LECTURE 2

Convexity and related notions

Last time:

- Goals and mechanics of the class
- notation
- entropy: definitions and properties
- mutual information: definitions and properties

Lecture outline

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

Reading: Scts. 2.6-2.8, 2.11.

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 $E_X[f(X)] = f(E_X[X]) \Rightarrow X = E[X]$.

Concavity of entropy

Let $f(x) = -x\log(x)$ then

$$\begin{aligned} f'(x) &= -x\log(e)\frac{1}{x} - \log(x) \\ &= -\log(x) - \log(e) \end{aligned}$$

and

$$f''(x) = -\log(e)\frac{1}{x} < 0$$

for $x > 0$.

$H(X) = \sum_{x\in|\mathcal{X}|} f(P_X(x))$

thus the entropy of X is concave in the value of $P_X(x)$ for every x.

Thus, consider two random variables, X_1 and X_2 with common $\mathcal{X}$. Then the random variable X defined over the same $\mathcal{X}$ such that $P_X(x) = \lambda P_{X_1}(x) + (1-\lambda)P_{X_2}(x)$ satisfies:

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

Maximum entropy

Consider any random variable X_1^1 on $\mathcal{X}$. For simplicity, consider $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ (we just want to use the elements of $\mathcal{X}$ as indices). Now consider X_2^1 a random variable such $P_{X_2^1}(x) = P_{X_1^1}(shift(x))$ where $shift$ denotes the cyclic shift on $(1, \ldots, \mathcal{X})$. Clearly $H(X_1^1) = H(X_2^1)$. Moreover, consider X_1^2 defined over the same $\mathcal{X}$ such that $P_{X_1^2}(x) = \lambda P_{X_1^1}(x) + (1-\lambda) P_{X_2^1}(x)$ then $H(X_1^2) \geq H(X_1^1)$.

We can show recursively with the obvious extension of notation that

$$H(X_1^n) \geq H(X_1^m)$$

$\forall n > m \geq 1$. Now $lim_{n\to\infty} P_{X_1^n}(x) = \frac{1}{|\mathcal{X}|}$ $\forall x \in \mathcal{X}$. Hence, the uniform distribution maximizes entropy and $H(X) \leq \log(|\mathcal{X}|)$.

Conditioning reduces entropy

$$\begin{aligned} & H(Y|X) = E_Z[H(Y|X = Z)] \\ = & -\sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) log_2[P_{Y|X}(y|x)] \end{aligned}$$

$P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x)$ hence by concavity $H(Y|X) \leq H(Y)$.

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

Independence bound:

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

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

Mutual information and transition probability

Let us call $P_{Y|X}(y|x)$ the transition probability from X to Y. Consider a r.v. Z that takes values 0 and 1 with probability θ and $1-\theta$ and s.t.

$$P_{Y|X,Z}(y|x,0) = P'_{Y|X}(y|x)$$

$$P_{Y|X,Z}(y|x,1) = P''_{Y|X}(y|x)$$

and Z is independent from X

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

and

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

hence

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

so

$$\theta I(X;Y|Z=0) + (1-\theta) I(X;Y|Z=1) \geq I(X;Y)$$

For a fixed input assignment, $I(X;Y)$ is convex in the transition probabilities

Mutual information and input probability

Consider a r.v. Z such that $P_{X|Z}(x|0) = P'(x)$, $P_{X|Z}(x|1) = P''(x)$, Z takes values 0 and 1 with probability θ and $1-\theta$ and Z and Y are conditionally independent, given X

$$I(Y;Z|X) = 0$$

and

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

so

$$I(X;Y|Z) \leq I(X;Y).$$

Mutual information is a concave function of the input probabilities.

Exercise: jamming game in which we try to maximize mutual information and jammer attempts to reduce it. What will the policies be?

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

Consequence: you cannot "undo" degradation

Consequence: Second Law of Thermodynamics

The conditional entropy $H(X_n|X_0)$ is non-decreasing as n increases for a stationary Markov process $X_0, \ldots, X_n$

Look at the Markov chain $X_0 \rightarrow X_{n-1} \rightarrow X_n$

DPT says

$I(X_0; X_{n-1}) \geq I(X_0; X_n)$

$H(X_{n-1}) - H(X_{n-1}|X_0) \geq H(X_n) - H(X_n|X_0)$

so $H(X_{n-1}|X_0) \leq H(X_n|X_0)$

Note: we still have that $H(X_n|X_0) \leq H(X_n)$.

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

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

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

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

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