EE514A – Information Theory I  
Winter 2012

Prof. Jeff Bilmes

University of Washington, Seattle  
Department of Electrical Engineering  
Winter Quarter, 2012  
http://j.ee.washington.edu/~bilmes/classes/ee514a_winter_2012/

Lecture 4 - Jan 12th, 2012

Reading

- Read chapters 1, and 2 in C&T.
- Read chapter 3 in C&T, as we’ll start going over it on Tuesday.
Announcements, Assignments, and Reminders

- Late policy: 10% every 24 hour period that you are late, and no more than 3 days late accepted.
- Lowest grade out of all HW grades is not counted towards final grade (so you can skip one HW with impunity).
- Homework 1 is posted, due next Monday at 11:45pm via our dropbox (https://catalyst.uw.edu/collectit/dropbox/karna/19164).
- Please do use our discussion board (https://catalyst.uw.edu/gopost/board/karna/25503/) for all questions, so that all will benefit from them being answered.

Class Road Map

- L1 (1/3): Overview, Entropy
- L2 (1/5): Props. Entropy, Mutual Information, KL-Divergence
- L3 (1/10): KL-Divergence, Jensen, properties, Data Proc. Inequality
- L5 (1/17): Fano, AEP
- L6 (1/19): snow
- L6 (1/24): AEP, source coding
- L7 (1/26): Method of Types
- L9 (2/2): HMMs, coding
- L10 (2/7): Coding, Kraft, Huffman
- L11 (2/9): Huffman, Arithmetic, midterm
- L12 (2/14): Midterm
- L13 (2/16):
- L14 (2/21):
- L15 (2/23):
- L16 (2/28):
- L17 (3/1):
- L18 (3/6):
- L19 (3/8):

Finals Week: March 12th–16th.
**H Relationships**

\[ H(X) = EI(x) = - \sum_x p(x) \log p(x) \]  
(1)

\[ H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y) \]  
(2)

\[ H(Y|X) = - \sum_{x,y} p(x, y) \log p(y|x) \]  
(3)

\[ H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \]  
(4)

\[ 0 \leq H(X) \leq \log n, \quad \text{where } n \text{ is } X's \text{ alphabet size.} \]  
(5)

---

**Review**

- **KL-D**: \( D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \)
- **MI**: \( I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x, y)||p(x)p(y)) \)
- **CMI**: \( I(X; Y|Z) = \sum_{x,y,z} p(x, y, z) \log \frac{p(x,y,z)}{p(x|z)p(y|z)} = H(X|Z) - H(X|Y, Z) \)
- Chain Rule MI: \( I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \)
- **Cond Rel Ent**: \( D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \)
- Chain Rule KL: \( D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \)
- Jensen: \( f \) convex \( \Rightarrow Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f(\sum_x xp(x)) \)
- **KL non-negative**: \( D(p||q) \geq 0 \), \( D(p||q) = 0 \Leftrightarrow p = q \).
- **MI non-negative**: \( I(X; Y) \geq 0 \), \( I(X; Y) = 0 \Leftrightarrow X \perp \!\!\!\!\perp Y \).
- Conditioning reduces entropy: \( H(X) \geq H(X|Y) \), \( H(X) = H(X|Y) \Leftrightarrow X \perp \!\!\!\!\perp Y \).
- **Indep. bound on H**: \( H(X_1, \ldots, X_N) \leq \sum_i H(X_i) \), equality iff all independent.
• Log-sum inequality. Given \((a_1, \ldots, a_n)\) and \((b_1, \ldots, b_n)\), with \(a_i \geq 0\) and \(b_i \geq 0\), we have
\[
\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{n} a_i \right) \log \sum_{i=1}^{n} \frac{a_i}{b_i}
\]
and we have equality iff \(a_i/b_i = c = \text{const.}\).

• \(D(p||q)\) is convex in the pair.

• Entropy \(H(p)\) is concave in \(p\).

