Information Theory

The 1st Module

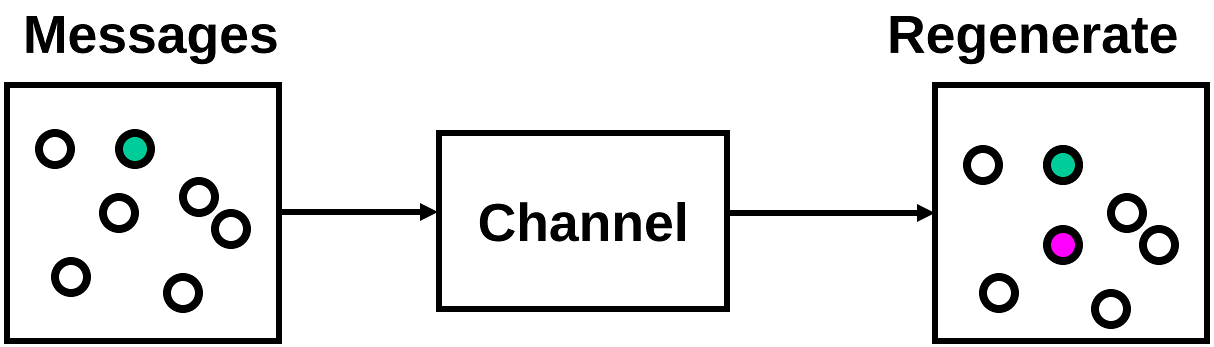
Play

Claude E. Shannon (1916-2001)

- Math/EE Bachelor from UMich (1936)
- MSEE and Math Ph.D. from MIT (1940)
- A landmark paper "Mathematical Theory of Communications" (1948)
  - Founder of Information Theory
  - Fundamental limits on communications
  - Information quantified as a logarithmic measure
- For more info on him, make a visit to http://www.bell-labs.com/news/2001/february/26/1.html

Play

Novel Perspective on Communications



- Communications: Transfer of information from a source to a receiver
- Messages (information) can have semantic meaning; but they are irrelevant for the design of a comm. system.
- What's important then?
  - A message is selected from a set of all possible messages and transmitted, and regenerated at the receiver.
  - The size of the message set has something to do with the amount of information.
- *The capacity of the channel is the maximum size of message set* that can be transferred over the channel and can be regenerated *almost error-free* at the receiver

The Size *M* of Message Set

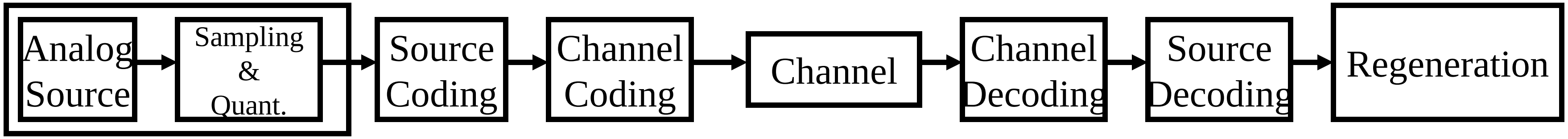
- Is the Amount of information
- *M* or any monotonic function of *M* can be used as a measure of information.
- His choice was the logarithmic function. Why?
  - If $M_1 > M_2$, $\log(M_1) > \log(M_2)$
  - When base 2, log2(M) is the number of memory cells.
  - We call the resulting unit "bits."
  - A four-bit register can represent a message set of size 2^4, and a three-bit register 2^3.
  - The amount of information is $\log_2(2^4) = 4$ bits and 3 bits.
  - This choice was made out of convenience; but considered appropriate (See the axiomatic definition of entropy in Cover & Thomas 1st Ed., Prob2.4)

Fundamental Limits on Communications Systems

- The Sampling and Modulation Theorem (Nyquist and Hartley 1928)
- Source and Channel Coding Theorem (Shannon)

- Can we define a quantity which measures the amount of information produced by a digital or an analog source?
- Rate Distortion and Source Coding Theorem:
  - "*n*-bit quantization": Distortion will increase if we reduce *n*.
  - *Source code* takes away redundancy in the source and reduces the number of bits required.

