

# EE 229A

VENKAT ANANTHARAM

*Davis Foote\**

*University of California, Berkeley*

*August 27<sup>th</sup>, 2015 – December 10<sup>th</sup>, 2015*

## CONTENTS

1	Toward a Calculus of Information	1
1.1	Defining Information . . . . .	1
1.2	Some Mechanics . . . . .	4
1.3	Divergence . . . . .	7
2	Entropy Rate	8

## 1 TOWARD A CALCULUS OF INFORMATION

### 1.1 Defining Information

1.1.1 NOTATION. For this class,  $\log$  refers to the logarithm in base 2.

1.1.2 DEFINITION. Let  $(p_1, \dots, p_m)$  be a probability distribution on  $\{1, \dots, m\}$ . The **entropy** of  $D$  is defined as

$$H(p_1, \dots, p_m) = - \sum_{i=1}^m p_i \log p_i = \sum_{i=1}^m p_i \log \frac{1}{p_i}$$

---

\*djfoote@berkeley.edu

*Lecture 1*  
*August 27<sup>th</sup>, 2015*  
*Lecture 2*  
*September 1<sup>st</sup>, 2015*

This can be intuitively decomposed into two components: the coefficient  $p_i$  is the probability with which symbol  $i$  occurs, and  $\log \frac{1}{p_i}$  is the information content inherent to symbol  $i$  as a member of the distribution  $(p_1, \dots, p_m)$ .

1.1.3 NOTATION. If  $X$  is a random variable taking values in a finite set  $\mathcal{X}$ , we write  $H(X)$  to mean  $H((p_X(x) : x \in \mathcal{X}))$ .

1.1.4 NOTATION. A somewhat dangerous but common convention is to drop the subscript in  $p_X(x)$  to just be  $p(x)$ .

The first and second derivatives of  $\log$  are:

1.  $\frac{d}{du} \log u = (\log u)' = (\log e) \frac{1}{u} > 0$  for  $u > 0$
2.  $\frac{d^2}{du^2} \log u = -\frac{\log e}{u^2} < 0$  for  $u > 0$

Since the derivative is nonincreasing this is a “concave” function.

1.1.5 DEFINITION. A **concave function** is the negative of a convex function.

1.1.6 DEFINITION. A **convex function** is a real-valued function on a convex set, such that

$$f(\eta x_1 + (1 - \eta)x_0) \leq \eta f(x_1) + (1 - \eta)f(x_0)$$

for all  $x_0, x_1$  in the domain and all  $\eta \in [0, 1]$

1.1.7 DEFINITION. A **convex set** is a subset  $C$  of  $\mathbb{R}^d$  (for some  $d \geq 1$ ) such that for all  $x_0, x_1 \in C$  and all  $\eta \in [0, 1]$ ,  $\eta x_1 + (1 - \eta)x_0 \in C$ .

Note that since the domain of a convex function must be a convex set, the definition given for convex function makes sense.

1.1.8 EXAMPLE. Consider a convex subset  $C$  of the real line. Given any two points in  $C$ , every point on the real line between these two points must also be in  $C$ .

1.1.9 THEOREM. *For a real-valued function  $f$  on a convex subset  $C$  of the real line, if  $f$  has a nonnegative second derivative it is convex. Hence if  $f$  has a nonpositive second derivative it is concave.*

TODO: include drawing of  $u \mapsto u \log u$

Need to first understand the limit as  $u$  goes to 0 from the right,  $\lim_{u \rightarrow 0^+} u \log u$ . In fact, this limit is 0 because

$$u \log u = -u \log \frac{1}{u} = \frac{\log \frac{1}{u}}{2^{\log \frac{1}{u}}}$$

Since the numerator is approaching infinity linearly in  $\log \frac{1}{u}$  and the denominator is approach-

ing infinity exponentially in  $\log \frac{1}{u}$ , the limit must be 0.

The first and second derivatives of  $u \log u$  are

- $\frac{d}{du} u \log u = (u \log u)' = \log u + \log e$ , which is negative if  $u < 1/e$  and positive when  $u > 1/e$
- $\frac{d^2}{du^2} u \log u (u \log u)'' = (\log e) \frac{1}{u} > 0$  if  $u > 0$ , so the function is convex.

The purpose of this is to get a feeling for  $H(p_1, \dots, p_m)$  as a function from probability distributions on  $\{1, \dots, m\}$  to real numbers.

1.1.10 DEFINITION. When  $m = 2$ ,

$$H(p, 1-p) = -p \log p - (1-p) \log(1-p)$$

This function is called the **binary entropy function**. The function is nonnegative, its maximum occurs at  $p = \frac{1}{2}$ , the derivative at 0 is  $+\infty$ , and the derivative at 1 is  $-\infty$ . The function is concave.

