COMP2610 / COMP6261 - Information Theory

Lecture 8: Some Fundamental Inequalities

Robert C. Williamson

Research School of Computer Science

Australian National University

14 August 2018

Last time

- Decomposability of entropy

- Relative entropy (KL divergence)

- Mutual information

Review

Relative entropy (KL divergence):

$$D_{\mathrm{KL}}(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$$

Mutual information:

$$\begin{aligned} I(X;Y) &= D_{\mathrm{KL}}(p(X,Y)\|p(X)p(Y)) \\ &= H(X)+H(Y)-H(X,Y) \\ &= H(X)-H(X|Y). \end{aligned}$$

- Average reduction in uncertainty in X when Y is known
- $I(X;Y) = 0$ when X, Y statistically independent

Conditional mutual information of X, Y given Z:

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

This time

Mutual information chain rule

Jensen's inequality

"Information cannot hurt"

Data processing inequality

Outline

Recall: Joint Mutual Information

Recall the mutual information between X and Y:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = I(Y;X).$$

We can also compute the mutual information between $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_M$:

$$\begin{aligned} I(X_1, \ldots, X_N; Y_1, \ldots, Y_M) &= H(X_1, \ldots, X_N) + H(Y_1, \ldots, Y_M) \\ &\quad - H(X_1, \ldots, X_N, Y_1, \ldots, Y_M) \\ &= I(Y_1, \ldots, Y_M; X_1, \ldots, X_N). \end{aligned}$$

Note that $I(X,Y;Z) \neq I(X;Y,Z)$ in general

- Reduction in uncertainty of X and Y given Z versus reduction in uncertainty of X given Y and Z

Chain Rule for Mutual Information

Let X, Y, Z be r.v. and recall that:

$$p(Z, Y) = p(Z|Y)p(Y)$$

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

$$\begin{aligned} I(X; Y, Z) &= I(Y, Z; X) && \text{symmetry} \\ &= H(Z, Y) - H(Z, Y|X) && \text{definition of mutual info.} \\ &= H(Z|Y) + H(Y) - H(Z|X, Y) - H(Y|X) && \text{entropy's chain rule} \\ &= \underbrace{H(Y) - H(Y|X)}_{I(Y;X)} + \underbrace{H(Z|Y) - H(Z|X, Y)}_{I(Z;X|Y)} \end{aligned}$$

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \quad \text{definition of mutual info and conditional mutual info}$$

Similarly, by symmetry:

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

Chain Rule for Mutual Information

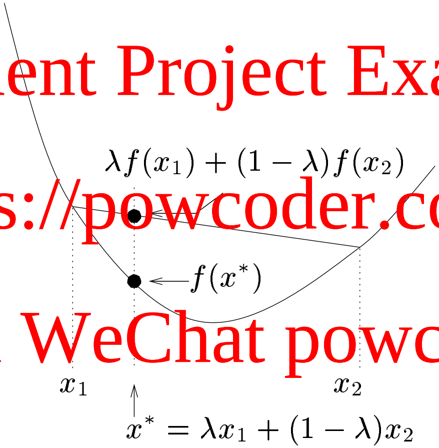
General form

For any collection of random variables $X_1, \ldots, X_N$ and Y:

$$
\begin{aligned}
I(X_1, \ldots, X_N; Y) &= I(X_1; Y) + I(X_2, \ldots, X_N; Y|X_1) \\
&= I(X_1; Y) + I(X_2; Y|X_1) + I(X_3, \ldots, X_N; Y|X_1, X_2) \\
&= \ldots \\
&= \sum_{i=1}^{N} I(X_i; Y|X_1, \ldots, X_{i-1}) \\
&= \sum_{i=1}^{N} I(Y; X_i|X_1, \ldots, X_{i-1}).
\end{aligned}
$$

Convex Functions:

Introduction



$0 \leq \lambda \leq 1$ (Figure from Mackay, 2003)

A function is convex $\smile$ if every chord of the function lies above the function

