Information Theory **Prof. I-Hsiang Wang**

Lecture Notes: Measurement of Information

September 16, 2015 **Scribe: Chi-Ning Chou**

1 Motivation & Intuition

Information theory is a theory concern about the quantification of information. So, in the very beginning, we have to find some ways to describe the quantities we care about. And before we define the mathematical objects such as entropy, mutual information and some related theorem such as data processing theorem, now let's think about what exactly do we want to describe in a communication system?

1.1 Single source

First, let's think about a single source, what do we want to quantify for such single source?

Since we model a source as a **random variable**, it has a corresponding support and a probability distribution over it. So what we want to capture might be

- How many possible outcomes come from this source?
- What's the probability for each outcome to appear?

Of course, we want more. For example, we hope the quantification has some good properties so that we can perform meaningful operations over it. And soon we will see how Shannon use **entropy** to capture all this ideas.

1.2 Two sources

Clearly that we not only consider a single source, we also care about the relation between two different sources. But what do we care?

- We might care about the **lost** between two source. For example, consider the transmission terminal X and the reception terminal Y. If there's some information missed in the middle, it's important to us. However, note that we still only care about **how much** information we lose, instead of what information we lose.

1.3 Network

In a whole communication network, i.e. source-channel scheme. What we care is whether there are some informations drops in the middle. Also, different from two sources scenario, the terminal here has some **sequential relation**. That is, the middle terminal somehow contains all the information from the other sides. Consider a simple example: we send message from terminal X to Z through Y. Since the message must go through Y, the amount of information we receive at Z about X cannot be greater than that of Y. Actually, this is the so called **data processing theorem**, which will be covered later.

2 The quatification fo information: Entropy

2.1 Intuitions and definition

We think of information as the **resolution of uncertainties**. Intuitively, we can divide this concept into two parts:

1. Number of outcomes.
2. Likeliness of each outcome.

Formally, here we want to find a information function f maps a source(random variable, etc.) to a information quantity. With the above observations, we also want this function f to obey some properties:

Intuition (entropy)

1. (**log function (Hartley information)**) The multiplication in number of outcomes transforms to an addition in the information quantity in a counting theory sense.
2. (**Expectation**) Average the information quantities in an expectation sense w.r.t different outcomes.

With such approach, we can further define **axioms** for information function. Moreover, we can find it's unique! (Or relax into Renyi entropy)

Thus for a random process X we can define its entropy to quantify the information it has as follow:

Definition 1 (entropy)
The entropy of a random process X over finite alphabet set $\mathcal{X}$ with probability measure P is

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

2.2 Well-defined of entropy function

Now we have the definition of entropy for random variable and it's clearly that once the support size if finite, the entropy function is well-defined. However, what if the size of support is countably infinite? Or, what if the source and channel are working on a continuous space?

For now, we first consider the **finite support** case for convenience.

2.3 Basic properties

Here, let's find out the possible value of the entropy of a random variable X that lies on finite support. Immediately, we have

- $H(X) \geq 0$
- $H(X) \leq |\mathcal{X}|$

Moreover, the minimizer and maximizer are

- $H(X) = 0 \Leftrightarrow X$ is deterministic
- $H(X) = |\mathcal{X}| \Leftrightarrow X$ is uniform on $\mathcal{X}$

To prove the upper bound, we should delegate the calculation to a dummy variable: $U := \frac{1}{p(X)}$ and utilize the **concavity** of log function then apply **Jenson's inequality**.

2.4 Variants of entropy definition

Only with the entropy definition for random variable is not quite enough. In real life, we usually consider a sequence of correlated events. Thus, it motivates us to define the entropy for a sequence of random variables and the entropy for a random variable conditioned on another.

Definition 2 (joint entropy)
For a sequence (vector) of random variables $X_1, ..., X_n$ over supports $\mathcal{X}_1, ..., \mathcal{X}_n$, the joint entropy of them is

$$H(X_1, ..., X_n) := \sum_{x_1 \in \mathcal{X}_1, ..., x_n \in \mathcal{X}_n} p(X_1 = x_1, ..., X_n = x_n) \log \frac{1}{p(X_1 = x_1, ..., X_n = x_n)}$$

Note that, this just average out the Hartley information of all possible events. That is, we can see that the entropy is only quantitatively related to the probability distribution over the possible events. Entropy does not care what's the really object.

