EE/Stats 376A: Information theory	**Winter 2017**
Lecture 3 — Jan 17	
Lecturer: David Tse	*Scribe: Aditya G, Shawn Hu, Saliel B, Akshay M, Vivek B.*

3.1 Outline

- Relative Entropy
- Jensen's Inequality
- Data compression

3.1.1 Readings

- Shannon: 5,6,7
- CT: 5.1-5.8

3.2 Recap

Let us start by recapping the definition of mutual information. The mutual information $I(X;Y)$ between two random variables X and Y can be defined in the following equivalent ways:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= E\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \\ &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \end{aligned}$$

3.3 Relative Entropy

We first state a theorem about mutual information.

Theorem 1. For random variables X, Y we have:

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

Proof. Proved later. □

In order to prove the above theorem, we will first express mutual information in terms of a more general non-negative quantity called relative entropy.

Definition 1. The *relative entropy* between two distributions p, q defined on $\mathcal{X}$ is :

$$\begin{aligned} D(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= E_{X \sim p}\left[\log \frac{p(X)}{q(X)}\right]. \end{aligned}$$

3.3.1 Properties

1. In general, relative entropy is asymmetric $\big(D(p||q) \neq D(q||p)\big)$, and does not satisfy the triangle inequality. Therefore, it is **not** a metric.
2. $D(p||p) = 0$.
3. $D(p||q) \geq 0$ for all distributions p, q with equality holding iff $p = q$.

Mutual information between two random variables X, Y can be expressed in terms relative entropy between their joint distribution $p_{X,Y}$ and the product of their marginal distributions $p_X . p_Y$

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) \cdot p_Y(y)} \\ &= D(p_{X,Y}||p_X \cdot p_Y). \end{aligned} \tag{3.1}$$

We will prove Property 3 using Jensen's inequality and thereby prove Theorem 1.

3.3.2 Jensen's inequality

A real-valued function is *convex*, if the line segment joining any two points on the function curve lies **above** or on the curve. Mathematically,

Definition 2. *Convexity*: A real-valued function $f(x)$ is said to be *convex* over an interval (c, d) if $\forall x_1, x_2 \in (c, d)$ and $0 \leq \lambda \leq 1$,

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

For a doubly differentiable function f, convexity is equivalent to

1. $f'(x)$ is non-decreasing.
2. $f''(x) \geq 0$.

Note: A function f is a *concave* function if $-f$ is a convex function.

Theorem 2. *Jensen's Inequality*: For a convex function f, and a random variable $\mathcal{X}$,

$$f(E[X]) \leq E[f(X)].$$

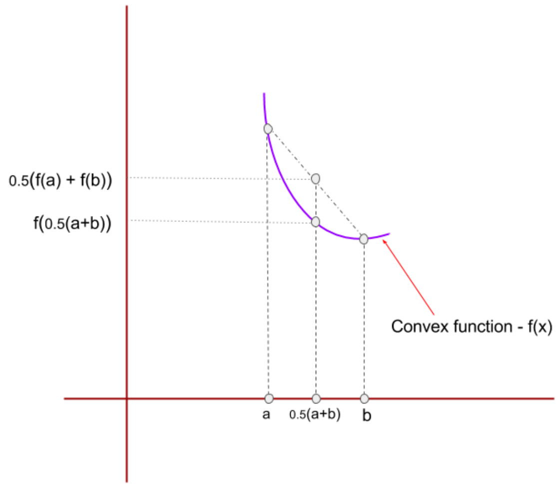

Figure 3.1: Illustration of Jensen's inequality.

Example : Consider a uniform random variable X defined on set $\{a, b\}$ and a convex function f as shown in Figure 3.1. By Jensen's inequality, $0.5\big(f(a) + f(b)\big) \geq f\big(0.5(a+b)\big)$, which can also be inferred from Figure 3.1.