Definition

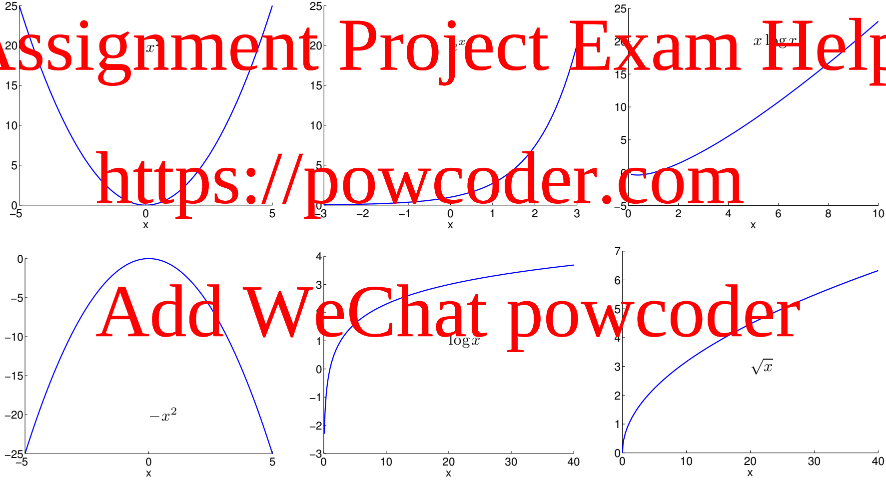
A function $f(x)$ is convex $\smile$ over (a, b) if for all $x_1, x_2 \in (a, b)$ and $0 \le \lambda \le 1$:

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

We say f is strictly convex $\smile$ if for all $x_1, x_2 \in (a, b)$ the equality holds only for $\lambda = 0$ and $\lambda = 1$.

Similarly, a function f is concave $\frown$ if $-f$ is convex $\smile$, i.e. if every chord of the function lies below the function.

Examples of Convex and Concave Functions



Verifying Convexity

Theorem (Cover & Thomas, Th 2.6.1)

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

This allows us to verify convexity or concavity.

Examples:

- x^2: $\displaystyle \frac{d}{dx}\left(\frac{d}{dx}(x^2)\right) = \frac{d}{dx}(2x) = 2$

- e^x: $\displaystyle \frac{d}{dx}\left(\frac{d}{dx}(e^x)\right) = \frac{d}{dx}(e^x) = e^x$
- $\sqrt{x}$, $x > 0$: $\displaystyle \frac{d}{dx}\left(\frac{d}{dx}(\sqrt{x})\right) = \frac{1}{2}\frac{d}{dx}\left(\frac{1}{\sqrt{x}}\right) = -\frac{1}{4}\frac{1}{\sqrt{x^3}}$

Convexity, Concavity and Optimization

If $f(x)$ is concave $\frown$ and there exists a point at which

$$\frac{df}{dx} = 0,$$

then $f(x)$ has a maximum at that point.

Note: the converse does not hold: if a concave $\frown$ $f(x)$ is maximized at some x, it is not necessarily true that the derivative is zero there.

- $f(x) = -|x|$: is maximized at $x = 0$ where its derivative is undefined

- $f(p) = \log p$ with $0 \leq p \leq 1$, is maximized at $p = 1$ where $\frac{df}{dp} = 1$
- Similarly for minimisation of convex functions

Jensen's Inequality for Convex Functions

Theorem: Jensen's Inequality

If f is a convex $\smile$ function and X is a random variable then:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Moreover, if f is strictly convex $\smile$, the equality implies that $X = \mathbb{E}[X]$ with probability 1, i.e. X is a constant.

In other words, for a probability vector $\mathbf{p}$,

$$f\left(\sum_{i=1}^{N} p_i x_i\right) \leq \sum_{i=1}^{N} p_i f(x_i).$$

Similarly for a concave $\frown$ function: $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$.