Definition 3 *conditional entropy The conditional entropy for a random variable X conditioned another random variable Y with value y is*

$$H(X|Y = y) := \sum_{x \in \mathcal{X}} p(X = x|Y = y) \log \frac{1}{p(X = x|Y = y)}$$

And the conditioned entropy for X to condition on Y is

$$H(X|Y) = \mathbb{E}_Y[H(X|Y = y)] = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(X = x, Y = y) \log \frac{1}{p(X = x|Y = y)}$$

After defining joint entropy and conditional entropy, we must want to derive some properties about them. And the following properties are all about the relation between the entropy that before and after joint or condition.

Properties of joint entropy

Theorem 4 (joint reduces entropy) *The joint entropy of random variable $X_1, ..., X_n$ is no larger than the sum of marginal entropy.*

$$H(X_1, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Note that the equality holds when $X_1, ..., X_n$ are independent.

Intuition (joint reduces entropy)

This theorem is consistent to our intuition that joint event will disclose some correlated outcomes.

Properties of conditional entropy

Theorem 5 (chain rule) *Suppose there are random variables $X_1, .., X_n$, then*

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

Intuition (chain rule)

Intuitively, we can think of the RHS as measuring $X_1, .., X_n$ **altogether** and think of LHS as **sequentially** measure $X_1, X_2, ..., X_n$. And the theorem tells us that there's no difference to measure the random sources altogether or measuring sequentially. Note that, we can even arbitrarily choose the order!

Theorem 6 (conditioning reduces entropy) *Let X, Y be two random variables, then*

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

Note that the equality happens when X and Y are independent.

Intuition (conditioning reduces entropy)

This one is quite intuitive as we think in the converse way: the amount of information (uncertainty) about X cannot increase after knowing some other things.

2.5 High probability set

The following theorem helps us get more intuition on the definition of entropy.

Theorem 7 (Cardinality of high probability set)
Let random process X over finite alphabet set $\mathcal{X}$ with probability measure P. $\forall \epsilon < 1$, we call a set A $\epsilon-$significant if $P(A) > 1-\epsilon$. Denote the smallest cardinality of a $\epsilon-$significant set containing length n elements as

$$s(n, \epsilon) := \min_{A: P(A) > 1-\epsilon} |A|$$

Then, we have

$$\lim_{n\to\infty} \frac{1}{n} \log s(n, \epsilon) = H(X)$$

With the this theorem, we can see that $H(X)$ is actually the **compression ratio** or **saving** of random process X. It gives us an intuition that when the length of message n grows close to infinity (in an asymptotic sense), there are only $nH(X)$ of message will frequently show up. Or, we can say that the probability for other strings to appear will converge to zero. (i.e. measure zero) Moreover, Prof. Wang says that the strings in the significant set will gradually have the same probability to show up. ($2^{-nH(X)}$) And this is just what we want in the source coding: uniformity. In some sense this result shows the possibility to **remove redundancy** and compress the data. The property mentioned above is called AEP (Asymptotic Equipartition Property).

2.6 Conclusion

Conclusion

All we need to bear in mind about entropy is as follow:

- **Intuition**: expectation of Hartley information.
- **Range**: $[0, \log|\mathcal{X}|]$, for finite support.
- **Minimizer & Maximizer**: deterministic and uniform distribution.
- **Concavity**: Jenson's inequality.
- **Conditioning reduces entropy**.
- **Sequentially measuring = Single measurement**: chain rule.
- **Best compression rate**: cardinality of high probability set, maximal savings.

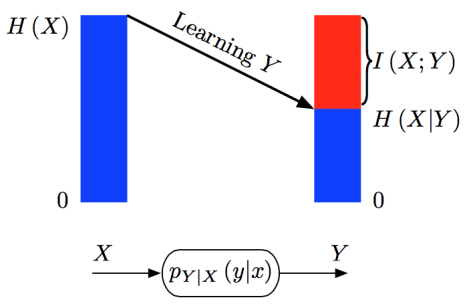
3 The difference of information: Mutual information

After quantifying a the amount of information contained in single source, now we want to compare the information amount among a pair of sources (X, Y). Intuitively, what we want to capture is the **difference** between X and Y. And directly from the definition of entropy, we can describe such difference as

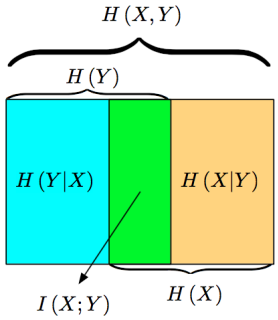
$$H(X) - H(X|Y)$$

, which is the amount of information that knowing X but not knowing Y.