- How about the size of message set that can be transferred over a noisy channel almost error-free?
- Channel Capacity and Channel Coding Theorem:
  - *Channel code* adds redundancy in order to gain protection against random error occurring in the channel



Uncertainty and Entropy

- Suppose a set of *n* possible outcomes, each having the probability of occurrence as $p_1, p_2, \ldots, p_n$.
- After a random experiment, we have an outcome.
- Then, we can say about the occurrence of an event.
- **Entropy is a measure of uncertainty** (randomness) on the occurrence of an event.
- We use logarithmic measures (non-negative)
  - $\log(1/p_i) \geq 0$,
- If $p_i < p_j$, then $\log(1/p_i) > \log(1/p_j)$.
  - Less probable event means larger uncertainty.
  - More probable event means smaller uncertainty.
  - The sure event has zero uncertainty.

Definition of Entropy

- Entropy is the average *measure* of uncertainty of a distribution, $p_1, p_2, \ldots, p_n$.

  $$H(p_1, p_2, \ldots, p_n) := \sum_{j=1}^{n} p_j \log(1/p_j)$$

Some Properties of Entropy

- Uncertainty = Amount of Information = The number of bits needed in representation
- More uncertain event carries more information.
- The sure event carries zero amount of information
  - A binary source generates "1" with probability 1. Then, the source produces zero amount of information, i.e., $\log(1/1) = 0$.
  - A binary source generates "1" and "0" with equal probability. Each event carries the same amount of information. Then, this source generates 1 bit of information.

Entropy of a RV

- Let X be a random variable with alphabet A = $\{x_1, x_2, \ldots, x_n\}$ and its probability mass function $p(x) = \Pr\{X=x_i \in A\}$
- We define entropy for r.v. X

$$H(X) := \sum_{x\in\mathcal{X}} p(x) \log(1/p(x))$$

  – Note that in fact this measure has nothing to do with the random variable X, but has everything to do with the distribution.
  – The range of X does not play any role in the calculation of H(X).
- When the base of the logarithm is 2, the unit is "bits."
- When the base is e, the unit is "nats."

H(X) is the Average Uncertainty (Information) of X

- Let’s take some examples
- Ex1) When X is binary
- Ex2) When X is quaternary

Entropy gives the largest lower bound on the number of bits required to represent the set of events

- Ex3) Average Information Content in English
- Assume all 26 letters occur equally likely from a source
  - $\mathbf{H = \log_2(26) = 4.7}$ bits/character

Entropy gives the largest lower bound on the number of bits required to represent the set of events

- Assume some distribution other than uniform
  - a, e, o, t with prob = 0.1
  - h, i, n, r, s with prob= 0.07
  - c, d, f, l, m, p, u, y with prob. = 0.02
  - b, g, j, k, q, v, w, x, z with prob. = 0.01
  - **H = 4.17** bits/character
- Thus, if there was a source generating letters according to this distribution (ignoring spaces, commas, etc), then the source's information rate is 4.17 bits per character.

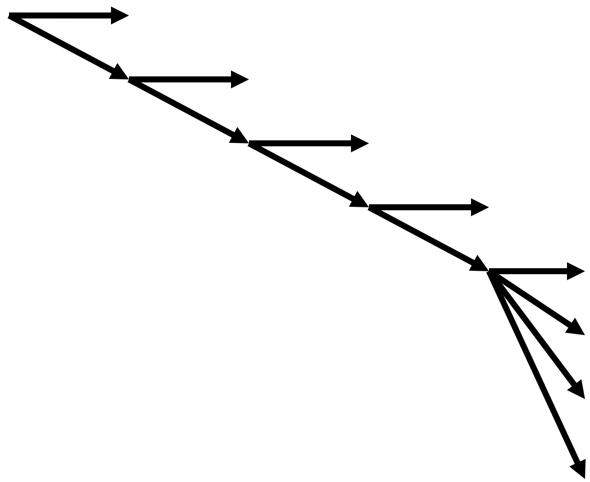
Entropy and Information

- Entropy is the minimum attainable average length of any binary description system.
  - I'll explain this with the next example.

