of “information distance”, i.e., if we think the distribution is $q$ but it is really $p$, what cost do we incur for this error.

Cost might take the form of compression inefficiency.
KL-Divergence Intuition

- Another fundamental relationship, here between two probability distributions, say \( p = (p_1, \ldots, p_n) \) and \( q = (q_1, \ldots, q_n) \) over the same alphabet size.
- Has important relationships to entropy and mutual information.
- How do we measure the “distance” between two probability distributions \( p \) and \( q \) in a way that is useful?
  - One (of many) ways \( D(p, q) = \sum_{i=1}^{n} (p_i - q_i)^2 \)
  - But we’d like a form of “information distance”, i.e., if we think the distribution is \( q \) but it is really \( p \), what cost do we incur for this error.
  - Cost might take the form of compression inefficiency.
- The KL-Divergence (equivalently “Kullbach-Leibler distance/divergence”, or the “Information Divergence”, or the “Information for discrimination”) satisfies these ideas.
Definition 3.3.1 (distance)

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a **distance** on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
**Definition 3.3.1 (distance)**

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a **distance** on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = d(y, x)$ (symmetry)
Definition 3.3.1 (distance)

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a distance on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, x) = 0$ (reflexivity)
Definition 3.3.1 (distance)

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a distance on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, x) = 0$ (reflexivity)
Distance

Definition 3.3.1 (distance)

Let \( S \) be a set. A function \( d : S \times S \rightarrow \mathbb{R} \) is called a distance on \( S \) if, for all \( x, y \in S \), we have:

- \( d(x, y) \geq 0 \) (non-negativity)
- \( d(x, y) = d(y, x) \) (symmetry)
- \( d(x, x) = 0 \) (reflexivity)

Definition 3.3.2 (metric)

Let \( S \) be a set. A function \( d : S \times S \rightarrow \mathbb{R} \) is called a metric on \( S \) if, for all \( x, y, z \in S \), we have:

- \( d(x, y) \geq 0 \) (non-negativity)
Distance

Definition 3.3.1 (distance)

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a distance on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, x) = 0$ (reflexivity)

Definition 3.3.2 (metric)

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a metric on $S$ if, for all $x, y, z \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = 0$ iff $x = y$ (identity of indiscernibles)

Semi-metric if we replace identity of indiscernibles with reflexivity.
Distance

Definition 3.3.1 (distance)
Let S be a set. A function d : S × S → ℝ is called a distance on S if, for all x, y ∈ S, we have:
- d(x, y) ≥ 0 (non-negativity)
- d(x, y) = d(y, x) (symmetry)
- d(x, x) = 0 (reflexivity)

Definition 3.3.2 (metric)
Let S be a set. A function d : S × S → ℝ is called a metric on S if, for all x, y, z ∈ S, we have:
- d(x, y) ≥ 0 (non-negativity)
- d(x, y) = 0 iff x = y (identity of indiscernibles)
- d(x, y) = d(y, x) (symmetry)
Definition 3.3.1 (distance)

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a **distance** on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, x) = 0$ (reflexivity)

Definition 3.3.2 (metric)

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a **metric** on $S$ if, for all $x, y, z \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = 0$ iff $x = y$ (identity of indiscernibles)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality)
Distance

**Definition 3.3.1 (distance)**

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a **distance** on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, x) = 0$ (reflexivity)

**Definition 3.3.2 (metric)**

Let $S$ be a set. A function $d : S \times S \rightarrow \mathbb{R}$ is called a **metric** on $S$ if, for all $x, y, z \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = 0$ iff $x = y$ (identity of indiscernibles)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality)
Distance

Definition 3.3.1 (distance)
Let $S$ be a set. A function $d : S \times S \to \mathbb{R}$ is called a distance on $S$ if, for all $x, y \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, x) = 0$ (reflexivity)

Definition 3.3.2 (metric)
Let $S$ be a set. A function $d : S \times S \to \mathbb{R}$ is called a metric on $S$ if, for all $x, y, z \in S$, we have:

- $d(x, y) \geq 0$ (non-negativity)
- $d(x, y) = 0$ iff $x = y$ (identity of indiscernibles)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, y) + d(y, z) \geq d(x, z)$ (triangle inequality)

Semi-metric if we replace identity of indiscernibles with reflexivity.
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:

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

**Definition 3.3.3 (KL-Divergence)**

$$D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (3.1)$$
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:

**Definition 3.3.3 (KL-Divergence)**

$$D(p\|q) \equiv \sum_x p(x) \log \frac{p(x)}{q(x)}$$  \hspace{1cm} (3.1)

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

**Definition 3.3.3 (KL-Divergence)**

$$D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}$$ (3.1)

- 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”).
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:

Definition 3.3.3 (KL-Divergence)

\[
D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}
\] (3.1)

- 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 over vectors

- KL-divergence can be generalized to distributions over vectors of random variables.
KL-Divergence over vectors

- KL-divergence can be generalized to distributions over vectors of random variables.
- Let \( p(x_1, \ldots, x_N) \) and \( q(x_1, \ldots, x_N) \) be two distributions over vector \((x_1, x_2, \ldots, x_N)\).
KL-Divergence can be generalized to distributions over vectors of random variables.

Let \( p(x_1, \ldots, x_N) \) and \( q(x_1, \ldots, x_N) \) be two distributions over vector \((x_1, x_2, \ldots, x_N)\).

Then we can define the KL-divergence between \( p \) and \( q \) as

\[
D(p\|q) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) \log \frac{p(x_1, x_2, \ldots, x_N)}{q(x_1, x_2, \ldots, x_N)}
\] (3.2)
KL-Divergence over vectors

- KL-divergence can be generalized to distributions over vectors of random variables.
- Let \( p(x_1, \ldots, x_N) \) and \( q(x_1, \ldots, x_N) \) be two distributions over vector \((x_1, x_2, \ldots, x_N)\).
- Then we can define the KL-divergence between \( p \) and \( q \) as

\[
D(p||q) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) \log \frac{p(x_1, x_2, \ldots, x_N)}{q(x_1, x_2, \ldots, x_N)} \quad (3.2)
\]

- So, like entropy, MI, etc. KL-divergence is a function of the probability values, not the values that the random variables take on.
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 \( p(y) = \sum_x p(x, y) \)

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., \( p(x)p(y) = \arg\max p'(x, y) \) s.t. \( p'(x, y) = p'(x)p'(y) \)

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

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

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

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

(3.3)

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

(3.4)

- 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 $Y$. 
Let $\mu_1(x, y) = p(x, y)$ and $\mu_2(x, y) = p(x)p(y)$ with 
\[ p(x) = \sum_y p(x, y) \text{ and } p(y) = \sum_x p(x, y) \]
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.
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.
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

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

(3.6)
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

\[
I(X; Y|Z = z) = \sum_{x,y} p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \tag{3.6}
\]

Note that this is identical to regular mutual information except in this case we are always conditioning on the event \( z \).
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

\[
I(X; Y | Z = z) = \sum_{x,y} p(x, y | z) \log \frac{p(x, y | z)}{p(x | z)p(y | z)}
\] (3.6)

- 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:

\[
I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\] (3.7)

we use different distributions, \( p(x, y) \to p(x, y | z), p(x) \to p(x | z) \), and \( p(y) \to p(y | z) \).
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.
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 3.3.4 (conditional mutual information)**

\[
I(X; Y|Z)
\]
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 3.3.4 (conditional mutual information)**

$$I(X; Y | Z) \triangleq \sum_z p(z) I(X; Y | Z = z)$$ (3.8)
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 3.3.4 (conditional mutual information)**

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

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

(3.8)