With this definition, we can have the following two intuitions about mutual information: the amount of inference and the level of dependency, while the first one can be easily recognized with the definition, the second one we be discussed later. For now, please remember these two high-level ideas.



(a) Amount of inference.



(b) The level of dependency.

Figure 1: The intuitions of mutual information.

Intuition (mutual information)

- Amount of **inference**: For example, in channel coding scheme, we want to estimate X with only knowledge of Y, Thus, we want to know how much information about X we can get from learning Y. See Figure 1a.
- The level of **dependency**: If $I(X;Y)$ is large, we can somehow think of they are highly correlated since $I(X;Y)$ is maximized as $X = Y$. On the contrary, if $I(X;Y)$, it means that X and Y are independent. As we analyze the formation of $I(X;Y)$

$$I(X;Y) = \mathbb{E}_{U,V}[\log \frac{1}{P_X(X)P_Y(Y)} - \log \frac{1}{P_{X,Y}(X,Y)}]$$

  We can see that $I(X;Y)$ actually measures how far X and Y are independent! See Figure 1b.

3.1 Basic properties

Now, we're going to derive some basic properties about mutual informations. Almost the same as entropy, will consider the range of mutual information, the maximizer, and the minimizer.

- $0 \leq I(X;Y) \leq \min(H(X), H(Y))$
- The maximizer is the situation that Y is a **deterministic** function of X and vice versa.
- The minimizer is the situation that X and Y are **independent**.

Other than these basic properties, there're also one thing worth mentioning: the identity (symmetry) of mutual information:

Theorem 8 (symmetry of mutual information)
Suppose X and Y are two random sources, then

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

Figure 1b gives an graphical intuition. And the second equality is by the chain rule of entropy. ($H(X,Y) = H(Y) + H(X|Y)$)

3.2 Conditional mutual information

Here, we consider the direct definition of mutual information:

Definition 9 (conditional mutual information) *Suppose X, Y, and Z are three random sources. Then the mutual information among X and Y conditioned on Z is defined as*

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

Intuitively, we can think of such Z or conditional variable as **a priori**. That is, the given assumption or configuration of the system. Sometimes it will be more convenient to work in such context.

Before we give another great intuition about conditional mutual information, let's first define the notion of Markov chain for our usage:

Definition 10 (Markov chain) *Suppose $X_1, X_2, ..., X_n$ are random sources. We say $X_1 - X_2 - ... - X_n$ is a Markov chain if*

$$Pr(X_1, ..., X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2)...P(X_n|X_{n-1})$$

That is, the joint probability is the same as sequentially measure the conditional probability over the previous source on this chain.

With the definition of Markov chain, we can immediately yield the following corollary about conditional mutual information:

Corollary 11 *Suppose X, Y, and Z are three random sources. If $I(X;Y|Z) = 0$, then $X - Z - Y$ forms a Markov chain.*

Intuitively, as Y learns nothing **more** with a priori Z, the uncertainty involves X, Y, and Z would not have the term $X|Y,Z$ or $Y|X,Z$ since they are independent with the appearance of Z.

Intuition (Markov chain)

Suppose $X_1 - X_2 - ... - X_n$ forms a Markov chain, it gives us the following meanings:

- (**Firewall**) For any $1 \leq i < j < k \leq n$, the information that X_i provides to X_k is all possessed by X_j.
- (**No addition information**) For and $1 \leq i < j < k < l \leq n$, The information amount of X_i provides to X_l is less than the information amount of X_j gives to X_k.

Last but not least, let's see the chain rule of mutual information

Theorem 12 (chain rule of mutual information) *Suppose X, $X_1, X_2, ..., X_n$ are random sources, then*

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

Similar to the chain rule of entropy, the theorem gives us an intuition that the total amount of information will be the same as we measure all together (LHS) or measure sequentially (RHS).

3.3 Data processing theorem

In the previous subsection, we mentioned the intuition about Markov chain and regard this property as blocking or filtering the information from previous sources. Now, we state it as a formal theorem.

Theorem 13 (data processing theorem) *Suppose X, Y, Z are three random sources, and $X - Y - Z$ forms a Markov chain. Then,*

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

Actually, $I(X;Y,Z) = I(X;Y)$.

