Class Road Map - IT-I

L1 (9/25): Overview, Communications, Information, Entropy
L2 (9/30): Entropy, Mutual Information, KL-Divergence, Jensen, Properties
L3 (10/2):
L4 (10/7):
L5 (10/9):
L6 (10/14):
L7 (10/16):
L8 (10/21):
L9 (10/23):

L10 (10/28):
L11 (10/30):
L12 (11/4):
LXX (11/6): In class midterm exam
LXX (11/11): Veterans Day holiday
L13 (11/13):
L14 (11/18):
L15 (11/20):
L16 (11/25):
L17 (11/27):
L18 (12/2):
L19 (12/4):
LXX (12/10): Final exam

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

- Homework 1 will be posted on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), later this evening, due early next week.
Last time we discussed a model of communication and what “information” might mean.
Last time we discussed a model of communication and what “information” might mean.

Intuitively, we gain more information when something unexpected happens, and gain less information when something expected happens.
Last time we discussed a model of communication and what “information” might mean.

Intuitively, we gain more information when something unexpected happens, and gain less information when something expected happens.

Entropy will characterize this mathematically and is useful for characterizing “information” even if it doesn’t capture all the information in a given source.
What is entropy?

- Events $E_k$ each occur with probability $p_k$. $p_k$ indicates the likelihood of the event $E_k$ happening.
- Shannon/Hartley information of event $E_k$ is $I(E_k) = \log(1/p_k)$, indicating:
  1. a measure of surprise of finding out $E_k$. If $p_k = 1$ ⇒ no surprise in finding out that $E_k$ occurred, while $p_k = 0$ ⇒ infinite surprise in finding out $E_k$.
  2. A measure if information gained in finding out $E_k$ (information gained is equal to surprise). $p_k = 1$ ⇒ No information is gained, while $p_k = 0$ ⇒ infinite information is gained.
  3. A measure of the “uncertainty” of $E_k$ (but really unexpectedness). Unexpectedness is the thing that determines interest, or information. Also, information required to resolve this particular unexpectedness.
  4. $I(E_k) = -\log p(E_k) = \text{the self information of that event, or that message. Why is it called self-information? We’ll soon see.}$
- All logs are base 2 (by default), so $\log \equiv \log_2$ unless otherwise stated. $\ln$ will be natural log.
**Entropy**

**Definition 2.2.1 (Entropy)**

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

$$H(X) \triangleq E \log \frac{1}{p(X)} = \sum_x p(x) \log \frac{1}{p(x)} = -\sum_x p(x) \log p(x) \quad (2.1)$$

- Entropy is in units of “bits” since logs are base 2 (units of “nats” if base $e$ logs).
- Measures the degree of uncertainty in a distribution.
- Measures the disorder or spread of a distribution.
- Measures the “choice” that a source has in choosing symbols according to the density (higher entropy means more choice).
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 the function H(p) with p on the x-axis and H(p) on the y-axis. The graph is a parabola opening downwards with the vertex at (0.5, 1).]

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

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

\[
H(X, Y) = - \sum_x \sum_y p(x, y) \log p(x, y) = E \log \frac{1}{p(X, Y)} \tag{2.1}
\]

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

\[
H(X_1, \ldots, X_N) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, \ldots, x_N) \log \frac{1}{p(x_1, \ldots, x_N)} \tag{2.2}
\]

\[
= E \log \frac{1}{p(x_1, \ldots, x_N)} \tag{2.3}
\]
The non-negativity of discrete Entropy

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

- Discrete since $X$ is a discrete random variable (i.e., $x \in \mathcal{X}$ where $\mathcal{X}$ is countable).

- Note $\lim_{\alpha \to 0} \alpha \log \alpha = 0$, hence if $p(x) = 0$, the entropy is uninfluenced.

- Also since $p(x) \geq 0$ and $\log 1/p(x) \geq 0$ discrete entropy is always non-negative $H(X) \geq 0$. 

- Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 2 - Sep 30th, 2019
L2 F11/68 (pg.13/244)
Conditional Entropy

- For two random variables $X, Y$ related via $p(x, y)$, knowing the event $X = x$ can change the entropy of $Y$. 

\[ H(Y | X = x) = -\sum_x p(x) \sum_y p(y | x) \log p(y | x) \] 

Averaging over all $x$, we get the conditional entropy $H(Y | X)$. 

\[ H(Y | X) = -\sum_x p(x) \sum_y p(y | x) \log p(y | x) = \mathbb{E} \log 1 / p(Y | X) \]
Conditional Entropy

- For two random variables $X, Y$ related via $p(x, y)$, knowing the event $X = x$ can change the entropy of $Y$.
- Event conditional entropy $H(Y|X = x)$

$$H(Y|X = x) = E \log \frac{1}{p(Y|X = x)}$$  \hspace{1cm} (2.1)

$$= - \sum_y p(y|x) \log p(y|x)$$  \hspace{1cm} (2.2)
Conditional Entropy

- For two random variables $X, Y$ related via $p(x, y)$, knowing the event $X = x$ can change the entropy of $Y$.
- Event conditional entropy $H(Y|X = x)$

$$H(Y|X = x) = E \log \frac{1}{p(Y|X = x)}$$

$$= - \sum_y p(y|x) \log p(y|x) \tag{2.2}$$

- Averaging over all $x$, we get the conditional entropy $H(Y|X)$.

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$

$$= - \sum_x p(x) \sum_y p(y|x) \log p(y|x) \tag{2.4}$$

$$= - \sum_{x,y} p(x, y) \log p(y|x) = E \log \frac{1}{p(Y|X)} \tag{2.5}$$
**Chain rule for Entropy**

### Proposition 2.3.1 (Chain Rule for Entropy)

\[
H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)
\]  
(2.6)

### Proof.

\[
- \log p(x, y) = - \log p(x) - \log p(y|x)
\]  
(2.7)

then take expected value of both sides to get result.

### Corollary 2.3.2

If \( X \perp \!\!\!\!\!\!\perp Y \) then \( H(X, Y) = H(X) + H(Y) \).
Proposition 2.3.3 (Chain Rule for Entropy)

\[ H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, X_2, \ldots, X_{i-1}) \]  \hspace{1cm} (2.8)
General Chain rule for Entropy

Proposition 2.3.3 (Chain Rule for Entropy)

\[ H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, X_2, \ldots, X_{i-1}) \quad (2.8) \]

Proof.

Use chain rule of conditional probability, i.e., that

\[ p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{i-1}) \quad (2.9) \]
General Chain rule for Entropy

Proposition 2.3.3 (Chain Rule for Entropy)

\[ H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, X_2, \ldots, X_{i-1}) \] (2.8)

Proof.

