INFORMATION THEORY & CODING

Week 3 : Inequalities

Dr. Rui Wang

Department of Electrical and Electronic Engineering

Southern Univ. of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

September 22, 2020

Review Summary

- **Definitions:**

$$
\begin{aligned}
H(X) &= E_p \log \frac{1}{p(X)} \\
H(X,Y) &= E_p \log \frac{1}{p(X,Y)} \\
H(X|Y) &= E_p \log \frac{1}{p(X|Y)} \\
I(X;Y) &= E_p \log \frac{p(X,Y)}{p(X)p(Y)} \\
D(p\|q) &= E_p \log \frac{p(X)}{q(X)} \\
I(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y).
\end{aligned}
$$

Review Summary

- **Chain rules:**

  Entropy:

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

  Mutual information:

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

  Relative entropy:

  $D\big(p(x,y) \| q(x,y)\big) = D\big(p(x) \| q(x)\big) + D\big(p(y|x) \| q(y|x)\big).$

Jensen's Inequality

Definition (Convexity)

A function $f(x)$ is said to be *convex* over an interval (a, b) if $\forall x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

A function f is called *strictly convex* if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition (Concavity)

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

A function is convex if it always lies below any chord. A function is concave if it always lies above any chord.

Jensen's Inequality

Definition (Convexity)

A function $f(x)$ is said to be *convex* over an interval (a, b) if for every $x_1 x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

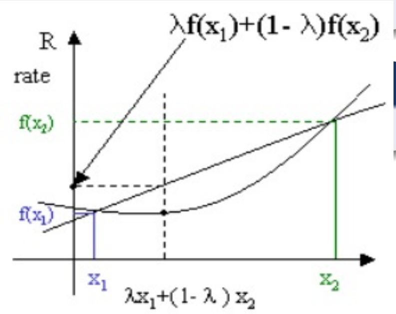
$$f\left(\lambda x_1 + (1-\lambda)x_2\right) \leq \lambda f(x_1) + (1-\lambda)f(x_2).$$

A function f is called *strictly convex* if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition (Concavity)

A function f is *concave* if $-f$ is co

A function is convex if it always lie
concave if it always lies above any

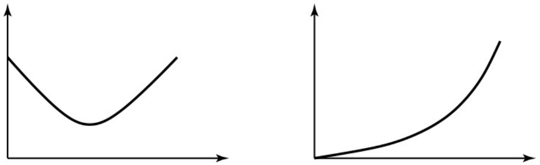


Jensen's Inequality

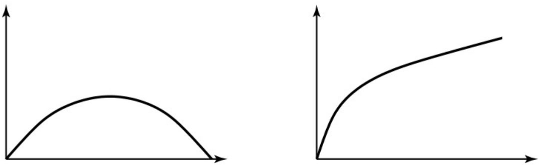
Example

$$f(x) = x^2, \quad |x|, \quad e^x, \quad x \log x \quad (x > 0)$$

$$g(x) = \log x, \quad \sqrt{x}, \quad (x \geq 0)$$



(a)



(b)

Jensen's Inequality

Theorem 2.6.2 (Jensen's Inequality)

If f is a convex function and X is a random variable,

$$E[f(X)] \geq f(E[X]).$$

Moreover, if f is strictly convex, $E[f(X)] = f(E[X])$ implies that $X = E[X]$ with probability 1 (i.e., X is a constant).

Proof.

By mathematical induction.

- $k = 2$:
  $p(x_1)f(x_1) + p(x_2)f(x_2) \geq f\left(p(x_1)x_1 + p(x_2)x_2\right)$.
- Hypothesis: $\sum_{i=1}^{k-1} p(x_i)f(x_i) \geq f\left(\sum_{i=1}^{k-1} p(x_i)x_i\right)$.
- Induction: $\sum_{i=1}^{k} p(x_i)f(x_i)$.

□

Information Inequality

Theorem 2.6.3 (*Information Inequality*)

Let $p(x)$, $q(x)$, $x \in X$, be two probability mass functions. Then

$$D(p\|q) \geq 0$$

with equality iff $p(x) = q(x)$ for all x.

Proof.