• \(I(X;Y)\) is concave in \(p(x)\) for fixed \(p(y|x)\).

• \(I(X;Y)\) is convex in \(p(y|x)\) for fixed \(p(x)\).

---

Given random variable \(X\), the entropy of (uncertainty in, average surprise in, information contained within, etc.) a random variable can be displayed using a 2D area, as given above.

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

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

Mutual Information and Entropy - Block Diagram

Another way of looking at the same relationships.
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, $I(X_1; X_2) = I(X_1; X_2) + I(X_1; X_2 | X_3)$

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

thus, $I(X_1; X_2; X_3) = I(X_1; X_2) - I(X_1; X_2 | X_3)$ 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, $I(X_1; X_2; X_3) = I(X_1; X_2) + I(X_3; X_2) - I(X_1, X_3; X_2)$
**Data Processing Inequality**

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

- Let’s view this as a picture:

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

### Image denoising, important problem in computer vision, big commercial market.

- High ISO images are noisy, but they are the only way to take pictures in low light with narrow aperture (meaning wide depth-of-field).

- Goal of image denoising is to remove the noise from the image and recover the original image.

- Example in our current context:

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

Definition 4.1
Random variables $X$, $Y$, and $Z$ form a Markov chain if $Z \perp \!\!\!\!\!\!\!\!\perp X | Y$. I.e.,

$$p(z, x | y) = p(z | y)p(x | y) \quad \forall x, y, z \quad (7)$$

- This means that

$$p(x, y, z) = p(z | x, y)p(y | x)p(x) = p(z | y)p(y | x)p(x)$$

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

![Graphs](image)

- Ex: If $Z = f(Y)$, then $X \to Y \to Z$ is true (i.e., $X, Y, Z$ form a Markov chain). $f(\cdot)$ can either random or deterministic. Key is that $X$ is irrelevant to determine $Z$ given $Y$.

- Notationally, when we state “$X \to Y \to Z$”, this means that we assert that $X, Y, Z$ form a Markov chain.

Data Processing Inequality

Theorem 4.2 (Data Processing Inequality)

If $X \to Y \to Z$ then

$$I(X; Y) \geq I(X; Z) \quad (8)$$

- So in the Markov chain, the “arrows” correspond to processing and the random variables correspond to data.

- The processing can be either random or deterministic.

- The data processing inequality says that as we perform further processing of a data source, we move away from it, in a Markov chain, and we can (only) lose information about the original source, as measured by mutual information.
Data Processing Inequality

### Proof of Data Processing Inequality

By the chain rule of mutual information:

\[
I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \tag{9}
\]

\[
= I(X; Z) + I(X; Y|Z) \tag{10}
\]

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

- **Example:** what if \(X = Y\)? What if \(H(Y|Z) = 0\)?
- Similarly \(I(Y; Z) \geq I(X; Z)\).
- Chain distance decreases information. How fast?
- **Corollary:** If \(Z = f(Y)\), then \(X \rightarrow Y \rightarrow Z\), or \(X \rightarrow Y \rightarrow f(Y)\). Thus, \(I(X; Y) \geq I(X; f(Y))\).

### Intuition

- Processing can only lose information about \(X\). When \(X\) is source and \(Y\) is receiver, no processing will increase information about \(X\).
- Consider pattern recognition: \(X\) is an object, \(Y\) is a list of features about the object, and \(f(Y)\) is further processing.
- Then any further processing can only reduce information about original object.
- Question: What then is image de-noising? Why do we see result as higher fidelity but it has less information about original image? Other examples: audio restoration.
- Another corollary: If \(X \rightarrow Y \rightarrow Z\), then \(I(X; Y|Z) \leq I(X; Y)\). I.e., 
  \[I(X; Y|Z) = H(X|Z) - H(X|Y, Z) \leq H(X) - H(X|Y)\].
- **Intuition:** Knowing \(Z\) reduces amount learnt between \(X\) and \(Y\)
Data Processing Inequality

