An introduction to Information Theory

Adrish Banerjee

Department of Electrical Engineering
Indian Institute of Technology Kanpur
Kanpur, Uttar Pradesh
India

July 18, 2016

Lecture #2A: Information Inequalities

Outline of the lecture

- Jensen's inequality

Jensen's inequality Log sum inequality Data Processing Lemma Fano's Lemma

Outline of the lecture

- Jensen's inequality
- Log sum inequality

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Outline of the lecture

- Jensen's inequality
- Log sum inequality
- Data processing Lemma

Jensen's inequality Log sum inequality Data Processing Lemma Fano's Lemma

Outline of the lecture

- Jensen's inequality
- Log sum inequality
- Data processing Lemma
- Fano Lemma

Outline

Jensen's inequality

- A function f is concave on a nonzero length interval I, if for each point $x_0 \in I$, there exists a real number c (may depend on x_0) such that

$$f(x) \leq f(x_0) + c(x - x_0),$$

for all $x \in I$.

Jensen's inequality

- A function f is concave on a nonzero length interval I, if for each point $x_0 \in I$, there exists a real number c (may depend on x_0) such that
$$f(x) \leq f(x_0) + c(x - x_0),$$
for all $x \in I$.
- Jensen's inequality: If f is concave, and X is a random variable taking values in I, then
$$E[f(X)] \leq f(E[X])$$

Jensen's inequality

- Function f is concave on I if and only if for every x_1 and x_2 in I.
$$\lambda f(x_1) + (1 - \lambda) f(x_2) \leq f[\lambda x_1 + (1 - \lambda) x_2]$$
for $0 < \lambda < 1$.

Jensen's inequality

- Function f is concave on I if and only if for every x_1 and x_2 in I.

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

  for $0 < \lambda < 1$.
- Define a discrete random variable X that takes values x_1 and x_2 with probabilities λ and $1-\lambda$ respectively.

Jensen's inequality

- Function f is concave on I if and only if for every x_1 and x_2 in I.

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

  for $0 < \lambda < 1$.
- Define a discrete random variable X that takes values x_1 and x_2 with probabilities λ and $1-\lambda$ respectively.
- Using Jensen's inequality, we get

$$E[f(X)] = \lambda f(x_1) + (1-\lambda) f(x_2) \leq f(E[X]) = f[\lambda x_1 + (1-\lambda) x_2]$$

Convex Function

- If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval

Jensen's inequality Log sum inequality Data Processing Lemma Fano's Lemma

Convex Function

- If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval
- Proof:

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Convex Function

- If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval
- Proof:
- We use the Taylor series expansion of the function around x_0:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \qquad (1)$$

where x^* lies between x_0 and x.

Jensen's inequality Log sum inequality Data Processing Lemma Fano's Lemma

Convex Function

- If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval
- Proof:
- We use the Taylor series expansion of the function around x_0:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \qquad (1)$$

where x^* lies between x_0 and x.

- By hypothesis, $f''(x^*) \geq 0$, and thus the last term is non negative for all x.

Convex Function

- If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval
- Proof:
- We use the Taylor series expansion of the function around x_0:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \tag{1}$$

  where x^* lies between x_0 and x.
- By hypothesis, $f''(x^*) \geq 0$, and thus the last term is non negative for all x.
- We let $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and take $x = x_1$, to obtain

$$f(x_1) \geq f(x_0) + f'(x_0)((1 - \lambda)(x_1 - x_2)) \tag{2}$$

Convex Function

- Similarly, taking $x = x_2$, we obtain

$$f(x_2) \geq f(x_0) + f'(x_0)(\lambda(x_2 - x_1)) \tag{3}$$

Convex Function

- Similarly, taking $x = x_2$, we obtain

$$f(x_2) \geq f(x_0) + f'(x_0)(\lambda(x_2 - x_1)) \tag{3}$$

- Multiplying (2) by λ and (3) by $1 - \lambda$ and adding, we obtain

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \tag{4}$$

Jensen's inequality

- Use Jensen's inequality to find the appropriate inequalities between

Jensen's inequality

- Use Jensen's inequality to find the appropriate inequalities between
  (i) $E[e^{-ax}]$ and $e^{-aE[X]}$, $a > 0$.

Jensen's inequality Log sum inequality Data Processing Lemma Fano's Lemma

Jensen's inequality

- Use Jensen's inequality to find the appropriate inequalities between
  (i) $E[e^{-ax}]$ and $e^{-aE[X]}$, $a > 0$.
  (ii) $E[\sqrt{X}]$ and $\sqrt{E[X]}$

Jensen's inequality

- Use Jensen's inequality to find the appropriate inequalities between
  (i) $E[e^{-ax}]$ and $e^{-aE[X]}$, $a > 0$.
  (ii) $E[\sqrt{X}]$ and $\sqrt{E[X]}$

(i)

$$\begin{aligned} f(x) &= e^{-ax} \\ f'(x) &= -ae^{-ax} \\ f''(x) &= a^2 e^{-ax} > 0, \end{aligned}$$

Thus the function $f(x) = e^{-ax}$ is strictly convex for $a > 0$. Thus Jensen's inequality implies that $E[e^{-ax}] \geq e^{-aE[X]}$, $a > 0$.