- Ex4) Suppose a race of 8 horses. The race was held in LA yesterday. We are here in Gwangju. There is a reporter in LA. The reporter can only make an binary answer—Yes or No—to our question. Now, knowing that the winning prob. of each horse is (1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64) respectively; which horse would you ask first to be the winning horse? The objective is to determine the winning horse as quickly as possible.
  - Note that the entropy is $H = 2$ bits.

Entropy and Information

p_i			Length
1/2	0	0	1
1/4	1	10	2
1/8	2	110	3
1/16	3	1110	4
1/64	4	111100	6
1/64	5	111101	6
1/64	6	111110	6
1/64	7	111111	6



- The map from the horse index to the binary sequence is a code.
- This coding strategy achieves the entropy bound.
- The average length = 1(1/2) + 2(1/4) + 3(1/8) + 4(1/16) + 6(1/64)*4 = 2 (which is the same as H = 2)
- What happens if the horse index, 0, 1, …,7, was used for the coding? How many bits would be needed then?

Joint Entropy and Conditional Entropy

- Joint Entropy: The joint entropy $H(X, Y)$ of a pair of discrete random variable (X, Y) with a joint distribution $p(x, y)$ is defined as

  $$H(X, Y) := - \sum_x \sum_y p(x, y) \log p(x, y)$$
  $$= - E\{\log p(X, Y)\}$$

- Conditional Entropy:

  $$H(Y \mid X) := - \sum_x \sum_y p(x, y) \log p(y \mid x)$$
  $$= - E\{\log p(Y|X)\}$$
  $$= - \sum_x p(x) H(Y \mid X = x)$$
  $$= - \sum_x p(x) \sum_y p(y|x) \log p(y|x)$$

Chain Rule: $H(X, Y) = H(X) + H(Y|X)$

- $H(X, Y) := - \sum_x \sum_y p(x, y) \log p(x, y)$

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

or similarly

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

Example

Y \ X	0	1
0	2/8	4/8
1	1/8	1/8

- $H(X) = 3/8 * \log_2(8/3) + 5/8* \log_2(8/5) = 0.9544$
- $H(Y) = 6/8 * \log_2(8/6) + 2/8*\log_2(8/2) = 0.8113$
- $H(Y|X) = \sum_x p(x)\, H(Y|X=x)$

  $= 3/8*H(Y|X=0)+5/8*H(Y|X=1)$

  $= 3/8*H(2/3, 1/3)+5/8*H(4/5, 1/5)$

  $= 3/8*0.9183+5/8*0.7219$

  $= 0.7955$
- $H(X, Y) = H(X) + H(Y|X) = 1.75$
- $H(X, Y) = - E\{\log p(X, Y)\}$

  $= 2/8*\log_2(4) + (4/8)*\log_2(2) + 2*1/8*\log_2(8)$

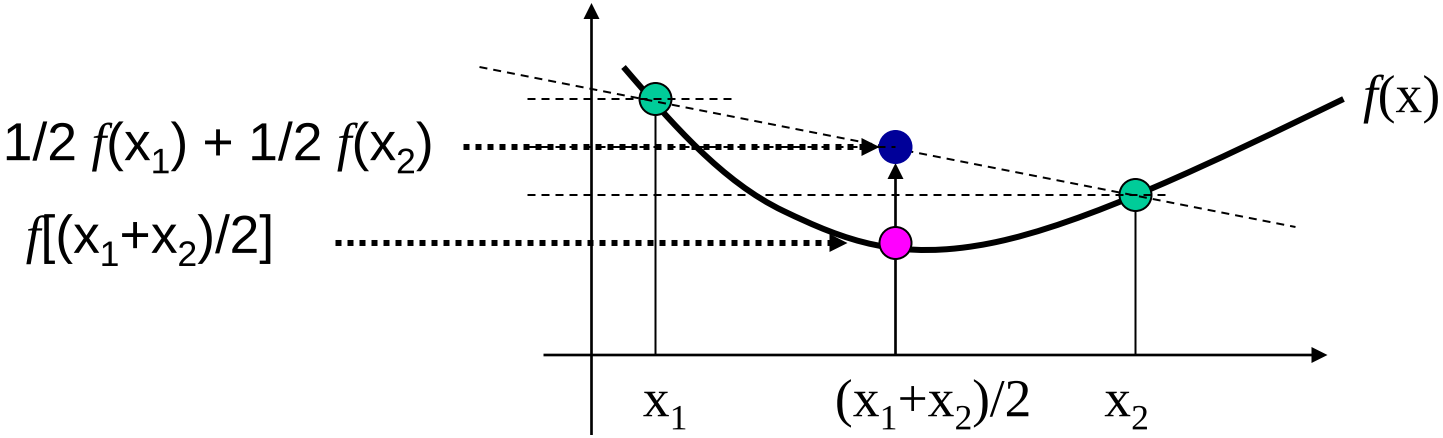
  $= 1/4*2 + 1/2 + 2*3/8 = 1 + 3/4 = 1.75$