Use chain rule of conditional probability, i.e., that

\[ p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{i-1}) \] (2.9)

then

\[ -\log p(x_1, x_2, \ldots, x_N) = -\sum_{i=1}^{N} \log p(x_i | x_1, x_2, \ldots, x_{i-1}) \] (2.10)

then take expected value of both sides to get result.
Convex analysis gives variational representation

$$\ln x = \min_{\lambda} \{\lambda x - \ln \lambda - 1\}$$  \hfill (2.11)

so for any $\lambda$, we have

$$\ln x \leq \lambda x - \ln \lambda - 1$$  \hfill (2.12)

and with $\lambda = 1$, we thus get

$$\ln x \leq x - 1$$  \hfill (2.13)
Max value of (discrete) Entropy

**Proposition 2.3.4**

Let $X \in \{x_1, x_2, \ldots, x_n\}$. Then $H(X) \leq \log n$, and equality is achieved iff $p(X = x_i) = \frac{1}{n}$ for all $i$. 

\[ H(X) - \log n = -\sum x p(x) \log p(x) - \sum x p(x) \log n \]

\[ \leq \log e \sum x p(x) \left[ \frac{1}{p(x)} - 1 \right] \]

\[ = \log e \left[ \frac{1}{n} - \sum x p(x) \right] = 0 \]
Proposition 2.3.4

Let \( X \in \{x_1, x_2, \ldots, x_n\} \). Then \( H(X) \leq \log n \), and equality is achieved iff \( p(X = x_i) = \frac{1}{n} \) for all \( i \).

Proof.

- Approach: show that \( H(X) - \log n \leq 0 \).
Max value of (discrete) Entropy

Proposition 2.3.4

Let \( X \in \{x_1, x_2, \ldots, x_n\} \). Then \( H(X) \leq \log n \), and equality is achieved iff \( p(X = x_i) = \frac{1}{n} \) for all \( i \).

Proof.

Approach: show that \( H(X) - \log n \leq 0 \).

\[ H(X) - \log n \]
**Proposition 2.3.4**

Let \( X \in \{x_1, x_2, \ldots, x_n\} \). Then \( H(X) \leq \log n \), and equality is achieved iff \( p(X = x_i) = \frac{1}{n} \) for all \( i \).

**Proof.**

Approach: show that \( H(X) - \log n \leq 0 \).

\[
H(X) - \log n = - \sum_x p(x) \log p(x) - \sum_x p(x) \log n \quad (2.14)
\]
Max value of (discrete) Entropy

Proposition 2.3.4

Let $X \in \{x_1, x_2, \ldots, x_n\}$. Then $H(X) \leq \log n$, and equality is achieved iff $p(X = x_i) = \frac{1}{n}$ for all $i$.

Proof.

Approach: show that $H(X) - \log n \leq 0$.

\[
H(X) - \log n = -\sum_{x} p(x) \log p(x) - \sum_{x} p(x) \log n
\]

\[
= \log_2 e \sum_{x} p(x) \ln \frac{1}{p(x)n}
\]
Max value of (discrete) Entropy

**Proposition 2.3.4**

Let $X \in \{x_1, x_2, \ldots, x_n\}$. Then $H(X) \leq \log n$, and equality is achieved iff $p(X = x_i) = \frac{1}{n}$ for all $i$.

**Proof.**

- Approach: show that $H(X) - \log n \leq 0$.

\[
H(X) - \log n = - \sum_x p(x) \log p(x) - \sum_x p(x) \log n \tag{2.14}
\]

\[
= \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n} \tag{2.15}
\]

\[
\leq \log e \sum_x p(x) \left[ \frac{1}{p(x)n} - 1 \right] \tag{2.16}
\]
### Proposition 2.3.4

Let $X \in \{x_1, x_2, \ldots, x_n\}$. Then $H(X) \leq \log n$, and equality is achieved iff $p(X = x_i) = \frac{1}{n}$ for all $i$.

### Proof.

**Approach:** show that $H(X) - \log n \leq 0$.

\[
H(X) - \log n = - \sum_x p(x) \log p(x) - \sum_x p(x) \log n \quad (2.14)
\]

\[
= \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n} \quad (2.15)
\]

\[
\leq \log e \sum_x p(x) \left[ \frac{1}{p(x)n} - 1 \right] \quad (2.16)
\]

\[
= \log e \left[ \sum_x \frac{1}{n} - \sum_x p(x) \right]
\]
Max value of (discrete) Entropy

**Proposition 2.3.4**

Let \( X \in \{x_1, x_2, \ldots, x_n\} \). Then \( H(X) \leq \log n \), and equality is achieved iff \( p(X = x_i) = \frac{1}{n} \) for all \( i \).

**Proof.**

- **Approach:** show that \( H(X) - \log n \leq 0 \).

\[
H(X) - \log n = - \sum_x p(x) \log p(x) - \sum_x p(x) \log n \tag{2.14}
\]

\[
= \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n} \tag{2.15}
\]

\[
\leq \log e \sum_x p(x) \left[ \frac{1}{p(x)n} - 1 \right] \tag{2.16}
\]

\[
= \log e \left[ \sum_x \frac{1}{n} - \sum_x p(x) \right] = 0 \tag{2.17}
\]
Max value of (discrete) Entropy

- Since $\ln z = z - 1$ when $z = 1$, the above becomes an equality at stationary point, i.e., when $\frac{1}{p(x)n} = 1$, or $p(x) = 1/n$ the uniform distribution.
Max value of (discrete) Entropy

- Since $\ln z = z - 1$ when $z = 1$, the above becomes an equality at stationary point, i.e., when $\frac{1}{p(x)n} = 1$, or $p(x) = 1/n$ the uniform distribution.

- Another way to see this is if $p_i = 1/n$, then $-\sum_i p_i \log p_i = -\sum_i \frac{1}{n} \log \frac{1}{n} = -\log \frac{1}{n} = \log n$. 
Max value of (discrete) Entropy

- Since \( \ln z = z - 1 \) when \( z = 1 \), the above becomes an equality at stationary point, i.e., when \( \frac{1}{p(x)n} = 1 \), or \( p(x) = 1/n \) the uniform distribution.

- Another way to see this is if \( p_i = 1/n \), then
  \[
  - \sum_i p_i \log p_i = - \sum_i \frac{1}{n} \log \frac{1}{n} = - \log \frac{1}{n} = \log n.
  \]

- Implications: entropy increases when the distribution becomes more uniform.
Since \( \ln z = z - 1 \) when \( z = 1 \), the above becomes an equality at stationary point, i.e., when \( \frac{1}{p(x)n} = 1 \), or \( p(x) = 1/n \) the uniform distribution.