(3.9)
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 3.3.4 (conditional mutual information)**

\[
I(X; Y|Z) \triangleq \sum_z p(z) I(X; Y|Z = z) \tag{3.8}
\]

\[
= \sum_z p(z) E_{p(x,y|z)} \log \frac{p(x, y|Z = z)}{p(x|Z = z)p(y|Z = z)} \tag{3.9}
\]

\[
= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \tag{3.10}
\]
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 3.3.4 (conditional mutual information)**

\[
I(X; Y|Z) \triangleq \sum_z p(z)I(X; Y|Z = z) \tag{3.8}
\]

\[
= \sum_z p(z)E_{p(x,y|z)} \log \frac{p(x, y|Z = z)}{p(x|Z = z)p(y|Z = z)} \tag{3.9}
\]

\[
= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \tag{3.10}
\]

\[
= E \left[ \log \frac{1}{p(x|z)} - \log \frac{1}{p(x|y, z)} \right] \tag{3.11}
\]
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 3.3.4 (conditional mutual information)**

\[
I(X; Y|Z) \triangleq \sum_z p(z) I(X; Y|Z = z) \tag{3.8}
\]

\[
= \sum_z p(z) E_{p(x,y|z)} \left( \log \frac{p(x, y|Z = z)}{p(x|Z = z)p(y|Z = z)} \right) \tag{3.9}
\]

\[
= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \tag{3.10}
\]

\[
= E \left[ \log \frac{1}{p(x|z)} - \log \frac{1}{p(x|y, z)} \right] \tag{3.11}
\]

\[
= H(X|Z) - H(X|Y, Z) \tag{3.12}
\]
Proposition 3.3.5

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

(3.13)

example: 
\[
I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1)
\]  

(3.14)

Proof.

\[
I(X_1, \ldots, X_N; Y) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_N|Y)
\]  

(3.15)
Proposition 3.3.5

\[ I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y | X_1, X_2, \ldots, X_{i-1}) \] (3.13)

example: \[ I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y | X_1) \] (3.14)

Proof.

\[ I(X_1, \ldots, X_N; Y) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_N | Y) \] (3.15)

\[ = \sum_i H(X_i | X_1, \ldots, X_{i-1}) - \sum_i H(X_i | X_1, \ldots, X_{i-1}, Y) \] (3.16)
Chain Rule for Mutual Information

**Proposition 3.3.5**

\[
I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \tag{3.13}
\]

Example:

\[
I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \tag{3.14}
\]

**Proof.**

\[
I(X_1, \ldots, X_N; Y) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_N|Y) \tag{3.15}
\]

\[
= \sum_i H(X_i|X_1, \ldots, X_{i-1}) - \sum_i H(X_i|X_1, \ldots, X_{i-1}, Y) \tag{3.16}
\]

\[
= \sum_i I(X_i; Y|X_1, \ldots, X_{i-1}) \tag{3.17}
\]
Definition 3.3.6

\[ D(p(y|x) || q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \] (3.18)

- Same as standard KL-divergence but now using conditional distribution.
Proposition 3.3.7

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

(3.19)
Proposition 3.3.7

\[ D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x)) \]  \hspace{1cm} (3.19)

Proof.

\[ D(p(x, y) \| q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \]  \hspace{1cm} (3.20)
Proposition 3.3.7

\[ D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x)) \]  (3.19)

Proof.

\[ D(p(x, y) \| q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \]  (3.20)

\[ = \sum_{x,y} p(x, y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \]  (3.21)
Proposition 3.3.7

\[ D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y|x) || q(y|x)) \] (3.19)

Proof.

\[ D(p(x, y) || q(x, y)) = \sum_{x, y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \] (3.20)

\[ = \sum_{x, y} p(x, y) \log \frac{p(y|x) p(x)}{q(y|x) q(x)} \] (3.21)

\[ = \sum_{x, y} p(x, y) \log \frac{p(y|x)}{q(y|x)} + \sum_{x, y} p(x, y) \log \frac{p(x)}{q(x)} \] (3.22)
Convex Functions

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

  (3.23)
ConVex Functions

- $f$ is said to be convex on $(a, b)$ if for all $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$,
  \[ f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \]  
  (3.23)

- 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 $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$,
  \[ f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2) \]  
  (3.23)
- Many convex functions, $f(x) = x^2$, or $f(x) = e^x$, or $f(x) = x \log x$, $x \geq 0$.
- Visualized:

![Visualized Convex Function](image)
ConVex Functions

- $f$ is said to be convex on $(a, b)$ if for all $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$,
  \[ f(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2) \]  
  (3.23)

- 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 $\lambda = 1$. 

---

Prof. Jeff Bilmes  
EE514a/Fall 2013/Information Theory I – Lecture 3 - Oct 3rd, 2013  
L3 F23/69 (pg.66/295)
Theorem 3.3.8 (Jensen)

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

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

(3.24)
Theorem 3.3.8 (Jensen)

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

$$ Ef(X) = \sum_x p(x) f(x) \geq f(EX) = f(\sum_x x p(x)) \quad (3.24) $$

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

Lemma 3.3.9

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  (3.25)

Proof.

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

\[ -D(p||q) \leq \log \left( \sum_{x \in A} q(x) \right) \leq \log \left( \sum_x q(x) \right) = \log 1 = 0 \]  (3.28)
KL Divergence is non-negative

**Lemma 3.3.9**

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \quad (3.25) \]

**Proof.**

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

\[ -D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (3.26) \]

(3.28)
KL Divergence is non-negative

**Lemma 3.3.9**

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \] (3.25)

**Proof.**

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

\[ -D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \] (3.26)
KL Divergence is non-negative

**Lemma 3.3.9**

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  
(3.25)

**Proof.**

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

\[ -D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \]  
(3.26)

\[ = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \]  
(3.28)
KL Divergence is non-negative

**Lemma 3.3.9**

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

(3.25)

**Proof.**

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

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

(3.26)

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

(3.27)

(3.28)
KL Divergence is non-negative

**Lemma 3.3.9**

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

**Proof.**

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

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

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

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

Lemma 3.3.9

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

(3.25)

Proof.

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

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

(3.26)

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

(3.27)

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

(3.28)
KL Divergence is non-negative

Lemma 3.3.9

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  (3.25)

Proof.

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

\[ -D(p||q) = -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \]  (3.26)

\[ = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \]  (3.27)

\[ = \log \left( \sum_{x \in A} q(x) \right) \leq \log \left( \sum_x q(x) \right) = \log 1 = 0 \]  (3.28)
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.
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.
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 \( Z = p(X)/q(X) \), so \( Z \) is a constant random variable.
KL Divergence is non-negative

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

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

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

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

Thus, if $p(x) = q(x)$ then $D(p||q) = 0$ and vice versa.
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 $Z = p(X)/q(X)$, so $Z$ is a constant random variable.

- The only valid constant, with $p$ and $q$ still being probability distributions is $Z = 1$ 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.
Mutual Information is non-negative

Proposition 3.3.10

\[ I(X; Y) \geq 0 \quad \text{and} \quad I(X; Y) = 0 \iff X \perp \!
\!
\perp Y \] (3.29)
Mutual Information is non-negative

Proposition 3.3.10

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

(3.29)

Proof.

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

(3.30)

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 $Y$. 
Mutual Information, more intuition

- So $I(X;Y)$ measures the “degree of dependence” between $X$ and $Y$.
- We have $0 \leq I(X;Y) \leq \min(H(X), H(Y))$. 

So $I(X; Y)$ measures the “degree of dependence” between $X$ and $Y$.

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

$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. 
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 $H(Y|X) = H(Y)$. 

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 3.3.11

$$H(X|Y) \leq H(X) \text{ and } H(X|Y) = H(X) \text{ iff } X \perp \!\!\!\!\!\!\! Y$$ (3.31)
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 3.3.11**

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

(3.31)
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 3.3.11

\[ H(X|Y) \leq H(X) \text{ and } H(X|Y) = H(X) \text{ iff } X \perp \! \! \! \! \perp Y \] (3.31)

Proof.

\[ 0 \leq I(X;Y) = H(X) - H(X|Y) \] (3.32)

- As mentioned, we could have $H(X|Y = y) > H(X)$, but (in the average case), $\sum_y p(y)H(X|Y = y) \leq H(X)$.
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 3.3.12

\[ H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \] (3.33)
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 3.3.12**

\[
H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \tag{3.33}
\]
Entropy of a set of random variables is highest when the random variables are independent - the least redundancy between them.