The units are [bit].

Max. entropy when uniform

- $H(X) \leq \log|\mathcal{X}|$, where $|\mathcal{X}|$ is the size of alphabet, with equality iff X is uniform over $\mathcal{X}$.
  - Non-uniform gives maximum entropy under a certain input criteria
  - cf) Gaussian distribution gives max. entropy under average energy constraint.
  - I owe you the proof of this statement, especially the only if part.

Jensen's Inequality



- For any $f(x)$ convex U, it is easy to see

$$\frac{1}{2} f(x_1) + \frac{1}{2} f(x_2) \geq f[(x_1+x_2)/2]$$

- This holds true for any distribution $p_1 + p_2 = 1$ such that

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

- For r.v. X and function f convex U,

$$E\{f(X)\} \geq f(E\{X\})$$

  – For strictly convex U $f(x)$, equality *iff* X is a constant

- What if a function is concave ∩?

Relative Entropy is Non-Negative!

- $D(p \| q)$ = Kullback Leibler Distance between two distributions p(z) and q(z)) or Relative Entropy

  $$:= \sum_z p(z) \log(p(z)/q(z))$$

- Suppose p(z) and q(z) are *strict positive* distributions (no zero probability masses). Let S_p and S_q denote their alphabets respectively.

- $$- D(p \| q) = \sum_{z \in S_p} p(z) \log[q(z)/p(z)]$$

  $$\leq \log\{\sum_{z \in S_p} p(z) [q(z)/p(z)]\}$$

  (log is strict concave $\cap$;thus equality only if p(z)/q(z) constant)

  $$= \log\{\sum_{z \in S_p} q(z)\}$$

  $$\leq \log\{\sum_{z \in S_q} q(z)\} = \log(1) = 0$$

- Thus, $D(p \| q) \geq 0$ with equality *iff* p(z) = q(z).
  - Is the equality iff part easy to prove?

Example on Relative Entropy

- Let $\mathcal{X} = \{0, 1\}$ and two distr.'s p(x) and q(x)
- $p(x=0) = 1- r$, $p(x=1) = r$
- $q(x=0) = 1- s$, $q(x=1) = s$
- $D(p \| q) = (1-r) \log[(1-r)/(1-s)] + r \log[r/s]$
- $D(q \| p) = (1-s) \log[(1-s)/(1-r)] + s \log[s/r]$
- Thus, $D(p \| q) \neq D(q \| p)$ in general
  - *Relative Entropy is not symmetric* in general
- Ex) when r = s, then $D(p\|q)=D(q\|p) = 0$
- Ex) when r = 1/2, s = 1/4, $D(p\|q) = 0.2075$, $D(q\|p) = 0.1887$

Relative Entropy is Non-Negative! (Other Approach)

- Suppose p(z) and q(z) are *strict positive* distributions (no zero probability masses). Let S_p and S_q denote their alphabets respectively.
- If the sum $\sum_{z\in S_p} p(z) \log(p(z)/q(z)) = 0$, then $p(z) = q(z)$ for all $z\in S_p$.
- Proof:

$$\sum_z p(z) \log(p(z)/q(z)) \geq \sum_z p(z)\,(1 - q(z)/p(z)) \quad \text{(Why?)}$$
$$= \sum_{z\in S_p} p(z) - \sum_{z\in S_p} q(z)$$
$$\geq (1 - 1) = 0 \quad \text{(Why?)}$$