Another way to see this is if \( p_i = 1/n \), then
\[
- \sum_i p_i \log p_i = - \sum_i \frac{1}{n} \log \frac{1}{n} = - \log \frac{1}{n} = \log n.
\]

Implications: entropy increases when the distribution becomes more uniform.

E.g., mixing. \( \lambda p_1 + (1 - \lambda)p_2 \), we have
\[
H(\lambda p_1 + (1 - \lambda)p_2) \geq \lambda H(p_1) + (1 - \lambda)H(p_2),
\]
entropy is concave.
What if we permute the probabilities themselves?
What if we permute the probabilities themselves?

I.e., let distribution \( p = (p_1, p_2, \ldots, p_n) \) be a discrete probability distribution and \( \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \) be a random permutation.
What if we permute the probabilities themselves?

I.e., let distribution \( p = (p_1, p_2, \ldots, p_n) \) be a discrete probability distribution and \( \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \) be a random permutation.

Let \( p_\sigma = (p_{\sigma_1}, p_{\sigma_2}, \ldots, p_{\sigma_n}) \) be a permutation of the distribution.
What if we permute the probabilities themselves?

I.e., let distribution \( p = (p_1, p_2, \ldots, p_n) \) be a discrete probability distribution and \( \sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \) be a random permutation.

Let \( p_\sigma = (p_{\sigma_1}, p_{\sigma_2}, \ldots, p_{\sigma_n}) \) be a permutation of the distribution.

How does \( H(p) = - \sum_i p_i \log p_i \) compare with \( H(p_\sigma) \)?
What if we permute the probabilities themselves?

I.e., let distribution $p = (p_1, p_2, \ldots, p_n)$ be a discrete probability distribution and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ be a random permutation.

Let $p_\sigma = (p_{\sigma_1}, p_{\sigma_2}, \ldots, p_{\sigma_n})$ be a permutation of the distribution.

How does $H(p) = -\sum_i p_i \log p_i$ compare with $H(p_\sigma)$?

Same, since $H(p) = H(p_\sigma) = -\sum_i p_{\sigma_i} \log p_{\sigma_i}$. 
Summary so far

\[
H(X) = EI(X) = -\sum_x p(x) \log p(x) \tag{2.18}
\]

\[
H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y) \tag{2.19}
\]

\[
H(Y|X) = -\sum_{x,y} p(x, y) \log p(y|x) \tag{2.20}
\]

\[
H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{2.21}
\]

and

\[
0 \leq H(X) \leq \log n, \quad \text{where } n \text{ is } X\text{’s alphabet size.} \tag{2.22}
\]
Why log?

- We defined the information in event \( \{X = x\} \) as
  \[
  I(\{X = x\}) = I(x) = \log \frac{1}{p(x)},
  \]
  but why log?
Why log?

- We defined the information in event \( \{ X = x \} \) as 
  \[ I(\{ X = x \}) = I(x) = \log \frac{1}{p(x)}, \]
  but why log?

- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.
Why log?

- We defined the information in event \( \{X = x\} \) as 
  \[ I(\{X = x\}) = I(x) = \log \frac{1}{p(x)}, \]
  but why log?

- Intuition says we want the information about an event to be
  inversely related to the probability, but there are many such
  relationships that might be useful.

- Other possible functions include 
  \[ I(x) = p(x)^{-1/n} \]
  for some \( n > 0 \).
Why log?

- We defined the information in event \( \{X = x\} \) as 
  \[ I(\{X = x\}) = I(x) = \log \frac{1}{p(x)}, \] 
  but why log?

- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.

- Other possible functions include 
  \[ I(x) = p(x)^{-1/n} \] 
  for some \( n > 0 \).

- Or any monotone non-decreasing non-negative concave function?
We defined the information in event \( \{ X = x \} \) as
\[ I(\{ X = x \}) = I(x) = \log \frac{1}{p(x)}, \]
but why log?

Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.

Other possible functions include
\[ I(x) = p(x)^{-1/n} \]
for some \( n > 0 \).

Or any monotone non-decreasing non-negative concave function?

Another example. \( I(x) = \) number of prime factors in \( \lceil 1/p(x) \rceil \).
Why \( \log \)?

- We defined the information in event \( \{X = x\} \) as
  \[ I(\{X = x\}) = I(x) = \log \frac{1}{p(x)}, \] but why log?

- Intuition says we want the information about an event to be
  inversely related to the probability, but there are many such
  relationships that might be useful.

- Other possible functions include \( I(x) = p(x)^{-1/n} \) for some \( n > 0 \).
- Or any monotone non-decreasing non-negative concave function?
- Another example. \( I(x) = \) number of prime factors in \( \lceil 1/p(x) \rceil \)
- But log is attractive for a number of reasons.
Why log?

- We defined the information in event \( \{ X = x \} \) as 
  \[ I(\{ X = x \}) = I(x) = \log \frac{1}{p(x)}, \] but why log?

- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.

- Other possible functions include 
  \[ I(x) = p(x)^{-1/n} \text{ for some } n > 0. \]

- Or any monotone non-decreasing non-negative concave function?

- Another example. \( I(x) = \) number of prime factors in \( \lceil 1/p(x) \rceil \)

- But log is attractive for a number of reasons.

- Khinchin’s uniqueness theorem: Assume a measure of information, over distributions, satisfies: (1) maximum value at uniform distribution, (2) Joint entropy of two random variables is sum of marginal plus conditional (chain rule), and (3) augmenting distribution by a zero-probability event makes no change. Only satisfying measure is entropy, up to multiplicative positive constant.
Why \(\log\): Khinchin’s uniqueness theorem

- For a distribution on \(n\) symbols with probabilities \(p = (p_1, p_2, \ldots, p_n)\), let \(H(p) = H(p_1, p_2, \ldots, p_n)\) be the entropy of that distribution.
Why $\log$?: Khinchin’s uniqueness theorem

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

- Consider any information measure, say $\mathcal{H}(p)$ on $p$, and consider the following three natural and desirable properties.
Why $\log$?: Khinchin’s uniqueness theorem

- For a distribution on $n$ symbols with probabilities $p = (p_1, p_2, \ldots, p_n)$, let $H(p) = H(p_1, p_2, \ldots, p_n)$ be the entropy of that distribution.
- Consider any information measure, say $\mathcal{H}(p)$ on $p$, and consider the following three natural and desirable properties.
  - $\mathcal{H}(p)$ takes its largest value when $p_i = 1/n$ for all $i$. 
Why log?: Khinchin’s uniqueness theorem

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

Consider any information measure, say $\mathcal{H}(p)$ on $p$, and consider the following three natural and desirable properties.

