Entropy, Relative Entropy, and Mutual Information

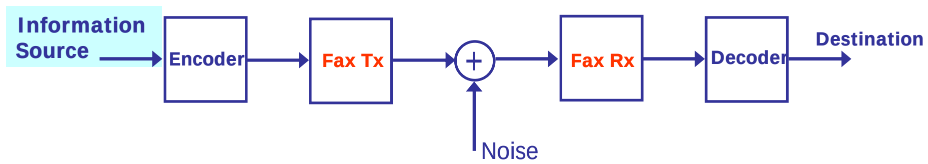
2009 2-2 Course

- Information Theory -

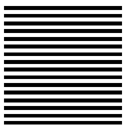
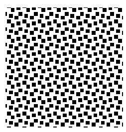
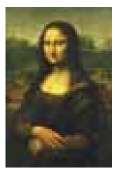
Tetsuo Asano and
Tad matsumoto
Email: {t-asano, matumoto}@jaist.ac.jp

Japan Advanced Institute of Science and Technology
Asahidai 1-1, Nomi, Ishikawa 923-1292, Japan
http://www.jaist.ac.jp

Preliminary Experiment



Send the following three pictures via fax to your friend:

Which picture required the longest time to send?
Why it required the longest time compared to the others?

Objectives of this Chapter

Observations:

(1) Encoder doesn't know the picture to be transmitted.
- It only "scans" the picture → Appearance of black and while (=pixels) is a random variable.

(2) The transmission time depends on the picture.
- The fax encoder analyzes "characteristic" of pictures, and use it when converting the "scanned" data.
- The shorter the transmission time, the better "suits" the picture to the encoding rule.

Objectives:

(1) Define information as a random variable,
(2) Define measures of information, uncertainty, closeness, and reduction of uncertainty, and
(3) Derive the relationships between the measures.

Outline

1. Information Measures
   - Entropy
   - Joint Entropy and Conditional Entropy
   - Kullback Leibler Distance (Relative Entropy)
2. Mutual Information
   - Chain Rules
3. Information Inequalities
   - Log Sum Inequality
4. Data Processing Inequality
5. Fano's Inequality

Entropy (1)

Definition 4.1.1: Entropy

The entropy $H(X)$ of a discrete random variable X is defined by:

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

where with the limit: $0 \log 0 = 0$

entropy $H(X)$ does NOT take negative values.

Note that the base of the logarithm is in many cases 2, with which entropy measure is measured in *bits*. However, it should not necessarily be always the case. If the base is *e*, the measure is *nats*.

Definition 4.1.2: Equivalent Description:

$$H(X) = -E_p[\log p(x)] = E_p\left[\log \frac{1}{p(x)}\right]$$

where E_p is the expectation with respect to the distribution p.

Entropy (2)

Property 4.1.1: Logarithm

Since $\log_b p(x) = \log_b a \log_a p(x)$,

$$H_b(X) = (\log_b a) H_a(X)$$

holds, where $H_y(X)=-p(X)\log_y p(X)$ with $y=a$ or b.

Example 4.1.1: Binary Random Variable

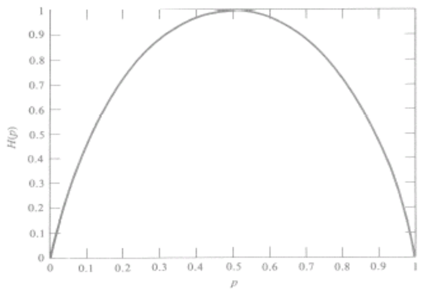
Let $X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$

Then, $H(X) = -p \log p - (1-p)\log(1-p)$

Sometimes, because of the definition above, the entropy is also denoted as $H(X) = H(\boldsymbol{p})$

Entropy (3)

The entropy $H(p)$ of the random variable in Example 4.1.1 is described as a function of p:



Observations:

(1) Entropy is 0 when p=0 or 1. This is reasonable, because the random variable is not random, and there is no uncertainty.

(2) Entropy is maximum when p=1/2. This is reasonable, because with p=1/2 the uncertainty is maximum.

Entropy (4)

Example 4.1.2:

Let

$$X = \begin{cases} a & \text{with probability } 1/2 \\ b & \text{with probability } 1/4 \\ c & \text{with probability } 1/8 \\ d & \text{with probability } 1/8 \end{cases}$$

Then, $H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4}$ *bits*

Exercise 4.1.1:

Consider a random variable which has a uniform distribution over 32 outcomes. To identify an outcome, we need a label that takes on 32 different values.