Entropy is maximum, when uniform distributed

- Proof: Let u(x) be uniform on $\mathcal{X}$

$$
\begin{aligned}
H(p) &= \textstyle\sum_x p(x) \log(1/p(x)) \\
&= \textstyle\sum_x p(x) \{\log(1/p(x)) + \log(u(x)) - \log(u(x))\} \\
&= -\textstyle\sum_x p(x) \log(u(x)) + \sum_x p(x) \{\log[u(x)/p(x)] \\
&= \log|\mathcal{X}| - D(p \,||\, u)
\end{aligned}
$$

Mutual Information is Non-Negative!

- $I(X; Y) := \sum_x \sum_y p(x, y) \log[p(x, y)/p(x)p(y)]$

  $= D(p(x, y) \| p(x)p(y))$

  ----- Distance between the joint and the product distribution.

  ----- Thus, Mutual Information is non-negative.

  $= E_{(x,y)}\{\log[p(X, Y)/p(X)p(Y)]\} \geq 0$

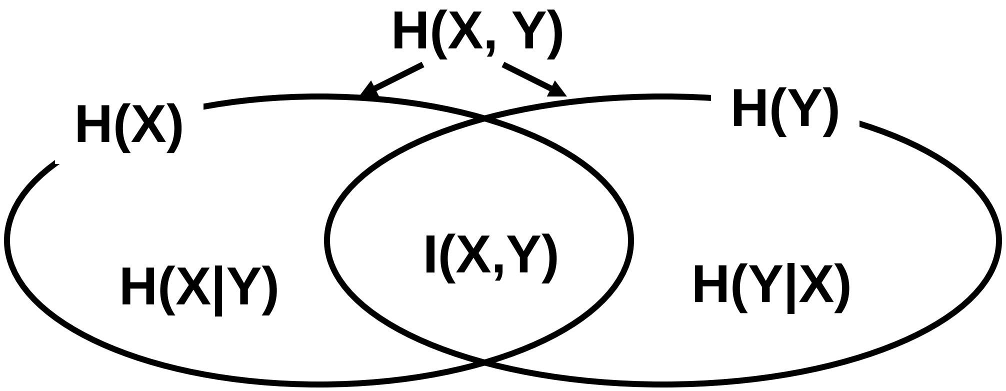
$I(X, Y) = H(X) – H(X | Y)$

- $I(X; Y) = \sum_{x\in X} \sum_{y \in Y} p(x, y) \log[p(x, y)/p(x)p(y)]$
  $= \sum_{x\in X} \sum_{y \in Y} p(x, y) \log[\, p(y)\, p(x |y)/p(x)\, p(y)\,]$
  $= \sum_{x\in X} \sum_{y \in Y} p(x, y) \{\log[p(x |y)] – \log[p(x)]\}$
  $= H(X) – H(X|Y)$
- Reduction in uncertainty of X due to the knowledge of Y
- Also, $I(X; Y) = H(Y) – H(Y|X)$
- How much can I tell about X knowing Y?
- How much can I tell about Y knowing X?
- $I(X; Y) = I(Y; X)$

Mutual Information?

- The measure of amount of information about X we can have knowing Y (vise versa).
  - Cf) Measure of correlation between X and Y, see P2.11.
- Ex) Suppose Y = X, then $H(X|Y) = 0$ (no uncertainty). → Self-mutual information is entropy.
  - Thus, knowing Y means knowing X exactly (the full information $H(X) = H(Y)$ is obtained)
- Ex) Suppose Y and X independent, then $H(X|Y) = H(X)$, then $I(X;Y) = H(X) – H(X) = 0$.
  - Knowing Y cannot tell anything about X.
  - Can you show that if $I(X; Y) = 0$, then X and Y independent?