Theorem 3. (Property 3) Relative entropy between two distributions p and q is non-negative:

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

Proof. We prove relative entropy is non-negative by applying Jensen's inequality to convex function $-\log(x)$ and random variable $\frac{q(X)}{p(X)}$,

$$\begin{aligned} D(p||q) &= E\left[-\log\frac{q(X)}{p(X)}\right] \\ \text{(using Jensen's inequality)} &\geq -\log E\left[\frac{q(X)}{p(X)}\right] \\ &= -\log\left(\sum_x p(x)\frac{q(x)}{p(x)}\right) \\ &= 0 \end{aligned} \tag{3.2}$$

□

Corollary 1. For random variables X and Y,

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

Interpretation in words - The nonnegativity of mutual information implies that **"on average"** the entropy of X conditioned on the observation $\{Y = y\}$ is equal to or lesser than the entropy of X (which intuitively makes sense).

Common pitfall : The above law (1) is applied to $H(X|Y)$, which is an averaged quantity: $H(X|Y) = \sum_y p(y)H(X|Y = y)$. However, $H(X|Y = y) \leq H(X)$ is not **necessarily**

true for all y, i.e., we could have cases where $H(X|Y=y) \geq H(X)$.

Using the nonnegativity property of relative entropy, we show that - among all possible distributions over a finite alphabet, the uniform distribution achieves the maximum entropy.

Consider random variable X defined on an alphabet $\mathcal{X}$ of size n. Let U be the uniform random variable defined on $\mathcal{X}$. Then,

Theorem 4. $H(X) \leq H(U)$

Proof.

$$\begin{aligned}
H(U) - H(X) &= \sum_x \frac{1}{n}\log n + \sum_x p(x)\log p(x) \\
&= \sum_x p(x)\log n + \sum_x p(x)\log p(x) \\
&= \sum_x p(x)\log\frac{p(x)}{1/n} \\
&= D(p||u) \qquad \text{(where } u \text{ is the uniform function)} \\
&\geq 0 \text{ (using Property 3)}.
\end{aligned}$$

□

3.4 Entropy and Data compression

Entropy is directly related to the fundamental limit of data compression. We consider two simple examples to get an intuition of the preceding statement:

1. For a sequence of i.i.d random variables $X_i \sim Bern(1/2)$, we need $n \times H(X_1) = n$ bits to encode $X_1, X_2, \cdots, X_n$.
2. However, for a sequence of i.i.d random variables $X_i \sim Bern(1/3)$, we need only $n \times H(X_1) \approx 0.918\, n$ bits to encode $X_1, X_2, \cdots, X_n$.

From the above two examples, we infer that the number of bits required to encode a sequence of i.i.d random variables depends on their entropy.

Rough Analysis: Consider a sequence of i.i.d random variables $X_i \sim Bern(p)$. The probability of a sequence $\{x_i\}$ with k ones and $n-k$ is

$$\begin{aligned}
p\big(x_1, x_2, \cdots, x_n\big) &= p^k(1-p)^{n-k} \\
&= 2^{k\log p + (n-k)\log(1-p)} \\
&= 2^{-n\left[\frac{k}{n}\log p + \left(1-\frac{k}{n}\right)\log(1-p)\right]} \\
\big(k \approx np \text{ by L.L.N}\big) &\approx 2^{-n\left[p\log p + (1-p)\log(1-p)\right]} \\
&= 2^{-nH(X_1)}.
\end{aligned}$$

So although there are 2^n possible sequences, the "typical" ones will have probability close to $2^{-nH(X_1)}$. Hence we can think of the source as having roughly $2^{nH(X_1)}$ typical sequences each with roughly the same probability. Thus, we need "roughly" $nH(X_1)$ bits to encode $\{X_i\}_{i=1}^{n} \sim Bern(p)$. This argument will be made rigorous in the next set of lecture notes.

Theorem 5. n i.i.d random variables distributed as $X \sim \mathcal{P}$, can be compressed using $nH(X)$ bits.

The proof uses weak law of large numbers and will be discussed in the next set of lecture notes.

Figure 3.2: Data compression.