Let $A = \{x : p(x) > 0\}$ be the support set of $p(x)$. Then

$$\begin{aligned}
-D(p\|q) &= -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \quad \text{(Jensen's Inequality)} \\
&= \log \sum_{x \in A} q(x) \\
&\leq \log \sum_{x \in \mathcal{X}} q(x) = 0
\end{aligned}$$

Corollaries

Corollary (*Nonnegativity of mutual information*)

For any two random variables, X, Y,

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

with equality iff X and Y are independent.

Corollary

$$D\big(p(y|x)\|q(y|x)\big) \geq 0,$$

with equality iff $p(y|x) = q(y|x)$ for all y and x such that $p(x) > 0$.

Corollary

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

with equality iff X and Y are conditionally independent given Z.

The maximum entropy distribution

Theorem 2.6.4

$H(X) \leq \log|\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of X, with equality iff X has a uniform distribution over $|\mathcal{X}|$.

Proof.

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

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

□

Conditioning reduces entropy

Theorem 2.6.5 (*Conditioning reduces entropy*)

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

with equality *iff* *X and Y are independent.*

Theorem 2.6.6 (*Independence bound on entropy*)

Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$, then

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

with equality *iff* *the X_i's are independent.*

Data-processing inequality

Definition (*Markov Chain*)

Random variables X, Y, Z are said to *form a Markov chain* in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y and Z form a Markov chain $X \to Y \to Z$ if the join probability mass function can be written as

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

- $X \to Y \to Z \Rightarrow p(x, z|y) = p(x|y)p(z|y)$
- $X \to Y \to Z \Rightarrow Z \to Y \to X$
- If $Z = f(Y)$, then $X \to Y \to Z$.

Data-processing inequality

Theorem 2.8.1 (*Data-processing inequality*)

If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

Proof.

By the chain rule, we expand $I(X;Y,Z)$ in two ways:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y). \end{aligned}$$

Since $X \to Y \to Z$, we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$, we have $I(X;Y) \geq I(X;Z)$. □

Corollaries

Corollary

In particular, if $Z = g(Y)$, we have $I(X;Y) \geq I(X;g(Y))$.

Corollary

If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$.

Fano's inequality

Problem 2.5 (*Zero conditional entropy*)

Show that if $H(X|Y) = 0$, then X is a function of Y, i.e., for all y with $p(y) > 0$, there is only one possible value of x with $p(x, y) > 0$.

Proof.

Assume that there exists an y, say y_0 and two different values of x, say x_1 and x_2 such that $p(y_0, x_1) > 0$ and $p(y_0, x_2) > 0$. Then $p(y_0) \geq p(y_0, x_1) + p(y_0, x_2) > 0$, and $p(x_1|y_0)$ and $p(x_2|y_0)$ are not equal to 0 or 1. Thus,

$$\begin{aligned} H(X|Y) &= -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \\ &\geq p(y_0) \left(-p(x_1|y_0) \log p(x_1|y_0) - p(x_2|y_0) \log p(x_2|y_0)\right) \\ &> 0 \end{aligned}$$

since $-t \log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for $t \neq 0, 1$, which is a contradiction to $H(X|Y) = 0$. □

Fano's inequality

- The conditional entropy of a random variable X given another random variable Y is zero ($H(X|Y) = 0$) iff X is a function of Y. Hence we can estimate X from Y with zero probability of error iff $H(X|Y) = 0$.
- We can estimate X with a low probability of error P_e only if the conditional entropy $H(X|Y)$ is small. *Fano's inequality* quantifies this idea.

Why do we need to related P_e to entropy $H(X|Y)$? When we have a communication system, we send X, but receive a corrupted version Y. We want to infer X from Y. Our estimate is $\hat{X}$ and we will make a mistake as

$$P_e = \Pr[\hat{X} \neq X]$$

Markov chain $X \to Y \to \hat{X}$.

Fano's inequality

Problem

A random variable Y is related to another random variable X with a distribution $p(x)$. From Y, we calculate a function $g(Y) = \hat{X}$, where $\hat{X}$ is an estimate of X and takes on values in $\hat{\mathcal{X}}$. We observe that $X \to Y \to \hat{X}$ forms a Markov chain. How to bound the estimate error probability $P_e = \Pr[\hat{X} \neq X]$?