1. $\mathcal{H}(p)$ takes its largest value when $p_i = 1/n$ for all $i$.
2. If we define the conditional information as

$$
\mathcal{H}(Y|X) \triangleq \sum_x p(x) \mathcal{H}(p(y_1|x), p(y_2|x), \ldots, p(y_n|x))
$$

(2.23)

$$
= \sum_x p(x) \mathcal{H}(Y|X = x),
$$

(2.24)

then we wish to have additivity in the following way

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

(2.25)
Why $\log$?: Khinchin’s uniqueness theorem

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

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

Theorem 2.3.5 (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

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

(2.26)

for \( \lambda \) a positive constant.

Thus, we get entropy for some logarithmic base.
Entrophy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).
What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

\[
\begin{array}{cc|ccccc}
  x & x_1 & x_2 & x_3 & x_4 & x_5 \\
  p(x) & 0.3 & 0.2 & 0.2 & 0.15 & 0.15 \\
\end{array}
\]
Entropy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:
Entropy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy: a. Is $X = x_5$?
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Consider the following fixed successive guessing strategy:  
a. Is $X = x_5$?  
b. Is $X = x_4$?
What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:

a. Is $X = x_5$?

b. Is $X = x_4$?

c. Is $X = x_3$?
Entrophy and The Guessing Game

What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0.3</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.2</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.2</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.15</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:

- a. Is $X = x_5$?
- b. Is $X = x_4$?
- c. Is $X = x_3$?
- d. Is $X = x_2$?
What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:  

- a. Is $X = x_5$?  
- b. Is $X = x_4$?  
- c. Is $X = x_3$?  
- d. Is $X = x_2$?  
- e. Is $X = x_1$?
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).
- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability
  \[
  \begin{array}{cccccc}
  x & x_1 & x_2 & x_3 & x_4 & x_5 \\
  p(x) & 0.3 & 0.2 & 0.2 & 0.15 & 0.15
  \end{array}
  \]
- Consider the following fixed successive guessing strategy: a. Is $X = x_5$? b. Is $X = x_4$? c. Is $X = x_3$? d. Is $X = x_2$? e. Is $X = x_1$?
- So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Consider the following fixed successive guessing strategy: a. Is $X = x_5$? b. Is $X = x_4$? c. Is $X = x_3$? d. Is $X = x_2$? e. Is $X = x_1$?

- So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.

- Average number of questions asked is $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^T = 3.35$. 
Entropy and The Guessing Game

- What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq X$ (the domain of the r.v. $X$).
- Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- Consider the following fixed successive guessing strategy: a. Is $X = x_5$? b. Is $X = x_4$? c. Is $X = x_3$? d. Is $X = x_2$? e. Is $X = x_1$?
- So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.
- Average number of questions asked is $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^T = 3.35$.
- If we first ask about $x_1$, then $x_2$ it is better, i.e., $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5)^T = 2.65.$
What is the best strategy to guess the value of a random variable with yes/no questions of the form “Is $X \in S$?” for some set $S \subseteq \mathcal{X}$ (the domain of the r.v. $X$).

Ex: Let $X \in \{x_1, x_2, x_3, x_4, x_5\} = \mathcal{X}$ have probability

<table>
<thead>
<tr>
<th>$x$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Consider the following fixed successive guessing strategy:

a. Is $X = x_5$?

b. Is $X = x_4$?

c. Is $X = x_3$?

d. Is $X = x_2$?

e. Is $X = x_1$?

So we ask 5 questions 30% of the time, 4 questions 20% of the time, etc.

Average number of questions asked is

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^T = 3.35.$$

If we first ask about $x_1$, then $x_2$ it is better, i.e.,

$$(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5)^T = 2.65.$$

Is this the best we can do?
Consider the following strategy.

- If $X \in \{x_2, x_3\}$:
  - Y: $x_2$ with probability 0.2
  - N: $x_3$ with probability 0.2

- If $X \in \{x_1\}$:
  - Y: $x_2$ with probability 0.3
  - N: $x_3$ with probability 0.2

- If $X \in \{x_4\}$:
  - Y: $x_4$ with probability 0.15
  - N: $x_5$ with probability 0.15

---

Average number of questions:

\[
2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3
\]

Note, $H(X) = 2.271$. In general, the average number of questions is always $\geq H(X)$. 
Consider the following strategy.

- Average number of questions
  \[2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3\]
Consider the following strategy.

![Diagram of a guessing game]

- Average number of questions
  \[2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3\]
- Note, \(H(X) = 2.271\)

**Entropy and The Guessing Game**

- \(X \in \{x_2, x_3\}\)
- \(X \in \{x_2, x_3\}\)
- \(X \in \{x_1\}\)
- \(X \in \{x_4\}\)
- \(X \in \{x_5\}\)
Consider the following strategy.

Average number of questions
\[ 2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3 \]

Note, \( H(X) = 2.271 \)

In general, the average number of questions is always \( \geq H(X) \).
Entropy and The Guessing Game

- Difference between the best & worst first question in terms of how it splits the distribution.
- Worst: $X = x_5$? $p(X = x_5) = 0.15$, $p(X \neq x_5) = 0.85$, and the entropy $H(0.15, 0.85) = 0.6098$.
- Best: $X \in \{x_2, x_3\}$? $p(X \in \{x_2, x_3\}) = 0.4$, $p(X \notin \{x_2, x_3\}) = 0.6$, and the entropy $H(0.4, 0.6) = 0.971$.
- In general, it is better to first ask questions who, when seen as a random variable, have higher entropy.
- This will quickly reduce the remaining uncertainty.
- Note the relationship to $H(Y|X) + H(X) = H(X,Y)$. If we ask a question with large $H(X)$, the residual uncertainty $H(Y|X)$ is made smaller.
Mutual Information Intuition

- Given two random variables $X$ and $Y$, how much information do they have about each other?
- If we know $X$ how much do we learn about $Y$? If we know $Y$, how much do we learn about $X$?
- If they are independent, $X \perp \! \! \! \! \! \! \! \! \! \! \! \! \perp Y$, then knowing $X$ should tell us nothing about $Y$ and vice versa.
- Since we now have a measure of information in a random source, $H(X)$, we can quantity how much information random variables have about each other, this is mutual information.
Event Mutual Information

- Given event \( \{X = x, Y = y\} \), we can ask for the information provided about event \( x \) from the fact that event \( y \) occurred.
- This can be quantified as:

\[
I(x; y) = \log \frac{p(x|y)}{p(x)} = \log \frac{1}{p(x)} - \log \frac{1}{p(x|y)} \quad \text{(2.27)}
\]

- First term: surprise that \( x \) occurred
- Second term: surprise that \( x \) occurred given that \( y \) occurred.
- Difference: difference of surprise, how much the surprise has changed between not knowing \( y \) and knowing \( y \).
- Note: \( p(x|x) = 1 \), so

\[
I(x; x) = \log \frac{1}{p(x)} - \log 1 = \log \frac{1}{p(x)} = I(x), \text{ so that } I(x) \text{ can be seen as a form of “self-information”}.
\]
The mutual information (MI) is the average amount of information that r.v. $X$ has about $Y$, and vice verse.