(1) How many labels are needed to uniquely identify the outcomes?
(2) Calculate the entropy of the random variable.
(3) Are the results consistent with each other?

Entropy (5)

Exercise 4.1.2: Horse Race

Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are given by

$$P_{horse} = \left[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right]$$

(1) How many labels are needed to uniquely identify the outcomes?
(2) Calculate the entropy of the random variable.
(3) Are the results consistent with each other? If not, how can we label the each horse to minimize the average length of the label?
(4) Assume that you don't know the result of the race, but someone else does. How many questions, in average, do you need to ask him in order to identify the winning horse? The answer has to be "yes-or-no".

Joint Entropy and Conditional Entropy (1)

Definition 4.1.3: Joint Entropy

The joint entropy $H(X, Y)$ of discrete random variables X and Y is defined by:

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

where E_p is the expectation with respect to the joint distribution p.

Note that $H(X,Y)$ does NOT take negative values.

Definition 4.1.4: Conditional Entropy

If discrete random variables X and Y follow the joint distribution $p(x, y)$, conditional entropy $H(Y|X)$ is defined as:

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

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

Joint Entropy and Conditional Entropy (2)

Theorem 4.1.1: Chain Rule

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

Proof:

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

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

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

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

Corollary 4.1.1: Chain Rule in Probability Domain

$$\log p(X,Y) = \log p(X) + \log p(Y|X)$$

Proof: Obvious from the conditional probability rule.

Joint Entropy and Conditional Entropy (3)

Corollary 4.1.2: $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$

Proof: Obvious from the chain rule.

Remark: $H(X|Y) \neq H(Y|X)$

However, $H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (= H(X,Y) : Mutual \quad Information)$

Example 4.1.3:

Let random variables $x, y \in \{1, \quad 2, \quad 3, \quad 4\}$ have the following joint distribution:

$$p(X=x, Y=y) = \begin{bmatrix} \frac{1}{8} & \frac{1}{16} & \frac{1}{32} & \frac{1}{32} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{32} & \frac{1}{32} \\ \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} \\ \frac{1}{4} & 0 & 0 & 0 \end{bmatrix}$$

(x: along rows →, y: down columns ↓)

JAIST

Joint Entropy and Conditional Entropy (4)

Example 4.1.3 (Continued):

Then, the marginal probability of X and Y are:

$$p(X=x)=\begin{bmatrix}\frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8}\end{bmatrix} \quad \text{and} \quad p(Y=y)=\begin{bmatrix}\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4}\end{bmatrix}$$

Hence, $H(X)=7/4$ *bits* and $H(Y)=2$ *bits*. The conditional entropy $H(X|Y)$ is:

$$H(X|Y)=\sum_{y=1}^{4} p(Y=y)H(X|Y=y)=\frac{11}{8} \quad \text{bits,}$$

and the conditional entropy $H(Y|X)$ is:

$$H(Y|X)=\sum_{x=1}^{4} p(X=x)H(Y|X=x)=\frac{13}{8} \quad \text{bits.}$$

The joint entropy $H(X,Y)$ is:

$$H(X,Y)=-\sum_{x=1}^{4}\sum_{y=1}^{4} p(X=x,Y=y)\log p(X=x,Y=y)=\frac{27}{8} \quad \text{bits.}$$

JAIST

Kullback Leibler Distance

Definition 4.1.5: Kullback Leibler Distance

The Kullback Leibler distance $D(p||q)$ between the two probability distribution functions $p(x)$ and $q(x)$ is defined as:

$$D(p\|q)=\sum_{x\in X} p(x)\log\frac{p(x)}{q(x)}=E_p\log\frac{p(X)}{q(X)}$$

with $0\log\frac{0}{q}=0 \quad , \quad p\log\frac{p}{0}=\infty$.

Property 4.1.2: Distance between Probability Distributions

(1) $D(p\|q)\neq D(q\|p) \quad in \quad general$

(2) $D(p||q)$ is non-negative, and is zero if and only if $p=q$ for *all* x.

(3) The Kullback Leibler distance is sometimes called *relative entropy*.

(4) If the probability distribution q, which is believed to be correct, is different from the true distribution p, we need

$$H(p)+D(p\|q)$$

bits on the average to describe the random variable following p.