TODO: include drawing of binary entropy function

1.1.11 NOTATION. For binary distributions, we often just write  $H(p)$ .

1.1.12 DEFINITION. The set of probability distributions on  $\{1, \dots, m\}$  can be visualized as a convex subset of  $\mathbb{R}^m$ , the **convex simplex** in  $\mathbb{R}^m$ .

1.1.13 EXAMPLE. For  $m = 3$ , this is the (filled in) triangle connecting the points  $(1, 0, 0)$ ,  $(0, 1, 0)$ , and  $(0, 0, 1)$ .

$H(p_1, \dots, p_m)$  viewed as a real-valued function on the unit simplex in  $\mathbb{R}^m$  is nonnegative (because each  $p_i \in [0, 1]$ ). It is also concave, because given  $\mathbf{p}^{(0)} = (p_1^{(0)}, \dots, p_m^{(0)})$  and  $\mathbf{p}^{(1)} = (p_1^{(1)}, \dots, p_m^{(1)})$  and  $\eta \in [0, 1]$ ,

$$\eta \mathbf{p}^{(1)} + (1-\eta) \mathbf{p}^{(0)}$$

TODO: finish derivation using concavity of  $-u \log u$

1.1.14 FACT.  $H(p_1, \dots, p_m)$  is invariant to permutations of the coordinates. This is  $m!$ -fold symmetry. Its maximum occurs at  $p_i = 1/m$  for  $1 \leq i \leq m$ . This should match intuition because the uniform distribution is the “most uncertain.”

1.1.15 NOTATION. Consider two random variables  $X$  and  $Y$ ;  $X$  taking values in  $\mathcal{X}$  and  $Y$  taking values in  $\mathcal{Y}$ . Both  $\mathcal{X}$  and  $\mathcal{Y}$  are finite sets. They have joint probability distribution

$$p_{XY}(x, y) = P(X = x, Y = y)$$

which we’ll abbreviate as  $p(x, y)$ .

1.1.16 DEFINITION. The **joint entropy**  $H(x, y) = -\sum_{x,y} p(x, y) \log p(x, y)$ .

1.1.17 EXAMPLE. Suppose  $\mathcal{X} = \{1, 2, 3, 4\}$ ,  $\mathcal{Y} = \{1, 2, 3\}$ , and  $p(x, y)$  is uniform on  $\{(x, y) \in \mathcal{X} \times \mathcal{Y} : x \geq y\}$ .

$$\begin{aligned} H(x, y) &= \log 9 \\ H(x) &= \frac{1}{9} \log 9 + \frac{2}{9} \log \frac{9}{2} + 2 \cdot \frac{3}{9} \log \frac{9}{3} = \log 9 - \frac{2}{9} \log 2 - \frac{2}{3} \log 3 \\ H(y) &\text{ is computed similarly.} \end{aligned}$$

Suppose that  $X$  and  $Y$  are not independent; that is, knowing  $X$  gives you some information about  $Y$ . Intuitively one should expect, then, that the total uncertainty between  $X$  and  $Y$  is less than the sum of their individual uncertainties.

1.1.18 NOTATION. Let us condition on the event  $\{Y = y\}$  for some  $y \in \mathcal{Y}$ . The conditional distribution of  $X$  given that  $Y = y$  is

$$p_{X|Y}(x|y), x \in \mathcal{X}$$

(which we'll abbreviate as  $p(x|y)$ ).

1.1.19 NOTATION. This conditional probability distribution has an entropy. According to the formula, this entropy is

$$-\sum_{x \in \mathcal{X}} p(x|y) \log p(x|y)$$

which we denote as  $H(X|Y = y)$ .

1.1.20 DEFINITION. Let  $p_Y(y) = P(Y = y) = \sum_{x'} p(x', y)$  be denoted  $p(y)$ . Then

$$\sum_{y \in \mathcal{Y}} p(y) H(X|Y = y)$$

is denoted  $H(X|Y)$  and is called the **conditional entropy** of  $X$  given  $Y$ .

## 1.2 Some Mechanics

1.2.1 THEOREM.  $H(X, Y) = H(Y) + H(X|Y)$

*Proof.*

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

□

Note that  $H(X|Y) \neq H(Y|X)$  in general.

**1.2.2 THEOREM** (Chain rule for entropy). *Suppose you are given  $n$  random variables  $X_1, \dots, X_n$ , each discrete and finite-valued. Then we have*

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \dots + H(X_n|X_1, X_2, \dots, X_{n-1})$$

*There are  $n!$  valid such formulae.*

This can be proved simply by induction using a similar derivation as that of the formula in the previous theorem.