**Definition 2.4.1 (mutual information)**

\[
I(X; Y) = E_{p(x,y)} \log \frac{p(x|y)}{p(x)} = E_{p(x,y)} \log \frac{p(x|y)p(y)}{p(x)p(y)} \tag{2.28}
\]

\[
= E_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)} = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{2.29}
\]
Proposition 2.4.2

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

(2.30)

Proof.

\[
E \log \frac{p(x|y)}{p(x)} = E \log \frac{1}{p(x)} - E \log \frac{1}{p(x|y)} = H(X) - H(X|Y) 
\]  

(2.31)

and the other

- By symmetry, we also have \( I(X; Y) = H(Y) - H(Y|X) \).
Proposition 2.4.2

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

(2.30)

Proof.

\[ E \log \frac{p(x|y)}{p(x)} = E \log \frac{1}{p(x)} - E \log \frac{1}{p(x|y)} = H(X) - H(X|Y) \]  

(2.31)

and the other

- By symmetry, we also have \( I(X; Y) = H(Y) - H(Y|X) \).
- Also, since \( H(X) \geq 0 \) and \( H(X|Y) \geq 0 \), we have \( I(X; Y) \leq \min(H(X), H(Y)) \).
Recall, chain rule of entropy: \( H(X, Y) = H(X) + H(Y|X) \)

MI: \( I(X; Y) = H(X) - H(X|Y) \)

Don’t get confused with the commas vs. semicolons!!

Given the above, we have

\[
I(X; Y) = H(X) + H(Y) - H(X, Y)
\]  

which is a simple instance of the inclusion-exclusion principle in combinatorics.

It is sometimes useful to visualize this relationship with pictures.
Given random variable $X$, the entropy of (uncertainty in, average surprise in, information contained within, etc.) a random variable can be displayed using a 2D area, as given above. The area of the set conveys the "degree of information".
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.
Given random variable $X$, the entropy of (uncertainty in, average surprise in, information contained within, etc.) a random variable can be displayed using a 2D area, as given above.

The area of the set conveys the “degree of information”
A way of looking at the relationships.

- $H(X, Y)$
- $H(X)$
- $H(Y)$
- $H(X|Y)$
- $H(Y|X)$
- $I(X;Y)$

Mutual Information and Entropy - Venn Diagram
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.
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
- Intuitively, how do we measure a form of “distance” (in an informal sense) between two probability distributions $p$ and $q$ in a way that is also useful?

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

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.
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.
- Intuitively, how do we measure a form of “distance” (in an informal sense) between two probability distributions \( p \) and \( q \) in a way that is also 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

- Intuitively, how do we measure a form of “distance” (in an informal sense) between two probability distributions \( p \) and \( q \) in a way that is also 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

Intuitively, how do we measure a form of “distance” (in an informal sense) between two probability distributions $p$ and $q$ in a way that is also 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.

- Intuitively, how do we measure a form of “distance” (in an informal sense) between two probability distributions $p$ and $q$ in a way that is also 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.
Distance

Definition 2.5.1 (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)
Distance

Definition 2.5.1 (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)
Distance

**Definition 2.5.1 (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)
Distance

Definition 2.5.1 (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 2.5.1 (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 2.5.2 (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)
### Distance

**Definition 2.5.1 (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 2.5.2 (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)
Distance

Definition 2.5.1 (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 2.5.2 (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)
Distance

Definition 2.5.1 (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 2.5.2 (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)
Definition 2.5.1 (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 2.5.2 (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)

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

**Definition 2.5.1 (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 2.5.2 (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.
KL-Divergence

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

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

\[
D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)}
\]  
(2.33)
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 2.5.3 (KL-Divergence)**

$$D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (2.33)$$

- It is like an expected log-odds ratio, weighted by $p$. 
Given two distributions $p(x)$ and $q(x)$ over the same alphabet, i.e., $p(x) = P_p(X = x)$ and $q(x) = P_q(X = x)$, then the KL-divergence is defined as follows:

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

(2.33)

- 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 2.5.3 (KL-Divergence)**

  \[
  \begin{align*}
  D(p||q) & \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} \\
  & \quad (2.33)
  \end{align*}
  \]

- 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 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

\[
D(p \| q) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) \log \frac{p(x_1, x_2, \ldots, x_N)}{q(x_1, x_2, \ldots, x_N)} \tag{2.34}
\]
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

\[
D(p||q) = \sum_{x_1, x_2, \ldots, x_N} p(x_1, x_2, \ldots, x_N) \log \frac{p(x_1, x_2, \ldots, x_N)}{q(x_1, x_2, \ldots, x_N)} \tag{2.34}
\]

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

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

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

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 $Y$. 
KL-Divergence and MI

Let \( \mu_1(x, y) = p(x, y) \) and \( \mu_2(x, y) = p(x)p(y) \) with
\[
p(x) = \sum_y p(x, y) \quad \text{and} \quad p(y) = \sum_x p(x, y)
\]
then
\[
D(\mu_1 || \mu_2) = \sum_{x,y} \mu_1(x, y) \log \frac{\mu_1(x, y)}{\mu_2(x, y)}
\]
\[
= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} = I(X; Y) \quad (2.36)
\]

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

Product of marginal distributions \( p(x) = \sum_y p(x, y) \) is a projection of \( p(x, y) \) down to the independent distribution. I.e.,
\[
p(x)p(y) = \arg\min_{p'(x,y)} D(p(x, y) || p'(x, y)) \quad (2.37)
\]
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.
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.
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

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

(2.38)
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

\[
I(X; Y|Z = z) = \sum_{x,y} \frac{p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}}{p(x|z)p(y|z)} \tag{2.38}
\]

- 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

\[
I(X; Y | Z = z) = \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}
\] (2.38)

- 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:

\[
I(X; Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\] (2.39)

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

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

$$I(X; Y|Z)$$
Conditional Mutual Information

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

Definition 2.5.4 (conditional mutual information)

$$I(X; Y|Z) \triangleq \sum_z p(z) I(X; Y|Z = z)$$ (2.40)
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 2.5.4 (conditional mutual information)**