**Proposition 3.3.12**

\[
\sum_{i=1}^{N} H(x_i | x_{-i}) \leq H(x_1, x_2, \ldots, x_N) \leq \sum_{i=1}^{N} H(x_i)
\]  

\((\text{3.33})\)

**Proof.**

\[
H(x_1, \ldots, x_N) = \sum_{i=1}^{N} H(x_i | x_1, \ldots, x_{i-1}) \leq \sum_{i=1}^{N} H(x_i)
\]

\((\text{3.34})\)
Independence Bounds on Entropy

Two variable instance of Proposition 3.3.12 is

\[ H(X_1, X_2) \leq H(X_1) + H(X_2) \]  \hspace{1cm} (3.35)

Note that equality in Equation 3.70 is achieved when all variables are mutually independent. I.e. when \( X_i \perp \perp X_j \) for all \( i, j \).
Two variable instance of Proposition 3.3.12 is

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

(3.35)

Note that equality in Equation 3.70 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. $I(X; Y|Z)$?
What about $I(X; Y)$ vs. $I(X; Y|Z)$?

If $X \independent Y|Z$ then $I(X; Y|Z) = 0$. For example, $X \independent Y|Z$ whenever $X \rightarrow Z \rightarrow Y$. Therefore, no general conditioning relationship for mutual information and conditional 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 $I(X; Y|Z) = 0$. 
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 $I(X; Y) > I(X; Y|Z)$.
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 $I(X; Y|Z) > 0$.
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 $X \to Z \to Y$.

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} (3.36)

\( H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y) \) \hspace{1cm} (3.37)

\( H(Y|X) = - \sum_{x,y} p(x, y) \log p(y|x) \) \hspace{1cm} (3.38)

\( H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \) \hspace{1cm} (3.39)

\( I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \) \hspace{1cm} (3.40)

0 \leq H(X) \leq \log n, \text{ where } n \text{ is } X\text{'s alphabet size.}
KL-Divergence & Cond. MI

Review and Venn

log sum

Data Proc. Inequality

Entropy & Thermo


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

Review and Venn

log sum

Data Proc. Inequality

Entropy & Thermo


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

Data Proc. Inequality

Entropy & Thermo


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: $D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$
## 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 \( \Rightarrow E f(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \)
KL-Divergence & Cond. MI

Review and Summary

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

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

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

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

MI non-negative: \( I(X; Y) \geq 0, \ I(X; Y) = 0 \iff X \perp\!\!\!\!\!\!\perp Y. \)
**KL-Divergence & Cond. MI**

**Review and Venn**

**log sum**

**Data Proc. Inequality**

**Entropy & Thermo**

**Suff. Stat.**

---

**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:** $D(p||q) \geq 0$, $D(p||q) = 0 \iff p = q$.  

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

- **Conditioning reduces entropy:** $H(X) \geq H(X|Y)$,  
  $H(X) = H(X|Y) \iff X \perp \perp Y$.  

---

**Prof. Jeff Bilmes**  
EE514a/Fall 2013/Information Theory I – Lecture 3 - Oct 3rd, 2013  
L3 F34/69 (pg.114/295)
### 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:** \[ f \text{ convex } \Rightarrow E f(X) = \sum_x p(x) f(x) \geq f(EX) = f(\sum_x xp(x)) \]

- **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:** \[ H(X_1, \ldots, X_N) \leq \sum_i H(X_i), \text{ equality iff all independent.} \]
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.

This is **not** set in the standard sense. Rather the area of the regions convey “degree of information.”
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 $Y$. 
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, \( I(X_1; X_2) = I(X_1; X_2|X_3) + I(X_1; X_2; X_3) \)
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 \( I(X_1; X_2) \leq I(X_1; X_2|X_3) \), 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
\( I(X_1; X_2) \not\geq I(X_1; X_2|X_3) \), but that neither is ever negative.

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

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

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

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

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

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

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 $I(X_1; X_2) \nless I(X_1; X_2|X_3)$, 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, $I(X_1; X_2; X_3) = I(X_1; X_2) + I(X_3; X_2) - I(X_1, X_3; X_2)$
\[ I(x_1; y) = \sum_{x, y} p(x, y) \log \frac{p(x_1, y)}{p(x_1)p(y)} \]

\[ I(x_1; x_2; x_3) = \sum_{x_1, x_2, x_3} p(x_1, x_2, x_3) \log \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)} \]
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

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

(3.41)

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

\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{3.41}
\]

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

\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{3.41}
\]

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

\[
\begin{align*}
\frac{d^2 f(t)}{dt^2} & = -\frac{1}{t} < 0 \\
\text{for all } t > 0
\end{align*}
\]

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

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:

$$
\sum_{i} \alpha_i f(t_i) \geq f\left( \sum_{i} \alpha_i t_i \right) \text{ with } \alpha_i \geq 0 \text{ and } \sum_{i} \alpha_i = 1. \quad (3.42)
$$
Proof of log-sum inequality.

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

\[
\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_i \alpha_i = 1. \tag{3.42}
\]

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

\[
\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i}
\]
Log-sum inequality

Proof of log-sum inequality.

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

\[
\sum_i \alpha_i f(t_i) \geq f(\sum_i \alpha_i t_i) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_i \alpha_i = 1. \quad (3.42)
\]

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

\[
\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \quad (3.45)
\]
Proof of log-sum inequality.

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

$$
\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_i \alpha_i = 1. \quad (3.42)
$$

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

$$
\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^n \alpha_i f(t_i) \quad (3.43)
$$
Proof of log-sum inequality.

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

\[
\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_i \alpha_i = 1. \quad (3.42)
\]

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

\[
\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^n \alpha_i f(t_i) \quad (3.43)
\]

\[
\geq f\left(\sum_i \alpha_i t_i\right)
\]
Log-sum inequality

Proof of log-sum inequality.

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

\[
\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (3.42)
\]

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

\[
\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^n \alpha_i f(t_i) \quad (3.43)
\]

\[
\geq f\left(\sum_i \alpha_i t_i\right) = \left(\sum_i \alpha_i t_i\right) \log \left(\sum_j \alpha_i t_i\right) \quad (3.44)
\]
Proof of log-sum inequality.

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

$$\sum_i \alpha_i f(t_i) \geq f(\sum_i \alpha_i t_i) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (3.42)$$

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

$$\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_i} \left( \frac{a_i}{b_i} \right) \log \frac{a_i}{b_i} = \sum_{i=1}^n \alpha_i f(t_i) \quad (3.43)$$

$$\geq f(\sum_i \alpha_i t_i) = \left( \sum_i \alpha_i t_i \right) \log \left( \sum_j \alpha_i t_i \right) \quad (3.44)$$

$$= \left( \sum_i \frac{a_i}{\sum_j b_j} \right) \log \left( \frac{\sum_i a_i}{\sum_j b_j} \right) \quad (3.45)$$
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

$$\left(\lambda p_1 + (1 - \lambda)p_2\right)(x) \log \left(\lambda p_1 + (1 - \lambda)p_2\right)(x) \leq \lambda p_1(x) \log p_1(x) + (1 - \lambda)p_2(x) \log p_2(x)$$

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

- $D(p \parallel q)$ 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.
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

\[
(\lambda p_1 + (1 - \lambda)p_2)(x) \log \frac{\lambda p_1 + (1 - \lambda)p_2(x)}{\lambda q_1 + (1 - \lambda)q_2(x)} \leq (\lambda p_1 || q_1) + (1 - \lambda)D(p_2 || q_2)
\] (3.46)
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

\[
(\lambda p_1 + (1 - \lambda)p_2)(x) \log \frac{(\lambda p_1 + (1 - \lambda)p_2)(x)}{(\lambda q_1 + (1 - \lambda)q_2)(x)} \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)}
\] (3.46)

\[
= \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)}
\] (3.47)
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

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

$$ \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1 - \lambda)p_2(x) \log \frac{(1 - \lambda)p_2(x)}{(1 - \lambda)q_2(x)} \quad (3.47) $$

$$ = \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1 - \lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)} \quad (3.48) $$
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

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

- 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.**

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |X| - \log |X| \quad (3.49)
\]
Entropy is concave in $p$

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

Proof.

