

# Entropy, Relative Entropy, and Mutual Information

Cover and Thomas

(presented by Mike Cafarella)

# Overview

- Main topics
  - Entropy and conditional entropy
  - Relative entropy
  - Mutual information
- Useful for analysis
  - Jensen's
  - Data processing
- Markov chains and thermodynamics
- Sufficient statistics
- Probably skip
  - Log sum inequality
  - Fano's inequality

# All About Entropy

- *Entropy* is a measure of uncertainty of a random variable
- Put another way, how many guesses to determine value of  $X$ ?
- $H(X) = -\sum_x p(x)\log p(x)$
- Or,  $H(X) = E_p \log \frac{1}{p(X)}$

Since  $0 \leq P(x) \leq 1$ ,  $\log(1/p(x)) \geq 0$ , so  $H(x) \geq 0$ .

Similarly, we also know that *joint entropy* is  $H(X, Y) = -E \log p(X, Y)$ .

# Conditional Entropy and the Chain Rule

- *Conditional entropy* is the entropy of a random variable given another
- $H(Y|X) = -E_{p(x,y)} \log p(Y|X)$
- The *chain rule states*  $H(X, Y) = H(X) + H(Y|X)$
- *Pf:*

$$\begin{aligned} H(X, Y) &= -\sum_x \sum_y p(x, y) \log p(x, y) \\ &= -\sum_x \sum_y p(x, y) \log p(x) p(y|x) \\ &= -\sum_x \sum_y p(x, y) \log p(x) - \sum_x \sum_y p(x, y) \log p(y|x) \\ &= -\sum_x p(x) \log p(x) - \sum_x \sum_y p(x, y) \log p(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

# Relative Entropy

- *Relative entropy*, also called the *Kullback Leibler distance*, measures distance between two distributions.
- Or, how much extra entropy guesses does it take to model distribution  $p$  with  $q$  instead?
  - Defined as expected log of likelihood ratios between  $p$  and  $q$ .
  - $D(p||q) = E_p \log \frac{p(X)}{q(X)}$
  - As with regular entropy,  $D(p||q) \geq 0$  (see pf below)
  - KL-distance is not symmetric, does not satisfy triangle ineq.
- Chain rule states:
$$D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

# Mutual Information

- *Mutual information* measures how much the entropy of one random variable is reduced by knowledge of another

- $I(X; Y) = E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}$

- We can relate  $I(X; Y)$  to entropy of those variables

$$I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

$$= \sum_{x,y} p(x, y) \log \frac{p(x|y)}{p(x)}$$

$$= - \sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(x, y) \log p(x|y)$$

$$= - \sum_x p(x) \log p(x) - (- \sum_{x,y} p(x, y) \log p(x|y))$$

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

## More Mutual Information

- We can say a few other things, too

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

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

$$I(X; X) = H(X) - H(X|X) = H(X)$$

- By chain rule (useful later):

$$I(X_1, X_2, \dots, X_n; Y) = \sum_x^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1)$$

# Jensen's Inequality

- If  $f$  is convex, and  $X$  is a random var, then  $E f(X) \geq f(EX)$
- $(-\log(\cdot))$  is convex
- We use Jensen's to prove  $D(p||q) \geq 0$ 
  - $$-D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$
  - $$= \sum_x p(x) \log \frac{q(x)}{p(x)}$$
  - $$\leq \log \sum_x p(x) \frac{q(x)}{p(x)} \text{ (using Jensen's)}$$
  - $$= \log \sum_x q(x)$$
  - $$= \log 1 = 0$$
- So,  $-D(p||q) \leq 0$
- A number of other properties result, as well
  - Mutual information is non-negative:  $I(X; Y) \geq 0$
  - Conditioning reduces entropy:  $H(X|Y) \leq H(X)$

# Data processing inequality

- Variables  $X, Y, Z$  form Markov chain in order  $X \rightarrow Y \rightarrow Z$
- That is,  $X$  and  $Z$  are conditionally independent given  $Y$ .
- DPI states that  $I(X; Y) \geq I(X; Z)$ . There is no “processing” of  $Y$  into  $Z$  that can increase information brought to bear on  $X$ .

- Pf:

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

We know  $I(X; Z|Y) = 0$  because of Markov property

Because  $I(X; Y|Z) \geq 0$ , it must be true that  $I(X; Y) \geq I(X; Z)$

# All About Thermodynamics

- Model a closed thermodynamic system as a Markov chain, with changing distribution over a number of *microstates*. We can use previous statements to prove:

- Relative entropy  $D(\mu_n || \mu'_n)$  decreases with  $n$

$$D(p(x_n, x_{n+1}) || q(x_n, x_{n+1}))$$

$$= D(p(x_n) || q(x_n)) + D(p(x_{n+1}|x_n) || q(x_{n+1}|x_n))$$

$$= D(p(x_{n+1}) || q(x_{n+1})) + D(p(x_n|x_{n+1}) || q(x_n|x_{n+1}))$$

Since  $D(p(x_{n+1}) || q(x_{n+1}|x_n)) = 0$ , and  $D(p(x_n|x_{n+1}) || q(x_n|x_{n+1}))$  is non-negative...

$$D(p(x_n) || q(x_n)) \geq D(p(x_{n+1}) || q(x_{n+1}))$$

$$D(\mu_n || \mu'_n) \geq D(\mu_{n+1} || \mu'_{n+1})$$

## More About Thermodynamics

- Relative entropy between a distribution at state  $n$  and a stationary distribution decreases with  $n$
- If  $\mu'_n$  is a stationary distribution  $\mu$ , then
$$D(\mu_n || \mu) \geq D(\mu_{n+1} || \mu)$$
- Entropy increases if stationary distribution is uniform

## Even more About Thermodynamics

- $H(X_n|X_1)$  increases with  $n$  for a stationary Markov process
- $H(X_n)$  is constant, because it's stationary.
- However,  $H(X_n|X_1)$  increases with  $n$  (uncertainty about future increases)

- Pf:

$$H(X_n|X_1) \geq H(X_n|X_1, X_2) \text{ (conditioning reduces entropy)}$$

$$= H(X_n|X_2) \text{ (Markov property: } X_n \text{ independent of } X_1 \text{ given } X_2)$$

$$= H(X_{n-1}|X_1) \text{ (because } X_n \text{ is stationary)}$$

$$\text{So, } H(X_n|X_1) \geq H(X_{n-1}|X_1).$$

## Sufficient statistics

- A family of distributions  $\{f_\theta(x)\}$  is indexed by  $\theta$ , and  $X$  is drawn from one.  $T(X)$  is a function of  $X$
- We have Markov property  $\theta \rightarrow X \rightarrow T(X)$
- By the DPI, we know  $I(\theta; T(X)) \leq I(\theta; X)$
- $T(X)$  is sufficient for  $\theta$  if it captures all the information that  $X$  does; that is, if  $I(\theta; T(X)) = I(\theta; X)$

**That's all!**

Thanks!