Relationships



- $I(X; Y) = H(X) – H(X|Y) = H(Y) – H(Y|X)$
- Thus, $I(X; Y) = H(X) + H(Y) – H(X, Y)$

  --- use $H(X, Y) = H(X) + H(Y|X)$

Conditioning reduces entropy

- $H(X|Y) \leq H(X)$, with equality *iff* X and Y independent
  - $I(X;Y) = H(X) - H(X|Y) \geq 0$
- cf) $I(X;Y) = 0$ *iff* X and Y independent.

Chain Rules

- Let $X_1, X_2, \ldots, X_n$ drawn from $p(x_1, x_2, \ldots, x_n)$. Then,

  $H(X_1, X_2) = H(X_1) + H(X_2|X_1)$

  $H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1)$

  $= H(X_1) + H(X_2|X_1) + H(X_3 |X_1,X_2)$

  …

  $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i |X_{i-1}, \ldots, X_1\}$

Watch out for the notation

Results in previous page lead to

- $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$
  with equality iff X_i are independent

Conditional Mutual Information

- $I(X; Y \mid Z) = H(X \mid Z) – H(Y \mid X, Z)$
  $= E\{\log[p(X, Y \mid Z)/p(X \mid Z)p(Y \mid Z)]\}$

- Can we say this?
  - $I(X; Y \mid Z) = 0$ IFF X and Y indep. given Z.

Chain Rule for Information

- $I(X_1, X_2, X_3; Y)$

$$= E\{\log[p(X_1, X_2, X_3, Y)/p(X_1, X_2, X_3)p(Y)]\}$$
$$= H(X_1, X_2, X_3) - H(X_1, X_2, X_3|Y)$$
$$= H(X_1)+H(X_2|X_1) + H(X_3|X_1, X_2)$$
$$- H(X_1|Y) - H(X_2|X_1, Y) - H(X_3|X_1,X_2,Y)$$
$$= I(X_1; Y) + I(X_2; Y|X_1) + I(X_3;Y| X_1, X_2)$$

In general, we have

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i ;Y |X_1, \ldots, X_{i-1})$$

Concavity of log: *Log Sum Inequality*

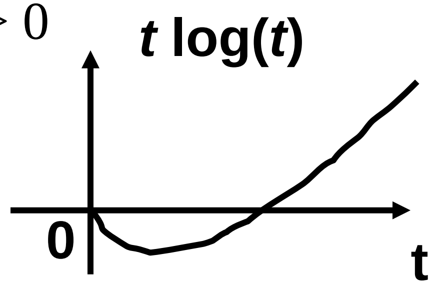
- For non-negative $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$

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

  with equality iff a_i/b_i constant.

  Note, sum of numbers $\geq$ a single number.

- Proof:
  - $f(t) = t \log t$, $t > 0$, is strictly convex ($f''(t) = 1/t > 0$ for $t > 0$)
  - Use the Jensen's Inequality: avg. of maps $\geq$ map of avg.
  - $\sum_{i=1}^{n} \alpha_i f(t_i) \geq f(\sum \alpha_i t_i)$ for $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$, $t_i > 0$
  - Substitute $\alpha_i = b_i / \sum_i b_i$, and $t_i = a_i/b_i$
  - Equality iff a_i/b_i constant



Use the Log Sum Inequality to show $D(p \| q) \geq 0$

- $D(p \| q) = \sum p(x) \log[p(x)/q(x)]$

  $\geq \sum p(x) \log[\sum p(x)/\sum q(x)]$

  $= 1 \log(1/1) = 0$

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

- Mixing distributions decreases the relative entropy
- Consider two pairs (p_1, q_1) and (p_2, q_2) of distributions
- Which one is bigger?
  - Avg. of relative entropies, $0.5(D(p_1\|q_1) + D(p_2\|q_2))$ – (1)
  - Relative entropy of avg. distribution: $D(0.5(p_1 + p_2) \| 0.5(q_1+q_2))$ – (2)
- (1)’: $p_1(x)\log(p_1(x)/q_1(x)) + p_2(x) \log[p_2(x)/q_2(x)]$
- (2)’: $(p_1(x) + p_2(x)) \log[(p_1(x)+p_2(x))/(q_1(x)+q_2(x))]$
- (1)’ $\geq$ (2)’ – the Log Sum Inequality
- Summing over all x, we have (1) $\geq$ (2)