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{3.49}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \tag{3.50}
\]
Entropy is concave in $p$

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

**Proof.**

\[
H(p) = -\sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{3.49}
\]

\[
= \log |\mathcal{X}| - \left(\sum_i p_i \log p_i + p_i \log |\mathcal{X}|\right) \tag{3.50}
\]

\[
= \log |\mathcal{X}| - \left(\sum_i p_i \log p_i - p_i \log \frac{1}{|\mathcal{X}|}\right) \tag{3.51}
\]
Entropy is concave in $p$

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

Proof.

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{3.49}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \tag{3.50}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i - p_i \log 1/|\mathcal{X}| \right) \tag{3.51}
\]

\[
= \log |\mathcal{X}| - D(p||u) \tag{3.52}
\]
Entropy is concave in $p$

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

**Proof.**

\[
H(p) = -\sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \quad (3.49)
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \quad (3.50)
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i - \frac{p_i}{|\mathcal{X}|} \log |\mathcal{X}| \right) \quad (3.51)
\]

\[
= \log |\mathcal{X}| - D(p||u) \quad (3.52)
\]

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

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

- 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.
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 showing $X$ sending to $Y$ through $p(y|x)$]
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.

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

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.
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.

$$X \xrightarrow{p(y|x)} Y$$

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.

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.

\[
\begin{array}{ccc}
X & \rightarrow & p(y|x) \\
& & \rightarrow \\
& & Y
\end{array}
\]

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.
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?

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

Let's 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](image)

**Information Source (state of nature)**

**True Amount Of Information In Source**

**Imperfect Observation Process**

**Further Processing**

**Refined Observed Process & Information**
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:

Information Source (state of nature)

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:

```
Xamount of information
```

```
Yamount of information
```

```
Zamount of information
```

- Question: Is it possible to obtain more information about a source given additional processing? **Unfortunately, no!**
Definition 3.6.1

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

$$p(z, x \mid y) = p(z \mid y)p(x \mid y) \quad \forall x, y, z \quad (3.53)$$

- This means that

$$p(x, y, z) = p(z \mid x, y)p(y \mid x)p(x) = p(z \mid y)p(y \mid x)p(x)$$
Definition 3.6.1

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

\[
p(z, x \mid y) = p(z \mid y)p(x \mid y) \quad \forall x, y, z
\] (3.53)

- 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 $Z \perp X|Y$. I.e.,

$$p(z, x|y) = p(z|y)p(x|y) \quad \forall x, y, z \quad (3.53)$$

- 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).
Definition 3.6.1

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

$$p(z, x | y) = p(z | y)p(x | y) \quad \forall x, y, z \tag{3.53}$$

- 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 random or deterministic. Key is that $X$ is irrelevant to determine $Z$ given $Y$. 
Definition 3.6.1

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

$$p(z, x \mid y) = p(z \mid y)p(x \mid y) \quad \forall x, y, z \tag{3.53}$$

- 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 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 3.6.2 (Data Processing Inequality)

\[ I(X; Y) \geq I(X; Z) \] (3.54)

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

\[
I(X; Y) \geq I(X; Z)
\]  (3.54)

- 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 \( X \to Y \to Z \) then

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

(3.54)

- 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

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} (3.55)
\[ = I(X; Z) + I(X; Y|Z) \]  \hspace{1cm} (3.56)
proof of data processing inequality.

By the chain rule of mutual information:

\[
I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \tag{3.55}
\]
\[
= I(X; Z) + I(X; Y|Z) \tag{3.56}
\]

Now, \(X \perp \!\!\!\!\!\!\!\!\!\!\!\perp Z|Y\) iff \(I(X; Z|Y) = 0\).
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} (3.55)

\[ = I(X; Z) + I(X; Y|Z) \]  \hspace{1cm} (3.56)

Now, \( X \perp\!\!\!\!\!\!\!\!\!\!\perp Z|Y \) 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:

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

(3.55)

(3.56)

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

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

\( \square \)
By the chain rule of mutual information:

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \]  \hspace{1cm} (3.55)
\[ = I(X; Z) + I(X; Y|Z) \]  \hspace{1cm} (3.56)

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 \( X = Y \)?
Data Processing Inequality

proof of data processing inequality.

By the chain rule of mutual information:

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

(3.55)  

(3.56)

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 \(H(Y|Z) = 0\)?
Data Processing Inequality

**proof of data processing inequality.**

By the chain rule of mutual information:

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$$  \hspace{1cm} (3.55)

$$= I(X; Z) + I(X; Y|Z)$$  \hspace{1cm} (3.56)

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

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

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

proof of data processing inequality.

By the chain rule of mutual information:

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \]  \hspace{2cm} (3.55)

\[ = I(X; Z) + I(X; Y|Z) \]  \hspace{2cm} (3.56)

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 \( H(Y|Z) = 0 \)?
- Similarly \( I(Y; Z) \geq I(X; Z) \).
- Chain distance decreases information.
proof of data processing inequality.

By the chain rule of mutual information:

\[ I(X; Y, Z) = I(X; Y) + I(X; Z | Y) \]
\[ = I(X; Z) + I(X; Y | Z) \]  \hspace{1cm} (3.55)
\[ = I(X; Z) + I(X; Y | Z) \]  \hspace{1cm} (3.56)

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 \( H(Y | Z) = 0 \)?
- Similarly \( I(Y; Z) \geq I(X; Z) \).
- Chain distance decreases information. How fast?
Data Processing Inequality

proof of data processing inequality.

By the chain rule of mutual information:

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) = I(X; Z) + I(X; Y|Z) \] (3.55) (3.56)

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 \( H(Y|Z) = 0 \)?
- 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 \( X \rightarrow Y \rightarrow f(Y) \).
proof of data processing inequality.

By the chain rule of mutual information:

\[
I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \tag{3.55}
\]
\[
= I(X; Z) + I(X; Y|Z) \tag{3.56}
\]

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

- Example: what if \( X = Y \)? What if \( H(Y|Z) = 0 \)?
- 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, \( I(X; Y) \geq I(X; f(Y)) \).
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.
- Then 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.
- 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.
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.
- 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 $I(X; Y|Z) \leq I(X; Y)$.
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.
- 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., \( I(X; Y|Z) = H(X|Z) - H(X|Y, Z) \leq H(X) - H(X|Y) \).
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.
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 $I(X; Y|Z) \geq I(X; Y)$.
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.
We have probably heard that entropy, in the universe, always increases. How does this relate to our entropy?
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).
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 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 unavailability 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

- Hot
- Cold
- Warm

There exists potential energy $T_{\text{emp.}}$ distribution over space.
Uneven distribution, smaller entropy same overall energy (1st law) less usable energy

Prof. Jeff Bilmes
Thermal energy in a closed system

- Less entropy
Thermal energy in a closed system

- Less entropy
- There exists potential energy
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- **Temp. distribution over space**

![Temperature distribution over space](image)
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space

Uneven distribution, smaller $H$
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space
  
  $T$
  
  space

- Uneven distribution, smaller $H$

- more entropy
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space

Uneven distribution, smaller $H$

- more entropy
- same overall energy (1st law)
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space

![Temperature distribution](image)

- Uneven distribution, smaller $H$

- More entropy
- Same overall energy (1st law)
- Less usable energy
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space

- Uneven distribution, smaller $H$

- more entropy
- same overall energy (1st law)
- less usable energy

Uniform temperature
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space

Uneven distribution, smaller $H$

- More entropy
- Same overall energy (1st law)
- Less usable energy
- Uniform temperature
- Temp. distribution over space
Thermal energy in a closed system

- Less entropy
- There exists potential energy
- Temp. distribution over space
  \[ T \]
  \[ \text{space} \]
  - Uneven distribution, smaller \( H \)

- More entropy
- Same overall energy (1st law)
- Less usable energy
- Uniform temperature
- Temp. distribution over space
  \[ T \]
  \[ \text{space} \]
  - Uniform distribution, big \( H \)
Entropy and thermodynamics

