LECTURE 2

Convexity and related notions

1 Handout: PS 1

Last time:

- Introduction
- Review of probability
- Entropy, joint entropy, conditional entropy
- Chain rule of entropy

Lecture outline

- Mutual Information.
- Convexity and concavity
- Jensen's inequality
- Positivity of mutual information
- Data processing theorem
- Fano's inequality

Reading: Scts. 2.3, 2.6-2.8, 2.11.

Quick Review

- Entropy

$$H(X) = -\sum_x P_X(x) \log P_X(x)$$

- $H(X) \geq 0$
- Uniform distribution, let $|\mathcal{X}| = n$

$$H(X) = \log n$$

- Chain Rule

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

- X, Y independent:

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

Question: $H(Y|X) = H(X|Y)$?

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

or equivalently

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

Definition: Mutual Information

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x,y) \log \left(\frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)} \right) \end{aligned}$$

The **average** amount of knowledge about X that one obtains by observing the value of Y.

Mutual Information and Communication Channels

Question what is $I(X;X)$?

Question If X and Y are independent, what is $I(X;Y)$?

Chain Rule for Mutual Information

Definition: Conditional Mutual Information

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

Chain Rule:

$$\begin{aligned} & I(X_1, X_2; Y) \\ = \; & H(X_1, X_2) - H(X_1, X_2|Y) \\ = \; & H(X_1) + H(X_2|X_1) - H(X_1|Y) - H(X_2|Y, X_1) \\ = \; & I(X_1; Y) + I(X_2; Y|X_1) \end{aligned}$$

By induction

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

Relative entropy

Relative entropy is a measure of the distance between two distributions, also known as the Kullback Leibler distance between PMFs $P_X(x)$ and $Q_X(x)$.

Definition:

$$D(P_X||Q_X) = \sum_{x\in\mathcal{X}} P_X(x)\log\left(\frac{P_X(x)}{Q_X(x)}\right)$$

in effect we are considering the log to be a r.v. of which we take the mean (note that we assume $0\log(\frac{0}{p}) = 0$ and $p\log(\frac{p}{0}) = \infty$

- Mutual information can be written as

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \\ &= D(P_{XY}||P_X P_Y) \end{aligned}$$

- Entropy written as relative entropy:

Let X take values in $\mathcal{X}$ with $|\mathcal{X}| = n$.

$$\begin{aligned} H(X) &= -\sum_{x\in\mathcal{X}} P_X(x) \log P_X(x) \\ &= -\sum_{x\in\mathcal{X}} P_X(x) \log \frac{P_X(x)}{1/n} + \log n \\ &= H(U) - D(P_X||P_U) \end{aligned}$$

where U is uniformly distributed over $\mathcal{X}$.

Convexity

Definition: a function $f(x)$ is convex over (a, b) iff $\forall x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$

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

and is strictly convex iff equality holds iff $\lambda = 0$ or $\lambda = 1$.

f is concave iff $-f$ is convex.

Convenient test: if f has a second derivative that is non-negative (positive) everywhere , then f is convex (strictly convex)

Jensen's inequality

if f is a convex function and X is a r.v., then

$$E_X[f(X)] \geq f(E_X[X])$$

if f is strictly convex, then $E_X[f(X)] = f(E_X[X]) \Rightarrow X = E[X]$.

Proof:

For two mass point distribution $P_X(x_i) = p_i, i = 1, 2$,

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2)$$

Induction.

Example:

$$\frac{1}{3}\log a + \frac{2}{3}\log b \qquad \log\left[\frac{a+2b}{3}\right]$$

Information Inequality

Theorem

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

, with equality if and only if $p(x) = q(x), \forall x$.

Proof:

$$\begin{aligned} -D(p||q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log \sum_x p(x) \frac{q(x)}{p(x)} \\ &= 0 \end{aligned}$$

Equality occurs only when $q(x) \propto p(x)$, which means $p = q$.

Tons of Good Stuff

Corollary 1

Uniform distribution is the most random. $H(X) \leq \log |\mathcal{X}|$.

since

$$H(X) = \log |\mathcal{X}| - D(P_X || P_U)$$

Corollary 2

Mutual Information is non-negative, $I(X;Y) \geq 0$.

since

$$I(X;Y) = D(P_{XY} || P_X P_Y)$$

Corollary 2.1
Conditioning reduces entropy, $H(X) \geq H(X|Y)$,

since

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

Question $H(Y) \geq H(Y|X = x)$??

Corollary 2.2
Independence bound

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

since

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

Concavity of entropy

Theorem:
Entropy $H(X)$ is concave in P_X. If X_1, X_2 are r.v.s defined on $\mathcal{X}$, with distribution $P_1(x), P_2(x)$, respectively. For any $\theta \in [0,1]$, consider a r.v. X with

$$P_X(x) = \theta P_1(x) + (1-\theta)P_2(x), \forall x$$

then

$$H(X) \geq \theta H(X_1) + (1-\theta)H(X_2)$$

Proof:
Let Z be binary r.v., with $P(Z=0)=\theta$. Let $X = X_1$ if $Z=0$, and $X = X_2$ if $Z=1$. All independent. Then

$$\begin{aligned} H(X) &\geq H(X|Z) \\ &= \theta H(X|Z=0) + (1-\theta)H(X|Z=1) \\ &= \theta H(X_1) + (1-\theta)H(X_2) \end{aligned}$$

Example The entropy of a binary r.v. is maximized by uniform distribution.