Concavity of Entropy

- Recall the proof that entropy is maximum when the distribution is uniform.
- Let u(x) be uniform on $\mathcal{X}$

$$
\begin{aligned}
H(p) &= \sum_x p(x) \log(1/p(x)) \\
&= \sum_x p(x) \{\log(1/p(x)) + \log(u(x)) - \log(u(x))\} \\
&= -\log(u(x)) + \sum_x p(x) \{\log[u(x)/p(x)] \\
&= \log|\mathcal{X}| - D(p \,||\, u)
\end{aligned}
$$

- Not only is entropy maximum for uniform distribution but also a concave function of p(x).

Concavity of Entropy (other approach)

- H(p) is a *concave* function of a distribution p(x)
- This means if you mix distributions, the entropy increases.
- Let $X_1 \sim p_1(x)$ and $X_2 \sim p_2(x)$
- Let $Z = X_\theta$ where $\theta = 1$ with prob. λ and 2 with $1-\lambda$
- Thus, the distr. of Z is $\lambda\, p_1(x) + (1 - \lambda)\, p_2(x)$
- We know $H(Z) \geq H(Z \mid \theta)$

  --- conditioning reduces entropy
- Thus, we have

  $$H[\lambda\, p_1(x) + (1 - \lambda)\, p_2(x)] \geq \lambda\, H[p_1(x)] + (1-\lambda)\, H[p_2(x)].$$

  – This shows $f(E) \geq E(f)$. Thus, entropy is a concave function of distribution.

Concavity of I(X; Y) over p(x) given p(y|x)

- I(X; Y) = H(Y) – H(Y|X)
- H(Y) is a concave function of p(y).
  - Note $p(y) = \sum p(x)\, p(y|x)$ is a linear function of p(x).
  - Thus, H(Y) is a concave function of p(x).
- $H(Y|X) = \sum p(x)\, H(Y|X = x)$, is a linear function of p(x).
- Thus, I(X; Y) is a concave function of p(x) given p(y|x).

Sequence of results so far

- Relative entropy is non negative. Proved!
- Relative entropy is zero IFF the two distributions are identical. Proved!
- Entropy H(X) is maximum with X ~ uniform distribution.
- Mutual information is a relative entropy.
- Mutual information is thus non negative.
- MI I(X; Y) = 0 IFF X and Y independent.
- Conditioning reduces entropy.
- Entropy is a concave function of distribution.
- MI I(X; Y) is a concave function of p(x) given p(y|x).

HW#1

- Cover & Thomas: Ch2: 1, 2, 5, 8, 12, 14, 18,
- Showing the convexity of $f(x) = e^x$ is easy. Use the Calculus: Take the derivatives twice and show that it's positive everywhere. Now, prove the convexity of $f(x)$ using the general convexity proving technique learned in this lecture.

- (Challenge; Optional) Consider arbitrary random variables X_1, X_2, and

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} N_1 \\ N_2 \end{pmatrix}$$

  where the matrix elements $[a_{ij}]$ are arbitrary non zero constants and N_1 and N_2 are independent random variables. Let's denote $\mathbf{X} := \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.
  Prove or disprove $I(\mathbf{X}; Y_1, Y_2) \le I(\mathbf{X}; Y_1) + I(\mathbf{X}; Y_2)$.

HW#1

- $I(X_1, X_2; Y)$ and $I(\mathbf{X}; Y)$. Are they different?
- Recall the HW#0 problem on the joint distribution of U and V.
  - (a) For the first case where $p_1 = 0.1$ and $p_2 = 0.2$, find the following measures: H(U), H(V), $H(U|\theta_1)$, $H(V|\theta_2)$, H(U|V), H(V|U), H(U, V), I(U; V), $I(U; \theta)$, $I(V; \theta)$.
  - (b) Repeat for $p_1=0.01$ and $p_2 = 0.02$.
  - (c) Note there is a notable change in I(U; V) between (a) and (b). Describe this change and make qualitative statements explaining the change. What would happen to I(U; V) when p_1 and p_2 approach zero? What would happen if they both approach 1/2.