Mutual Information (1)

Definition 4.2.1: Mutual Information

Consider two random variables X and Y with a joint probability distribution function $p(x,y)$ and the marginal distributions $p(x)$ and $p(y)$. The mutual Information $I(X,Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, i.e.,

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

$$= D\{p(x,y)\|p(x)p(y)\} = E_{p(x,y)}\left(\log\frac{p(X,Y)}{p(X)p(Y)}\right)$$

Theorem 4.2.1: Entropy and Mutual Information

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

Observation:

The mutual information $I(X,Y)$ is the reduction in the uncertainty of X by knowing Y.

Exercise: Give a proof for Theorem 4.2.1.

Mutual Information (2)

Corollary 4.2.1: $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y;X)$

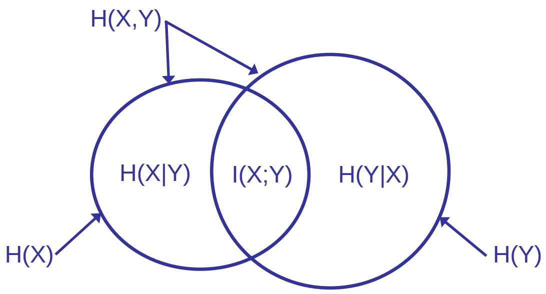
Proof: Obvious from the definition.

Corollary 4.2.2: $I(X;Y) = H(X) + H(Y) - H(X,Y)$

Proof: Obvious from the definition.

Corollary 4.2.3: $I(X;X) = H(X) - H(X|X) = H(X)$

Proof: Obvious from the definition.



Mutual Information (3)

Theorem 4.2.2: Extension of the Chain Rule

Let $X_1, X_2, \ldots, X_n$ be random variables drawn from the joint distribution $p(x_1, x_2, \ldots, x_n)$. Then,

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

Proof:

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

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

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

By continuing the process, we have:

$$H(X_1, X_2, \cdots, X_n) = H(X_1) + H(X_2 | X_1) + \cdots + H(X_n | X_{n-1}, \cdots, X_1)$$

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

Mutual Information (4)

Definition 4.2.2: Conditional Mutual Information

Conditional mutual information of random variable X and Y , given Z is defined by:

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

$$= E_{p(x,y,z)}\left(\log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)} \right)$$

Theorem 4.2.3: Chain Rule for Mutual Information

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \cdots, X_1)$$

Proof:

$$I(X_1, X_2, \cdots, X_n; Y) = H(X_1, X_2, \cdots, X_n) - H(X_1, X_2, \cdots, X_n | Y)$$

$$= \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1, Y)$$

$$= \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \cdots, X_{i-1})$$

Mutual Information (5)

Theorem 4.2.4: Non-Negativity of Mutual Information

$$I(X;Y) \geq 0$$

with equality if and only if X and Y are independent.

Proof: $I(X;Y) = D(p(x,y) \| p(x)p(y)) \geq 0$

with equality if and only if $p(x,y) = p(x)\,p(y)$, i.e., X and Y are independent.

Theorem 4.2.5: Non-Negativity of Conditional Mutual Information

$$I(X;Y|Z) \geq 0$$

with equality if and only if X and Y are conditionally independent given Z.

Proof: Obvious from the definition of the conditional mutual information.

Mutual Information (6)

Theorem 4.2.6: Entropy with Uniform Distribution

Let $|\mathcal{X}|$ denote the number of the elements in a set X. Then,

$$H(X) \leq \log|\mathcal{X}|$$

with equality if and only of x has a uniform distribution over X.

Proof: Let $u(x)=1/|\mathcal{X}|$ be the uniform probability distribution function over X, and let $p(x)$ be the probability distribution for X. Then,

$$D(p\|u) = \sum p(x)\log\frac{p(x)}{u(x)} = \log|\mathcal{X}| - H(X)$$

Hence, by the non-negativity of relative entropy,

$$0 \leq D(p\|u) = \log|\mathcal{X}| - H(X)$$

JAIST

Mutual Information (7)

Theorem 4.2.7: Conditioning reduces entropy

$$H(X|Y) \le H(X)$$

holds, with equality if and only if X and Y are independent.