- If $X \rightarrow Y \rightarrow Z$, then $I(X; Y|Z) \leq I(X; Y)$, as we just saw.
- Recall, if $X \rightarrow Z \leftarrow Y$, then $I(X; Y|Z) \geq I(X; Y)$.
- E.g., $X \perp \!\!\!\perp Y$ and $Z = X + Y$, the example we saw earlier.
- So, the relationship between $I(X; Y|Z)$ and $I(X; Y)$ depends on the underlying “causal” relationship between the variables.

Entropy Always Increases

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

- **Claim:** When there exists a transaction, the entropy always increases.
- **Consequence:** someday universe will be a gray blur of uniformly distributed matter.
- Model closed system as time-homogeneous 1st-order Markov chain,
  
  $p(x_{n+1}|x_n)$

- $X_n$ is a random variable at time $n$
- We have $\ldots \rightarrow x_{n-1} \rightarrow x_n \rightarrow x_{n+1} \rightarrow \ldots$
- The question we wish to study is: What happens to $H(X_n)$ as $n$ increases.
- Given two starting distributions $p_n(x_n)$ and $q_n(x_n)$, then $D(p_n(x_n)\|q_n(x_n))$ won’t increase with $n$. Why?
- We’ll write $p(x_n)$ for $p_n(x_n)$, also $q(x_n)$ for $q(x_n)$.
- Joint of two successive variables, via the homogeneous dynamics, is:
  
  $p(x_n, x_{n+1}) = p(x_{n+1}|x_n)p_n(x_n)$ (11)
  
  $q(x_n, x_{n+1}) = p(x_{n+1}|x_n)q_n(x_n)$ (12)

---

**Thermal energy in a closed system**

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

- Uneven distribution, smaller $H$

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

- Uniform distribution, big $H$
Entropy and thermodynamics

Then we have:

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

\[ = 0 \] (13)

\[ = D(p(x_{n+1}) || q(x_{n+1})) + D(p(x_{n} | x_{n+1}) || q(x_{n} | x_{n+1})) \geq 0 \] (14)

This means that

\[ D(p(x_n) || q(x_n)) \geq D(p(x_{n+1}) || q(x_{n+1})) \] (15)

This means that

\[ D(p(x_n) || q(x_n)) \geq D(p(x_{n+1}) || q(x_{n+1})) \] (16)

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

Then \( p(x_n) \) approaches this stationary distribution in the sense

\[ D(p(x_n) || q) \geq D(p(x_{n+1}) || q) \]. We cannot move farther away from any stationary distribution.
Entropy and thermodynamics

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

$$D(p(x_n)||u) = -H(X_n) + \log n \geq -H(X_{n+1}) + \log n$$  \hspace{1cm} (18)

which means that $H(X_{n+1}) \geq H(X_n)$ and indeed entropy can never decrease (can only increase) as expected in statistical physics.
- On the other hand, suppose the stationary distribution $q$ is a non-uniform distribution, and in fact, $H(q)$ could be very small.
- Then, entropy of the Markov chain could decrease over time.
- Thus, the increase/decrease of entropy over time is entirely dependent on the transition distribution, $p(x_{n+1}|x_n)$.

Statistics $T$

- Let $X_1, X_2, \ldots, X_N$, $X_i \in \{0, 1\}$ be i.i.d. sequence of coin tosses, $p(X=H) = \theta = 1 - P(X=T)$.
- Let $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$ count the number of heads.
- $T$ is said to be a statistic of the sample.
- In general, a statistic is some function of a collection of random variables (e.g., an empirical mean, an empirical variance, or an empirical max of a sample, etc.).
- A statistic is itself a r.v. with a mean, variance, etc.
- Good statistics contain useful information about the sample, while bad statistics don’t (e.g., $T(X_1, \ldots, X_N) = X_1$).
- Statistics are sometimes called “features” in pattern recognition and machine learning.
Bernoulli Trials

- Now consider the above counting statistic in a probability.