Mutual information and input distribution

Theorem For a fixed transition probabilities $P_{Y|X}$, $I(X;Y)$ is a concave function of P_X.

Proof Construct X_1, X_2, X, Z as in the previous proof. Consider

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

Condition on X, Y and Z are independent, $I(Y;Z|X) = 0$. Thus

$$\begin{aligned} I(X;Y) &\geq I(X;Y|Z) \\ &= \theta I(X;Y|Z=0) + (1-\theta) I(X;Y|Z=1) \\ &= \theta I(X_1;Y) + (1-\theta) I(X_2;Y) \end{aligned}$$

Mutual information and transition probability

Theorem For a fixed input distribution P_X, $I(X;Y)$ is convex in $P_{Y|X}$.

Proof Consider a random variable X, and two channels with $P_1(y|x)$ and $P_2(y|x)$. When feed with X, the outputs of the two channels are denoted as Y_1, Y_2.

Now let one channel be chosen randomly according to a binary r.v. Z that is independent of X, and denote the output as Y.

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

where $I(X;Z) = 0$. Thus

$$\begin{aligned} I(X;Y) &\leq I(X;Y|Z) \\ &= \theta I(X;Y_1) + (1-\theta) I(X;Y_2) \end{aligned}$$

Summary

- Entropy $H(p)$ is a **concave** function of p.
- Mutual information $I(X;Y)$ is a **concave** function of P_X for fixed $P_{Y|X}$.
- $I(X;Y)$ is a **convex** function of $P_{Y|X}$ for fixed P_X.

Markov chain

Markov chain:

random variables X, Y, Z form a Markov chain in that order $X \to Y \to Z$ if the joint PMF can be written as

$$P_{X,Y,Z}(x, y, z) = P_X(x)P_{Y|X}(y|x)P_{Z|Y}(z|y).$$

Markov chain

Consequences:

- $X \to Y \to Z$ iff X and Z are conditionally independent given Y

$$
\begin{aligned}
& P_{X,Z|Y}(x,z|y) \\
= \; & \frac{P_{X,Y,Z}(x,y,z)}{P_Y(y)} \\
= \; & \frac{P_{X,Y}(x,y)}{P_Y(y)} P_{Z|Y}(z|y) \\
= \; & P_{X|Y}(x|y) P_{Z|Y}(z|y)
\end{aligned}
$$

  so Markov implies conditional independence and vice versa

- $X \to Y \to Z \Leftrightarrow Z \to Y \to X$ (see above LHS and last RHS)

Data Processing Theorem

If $X \rightarrow Y \rightarrow Z$ then $I(X;Y) \geq I(X;Z)$

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

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

X and Z are conditionally independent given Y, so $I(X;Z|Y) = 0$

hence $I(X;Z) + I(X;Y|Z) = I(X;Y)$ so $I(X;Y) \geq I(X;Z)$ with equality iff $I(X;Y|Z) = 0$

note: $X \rightarrow Z \rightarrow Y \Leftrightarrow I(X;Y|Z) = 0$ Y depends on X only through Z

Consequence: you cannot "undo" degradation

Fano's lemma

Suppose we have r.v.s X and Y, Fano's lemma bounds the error we expect when estimating X from Y

We generate an estimator of X that is $\widehat{X} = g(Y)$.

Probability of error $P_e = Pr(\widehat{X} \neq X)$

Indicator function for error $\mathbf{E}$ which is 1 when $X = \widehat{X}$ and 0 otherwise. Thus, $P_e = P(\mathbf{E} = 0)$

Fano's lemma:

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

Proof of Fano's lemma

$$
\begin{aligned}
& H(\mathbf{E}, X|Y) \\
= \; & H(X|Y) + H(\mathbf{E}|X, Y) \\
= \; & H(X|Y)
\end{aligned}
$$

$$
\begin{aligned}
& H(\mathbf{E}, X|Y) \\
= \; & H(\mathbf{E}|Y) + H(X|\mathbf{E}, Y)
\end{aligned}
$$

$$H(\mathbf{E}|Y) \leq H(\mathbf{E})$$

$$
\begin{aligned}
& H(X|\mathbf{E}, Y) \\
= \; & P_e H(X|\mathbf{E} = O, Y) + (1 - P_e) H(X|\mathbf{E} = 1, Y) \\
= \; & P_e H(X|\mathbf{E} = O, Y) \\
\leq \; & P_e H(X|\mathbf{E} = O) \\
\leq \; & P_e \log(|\mathcal{X}| - 1)
\end{aligned}
$$