Proof:

$$0 \le I(X;Y) = H(X) - H(X|Y)$$

Theorem 4.2.8: Independence Bound on entropy

Let $X_1, X_2, \ldots, X_n$ be random variables drawn from the joint distribution $p(x_1, x_2, \ldots, x_n)$. Then,

$$H(X_1, X_2, \cdots, X_n) \le \sum_{i=1}^{n} H(X_i)$$

Proof:

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

JAIST

Information Inequalities (1)

Theorem 4.3.1: Log Sum Inequality

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

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

with equality if and only if $\frac{a_i}{b_i} = constant \quad for \quad any \quad i$

Proof: The function $f(t)=t \log t$ is strictly convex, since $f''(t)=(1/t)\log e>0$ for all positive t. Therefore, by Jensen's inequality, we have

$$\sum \alpha_i f(t_i) \ge f\left(\sum \alpha_i t_i\right) \quad for \quad \alpha_i \ge 0, \quad \sum_i \alpha_i = 1$$

Setting $\alpha_i = b_i / \sum_{j=1}^{n} b_j \quad and \quad t_i = a_i / b_i$,

We have:

$$\sum \frac{a_i}{\sum_{j=1}^{n} b_j} \log \frac{a_i}{b_i} \ge \sum \frac{a_i}{\sum_{j=1}^{n} b_j} \log \sum \frac{a_i}{\sum_{j=1}^{n} b_j}$$

Information Inequalities (2)

Theorem 4.3.2: Convexity of Relative Entropy

$D(p||q)$ is convex in the pair (p,q), i.e., if (p_1,q_1), (p_2,q_2), are two pairs of the probability distribution of $x \in X$,

$$D\left(\lambda p_1 + (1-\lambda) p_2 \| \lambda q_1 + (1-\lambda) q_2\right) \le \lambda D\left(p_1 \| q_1\right) + (1-\lambda) D\left(p_2 \| q_2\right), \quad 0 \le \lambda \le 1$$

Proof: Applying the log sum inequality to a term on LHS,

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

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

Summing this over all $x \in X$, we obtain the equation shown in the theorem.

Definition 4.3.1: Concavity

A function f is concave, if $-f$ is convex.

Information Inequalities (3)

Theorem 4.3.3: Concavity of Entropy

$H(p)$ is a concave function of p.

Proof: $H(p) = \log|X| - D\left(p \| u\right)$

where u is the uniform distribution having cardinality $|X|$.
Then the concavity of $H(p)$ is obvious.

Theorem 4.3.4: Concavity of Mutual Information

Let X and Y be random variables having a joint distribution $p(x,y)= p(x)\ p(y|x)$.

Mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$, and a convex function of $p(y|x)$ for fixed $p(x)$.

Proof: $I(X,Y) = H(Y) - H(Y \mid X) = H(Y) - \sum_x p(x) H(Y \mid X = x)$

Information Inequalities (4)

Theorem 4.3.4 (Proof continued-1):

If $p(y|x)$ is fixed, then $H(Y|X)$ is fixed.

Also, because $p(y|x)$ is fixed, $p(y)$ is a linear function of $p(x)$.
Since $H(Y)$ is a concave function of $p(y)$, it is also a concave function of $p(x)$.

Now, let's consider two different conditional probability distribution $p_1(y|x)$ and $p_2(y|x)$, with which the corresponding joint probabilities are: $p_1(y,x)=p(x)\,p_1(y|x)$ and $p_2(y,x)=p(x)\,p_2(y|x)$, and their marginals being $p(x)$ and $p_1(y)$, and $p(x)$ and $p_2(y)$, respectively.

Consider conditional, conditional, and marginal distributions:

$$p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)\,p_2(y|x)$$

$$p_\lambda(y,x) = \lambda p_1(y,x) + (1-\lambda)\,p_2(y,x)$$

$$p_\lambda(y) = \lambda p_1(y) + (1-\lambda)\,p_2(y)$$

Information Inequalities (5)

Theorem 4.3.4 (Proof continued-2):

Define $q_\lambda(x,y) = p(x)\,p_\lambda(y)$

Obviously, $q_\lambda(x,y) = \lambda q_1(x,y) + (1-\lambda)q_2(x,y)$

Since the mutual information is Kullback Leibler distance between the joint distribution and the product of the marginals,

$$I(X;Y) = D\left(p_\lambda \| q_\lambda\right)$$

and since the Kullback Leibler distance (=relative entropy) is a convex function of (p, q), it follows that the mutual information is a convex function of the conditional distribution.

Data Processing Inequality (1)

Definition 4.4.1: X→Y→Z

Random variables *X*, *Y*, *Z* are said to form a Markov Chain in the order X→Y→Z, if the joint probability follows:

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

Property 4.4.1:

(1) X→Y→Z if and only if *X* and *Z* are conditionally independent of *Y given*, i.e.,

$$p(x,z|y)=\frac{p(x,y,z)}{p(y)}=\frac{p(x,y)p(z|y)}{p(y)}=p(x|y)p(z|y)$$

(2) X→Y→Z implies Z→Y→X.