Claim: When there exists a transaction, the entropy always increases.
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.
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,** $p(x_{n+1}|x_n)$
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 $n$**
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 \( \cdots \rightarrow x_{n-1} \rightarrow x_n \rightarrow x_{n+1} \rightarrow \cdots \)
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.
Entropic 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 \( \cdots \to x_{n-1} \to x_n \to x_{n+1} \to \cdots \)
- 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 
  \[ D(p_n(x_n) || q_n(x_n)) \] won’t increase with \( n \).
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 \( D(p_n(x_n)||q_n(x_n)) \) won’t increase with \( n \). Why?
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 $D(p_n(x_n) || q_n(x_n))$ won’t increase with $n$. Why?
- **We’ll write** $p(x_n)$ for $p_n(x_n)$, also $q(x_n)$ for $q(x_n)$.
**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
  \[ D(p_n(x_n) \| q_n(x_n)) \] won’t increase with \( n \). Why?
- **We’ll write** \( p(x_n) \) for \( p_n(x_n) \), also \( q(x_n) \) for \( q(x_n) \).
- **Joint of two successive variables, via the homogeneous dynamics, is:**
  \[ p(x_n, x_{n+1}) = p(x_{n+1}|x_n)p_n(x_n) \quad (3.57) \]
  \[ q(x_n, x_{n+1}) = p(x_{n+1}|x_n)q_n(x_n) \quad (3.58) \]
Assume two stochastic processes with the same dynamics, but two distinct initial conditions.
Entropy and thermodynamics

- Assume two stochastic processes with the same dynamics, but two distinct initial conditions.
- Then we have:

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

(3.59)

\[ D(p(x_n, x_{n+1}) \| q(x_n, x_{n+1})) + D(p(x_{n+1}|x_n) \| q(x_{n+1}|x_n)) \geq 0 \]

(3.60)

(3.61)
Entropy and thermodynamics

- Assume two stochastic processes with the same dynamics, but two distinct initial conditions.
- Then we have:

\[ D(p(x_n, x_{n+1}) \| q(x_n, x_{n+1})) \]
\[ = D(p(x_n) \| q(x_n)) + D(p(x_{n+1}|x_n) \| q(x_{n+1}|x_n)) \]  \hspace{1cm} (3.59)

\[ \geq 0 \]  \hspace{1cm} (3.60)

\[ D(p(x_n) \| q(x_n)) \geq D(p(x_{n+1}|x_n) \| q(x_{n+1}|x_n)) \]  \hspace{1cm} (3.62)
Assume two stochastic processes with the same dynamics, but two distinct initial conditions.

Then we have:

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

\[
= 0
\]

(3.59)

(3.60)

(3.61)
Entropy and thermodynamics

- Assume two stochastic processes with the same dynamics, but two distinct initial conditions.

- Then we have:

\[
D(p(x_n, x_{n+1}) \parallel q(x_n, x_{n+1}))
= D(p(x_n) \parallel q(x_n)) + D(p(x_{n+1} | x_n) \parallel q(x_{n+1} | x_n))
= 0
= D(p(x_{n+1}) \parallel q(x_{n+1})) + D(p(x_n | x_{n+1}) | q(x_n | x_{n+1}))
\]  

(3.59)  
(3.60)  
(3.61)
Assume two stochastic processes with the same dynamics, but two distinct initial conditions.

Then we have:

\[
D(p(x_n, x_{n+1}) \| q(x_n, x_{n+1})) = D(p(x_n) \| q(x_n)) + D(p(x_{n+1}|x_n) \| q(x_{n+1}|x_n)) = 0
\]

\[
= D(p(x_{n+1}) \| q(x_{n+1})) + D(p(x_n|x_{n+1}) \| q(x_n|x_{n+1})) \geq 0
\]
Assume two stochastic processes with the same dynamics, but two distinct initial conditions.

Then we have:

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

\begin{align*}
  &= D(p(x_n) || q(x_n)) + D(p(x_{n+1}|x_n) || q(x_{n+1}|x_n)) \\
  &= 0 \\
  &= D(p(x_{n+1}) || q(x_{n+1})) + D(p(x_n|x_{n+1}) | q(x_n|x_{n+1})) \\
  &\geq 0
\end{align*}

This means that

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

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

(3.63)
This means that

\[ D(p(x_n) \| q(x_n)) \geq D(p(x_{n+1}) \| q(x_{n+1})) \]  \hspace{1cm} (3.63)

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 \( n \).
This means that

\[ D(p(x_n) \| q(x_n)) \geq D(p(x_{n+1}) \| q(x_{n+1})) \]  \hspace{1cm} (3.63)

Consider \( q(x) \) to be a stationary distribution of the Markov process, i.e., 
\[ q(x) = \sum_y p(x \mid y) q(y) \] 
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 $u(x)$.
Consider the stationary distribution \( q(x) \) is the uniform distribution \( u(x) \).

Then

\[
D(p(x_n) || u) = -H(X_n) + \log n \geq -H(X_{n+1}) + \log n \quad (3.64)
\]

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

Then

$$D(p(x_n) \| u) = -H(X_n) + \log n \geq -H(X_{n+1}) + \log n \quad (3.64)$$

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

Then

$$D(p(x_n) || u) = -H(X_n) + \log n \geq -H(X_{n+1}) + \log n \quad (3.64)$$

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

Then

\[
D(p(x_n) \| u) = -H(X_n) + \log n \geq -H(X_{n+1}) + \log n \tag{3.64}
\]

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, \( p(x_{n+1} | x_n) \).
Let $X_1, X_2, \ldots, X_N$, $X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X = H) = \theta = 1 - P(X = T)$. 
Let $X_1, X_2, \ldots, X_N, X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X = H) = \theta = 1 - P(X = T)$.

Let $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$ count the number of heads.
Let $X_1, X_2, \ldots, X_N$, $X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X = H) = \theta = 1 - P(X = T)$.

Let $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$ count the number of heads.

$T$ is said to be a statistic of the sample.
Let $X_1, X_2, \ldots, X_N, X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X = H) = \theta = 1 - P(X = T)$.

Let $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$ count the number of heads.

$T$ is said to be a statistic of the sample.

In general, a statistic is some function of a collection of random variables (e.g., an empirical mean, an empirical variance, or an empirical max of a sample, etc.).
Statistics $T$

- Let $X_1, X_2, \ldots, X_N, X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X = H) = \theta = 1 - P(X = T)$.
- Let $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$ count the number of heads.
- $T$ is said to be a statistic of the sample.
- In general, a statistic is some function of a collection of random variables (e.g., an empirical mean, an empirical variance, or an empirical max of a sample, etc.).
- A statistic is itself a r.v. with a mean, variance, etc.
Let $X_1, X_2, \ldots, X_N, X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X = H) = \theta = 1 - P(X = T)$.

Let $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$ count the number of heads.

$T$ is said to be a statistic of the sample.

In general, a statistic is some function of a collection of random variables (e.g., an empirical mean, an empirical variance, or an empirical max of a sample, etc.).

A statistic is itself a r.v. with a mean, variance, etc.

Good statistics contain useful information about the sample, while bad statistics don't (e.g., $T(X_1, \ldots, X_N) = X_1$).
Let $X_1, X_2, \ldots, X_N, X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X = H) = \theta = 1 - P(X = T)$.

Let $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$ count the number of heads.

$T$ is said to be a statistic of the sample.

In general, a statistic is some function of a collection of random variables (e.g., an empirical mean, an empirical variance, or an empirical max of a sample, etc.).

A statistic is itself a r.v. with a mean, variance, etc.

Good statistics contain useful information about the sample, while bad statistics don’t (e.g., $T(X_1, \ldots, X_N) = X_1$).

Statistics are sometimes called “features” in pattern recognition and machine learning.
Bernoulli Trials

- Now consider the above counting statistic in a probability.

\[
p(x_1, \ldots, x_N | T(x_1, \ldots, x_N), \theta) = p(x_1, \ldots, x_N | T(x_1, \ldots, x_N))
\]

\[
= \begin{cases} 
\frac{1}{\binom{N}{k}} & \sum_i x_i = k \\
0 & \text{else}
\end{cases}
\]

(3.65)
Bernoulli Trials

- Now consider the above counting statistic in a probability.

\[ p(x_1, \ldots, x_N | T(x_1, \ldots, x_N), \theta) = p(x_1, \ldots, x_N | T(x_1, \ldots, x_N)) \]

\[ = \begin{cases} 
\frac{1}{\binom{N}{k}} & \sum_i x_i = k \\
0 & \text{else}
\end{cases} \tag{3.65} \]

- Once we know the statistic, probability of the sequence is expressible without referring back to \( \theta \).
Now consider the above counting statistic in a probability.

\[ p(x_1, \ldots, x_N | T(x_1, \ldots, x_N), \theta) = p(x_1, \ldots, x_N | T(x_1, \ldots, x_N)) \]

\[
\begin{cases} 
\frac{1}{\binom{N}{k}} \sum_i x_i = k \\
0 \quad \text{else}
\end{cases}
\]

Once we know the statistic, probability of the sequence is expressible without referring back to \( \theta \).