The reason why data processing theorem sis important is that it provides a quantitative way to examine the communication system. See Figure 2. Here W is the source terminal and $\hat{W}$ is the receiving terminal while X and Y are middle message before and after the noise channel respectively. By checking the conditional probability, we can see that $W - X - Y - \hat{W}$ forms a Markov chain. As a result, data processing theorem can provide us a quantitative inequality in the form of mutual information which can help us derive th fundamental limit of coding rate and some impossibility results.

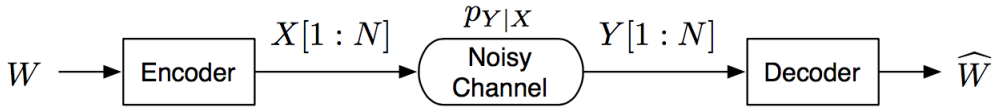


Figure 2: Data processing theorem and communication system.

3.4 Concavity and Convexity of mutual information

The following theorem tells us the concavity and convexity of mutual information, which is useful in computing channel capacity and rate distortion function.

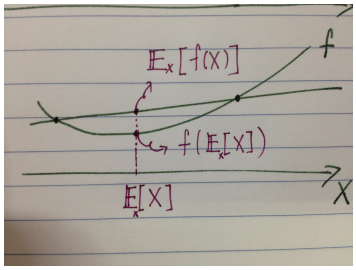
Theorem 14 (concavity and convexity of mutual information) *Suppose X and Y are two random sources and $p_{X,Y} = p_{X|Y}p_Y$, then*

- *If $p_{X|Y}$ is fixed, then $I(X;Y)$ is a* **concave** *function of p_X.*
- *If p_X is fixed, then $I(X;Y)$ is a* **convex** *function of $p_{X|Y}$.*

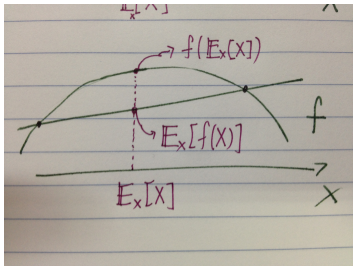
Intuition (concavity and convexity of mutual information)

The intuition is simple. Consider the following:

- $H(X)$ is a **concave** function of p_X.
- $H(X|Y)$ is a **convex** function of $p_{X|Y}$.



(a) Convex function.



(b) Concave function

Figure 3: Graphical intuition for Jenson's inequality.

3.5 Conclusion

Conclusion

- Two intuitions of mutual information:
  - The amount of inference.
  - The level of dependency.
- As $I(X;Y|Z) = 0$, $X - Z - Y$ forms a Markov chain.
- Conditioning **does not always** reduces the mutual information.
- Chain rule: Measure altogether = sequentially measuring.
- Data processing theorem: the latter terminal contains all the information of the previous ones.
- Concavity and convexity of mutual information:
  - If $p_{X|Y}$ is fixed, then $I(X;Y)$ is a **concave** function of p_X.
  - If p_X is fixed, then $I(X;Y)$ is a **convex** function of $p_{X|Y}$.

4 KL-Divergence

4.1 Definition and Intuitions

The definition of KL-divergence, also known as relative entropy, is as follow:

Definition 15 (KL-divergence) *Let p, q be two p.m.f.'s of a random variable X, then the KL-divergence between p and q is*

$$D_{KL}(p||q) := \mathbf{E}_p(\log \frac{p(X)}{q(X)})$$

Note that we should take care of the situation that $\log \frac{p(X)}{q(X)} = \log \frac{\cdot}{0}$.

Intuitively, we can reformulate the KL-divergence as

$$D_{KL}(p||q) = \mathbf{E}_p[\log \frac{1}{q(X)} - \log \frac{1}{p(X)}]$$

, which can be seen as the difference of measuring the information quantity of q instead of that of p over the measure p. Somehow, it's the information loss as we using thought of using p but actually facing q.

Another intuition is to consider the mutual information:

$$I(X;Y) = D_{KL}(p_{X,Y}||p_X p_Y)$$

, which is the information gain from X and Y are independent or not.

4.2 Properties

KL-Divergence is not a metric or even a semimetric!