Jensen's Inequality for Convex Functions

Proof by Induction

(1) $K = 2$:

- Two-state random variable $X \in \{x_1, x_2\}$

- With $\mathbf{p} = (p_1, p_2) = (p_1, 1 - p_1)$
- $0 \le p \le 1$

we simply follow the definition of convexity:

$$\underbrace{p_1 f(x_1) + p_2 f(x_2)}_{\mathbb{E}[f(X)]} \ge f(\underbrace{p_1 x_1 + p_2 x_2}_{\mathbb{E}[X]})$$

Jensen's Inequality for Convex Functions

Proof by Induction — Cont'd

(2) $(K-1) \rightarrow K$: Assuming the theorem is true for distributions with $K-1$ states, and writing: $p_i' = p_i/(1-p_K)$ for $i = 1, \ldots, K-1$:

$$\sum_{i=1}^{K} p_i f(x_i) = p_K f(x_K) + (1-p_K) \sum_{i=1}^{K-1} p_i' f(x_i)$$

$$\geq p_K f(x_K) + (1-p_K) f\left(\sum_{i=1}^{K-1} p_i' x_i\right) \quad \text{Induction hypothesis}$$

$$\geq f\underbrace{\left(p_K x_K + (1-p_K)\sum_{i=1}^{K-1} p_i' x_i\right)}_{\sum_{i=1}^{K} p_i x_i} \quad \text{definition of convexity}$$

$$\sum_{i=1}^{K} p_i f(x_i) \geq f\left(\sum_{i=1}^{K} p_i x_i\right) \Rightarrow \mathbb{E}[f(X)] \geq f(\mathbb{E}[x]) \quad \text{equality case?}$$

Jensen's Inequality Example: The AM-GM Inequality

Recall that for a concave $\frown$ function: $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$.

Consider $X \in \{x_1, \ldots, x_N\}$, $X \geq 0$ with uniform probability distribution $\mathbf{p} = (\frac{1}{N}, \ldots, \frac{1}{N})$ and the strictly concave $\frown$ function $f(x) = \log x$

$$\frac{1}{N}\sum_{i=1}^{N} \log x_i \leq \log\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)$$

$$\log\left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}} \leq \log\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)$$

$$\left(\prod_{i=1}^{N} x_i\right)^{\frac{1}{N}} \leq \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sqrt[N]{x_1 x_2 \ldots x_N} \leq \frac{x_1 + x_2 \ldots + x_N}{N}$$

Gibbs' Inequality

Theorem

The relative entropy (or KL divergence) between two distributions $p(X)$ and $q(X)$ with $X \in \mathcal{X}$ is non-negative:

$$D_{\mathrm{KL}}(p \| q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Gibbs' Inequality

Proof (1 of 2)

Recall that: $D_{\text{KL}}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{p(X)}\left[\log \frac{p(X)}{q(X)}\right]$

Let $\mathcal{A} = \{x : p(x) > 0\}$. Then:

$$
\begin{aligned}
-D_{\text{KL}}(p\|q) &= \sum_{x \in \mathcal{A}} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x \in \mathcal{A}} p(x) \frac{q(x)}{p(x)} && \text{Jensen's inequality} \\
&= \log \sum_{x \in \mathcal{A}} q(x) \\
&\leq \log \sum_{x \in \mathcal{X}} q(x) \\
&= \log 1 \\
&= 0
\end{aligned}
$$

Gibbs' Inequality

Proof (2 of 2)

Since $\log u$ is strictly convex we have equality if $\dfrac{q(x)}{p(x)} = c$ for all x. Then:

$$\sum_{x \in \mathcal{A}} q(x) = c \sum_{x \in \mathcal{A}} p(x) = c$$

Also, the last inequality in the previous slide becomes equality only if:

$$\sum_{x \in \mathcal{A}} q(x) = \sum_{x \in \mathcal{X}} q(x).$$

Therefore $c = 1$ and $D_{\mathrm{KL}}(p \| q) = 0 \Leftrightarrow p(x) = q(x)$ for all x .

Alternative proof: Use the fact that $\log x \leq x - 1$.

Non-Negativity of Mutual Information

Corollary

For any two random variables X, Y:

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

with equality if and only if X and Y are statistically independent.

Proof: We simply use the definition of mutual information and Gibbs' inequality:

$$I(X; Y) = D_{\mathrm{KL}}(p(X, Y) \| p(X)p(Y)) \geq 0,$$

with equality if and only if $p(X, Y) = p(X)p(Y)$, i.e. X and Y are independent.