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

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

(19)

- Once we know the statistic, probability of the sequence is expressible without referring back to \( \theta \).
- In other words, we have that \( X_1:N \perp \perp \theta | T(X_1:N) \).
- Markov chain (A): \( \theta \rightarrow T(X_1:N) \rightarrow X_1:N \)
- We also know that \( T(X_1:N) \) is a function of \( X_1:N \).
- (B) \( \theta \rightarrow X_1:N \rightarrow T(X_1:N) \).

DPI and Statistics

- (A): \( \theta \rightarrow T(X_1:N) \rightarrow X_1:N \)
- BY DPI, (A) \( \Rightarrow I(\theta; T(X_1:N)) \geq I(\theta; X_1:N) \)
- (B): \( \theta \rightarrow X_1:N \rightarrow T(X_1:N) \).
- BY DPI, (B) \( \Rightarrow I(\theta; X_1:N) \geq I(\theta; T(X_1:N)) \)
- Thus, (A) & (B) \( \Rightarrow I(\theta; X_1:N) = I(\theta; T(X_1:N)) \), and no information is lost about \( \theta \) in going from \( X_1:N \) to \( T(X_1:N) \).
- Such a statistic is called sufficient.

**Definition 6.1 (Sufficient Statistic)**

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

\[ X \perp \perp \theta | T(X) \].  

(20)

Alternatively, if the data processing inequality achieves equality.

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

- Traditional definition

**Definition 6.2 (Sufficient Statistic)**

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

\[ p(x_{1:N}|\theta) = g(T, \Theta)h(x_{1:N}) \]  \hfill (21)

- Compare with a definition of conditional independence.

**Definition 6.3 (Conditional Independence)**

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

\[ p(a, b, c) = g(a, c)h(b, c) \]  \hfill (22)

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

**Sufficiency of the “type” of the sample**

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

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

\[ P_{x_{1:N}} \triangleq \left( \frac{N(a_1|x_{1:N})}{N}, \frac{N(a_2|x_{1:N})}{N}, \ldots, \frac{N(a_D|x_{1:N})}{N} \right) \]  \hfill (23)

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

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

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

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

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

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

(24)

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

(25)

- So, $X_{1:N} \perp \theta | P_{x_{1:N}}$ and the type $P_{x_{1:N}}$ is sufficient.

Binary case, sufficiency of the “type”

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

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

(26)

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

$$p(\theta, T(x_{1:N}) = k) = \sum_{x_{1:N}} p(x_{1:N}, T(x_{1:N}) = k, \theta)$$

(27)

$$= \sum_{x_{1:N} : T(x_{1:N}) = k} p(x_{1:N}, T(x_{1:N}) = k, \theta)$$

(28)

$$= \left( \begin{array}{c} N \\ k \end{array} \right) p(0)^{N-k} p(1)^k$$

(29)
The joint probability

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

(30)

The marginal

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

(31)

So

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

(32)

which is the binary r.v. case of Equation 24.

**Minimal Sufficient Statistic**

**Definition 6.4**

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

\[ \theta \rightarrow T(X) \rightarrow U(X) \rightarrow X \]  

(33)

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

- Consider the following situation where we send $X$ through a noisy channel, receive $Y$, and do further processing.

  \[
  X \xrightarrow{\text{Noisy Channel}} Y \xrightarrow{\text{processing } g(\cdot)} \hat{X}
  \]

  $\hat{X}$ is an estimate of $X$.

- An error if $X \neq \hat{X}$. How do we measure the error? With probability, $P_e \triangleq p(X \neq \hat{X})$.

- Key insight: Intuitively, conditional entropy should tell us something about the error possibilities, i.e., how well we can do.

- Suppose $H(X|Y) = 0$. Would this mean high error is necessary? No.