Definition 16 (metric, quasimetric, semimetric, premetric, pseudometric) *Let $d : X \times X \to \mathbf{R}$ be a function. Consider the following properties, $\forall x, y, z \in X$*

1. *(***non-negativity***)* $d(x, y) \geq 0$
2. *(***coincidence axiom***)* $d(x, y) = 0 \Leftrightarrow x = y$
3. *(***symmetry***)* $d(x, y) = d(y, x)$
4. *(***triangle inequality***)* $d(x, y) + d(y, z) \geq d(x, z)$

Now, we say d is a

- **metric** *if all of the conditions are hold.*
- **quasimetric** *if we drop the symmetry condition.*
- **semimetric** *if we drop the triangle inequality.*
- **divergence (premetric)** *if we drop both symmetry and triangle inequality.*
- **pseudometric** *if we allow $x \neq y$ such that $d(x, y) = 0$.*

And with some careful verification, we can show that KL-divergence only satisfies the first two conditions. That is, KL-divergence is only a **divergence (premetric)**! For completeness, here we simply derive the non-negativity of KL-divergence:

$$D_{KL}(p||q) = \mathbf{E}_p[-\log \frac{q(X)}{p(X)}] \geq -\log(\mathbf{E}_p[\frac{q(X)}{p(X)}]) = -\log \sum_{x \in \mathcal{X}} q(x) = 0$$

Also, KL-divergence has some convex property as follow:

Theorem 17 (convexity of KL-divergence) *$D_{KL}(p||q)$ is convex w.r.t. pair (p, q).*

The proof will be discussed in later chapter.

4.3 Operational meaning of KL-divergence

KL-divergence has a special operational meaning in hypothesis testing in the form of Chernoff-Stein lemma. Here, we first state the definition of a hypothesis as follow:

Suppose $X^n = (X_1, X_2, ..., X_n)$, where $X_i \sim p_X \forall i$. Now we want to decide whether $pX = p_0$ or $p_X = p_1$, which refer to null hypothesis H_0 and alternative hypothesis H_a. And what we care about is two kind of error:

- (**False alarm**): $P_F^{(n)} := \Pr[H_0|p_X = p_1]$
- (**Miss detection**): $P_M^{(n)} := Pr[H_1|p_X = p_0]$

And our philosophy here is to control the false alarm probability (say $P_F^{(n)} \leq \alpha$), and minimize the miss detection probability $P_M^{(n)}$ since we have different favor in two different kinds of errors. And what Chernoff-Stein lemma tells us is that the minimal miss detection probability of any control level α will converge to $2^{-nD_{KL}(p_0||p_1)}$. the following state the lemma formally:

Lemma 18 (Chernoff-Stein lemma) *Define* $\beta^{(n)} := \min_{P_F^{(n)} \leq \alpha} P_M^{(n)}$, *then*

$$\lim_{n\to\infty} \frac{1}{n} \log \beta^{(n)} = -D_{KL}(p_0||p_1)$$

4.4 Conclusion

Conclusion

- KL-divergence describes the amount of wrong information using the distribution p to measure another distribution q.

$$D_{KL}(p||q) = \mathbf{E}_p[\log \frac{1}{q(X)} - \log \frac{1}{p(X)}]$$

- KL-divergence relates the mutual information as the difference between joint distribution and the multiplicative of marginal distribution.

$$I(X;Y) = D(p_{X,Y}||p_X p_Y)$$

- KL-divergence does not satisfy the symmetry and triangle inequality condition.
- $D_{KL}(p||q)$ is convex wr.r.t. (p, q).
- Chernoff-Stein lemma

$$\lim_{n\to\infty} \frac{1}{n} \log \beta^{(n)} = -D_{KL}(p_0||p_1)$$

A Inequalities

A.1 Log sum inequality

Theorem 19 (log sum inequality) *Suppose* $a_1, ..., a_n, b_1, ..., b_n \geq 0$, *then*

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

Note that here a_i and b_i only need to be **positive**. Intuitively, the inequality first divide each term with the sum then use the non-negativity of KL-divergence to show the result.

A.2 Information inequality

Information inequality provides an upper and a lower bound for log function.

Theorem 20 (information inequality) *For any base $b > 0$ and any $\xi > 0$,*

$$(1 - \frac{1}{\xi}) \log_b e \leq \log_b \xi \leq (\xi - 1) \log_b e$$

Intuitively, information inequality provides a **linear** upper bound and an **one minus reciprocal** lower bound.

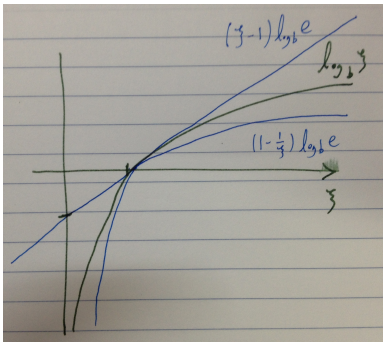


Figure 4: Information inequality.