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

(2.40)
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 2.5.4 (conditional mutual information)**

\[
I(X; Y|Z) \triangleq \sum_z p(z) I(X; Y|Z = z) \\
= \sum_z p(z) E_{p(x,y|z)} \log \frac{p(x, y|Z = z)}{p(x|Z = z)p(y|Z = z)} \\
= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}
\]
Conditional Mutual Information

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

**Definition 2.5.4 (conditional mutual information)**

$$I(X; Y|Z) \triangleq \sum_z p(z)I(X; Y|Z = z)$$  \hspace{1cm} (2.40)

$$= \sum_z p(z)E_{p(x,y|z)} \log \frac{p(x, y|Z = z)}{p(x|Z = z)p(y|Z = z)}$$ \hspace{1cm} (2.41)

$$= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}$$ \hspace{1cm} (2.42)

$$= E \left[ \log \frac{1}{p(x|z)} - \log \frac{1}{p(x|y,z)} \right]$$ \hspace{1cm} (2.43)
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 2.5.4 (conditional mutual information)**

\[
I(X; Y|Z) \triangleq \sum_z p(z) I(X; Y|Z = z) \tag{2.40}
\]

\[
= \sum_z p(z) E_{p(x,y|z)} \log \frac{p(x, y|Z = z)}{p(x|Z = z)p(y|Z = z)} \tag{2.41}
\]

\[
= \sum_{x,y,z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \tag{2.42}
\]

\[
= E \left[ \log \frac{1}{p(x|z)} - \log \frac{1}{p(x|y, z)} \right] \tag{2.43}
\]

\[
= H(X|Z) - H(X|Y, Z) \tag{2.44}
\]
Proposition 2.5.5

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

example:  \[ I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \]
Proposition 2.5.5

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

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

Proof.

\[ I(X_1, \ldots, X_N; Y) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_N|Y) \]
Proposition 2.5.5

\[ I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \] (2.45)

example: \[ I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1) \] (2.46)

Proof.

\[ I(X_1, \ldots, X_N; Y) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_N|Y) \] (2.47)

\[ = \sum_i H(X_i|X_1, \ldots, X_{i-1}) - \sum_i H(X_i|X_1, \ldots, X_{i-1}, Y) \] (2.48)
Proposition 2.5.5

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

(2.45)

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

(2.46)

Proof.

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

(2.47)

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

(2.48)

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

(2.49)
Definition 2.5.6

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

- Same as standard KL-divergence but now using conditional distribution.
Definition 2.5.6

\[ D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \tag{2.50} \]

- Same as standard KL-divergence but now using conditional distribution.
- We take the expected value of \( \log \frac{p(y|x)}{q(y|x)} \) w.r.t. \( p(x,y) \) even though \( D(p(y|x)||q(y|x)) \) mentions only \( p(y|x) \).
Chain Rule for KL-divergence

Proposition 2.5.7

\[ D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y|x) || q(y|x)) \] (2.51)
Proposition 2.5.7

\[ D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y \| x) \| q(y \| x)) \] (2.51)

Proof.

\[ D(p(x, y) \| q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \] (2.52)
Chain Rule for KL-divergence

Proposition 2.5.7

\[ D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \]  
(2.51)

Proof.

\[ D(p(x, y)||q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \]  
(2.52)

\[ = \sum_{x,y} p(x, y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \]  
(2.53)
Proposition 2.5.7

\[ D(p(x, y) \| q(x, y)) = D(p(x) \| q(x)) + D(p(y|x) \| q(y|x)) \] (2.51)

Proof.

\[ D(p(x, y) \| q(x, y)) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{q(x, y)} \] (2.52)

\[ = \sum_{x,y} p(x, y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \] (2.53)

\[ = \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} + \sum_{x,y} p(x, y) \log \frac{p(x)}{q(x)} \] (2.54)
ConVex Functions

- $f$ is said to be convex on $(a, b)$ if for all $x_1, x_2 \in (a, b), 0 \leq \lambda \leq 1$,
  \[
  f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)
  \]  
  (2.55)
ConVex Functions

- $f$ is said to be convex on $(a, b)$ if for all $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$,
  \[ f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \]  
  (2.55)

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

Prof. Jeff Bilmes
**Convex Functions**

- $f$ is said to be convex on $(a, b)$ if for all $x_1, x_2 \in (a, b), 0 \leq \lambda \leq 1$,
  
  $f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$  \hspace{1cm} (2.55)

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

- Visualized:
**ConVex Functions**

- \( f \) is said to be convex on \((a, b)\) if for all \( x_1, x_2 \in (a, b), \ 0 \leq \lambda \leq 1, \)
  \[
  f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \quad (2.55)
  \]

- 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 \( \lambda = 1.\)
Jensen’s Inequality

**Theorem 2.5.8 (Jensen)**

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

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

(2.56)
Theorem 2.5.8 (Jensen)

Let \( f \) be a convex function and \( X \) a random variable, then

\[
Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f\left(\sum_x xp(x)\right) \quad (2.56)
\]

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 2.5.9

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  \hspace{1cm} (2.57)

Proof.

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

\[ -D(p||q) \]  \hspace{1cm} (2.58)
Lemma 2.5.9

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  \hspace{1cm} (2.57)

Proof.

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

\[ -D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} \]  \hspace{1cm} (2.58)
KL Divergence is non-negative

Lemma 2.5.9

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  \hspace{1cm} (2.57)

Proof.

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

\[ -D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \]  \hspace{1cm} (2.58)
Lemma 2.5.9

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

(2.57)

Proof.

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

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

(2.58)

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

(2.60)
KL Divergence is non-negative

**Lemma 2.5.9**

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  \hspace{1cm} (2.57)

**Proof.**

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

\[ -D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \]  \hspace{1cm} (2.58)

\[ = \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \]  \hspace{1cm} (2.59)

\[ (2.60) \]
KL Divergence is non-negative

**Lemma 2.5.9**

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

(2.57)

**Proof.**

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

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

(2.58)

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

(2.59)

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

(2.60)
KL Divergence is non-negative

Lemma 2.5.9

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \]  \hspace{1cm} (2.57)

**Proof.**

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

\[
-D(p||q) = - \sum_x p(x) \log \frac{p(x)}{q(x)} = - \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \]  \hspace{1cm} (2.58)

\[
= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \]  \hspace{1cm} (2.59)

\[
= \log \left( \sum_{x \in A} q(x) \right) \leq \log \left( \sum_x q(x) \right) \]  \hspace{1cm} (2.60)
KL Divergence is non-negative

Lemma 2.5.9

\[ D(p||q) \geq 0 \text{ with equality iff } p(x) = q(x) \text{ for all } x \tag{2.57} \]

Proof.