Jensen's inequality

(ii)

$$\begin{aligned} f(x) &= \sqrt{x} \\ f'(x) &= \frac{1}{2}x^{-1/2} \\ f''(x) &= \frac{1}{2} \cdot \frac{(-1)}{2}x^{-3/2} < 0, \ \forall x > 0 \end{aligned}$$

Thus the function $f(x) = \sqrt{x}$ is strictly concave for $x > 0$. Thus Jensen's inequality implies that $E[\sqrt{X}] \leq \sqrt{E[X]}$.

Outline

Log sum inequality

- For any nonnegative numbers $a_1, a_2, \cdots$ and $b_1, b_2, \cdots$ such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$.,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i}$$

Log sum inequality

- For any nonnegative numbers $a_1, a_2, \cdots$ and $b_1, b_2, \cdots$ such that $\sum_i a_i < \infty$ and $0 < \sum_i b_i < \infty$.,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i}$$

- Let $a_i' = a_i / \sum_j a_j$ and $b_i' = b_i / \sum_j b_j$. Using divergence inequality

$$\begin{aligned} 0 &\leq \sum_i a_i' \log \frac{a_i'}{b_i'} \\ &= \sum_i \frac{a_i}{\sum_j a_j} \log \frac{a_i / \sum_j a_j}{b_i / \sum_j b_j} \\ &= \frac{1}{\sum_j a_j} \left[\sum_i a_i \log \frac{a_i}{b_i} - \left(\sum_i a_i \right) \log \frac{\sum_j a_j}{\sum_j b_j} \right] \end{aligned}$$

Properties of relative entropy

- $D(p||q)$ is convex in the pair (p, q), i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

$$D((\lambda p_1 + (1-\lambda)p_2)||(\lambda q_1 + (1-\lambda)q_2)) \leq \lambda D(p_1||q_1) + (1-\lambda) D(p_2||q_2)$$

for all $0 \leq \lambda \leq 1$

Properties of relative entropy

- $D(p||q)$ is convex in the pair (p, q), i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

  $$D((\lambda p_1+(1-\lambda)p_2)||(\lambda q_1+(1-\lambda)q_2)) \leq \lambda D(p_1||q_1)+(1-\lambda)D(p_2||q_2)$$

  for all $0 \leq \lambda \leq 1$
- Proof: Applying log sum inequality, we get

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

Properties of relative entropy

- $D(p||q)$ is convex in the pair (p, q), i.e., if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

  $$D((\lambda p_1+(1-\lambda)p_2)||(\lambda q_1+(1-\lambda)q_2)) \leq \lambda D(p_1||q_1)+(1-\lambda)D(p_2||q_2)$$

  for all $0 \leq \lambda \leq 1$
- Proof: Applying log sum inequality, we get

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

- Summing over x , we get the desired inequality.

Properties of entropy

- $H(p)$ is concave in p.

Properties of entropy

- $H(p)$ is concave in p.
- Proof:

$$H(p) = \log L - D(p||u)$$

where u is the uniform distribution on L outcomes.

Properties of entropy

- $H(p)$ is concave in p.
- Proof:

$$H(p) = \log L - D(p||u)$$

  where u is the uniform distribution on L outcomes.
- The concavity of H follows from the convexity of D.

Outline

Data Processing Lemma

- When X, Y, Z is a Markov chain, then

$$\begin{aligned} I(X;Z) &\leq I(X;Y) \\ I(X;Z) &\leq I(Y;Z) \end{aligned}$$

Data Processing Lemma

- *Proof:* Since $X \to Y \to Z$,

$$\begin{aligned} p(z/xy) &= p(z/y) \\ H(Z|XY) &= H(Z|Y) \end{aligned}$$

Data Processing Lemma

- *Proof:* Since $X \to Y \to Z$,

$$\begin{aligned} p(z/xy) &= p(z/y) \\ H(Z|XY) &= H(Z|Y) \end{aligned}$$

- Hence

$$\begin{aligned} I(Y;Z) &= H(Z) - H(Z|Y) \\ &= H(Z) - H(Z|XY) \\ &\geq H(Z) - H(Z|X) \\ &= I(X;Z) \end{aligned}$$

Data Processing Lemma

- We need to show Z, Y, X (the reverse of the sequence X, Y, Z) also form a Markov chain.

$$\begin{aligned} H(XYZ) &= H(Y) + H(X|Y) + H(Z|XY) \\ &= H(Y) + H(X|Y) + H(Z|Y) \end{aligned}$$

Data Processing Lemma

- We need to show Z, Y, X (the reverse of the sequence X, Y, Z) also form a Markov chain.

$$\begin{aligned} H(XYZ) &= H(Y) + H(X|Y) + H(Z|XY) \\ &= H(Y) + H(X|Y) + H(Z|Y) \end{aligned}$$