(3) If *Z=f(Y)*, then X→Y→Z.

Data Processing Inequality (2)

Theorem 4.4.1: Data Processing Inequality

If X→Y→Z, then $I(X;Y)\ge I(X;Z)$

Proof: Since $I(X;Y,Z)=I(X;Z)+I(X;Y|Z)=I(X;Y)+I(X;Z|Y)$

However, since *X* and *Z* are conditionally independent of *Y* given,
$I(X;Z|Y)=0$

Since $I(X;Y|Z)\ge 0$, we have: $I(X;Y)\ge I(X;Z)$.

Corollary 4.4.1:

$I(Y;Z)\ge I(X;Z)$ Proof: Obvious from the definition.

Corollary 4.4.2:

If X→Y→Z=*f*(Y), then $I(X;Y)\ge I(X;f(Y))$ Proof: Obvious from the definition.

Function does not increase information!!!

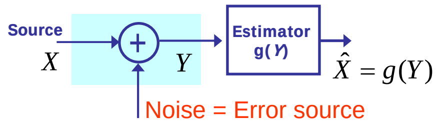
Corollary 4.4.3:

If X→Y→Z, then $I(X;Y|Z)\le I(X;Y)$ Proof: Obvious from the definition.

Observation reduces the dependence of the random variables!!!

Fano's Inequality (1)

Consider a very simple communication system described below:



From the received signal *Y*, the estimator *g*(*Y*) estimates the transmitted source *X*.
Faon's inequality specifies the bound of the probability of error: $P_e = \Pr(\hat{X} \neq X)$

Theorem 4.5.1: Fano's Inequality

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

Proof: Define variable corresponding to the error event:

$$E = \begin{cases} 1 & if \quad \hat{X} \neq X \\ 0 & if \quad \hat{X} = X \end{cases}$$

Using the chain rule for entropies, we can expand *H*(*E*,*X*|*Y*) in the following two ways:

Fano's Inequality (2)

Theorem 4.5.1 (Proof continued-1):

$$H(E, X|Y) = H(X|Y) + \underbrace{H(E|X,Y)}_{=0} = \underbrace{H(E|Y)}_{\leq H(P_e)} + \underbrace{H(X|E,Y)}_{\leq P_e \log(|\mathcal{X}|-1)}$$

(a) (b) (c)

(a) Since *E* is a function of *X* and *g*(*Y*), if the receiver knows *X*, *Y* and *g*(*Y*) as conditions of *E*, there is no uncertainty on *E*, and hence this term is zero.
(b) Since conditioning reduces entropy, $H(E|Y) \leq H(E) = H(P_e)$
(c) This term is bounded as:

$$H(X|E,Y) = Prob(E=0)\underbrace{H(X|Y,E=0)}_{=0} + Prob(E=1)\underbrace{H(X|Y,E=1)}_{\leq \log(|\mathcal{X}|-1)}$$

(d) (e)

(d) Since given *E=0, X=g*(*Y*) is known. Therefore, there is no uncertainty.
(e) Given *E=1,* we can upper-bound the conditional entropy by the logarithm of the remaining outcomes |*X*|-1.

Combining these results, we obtain Fano's inequality.

Fano's Inequality (3)

Corollary 4.5.1: Wider Sense of Fano's Inequality

$$1 + P_e \log(|\mathcal{X}|) \geq H(X|Y) \quad \text{or} \quad P_e \geq \frac{H(X|Y) - 1}{\log|\mathcal{X}|}$$

Proof: In the proof of Theorem 4.1.16, (b) and (e) can further be upper-bounded by

$$H(X|Y) \leq \underbrace{H(E|Y)}_{\leq H(P_e) \leq 1} + \underbrace{H(X|E,Y)}_{\leq P_e \log(|\mathcal{X}|-1)}$$

(b') (c)

$$H(X|E,Y) \leq P_e \log(|\mathcal{X}|-1) \leq P_e \log(|\mathcal{X}|)$$

(e) (e')

respectively.

Summary

We have made a journey on the topics:

1. Information Measures
   - Entropy
   - Joint Entropy and Conditional Entropy
   - Kullback Leibler Distance (Relative Entropy)
2. Mutual Information
   - Chain Rules
3. Information Inequalities
   - Log Sum Inequality
4. Data Processing Inequality
5. Fano's Inequality