Conditioning Reduces Entropy

Information Cannot Hurt — Proof

Theorem

For any two random variables X, Y,

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

with equality if and only if X and Y are independent.

Proof: We simply use the non-negativity of mutual information:

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

with equality if and only if $p(X, Y) = p(X)p(Y)$, i.e X and Y are independent.

Data are helpful, they don't increase uncertainty on average.

Conditioning Reduces Entropy

Information Cannot Hurt — Example (from Cover & Thomas, 2006)

Let X, Y have the following joint distribution:

$p(X, Y)$		X = 1	X = 2
Y	1	0	3/4
	2	1/8	1/8

$$\mathbf{p}(X) = (1/8, 7/8)$$
$$\mathbf{p}(Y) = (3/4, 1/4)$$
$$\mathbf{p}(X|Y = 1) = (0, 1)$$
$$\mathbf{p}(X|Y = 2) = (1/2, 1/2)$$

$H(X) \approx 0.544$ bits $\quad H(X|Y = 1) = 0$ bits $\quad H(X|Y = 2) = 1$ bit

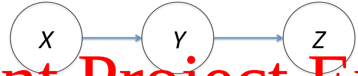
We see that in this case $H(X|Y = 1) < H(X)$, $H(X|Y = 2) > H(X)$.

However, $H(X|Y) = \sum_{y \in \{1,2\}} p(y) H(X|Y = y) = \frac{1}{4} = 0.25$ bits $< H(X)$

$H(X|Y = y_k)$ may be greater than $H(X)$ but the average: $H(X|Y)$ is always less or equal to $H(X)$.

Information cannot hurt on average

Markov Chain



Definition

Random variables X, Y, Z are said to form a Markov chain in that order (denoted by $X \to Y \to Z$) if their joint probability distribution can be written as:

$$p(X, Y, Z) = p(X)p(Y|X)p(Z|Y)$$

Consequences:

- $X \to Y \to Z$ if and only if X and Z are conditionally independent given Y.
- $X \to Y \to Z$ implies that $Z \to Y \to X$.
- If $Z = f(Y)$, then $X \to Y \to Z$

Data-Processing Inequality

Definition

Theorem

if $X \to Y \to Z$ then: $I(X;Y) \geq I(X;Z)$

- X is the state of the world, Y is the data gathered and Z is the processed data

- No “clever” manipulation of the data can improve the inferences that can be made from the data
- No processing of Y, deterministic or random, can increase the information that Y contains about X

Data-Processing Inequality

Proof

Recall that the chain rule for mutual information states that:

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

Therefore:

$$I(X;Y) + \underbrace{I(X;Z|Y)}_{0} = I(X;Z) + I(X;Y|Z) \quad \text{Markov chain assumption}$$

$$I(X;Y) = I(X;Z) + I(X;Y|Z) \quad \text{but } I(X;Y|Z) \geq 0$$

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

Corollary

In particular, if $Z = g(Y)$ we have that:

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

Proof: $X \to Y \to g(Y)$ forms a Markov chain.

Functions of the data Y cannot increase the information about X

Data-Processing Inequality

Observation of a “Downstream” Variable

Corollary

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

Proof: We use again the chain rule for mutual information:

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

Therefore:

$$I(X;Y) + \underbrace{I(X;Z|Y)}_{0} = I(X;Z) + I(X;Y|Z) \quad \text{Markov chain assumption}$$

$$I(X;Y|Z) = I(X;Y) - I(X;Z) \quad \text{but } I(X;Z) \geq 0$$

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

The dependence between X and Y cannot be increased by the observation of a “downstream” variable.

Summary & Conclusions

- Chain rule for mutual information

- Convex Functions

- Jensen's inequality, Gibbs' inequality

- Important inequalities regarding information, inference and data processing

- Reading: Mackay §2.6 to §2.10, Cover & Thomas §2.5 to §2.8

Next time

- Law of large numbers

- Markov's inequality

- Chebychev's inequality