- $H(XYZ)$ can also be written as

$$H(XYZ) = H(Y) + H(Z|Y) + H(X|ZY)$$

Data Processing Lemma

- We need to show Z, Y, X (the reverse of the sequence X, Y, Z) also form a Markov chain.

$$\begin{aligned} H(XYZ) &= H(Y) + H(X|Y) + H(Z|XY) \\ &= H(Y) + H(X|Y) + H(Z|Y) \end{aligned}$$

- $H(XYZ)$ can also be written as

$$H(XYZ) = H(Y) + H(Z|Y) + H(X|ZY)$$

- Comparing the equations we get

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

which is equivalent to

$$p(x/yz) = p(x/y)$$

Outline

Fano's Lemma

- Let $\hat{U}$ be the estimate of the random variable U. We define the probability of error

$$P_e = P(\hat{U} \neq U)$$

Fano's Lemma

- Let $\hat{U}$ be the estimate of the random variable U. We define the probability of error
$$P_e = P(\hat{U} \neq U)$$
- Let U and $\hat{U}$ are L-ary random variables and the error probability P_e is defined as above, then
$$H(P_e) + P_e \log_2(L-1) \geq H(U|\hat{U})$$
where uncertainty $H(U|\hat{U})$ is in bits.

Jensen's inequality Log sum inequality Data Processing Lemma Fano's Lemma

Fano's Lemma

- Let $\hat{U}$ be the estimate of the random variable U. We define the probability of error
$$P_e = P(\hat{U} \neq U)$$
- Let U and $\hat{U}$ are L-ary random variables and the error probability P_e is defined as above, then
$$H(P_e) + P_e \log_2(L-1) \geq H(U|\hat{U})$$
where uncertainty $H(U|\hat{U})$ is in bits.
- The equality holds if and only if the probability of error given $\hat{U} = u$ is the same for all u, and when there is an error, the $L-1$ erroneous values of U are always equally likely.

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Fano's Lemma

- Define a random variable Z as an indicator random variable for an error

$$Z = \begin{cases} 0 & \text{when } \hat{U} = U \\ 1 & \text{when } \hat{U} \neq U \end{cases}$$

Jensen's inequality Log sum inequality Data Processing Lemma Fano's Lemma

Fano's Lemma

- Define a random variable Z as an indicator random variable for an error

$$Z = \begin{cases} 0 & \text{when } \hat{U} = U \\ 1 & \text{when } \hat{U} \neq U \end{cases}$$

- This implies

$$H(Z) = H(P_e)$$

Adrish Banerjee Department of Electrical Engineering Indian Institute of Technology Kanpur Kanpur, Uttar Pradesh India
An introduction to Information Theory

Fano's Lemma

- Define a random variable Z as an indicator random variable for an error

$$Z = \begin{cases} 0 & \text{when } \hat{U} = U \\ 1 & \text{when } \hat{U} \neq U \end{cases}$$

- This implies

$$H(Z) = H(P_e)$$

- We note that

$$\begin{aligned} H(UZ|\hat{U}) &= H(U|\hat{U}) + H(Z|U\hat{U}) \\ &= H(U|\hat{U}) \end{aligned}$$

Fano's Lemma

- Define a random variable Z as an indicator random variable for an error

$$Z = \begin{cases} 0 & \text{when } \hat{U} = U \\ 1 & \text{when } \hat{U} \neq U \end{cases}$$

- This implies

$$H(Z) = H(P_e)$$

- We note that

$$\begin{aligned} H(UZ|\hat{U}) &= H(U|\hat{U}) + H(Z|U\hat{U}) \\ &= H(U|\hat{U}) \end{aligned}$$

- Thus

$$\begin{aligned} H(U|\hat{U}) &= H(UZ|\hat{U}) \\ &= H(Z|\hat{U}) + H(U|\hat{U}Z) \\ &\leq H(Z) + H(U|\hat{U}Z) \end{aligned}$$

Fano's Lemma

- Since

$$H(U|\hat{U}, Z = 0) = 0$$

and

$$H(U|\hat{U}, Z = 1) \leq \log_2(L - 1)$$

Fano's Lemma

- Since

$$H(U|\hat{U}, Z = 0) = 0$$

and

$$H(U|\hat{U}, Z = 1) \leq \log_2(L - 1)$$

- Thus

$$\begin{aligned} H(U|\hat{U}Z) &\leq P(Z = 1)\log_2(L - 1) \\ &= P_e \log_2(L - 1) \end{aligned}$$

Fano's Lemma

- Since
$$H(U|\hat{U}, Z = 0) = 0$$
and
$$H(U|\hat{U}, Z = 1) \leq \log_2(L - 1)$$
- Thus
$$\begin{aligned} H(U|\hat{U}Z) &\leq P(Z = 1)\log_2(L - 1) \\ &= P_e \log_2(L - 1) \end{aligned}$$
- Hence
$$H(P_e) + P_e \log_2(L - 1) \geq H(U|\hat{U})$$