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

\[
-D(p||q) = -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \tag{2.58}
\]

\[
= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \leq \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \tag{2.59}
\]

\[
= \log \left( \sum_{x \in A} q(x) \right) \leq \log \left( \sum_x q(x) \right) = \log 1 = 0 \tag{2.60}
\]
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.
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.
Recall from a few slides ago: if $f$ is strictly convex, then 
$\{Ef(Z) = f(E(Z))\} \Rightarrow \{Z = EZ\}$ meaning $Z$ is a constant random variable.

Note that $\log x$ is strictly concave.

Thus, equality in 
$\sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} = \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right)$
means $Z = EZ$ with $Z = q(X)/p(X)$, so $Z$ is a constant random variable.
Recall from a few slides ago: if $f$ is strictly convex, then 
\[
\{ Ef(Z) = f(E(Z)) \} \Rightarrow \{ Z = EZ \}
\]
meaning $Z$ is a constant random variable.

Note that $\log x$ is strictly concave.

Thus, equality in
\[
\sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} = \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right)
\]
means $Z = EZ$ with $Z = q(X)/p(X)$, so $Z$ is a constant random variable.

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

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

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

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

Note that \( \log x \) is strictly concave.

Thus, equality in \( \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} = \log \left( \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \right) \) means \( Z = EZ \) with \( Z = q(X)/p(X) \), so \( Z \) is a constant random variable.

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

Thus, if \( p(x) = q(x) \) then \( D(p || q) = 0 \) and vice versa.

We’ll use this theorem to prove important properties about mutual information.
Proposition 2.5.10

\[ I(X; Y) \geq 0 \text{ and } I(X; Y) = 0 \iff X \perp \!\!\!\!\!\!\!\!\!\perp Y \]
Proposition 2.5.10

\[ I(X; Y) \geq 0 \quad \text{and} \quad I(X; Y) = 0 \iff X \perp \! \! \! \perp Y \quad (2.61) \]

Proof.

\[ I(X; Y) = D(p(x, y) \| p(x)p(y)) \geq 0 \quad (2.62) \]

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

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

$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. 
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 never increase entropy

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

**Proposition 2.5.11**

\[ H(X|Y) \leq H(X) \text{ and } H(X|Y) = H(X) \iff X \perp \!\!\!\!\!\!\perp Y \]  \hspace{1cm} (2.63)
Conditioning can never increase entropy

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

**Proposition 2.5.11**

\[ H(X|Y) \leq H(X) \text{ and } H(X|Y) = H(X) \text{ iff } X \perp \!\!\!\!\!\! Y \] (2.63)

Proof.

0 \leq I(X;Y) = H(X) - H(X|Y) (2.64)
Conditioning can never increase entropy

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

**Proposition 2.5.11**

\[
H(X|Y) \leq H(X) \quad \text{and} \quad H(X|Y) = H(X) \quad \text{iff} \quad X \perp \!\!\!\!\!\!\perp Y \quad (2.63)
\]

**Proof.**

\[
0 \leq I(X;Y) = H(X) - H(X|Y) \quad (2.64)
\]

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

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

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

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

- Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions).
- Let $T$ be an independent random shuffle operation (permutation), i.e., $T \perp X$.
- Then $H(TX) \geq H(X)$.
- Follows since 
  \[ H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X) \]
  (conditioning reduces entropy, as we just learned).
Shuffles

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

- Shuffles. $X$ is a random variable indicating positions of a deck of cards (i.e., $X = x$ is one set of positions)
- Let $T$ be an independent random shuffle operation (permutation), i.e., $T \perp \perp X$.
- Then $H(TX) \geq H(X)$.
- Follows since $H(TX) \geq H(TX|T) = H(T^{-1}TX|T) = H(X|T) = H(X)$ (conditioning reduces entropy, as we just learned).
- Consider an example: $X$ is low entropy, one set of card positions very likely, the rest very unlikely. How could $T$ increase entropy?
- Question: What if $T$ is deterministic? Would this contradict the above result (yes or now)? Why?
Additive Independence Bounds on Entropy

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

**Proposition 2.5.12**

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

(2.65)
Entropy of a set of random variables is highest when the random variables are independent - the least redundancy between them.

**Proposition 2.5.12**

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

(2.65)
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 2.5.12

\[ H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \]  \hspace{1cm} (2.65)

Proof.

\[ H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{N} H(X_i) \]  \hspace{1cm} (2.66)
Independence Bounds on Entropy

Two variable instance of Proposition 2.5.12 is

\[ H(X_1, X_2) \leq H(X_1) + H(X_2) \] (2.67)
Independence Bounds on Entropy

- Two variable instance of Proposition 2.5.12 is

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

(2.67)

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

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

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

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

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

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

Thus, we can have $I(X; Y) > I(X; Y|Z)$.
What about $I(X; Y)$ vs. $I(X; Y|Z)$?

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

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

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

On the other hand, if $Z = X + Y$ and $X \perp \perp Y$ then $I(X; Y) = 0$ but $I(X; Y|Z) > 0$.
Conditioning and Mutual Information

- What about $I(X; Y)$ vs. $I(X; Y|Z)$?
- If $X \perp Y|Z$ then $I(X; Y|Z) = 0$. For example, $X \perp Y|Z$ whenever $X \rightarrow Z \rightarrow Y$.
- Alternatively, if $Z = Y$, then $I(X; Y|Z) = 0$.
- Thus, we can have $I(X; Y) > I(X; Y|Z)$.
- On the other hand, if $Z = X + Y$ and $X \perp Y$ then $I(X; Y) = 0$ but $I(X; Y|Z) > 0$.
- Thus, no general conditioning relationship for mutual information and conditional mutual information.
$H(X) = EI(x) = - \sum_x p(x) \log p(x)$ (2.68)

$H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y)$ (2.69)

$H(Y|X) = - \sum_{x,y} p(x, y) \log p(y|x)$ (2.70)

$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ (2.71)

$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$ (2.72)

$0 \leq H(X) \leq \log n$, where $n$ is $X$’s alphabet size. (2.73)
**Review and Summary**

**KL-D:**

\[ D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \]
Review and Summary

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

- **MI**: 
  \[ I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \]
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:** \( I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \)
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 \( \Rightarrow Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \)
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 $\Rightarrow Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x))$
- **KL non-negative:** $D(p||q) \geq 0$, $D(p||q) = 0 \iff p = q$. 