- Suppose $H(X|Y)$ is large. Would this mean high error is necessary? Yes.

- Can we more formally relate $H(X|Y)$ and $P_e$?

Fano’s Inequality

**Theorem 7.1**

\[
H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y)
\]

- So $P_e = 0$ requires that $H(X|Y) = 0$!

- Note, the theorem simplifies (and implies)
  \[
  1 + P_e \log(|\mathcal{X}|) \geq H(X|Y), \quad \text{or}
  \]
  \[
P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}
\]

  yielding a lower bound on the error.

- This will be used to prove the converse to Shannon’s coding theorem, i.e., that any code with probability of error $\to 0$ as the blocklength increases must have a rate $R < C =$ the capacity of the channel (to be defined).
Fano Proof

Proof.

- First, define an error indicator variable $E$
  \[ E = \begin{cases} 
  1 & \text{if } \hat{X} \neq X \\
  0 & \text{if } \hat{X} = X 
\end{cases} \quad (36) \]

- Then use entropy chain rule twice
  \[ H(E, X | \hat{X}) = H(X | \hat{X}) + H(E | X, \hat{X}) \]
  \[ = H(E | \hat{X}) + H(X | E, \hat{X}) \]
  \[ \leq H(E = H(P_e)) + H(X | E, \hat{X}) \leq P_e \log |X| \quad (37) \]

- This last bound follows from
  \[ H(X | \hat{X}, E) = p(E = 0)H(X | \hat{X}, E = 0) + p(E = 1)H(X | \hat{X}, E = 1) \]
  \[ = (1 - P_e)0 + P_e H(X | \hat{X}, E = 1) \leq P_e \log |X - 1| \quad (38) \]

Fano Proof cont.

Proof.

- Note, if $E = 1$ we know that $X \neq \hat{X}$, so there are $X - 1$ options left, with maximum entropy $\log |X - 1|$.
- Using the above results, we get
  \[ H(P_e) + P_e \log |X - 1| \geq H(X | \hat{X}) \quad (41) \]

- Now, $X \rightarrow Y \rightarrow \hat{X}$ is a Markov chain and we can use the data processing inequality.
- Thus, we have that $I(X; Y) \geq I(X; \hat{X})$ or
  \[ H(X) - H(X | Y) \geq H(X) - H(X | \hat{X}) \]
  \[ \text{Thus, we get } H(X | \hat{X}) \geq H(X | Y) \text{ and} \]
  \[ H(P_e) + P_e \log |X - 1| \geq H(X | \hat{X}) \geq H(X | Y) \quad (42) \]
An interesting bound on probability of equality

**Lemma 7.2**

Let $X, X'$ be two independent r.v.s with $X \sim p(x)$ and $X' \sim r(x)$, with $x, x' \in \mathcal{X}$ (same alphabet). Then

$$p(X = X') \geq \max \left( 2^{-H(p) - D(p||r)}, 2^{-H(r) - D(r||p)} \right)$$  \hspace{1cm} (43)

**Proof.**

$$2^{-H(p) - D(p||r)} = 2^{-\sum_x p(x) \log p(x) + \sum_x p(x) \log \frac{r(x)}{p(x)}}$$  \hspace{1cm} (44)

$$= 2^{\sum_x p(x) \log r(x)}$$  \hspace{1cm} (45)

$$\leq \sum_x p(x) 2^{\log r(x)}$$  \hspace{1cm} (46)

$$= \sum_x p(x) r(x)$$  \hspace{1cm} (47)

$$= p(X = X')$$  \hspace{1cm} (48)

Thus, taking $p(x) = r(x)$, the probability that two i.i.d. random variables $X, X'$ are the same can be bounded as follows:

$$P(X = X') \geq 2^{-H(p)}$$  \hspace{1cm} (49)

Many other probabilistic quantities can be bounded in terms of entropic quantities as we will see throughout the course.