In other words, we have that \( X_{1:N} \perp \theta | T(X_{1:N}) \).
Now consider the above counting statistic in a probability.

\[
p(x_1, \ldots, x_N | T(x_1, \ldots, x_N), \theta) = p(x_1, \ldots, x_N | T(x_1, \ldots, x_N))
\]

\[
= \begin{cases} 
\frac{1}{\binom{N}{k}} \sum_i x_i = k \\
0 \quad \text{else}
\end{cases}
\]  

\[ (3.65) \]

Once we know the statistic, probability of the sequence is expressible without referring back to \( \theta \).

In other words, we have that \( X_{1:N} \perp \perp \theta | T(X_{1:N}) \).

Markov chain (A): \( \theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N} \)
Now consider the above counting statistic in a probability.

\[
p(x_1, \ldots, x_N | T(x_1, \ldots, x_N), \theta) = p(x_1, \ldots, x_N | T(x_1, \ldots, x_N))
= \begin{cases} 
\frac{1}{\binom{N}{k}} \sum_i x_i = k \\
0 \quad \text{else}
\end{cases}
\]

(3.65)

Once we know the statistic, probability of the sequence is expressible without referring back to \( \theta \).

In other words, we have that \( X_{1:N} \perp \perp \theta | T(X_{1:N}) \).

Markov chain (A): \( \theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N} \)

We also know that \( T(X_{1:N}) \) is a function of \( X_{1:N} \).
Now consider the above counting statistic in a probability.

\[
p(x_1, \ldots, x_N | T(x_1, \ldots, x_N), \theta) = p(x_1, \ldots, x_N | T(x_1, \ldots, x_N))
\]

\[
= \begin{cases} 
\frac{1}{\binom{N}{k}} \sum_i x_i = k \\
0 & \text{else}
\end{cases}
\]

Once we know the statistic, probability of the sequence is expressible without referring back to \(\theta\).

In other words, we have that \(X_{1:N} \perp \perp \theta | T(X_{1:N})\).

Markov chain (A): \(\theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N}\)

We also know that \(T(X_{1:N})\) is a function of \(X_{1:N}\).

(B) \(\theta \rightarrow X_{1:N} \rightarrow T(X_{1:N})\).
DPI and Statistics

- (A): \( \theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N} \)
DPI and Statistics

- (A): $\theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N}$
- BY DPI, (A) $\Rightarrow I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$
DPI and Statistics

(A): $\theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N}$

BY DPI, $(A) \Rightarrow I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$

(B): $\theta \rightarrow X_{1:N} \rightarrow T(X_{1:N})$.
DPI and Statistics

(A): $\theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N}$

BY DPI, (A) $\Rightarrow I(\theta ; T(X_{1:N})) \geq I(\theta ; X_{1:N})$

(B): $\theta \rightarrow X_{1:N} \rightarrow T(X_{1:N})$

BY DPI, (B) $\Rightarrow I(\theta ; X_{1:N}) \geq I(\theta ; T(X_{1:N}))$

Thus, (A) & (B) $\Rightarrow I(\theta ; X_{1:N}) = I(\theta ; T(X_{1:N}))$, and no information is lost about $\theta$ in going from $X_{1:N}$ to $T(X_{1:N})$. Such a statistic is called sufficient.

**Definition 3.8.1 (Sufficient Statistic)**
A function $T(\cdot)$ is said to be sufficient for parameter $\theta$ governing the distribution of $X$ if $X \perp\perp \theta | T(X)$. (3.66)

Alternatively, if the data processing inequality achieves equality.
DPI and Statistics

- (A): $\theta \to T(X_{1:N}) \to X_{1:N}$
  - BY DPI, (A) $\implies I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$
- (B): $\theta \to X_{1:N} \to T(X_{1:N})$.
  - BY DPI, (B) $\implies I(\theta; X_{1:N}) \geq I(\theta; T(X_{1:N}))$
- Thus, (A) & (B) $\implies I(\theta; X_{1:N}) = I(\theta; T(X_{1:N}))$, and no information is lost about $\theta$ in going from $X_{1:N}$ to $T(X_{1:N})$. 

[Definition 3.8.1 (Sufficient Statistic)]

A function $T(\cdot)$ is said to be sufficient for parameter $\theta$ governing the distribution of $X$ if $X \perp \perp \theta | T(X)$. (3.66)

Alternatively, if the data processing inequality achieves equality.
DPI and Statistics

(A): $\theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N}$

BY DPI, (A) $\Rightarrow I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$

(B): $\theta \rightarrow X_{1:N} \rightarrow T(X_{1:N})$.

BY DPI, (B) $\Rightarrow I(\theta; X_{1:N}) \geq I(\theta; T(X_{1:N}))$

Thus, (A) & (B) $\Rightarrow I(\theta; X_{1:N}) = I(\theta; T(X_{1:N}))$, and no information is lost about $\theta$ in going from $X_{1:N}$ to $T(X_{1:N})$.

Such a statistic is called sufficient.
DPI and Statistics

- (A): $\theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N}$
  - BY DPI, (A) $\Rightarrow I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$
- (B): $\theta \rightarrow X_{1:N} \rightarrow T(X_{1:N})$
  - BY DPI, (B) $\Rightarrow I(\theta; X_{1:N}) \geq I(\theta; T(X_{1:N}))$
  - Thus, (A) & (B) $\Rightarrow I(\theta; X_{1:N}) = I(\theta; T(X_{1:N}))$, and no information is lost about $\theta$ in going from $X_{1:N}$ to $T(X_{1:N})$.
  - Such a statistic is called sufficient.

**Definition 3.8.1 (Sufficient Statistic)**

A function $T(\cdot)$ is said to be **sufficient** for parameter $\theta$ governing the distribution of $X$ if

$$X \perp \theta|T(X).$$

(3.66)

Alternatively, if the data processing inequality achieves equality.
DPI and Statistics

- (A): $\theta \rightarrow T(X_{1:N}) \rightarrow X_{1:N}$
  - BY DPI, (A) $\Rightarrow I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$
- (B): $\theta \rightarrow X_{1:N} \rightarrow T(X_{1:N})$
  - BY DPI, (B) $\Rightarrow I(\theta; X_{1:N}) \geq I(\theta; T(X_{1:N}))$
  - Thus, (A) & (B) $\Rightarrow I(\theta; X_{1:N}) = I(\theta; T(X_{1:N}))$, and no information is lost about $\theta$ in going from $X_{1:N}$ to $T(X_{1:N})$.
  - Such a statistic is called sufficient.

**Definition 3.8.1 (Sufficient Statistic)**

A function $T(\cdot)$ is said to be **sufficient** for parameter $\theta$ governing the distribution of $X$ if

$$X \perp\!\!\!\perp \theta|T(X). \quad (3.66)$$

Alternatively, if the data processing inequality achieves equality.

- Sufficient statistics use to estimate parameters from data: in the limit of infinite data, one estimates exactly (asymptotic consistency).
Sufficient Statistic

- Traditional definition

\[ T(\cdot) \text{ is sufficient for } \theta \iff \text{the probability} \]
\[ p(x_1:N|\theta) \text{ can be written as the product:} \]
\[ p(x_1:N|\theta) = g(T, \Theta) h(x_1:N) \quad (3.67) \]

Compare with a definition of conditional independence.

\[ \text{Definition 3.8.3 (Conditional Independence)} \]
\[ \text{Given three random variables } A, B, C, \text{ we have that} \]
\[ A \perp \perp B | C \iff \text{there exist functions} \]
\[ g \text{ and } h \text{ such that} \]
\[ p(a, b, c) \text{ can be written:} \]
\[ p(a, b, c) = g(a, c) h(b, c) \quad (3.68) \]

Compare: 
Set \( C \leftarrow T \), \( B \leftarrow X_1:N \), and \( A \leftarrow \theta \).