Fano's inequality

Theorem 2.11.1

For any estimator $\hat{X}$ *such that* $X \to Y \to \hat{X}$ *, with* $P_e = \Pr\{X \neq \hat{X}\}$, *we have*

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

This inequality can be weakened to

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

or

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

Fano's inequality

Proof.

Define an error random variable as

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

Using the chain rule for entropies to expand $H(E, X|\hat{X})$ in two different ways, we have

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

Since conditioning reduces entropy, $H(E|\hat{X}) \leq H(E) = H(P_e)$. Since E is a function of X and $\hat{X}$, the conditional entropy $H(E|X, \hat{X})$ is equal to 0. We now look at $H(X|E, \hat{X})$. By the equation $H(X|Y) = \sum_y p(y)H(X|Y = y)$, we have

$$H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{\Pr[\hat{X} = \hat{x}, E = 0]H(X|\hat{X} = \hat{x}, E = 0) + \Pr[\hat{X} = \hat{x}, E = 1]H(X|\hat{X} = \hat{x}, E = 1)\}.$$

□

Fano's inequality

Proof.

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

$$H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{\Pr[\hat{X} = \hat{x}, E = 0] H(X|\hat{X} = \hat{x}, E = 0)$$
$$+ \Pr[\hat{X} = \hat{x}, E = 1] H(X|\hat{X} = \hat{x}, E = 1)\}.$$

By definition of E, X is conditionally deterministic given $\hat{X} = \hat{x}$ and $E = 0$, then $H(X|\hat{X} = \hat{x}; E = 0) = 0$. If $\hat{X} = \hat{x}$ and $E = 1$, then X must take a value in the set $\{x \in \mathcal{X} : x \neq x\hat{x}\}$ which contains $|\mathcal{X}| - 1$ elements. Then $H(X|\hat{X} = \hat{x}, E = 1) \leq \log(|\mathcal{X}| - 1)$.

$$\begin{aligned} H(X|E, \hat{X}) &\leq \sum_{\hat{x} \in \mathcal{X}} \Pr[\hat{X} = \hat{x}, E = 1] \log(|\mathcal{X}| - 1) \\ &= \Pr[E = 1] \log(|\mathcal{X}| - 1) \\ &= P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

□

Fano's inequality

Proof.

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

$$H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{\Pr[\hat{X} = \hat{x}, E = 0] H(X|\hat{X} = \hat{x}, E = 0)$$
$$+ \Pr[\hat{X} = \hat{x}, E = 1] H(X|\hat{X} = \hat{x}, E = 1)\}.$$

$$\begin{aligned} H(X|E, \hat{X}) &\leq \sum_{\hat{x} \in \mathcal{X}} \Pr[\hat{X} = \hat{x}, E = 1] \log(|\mathcal{X}| - 1) \\ &= \Pr[E = 1] \log(|\mathcal{X}| - 1) \\ &= P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

By the data-processing inequality, we have $I(X; \hat{X}) \leq I(X; Y)$ and therefore $H(X|\hat{X}) \geq H(X|Y)$. □

Corollary

Corollary

For any two random variables X and Y, let $p = Pr(X \neq Y)$.

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

Proof.

Let $\hat{X} = Y$ in Fano's inequality. □

Fano's inequality

Remark

Suppose that there is no knowledge of Y. Thus, X must be guessed without any information. Let $X \in \{1, 2, \ldots, m\}$ and $p_1 \geq p_2 \geq \cdots \geq p_m$. Then the best guess of X is $\hat{X} = 1$ and the resulting probability of error is $P_e = 1 - p_1$. Fano's inquality becomes

$$H(P_e) + P_e \log(m-1) \geq H(X).$$

The probability mass function

$$(p_1, p_2, \cdots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \cdots, \frac{P_e}{m-1}\right)$$

achieves this bound with equality.

Applications of Fano's inequality

- Prove converse in many theorems (including channel capacity)

- Compressed sensing signal model

$$y = Ax + w$$

where $A \in \mathcal{R}^{M \times d}$: projection matrix for dimension reduction. Signal x is sparse. Want to estimate x from y.

Reading & Homework

Reading : Whole Chapter 2

Homework : Problems 2.13, 2.15, 2.32, 2.35