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 2 - Sep 30th, 2019
L2 F55/68 (pg.187/244)
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: \( I(X; Y) \geq 0, I(X; Y) = 0 \iff X \perp \perp Y. \)
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:** $I(X; Y) \geq 0$, $I(X; Y) = 0 \iff X \perp \perp Y$.
- **Conditioning reduces entropy:** $H(X) \geq H(X|Y)$, $H(X) = H(X|Y) \iff X \perp \perp Y$. 
Review and Summary

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

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

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

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: \( H(X) \geq H(X|Y), \ H(X) = H(X|Y) \iff X \perp \perp Y \).

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

The area of the set conveys the “degree of information”
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 
\[ I(X_1; X_2) \geq I(X_1; X_2 | X_3) \]
but that neither is ever negative.
Note in the diagram, \( I(X_1; X_2) = I(X_1; X_2 | X_3) + I(X_1; X_2; X_3) \)

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

Thus, \( I(X_1; X_2; X_3) \triangleq I(X_1; X_2) - I(X_1; X_2 | X_3) \) can be negative.
Note in the diagram, \( I(X_1; X_2) = I(X_1; X_2|X_3) + I(X_1; X_2; X_3) \)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

\(-I(X_1; X_2; X_3)\) called the EAR (explaining away residual) measure in pattern recognition, and “synergy” in neuroscience. Also,
\[ I(X_1; X_2; X_3) = I(X_1; X_2) + I(X_3; X_2) - I(X_1, X_3; X_2) \]
Log-sum inequality

**Theorem 2.7.1**

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

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

and we have equality iff \(a_i/b_i = c = \text{const.}\)
Theorem 2.7.1

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

\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.74}
\]

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 \(0 \log 0/0 = 0\).
Log-sum inequality

Theorem 2.7.1

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

\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}
\] (2.74)

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.

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:

$$\sum_i \alpha_i f(t_i) \geq f(\sum_i \alpha_i t_i) \text{ with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (2.75)$$
Log-sum inequality

Proof of log-sum inequality.

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

$$\sum_{i} \alpha_i f(t_i) \geq f(\sum_{i} \alpha_i t_i) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_{i} \alpha_i = 1. \quad (2.75)$$

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

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

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

\[
\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with} \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_i \alpha_i = 1. \quad (2.75)
\]

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

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

(2.76)
Log-sum inequality

Proof of log-sum inequality.

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

$$
\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \tag{2.75}
$$

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

$$
\sum_{i=1}^{n} \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \alpha_i f(t_i) \tag{2.76}
$$
Proof of log-sum inequality.

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

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (2.75)$$

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

$$\sum_{i=1}^{n} \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \alpha_i f(t_i) \quad (2.76)$$

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

Proof of log-sum inequality.

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

\[
\sum_i \alpha_i f(t_i) \geq f(\sum_i \alpha_i t_i) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1.
\] (2.75)

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

\[
\sum_{i=1}^{n} \frac{a_i}{\sum_{j} b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \frac{b_i}{\sum_{j} b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^{n} \alpha_i f(t_i)
\] (2.76)

\[
\geq f(\sum_i \alpha_i t_i) = \left(\sum_i \alpha_i t_i\right) \log \left(\sum_j \alpha_j t_j\right)
\] (2.77)
Proof of log-sum inequality.

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

$$\sum_i \alpha_i f(t_i) \geq f(\sum_i \alpha_i t_i) \quad \text{with } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1. \quad (2.75)$$

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

$$\sum_{i=1}^n \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \sum_{i=1}^n \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \sum_{i=1}^n \alpha_i f(t_i) \quad (2.76)$$

$$\geq f(\sum_i \alpha_i t_i) = \left( \sum_i \alpha_i t_i \right) \log \left( \sum_j \alpha_i t_i \right) \quad (2.77)$$

$$= \left( \sum_i \frac{a_i}{\sum_j b_j} \right) \log \left( \frac{\sum_i a_i}{\sum_j b_j} \right) \quad (2.78)$$
Convexity of $D(p||q)$ in the pair

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

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

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

\[ (\lambda p_1 + (1-\lambda)p_2)x \log (\lambda p_1 + (1-\lambda)p_2) \]

\[ \leq \lambda p_1 \log p_1 \]

\[ + (1-\lambda) p_2 \log p_2 \]

And then sum over $x$. 

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 2 - Sep 30th, 2019
L2 F64/68 (pg.216/244)
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 $D(p\|q)$ 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.
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

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

and then sum over $x$. 

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 2 - Sep 30th, 2019
L2 F64/68(pg.219/244)
**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

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

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

(2.79)
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

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

(2.79)
Convexity of $D(p||q)$ in the pair

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

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

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

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

(2.79) (2.80) (2.81)
Convexity of $D(p||q)$ in the pair

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

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

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

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

(2.79) - (2.81)

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

$$H(p)$$
We saw this before, mixing distributions can only increase entropy relative to the same mixture of the entropies.

Proof.

\[ H(p) = - \sum_i p_i \log p_i \]
We saw this before, mixing distributions can only increase entropy relative to the same mixture of the entropies.

Proof.

\[
H(p) = -\sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{2.82}
\]
Entropy is concave in \( p \)

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

**Proof.**

\[
H(p) = -\sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{2.82}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \tag{2.83}
\]
We saw this before, mixing distributions can only increase entropy relative to the same mixture of the entropies.

Proof.

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \quad (2.82)
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \quad (2.83)
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i - p_i \log 1/|\mathcal{X}| \right) \quad (2.84)
\]
Entropy is concave in $p$

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

Proof.

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

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

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

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

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

**Proof.**

\[
H(p) = - \sum_i p_i \log p_i = \sum_i (-p_i \log p_i) + \log |\mathcal{X}| - \log |X| \tag{2.82}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i + p_i \log |\mathcal{X}| \right) \tag{2.83}
\]

\[
= \log |\mathcal{X}| - \left( \sum_i p_i \log p_i - p_i \log 1/|\mathcal{X}| \right) \tag{2.84}
\]

\[
= \log |\mathcal{X}| - D(p||u) \tag{2.85}
\]

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

- Let \((X, Y)\) be a joint r.v. space, so \(p(x, y) = p(y|x)p(x)\).
Consequences for MI

- Let \((X, Y)\) be a joint r.v. space, so \(p(x, y) = p(y|x)p(x)\).
- Then \(I(X; Y)\) is a concave function of \(p(x)\) for fixed \(p(y|x)\).
Consequences for MI

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

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

- Let \((X, Y)\) be a joint r.v. space, so \(p(x, y) = p(y|x)p(x)\).
- Then \(I(X; Y)\) is a \textbf{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 \textbf{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 \(I_{p(y|x)}(X; Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)}\),

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

\[ X \xrightarrow{p(y|x)} Y \]
MI and communications and convexity

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

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

- 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.
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.

- 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.