Then 
\[ h(b, c) = h'(b) \text{ for } h'(b) = h(T(b), b). \]
Sufficient Statistic

- Traditional definition

**Definition 3.8.2 (Sufficient Statistic)**

$T(\cdot)$ is sufficient for $\theta$ iff the probability $p(x_{1:N}|\theta)$ can be written as the product:

$$p(x_{1:N}|\theta) = g(T, \Theta)h(x_{1:N})$$ (3.67)
Sufficient Statistic

- Traditional definition

**Definition 3.8.2 (Sufficient Statistic)**

\( T(\cdot) \) is sufficient for \( \theta \) iff the probability \( p(x_{1:N}|\theta) \) can be written as the product:

\[
p(x_{1:N}|\theta) = g(T, \Theta) h(x_{1:N})
\]  

(3.67)

- Compare with a definition of conditional independence.
Sufficient Statistic

- Traditional definition

**Definition 3.8.2 (Sufficient Statistic)**

$T(\cdot)$ is sufficient for $\theta$ iff the probability $p(x_1:N|\theta)$ can be written as the product:

$$p(x_1:N|\theta) = g(T, \Theta)h(x_1:N)$$  \hspace{1cm} (3.67)

- Compare with a definition of conditional independence.

**Definition 3.8.3 (Conditional Independence)**

Given three random variables $A, B, C$, we have that $A \perp B|C$ iff there exists functions $g$ and $h$ such that $p(a, b, c)$ can be written:

$$p(a, b, c) = g(a, c)h(b, c)$$  \hspace{1cm} (3.68)
Sufficient Statistic

- Traditional definition

**Definition 3.8.2 (Sufficient Statistic)**

$T(\cdot)$ is sufficient for $\theta$ iff the probability $p(x_{1:N}|\theta)$ can be written as the product:

$$p(x_{1:N}|\theta) = g(T, \Theta)h(x_{1:N})$$

(3.67)

- Compare with a definition of conditional independence.

**Definition 3.8.3 (Conditional Independence)**

Given three random variables $A, B, C$, we have that $A \perp B|C$ iff there exists functions $g$ and $h$ such that $p(a, b, c)$ can be written:

$$p(a, b, c) = g(a, c)h(b, c)$$

(3.68)

- Compare: Set $C \leftarrow T$, $B \leftarrow X_{1:N}$, and $A \leftarrow \theta$. 
Sufficient Statistic

- Traditional definition

**Definition 3.8.2 (Sufficient Statistic)**

$T(\cdot)$ is sufficient for $\theta$ iff the probability $p(x_{1:N}|\theta)$ can be written as the product:

$$p(x_{1:N}|\theta) = g(T, \Theta)h(x_{1:N})$$  \hfill (3.67)

- Compare with a definition of conditional independence.

**Definition 3.8.3 (Conditional Independence)**

Given three random variables $A, B, C$, we have that $A \perp\!\perp B|C$ iff there exists functions $g$ and $h$ such that $p(a, b, c)$ can be written:

$$p(a, b, c) = g(a, c)h(b, c)$$  \hfill (3.68)

- Compare: Set $C \leftarrow T$, $B \leftarrow X_{1:N}$, and $A \leftarrow \theta$. Then $h(b, c) = h'(b)$ for $h'(b) = h(T(b), b)$. 

Let $X_1, X_2, \ldots, X_N \equiv X_{1:N}$ be a length-$N$ sample of a D-ary discrete random variable. So $x_i \in D_X$ and alphabet size $D = |D_X|$, and $D_X = (a_1, a_2, \ldots, a_D)$. 

Let $X_1, X_2, \ldots, X_N \equiv X_{1:N}$ be a length-$N$ sample of a D-ary discrete random variable. So $x_i \in D_X$ and alphabet size $D = |D_X|$, and $D_X = (a_1, a_2, \ldots, a_D)$.
Sufficiency of the “type” of the sample

Let $X_1, X_2, \ldots, X_N \equiv X_{1:N}$ be a length-$N$ sample of a $D$-ary discrete random variable. So $x_i \in D_X$ and alphabet size $D = |D_X|$, and $D_X = (a_1, a_2, \ldots, a_D)$.

Define a statistic which is the empirical histogram of this sample.

$$P_{x_{1:N}} \triangleq \left( \frac{N(a_1|x_{1:N})}{N}, \frac{N(a_2|x_{1:N})}{N}, \ldots, \frac{N(a_D|x_{1:N})}{N} \right)$$ (3.69)

where $N(a_i|x_{1:N})$ counts occurrence of symbol $a_i$ in sample $x_{1:N}$. 
Sufficiency of the “type” of the sample

Let $X_1, X_2, \ldots, X_N \equiv X_{1:N}$ be a length-$N$ sample of a $D$-ary discrete random variable. So $x_i \in D_X$ and alphabet size $D = |D_X|$, and $D_X = (a_1, a_2, \ldots, a_D)$.

Define a statistic which is the empirical histogram of this sample.

$$P_{x_{1:N}} \triangleq \left( \frac{N(a_1|x_{1:N})}{N}, \frac{N(a_2|x_{1:N})}{N}, \ldots, \frac{N(a_D|x_{1:N})}{N} \right)$$

(3.69)

where $N(a_i|x_{1:N})$ counts occurrence of symbol $a_i$ in sample $x_{1:N}$.

This is a histogram, or type, of the sample.
Sufficiency of the “type” of the sample

Let $X_1, X_2, \ldots, X_N \equiv X_{1:N}$ be a length-$N$ sample of a $D$-ary discrete random variable. So $x_i \in D_X$ and alphabet size $D = |D_X|$, and $D_X = (a_1, a_2, \ldots, a_D)$.

Define a statistic which is the empirical histogram of this sample.

$$P_{x_{1:N}} \triangleq \left( \frac{N(a_1|x_{1:N})}{N}, \frac{N(a_2|x_{1:N})}{N}, \ldots, \frac{N(a_D|x_{1:N})}{N} \right)$$  \hspace{1cm} (3.69)

where $N(a_i|x_{1:N})$ counts occurrence of symbol $a_i$ in sample $x_{1:N}$.

This is a histogram, or type, of the sample.

It is also a statistic since it a function of the sample.
Sufficiency of the “type” of the sample

- Let $X_1, X_2, \ldots, X_N \equiv X_{1:N}$ be a length-$N$ sample of a $D$-ary discrete random variable. So $x_i \in D_X$ and alphabet size $D = |D_X|$, and $D_X = (a_1, a_2, \ldots, a_D)$.

- Define a statistic which is the empirical histogram of this sample.

$$P_{x_{1:N}} \triangleq \left( \frac{N(a_1|x_{1:N})}{N}, \frac{N(a_2|x_{1:N})}{N}, \ldots, \frac{N(a_D|x_{1:N})}{N} \right)$$  \hspace{1cm} (3.69)

where $N(a_i|x_{1:N})$ counts occurrence of symbol $a_i$ in sample $x_{1:N}$.

- This is a histogram, or type, of the sample.

- It is also a statistic since it a function of the sample.

- Is it sufficient?
Sufficiency of the “type” of the sample

- Let $N_i = N(a_i|x_1:N)$ as shorthand.
Let $N_i = N(a_i|x_{1:N})$ as shorthand.

Is the type sufficient?

\[
p(x_{1:n}|P_{x_{1:N}}, \theta) = \begin{cases} 
\frac{1}{N} & \text{if } \forall i, N(a_i|x_{1:N}) = NP_{x_{1:N}}(a_i) \\
0 & \text{else}
\end{cases}
\]

\[= p(x_{1:n}|P_{x_{1:N}}) \quad \text{(3.70)}
\]
Sufficiency of the “type” of the sample

- Let $N_i = N(a_i|x_{1:N})$ as shorthand.
- Is the type sufficient?

$$p(x_{1:n}|P_{x_{1:N}}, \theta) = \begin{cases} \frac{1}{\binom{N}{N_1 N_2 \ldots N_D}} & \text{if } \forall i, N(a_i|x_{1:N}) = NP_{x_{1:N}}(a_i) \\ 0 & \text{else} \end{cases}$$

[Equation 3.70]

$$= p(x_{1:n}|P_{x_{1:N}})$$

[Equation 3.71]

- So, $X_{1:N} \perp \theta|P_{x_{1:N}}$ and the type $P_{x_{1:N}}$ is sufficient.
Binary case, sufficiency of the “type”

- $X_i \in \{0, 1\}$, $T(x_1:N) = \text{number of ones in } x_{1:N}$. 
Binary case, sufficiency of the “type”

- $X_i \in \{0, 1\}$, $T(x_{1:N}) =$ number of ones in $x_{1:N}$.
- The joint probability

$$p(x_{1:N}, T(x_{1:N}), \theta) = \prod_{a \in D_X} p(a)^{N(a|x_{1:N})}$$  \hspace{1cm} (3.72)
Binary case, sufficiency of the “type”

- \( X_i \in \{0, 1\}, \ T(x_1:N) = \text{number of ones in } x_{1:N} \).
- The joint probability

\[
p(x_1:N, T(x_1:N), \theta) = \prod_{a \in D_X} p(a)^{N(a|x_1:N)} = p(0)^{N(0|x_1:N)}p(1)^{N(1|x_1:N)} \tag{3.72}
\]
Binary case, sufficiency of the “type”

- $X_i \in \{0, 1\}$, $T(x_{1:N}) =$ number of ones in $x_{1:N}$.
- The joint probability

$$p(x_{1:N}, T(x_{1:N}), \theta) = \prod_{a \in \mathcal{D}_X} p(a)^{N(a|x_{1:N})} = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \quad (3.72)$$

- Event $\{x_{1:N}, T(x_{1:n}) = k\}$ when $k$ is true number of ones in $x_{1:N}$ is same as event $\{x_{1:n}\}$. Event $\{x_{1:N}, T(x_{1:n}) = k\}$ when $k$ is not number of ones in $x_{1:N}$ is impossible (zero probability).
Binary case, sufficiency of the “type”

- \( X_i \in \{0, 1\} \), \( T(x_{1:N}) = \) number of ones in \( x_{1:N} \).

- The joint probability

\[
p(x_{1:N}, T(x_{1:N}), \theta) = \prod_{a \in D_X} p(a)^{N(a|x_{1:N})} = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \tag{3.72}
\]

- Event \( \{x_{1:N}, T(x_{1:n}) = k\} \) when \( k \) is true number of ones in \( x_{1:N} \) is same as event \( \{x_{1:n}\} \). Event \( \{x_{1:N}, T(x_{1:n}) = k\} \) when \( k \) is not number of ones in \( x_{1:N} \) is impossible (zero probability).

- The marginal \( p(\theta, T(x_{1:N}) = k) \) has expression:

\[
p(\theta, T(x_{1:N}) = k) = \sum_{x_{1:N}} p(x_{1:N}, T(x_{1:N}) = k, \theta) \tag{3.73}
\]
Binary case, sufficiency of the "type"

- $X_i \in \{0, 1\}$, $T(x_{1:N}) =$ number of ones in $x_{1:N}$.
- The joint probability
  \[
  p(x_{1:N}, T(x_{1:N}), \theta) = \prod_{a \in D_X} p(a)^{N(a|x_{1:N})} = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \tag{3.72}
  \]
- Event $\{x_{1:N}, T(x_{1:n}) = k\}$ when $k$ is true number of ones in $x_{1:N}$ is same as event $\{x_{1:n}\}$. Event $\{x_{1:N}, T(x_{1:n}) = k\}$ when $k$ is not number of ones in $x_{1:N}$ is impossible (zero probability).
- The marginal $p(\theta, T(x_{1:N}) = k)$ has expression:
  \[
  p(\theta, T(x_{1:N}) = k) = \sum_{x_{1:N}} p(x_{1:N}, T(x_{1:N}) = k, \theta) \tag{3.73}
  \]
  \[
  = \sum_{x_{1:N}: T(x_{1:N}) = k} p(x_{1:N}, T(x_{1:N}) = k, \theta) \tag{3.74}
  \]
  \[
  = \sum_{x_{1:N}: T(x_{1:N}) = k} p(x_{1:N}, T(x_{1:N}) = k, \theta) \tag{3.75}
  \]
Binary case, sufficiency of the “type”

- $X_i \in \{0, 1\}$, $T(x_{1:N}) =$ number of ones in $x_{1:N}$.
- The joint probability
  \[
p(x_{1:N}, T(x_{1:N}), \theta) = \prod_{a \in D_X} p(a)^{N(a|x_{1:N})} = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \tag{3.72}
  \]
- Event \( \{x_{1:N}, T(x_{1:n}) = k\} \) when \( k \) is true number of ones in \( x_{1:N} \) is same as event \( \{x_{1:n}\} \). Event \( \{x_{1:N}, T(x_{1:n}) = k\} \) when \( k \) is not number of ones in \( x_{1:N} \) is impossible (zero probability).
- The marginal $p(\theta, T(x_{1:N}) = k)$ has expression:
  \[
p(\theta, T(x_{1:N}) = k) = \sum_{x_{1:N}} p(x_{1:N}, T(x_{1:N}) = k, \theta) \tag{3.73}
  \]
  \[
  = \sum_{x_{1:N}:T(x_{1:N})=k} p(x_{1:N}, T(x_{1:N}) = k, \theta) \tag{3.74}
  \]
  \[
  = \binom{N}{k} p(0)^{N-k} p(1)^k \tag{3.75}
  \]
Binary case, sufficiency of the “type”

- The joint probability

\[
p(x_{1:N}, T(x_{1:N}), \theta) = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})}
\]  

(3.76)
Binary case, sufficiency of the “type”

- The joint probability

\[ p(x_{1:N}, T(x_{1:N}), \theta) = p(0)^N(0|x_{1:N})p(1)^N(1|x_{1:N}) \] (3.76)

- The marginal

\[ p(\theta, T(x_{1:N}) = k) = \binom{N}{k}p(0)^{N-k}p(1)^k \] (3.77)
The joint probability

$$p(x_{1:N}, T(x_{1:N}), \theta) = p(0)^N(0|x_{1:N})p(1)^N(1|x_{1:N})$$  \hspace{1cm} (3.76)

The marginal

$$p(\theta, T(x_{1:N}) = k) = \binom{N}{k}p(0)^{N-k}p(1)^k$$  \hspace{1cm} (3.77)

So

$$p(x_{1:N}|T, \Theta) = \frac{p(x_{1:N}, T, \Theta)}{p(T, \Theta)} = \begin{cases} \frac{1}{\binom{N}{k}} & \text{if } \sum_i x_i = k \\ 0 & \text{else} \end{cases}$$  \hspace{1cm} (3.78)

which is the binary r.v. case of Equation 3.70.
Definition 3.8.4

A statistic $T(X)$ is a minimal sufficient statistic relative to $\{p_\theta(x)\}$ if it is a function of every other sufficient statistic $U$. Interpreting this in terms of the data-processing inequality, this implies that

$$\theta \rightarrow T(X) \rightarrow U(X) \rightarrow X$$

(3.79)

- I.e., we know, from the definition of $T$ minimal, and any other sufficient statistic $U$ that $\theta \rightarrow X_{1:N} \rightarrow U(X_{1:N}) \rightarrow T(X_{1:N})$. 
Definition 3.8.4

A statistic $T(X)$ is a minimal sufficient statistic relative to $\{p_\theta(x)\}$ if it is a function of every other sufficient statistic $U$. Interpreting this in terms of the data-processing inequality, this implies that

$$\theta \rightarrow T(X) \rightarrow U(X) \rightarrow X \quad (3.79)$$

- I.e., we know, from the definition of $T$ minimal, and any other sufficient statistic $U$ that $\theta \rightarrow X_{1:N} \rightarrow U(X_{1:N}) \rightarrow T(X_{1:N})$.
- The fact that it is a statistic, however, means that $p(X|T, U, \theta) = p(X|T, U) = p(X|U)$ meaning $T$ is, for all intents and purposes, the minimal statistic replacement for $\theta$ in computing the probability.