Suppose we are given two discrete, finite-valued random variables  $X$  and  $Y$ . The expression  $H(X) - H(X|Y)$  seems to capture the amount by which the uncertainty about  $X$  is reduced (on average) when learning  $Y$ .

**1.2.3 DEFINITION.** This quantity  $H(X) - H(X|Y)$  is denoted  $I(X; Y)$ . It is also often written as  $H(X|Y)$ . It is called the **mutual information** between  $X$  and  $Y$ .

**1.2.4 THEOREM.**  $I(X; Y) = I(Y; X)$ .

*Proof.*

$$\begin{aligned}
I(X; Y) &\triangleq H(X) - H(X|Y) \\
&= - \sum_{x,y} p(x, y) \log p(x) + \sum_{x,y} p(y) p(x|y) \log p(x|y) \\
&= \sum_{x,y} \log p(x, y) \frac{p(x, y)}{p(x)p(y)} \\
&= I(Y; X)
\end{aligned}$$

where the last equality holds because the expression in the second-to-last line is symmetric with respect to  $x$  and  $y$ .  $\square$

**1.2.5 DEFINITION.** Given three random variables  $X$ ,  $Y$ , and  $Z$  (all discrete, finite), the **conditional mutual information** between  $X$  and  $Y$  conditioned on  $Z = z$  is

$$\begin{aligned} I(X; Y | Z = z) &\triangleq H(X | Z = z) - H(X | Y = y, Z = z) \\ &= \sum_{x,y} p(x, y | z) \log \frac{p(x, y | z)}{p(x | z)p(y | z)} \end{aligned}$$

We then define the **conditional mutual information** between  $X$  and  $Y$  given  $Z$  as

$$I(X; Y | Z) \triangleq \sum_z p(z) I(X; Y | Z = z)$$

Note that

$$\begin{aligned} I(X; Y | Z = z) &= \sum_x p(x | z) \log \frac{1}{p(x | z)} + \sum_y p(y | z) \sum_x p(x | y, z) \log p(x | y, z) \\ &= H(X | Z = z) - H(X | Y, Z = z) \end{aligned}$$

where  $H(X | Y, Z = z) \triangleq \sum_y P(Y = y | Z = z) H(X | Y = y, Z = z)$ .

**1.2.6 THEOREM (Chain rule for information).**  $I(X, Y_1, Y_2, \dots, Y_n) = I(X; Y_1) + I(X; Y_2 | Y_1) + \dots + I(X; Y_n | Y_1, Y_2, \dots, Y_{n-1})$ .

**1.2.7 NOTATION.**

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E} \left[ \log \frac{1}{p_X(X)} \right]$$

This is also written by abuse of notation as

$$H(X) = \mathbb{E} \left[ \log \frac{1}{p(X)} \right]$$

Similarly,

$$H(X | Y) = \mathbb{E} \left[ \log \frac{1}{p(X | Y)} \right]$$

**1.2.8 THEOREM (Chain rule for mutual information).**

$$I(X; Y_1, \dots, Y_n) = I(X; Y_1) + I(X; Y_2 | Y_1) + I(X; Y_3 | Y_1, Y_2) + \dots + I(X; Y_n | Y_1, \dots, Y_{n-1})$$

There are  $n!$  such formulae.

The proof of this theorem simply involves splitting each  $I$  into a difference of entropies and following algebraically.

### 1.3 Divergence

#### 1.3.1 DEFINITION.

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

where  $(p(x), x \in \mathcal{X})$  and  $(q(x), x \in \mathcal{X})$  are two probability distributions on the same finite set  $\mathcal{X}$ . This quantity has several names:

- The **relative entropy** of  $p$  with respect to  $q$ .
- The **information divergence** (or **divergence**) of  $p$  from  $q$ .
- The **Kullback-Liebler distance** (or **information distance**) of  $p$  from  $q$ .

This measure should be interpreted as a sort of distance measure between  $p$  and  $q$  (see below).

#### 1.3.2 FACT.

$$I(X; Y) = D(p(x, y) \| p(x)p(y))$$

This is true because

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

Some facts about  $D(p\|q)$ :

- $D(p\|q)$  can equal  $+\infty$ . This happens if and only if there is some  $x \in \mathcal{X}$  for which  $q(x) = 0$  but  $p(x) > 0$ . The intuition here is that if something can show up for  $p$  and not  $q$  then we have an infinite ability to determine that we are dealing with  $p$  and not  $q$  if that value appears. Note that this cannot happen with  $I(X; Y)$ .
- $D(p\|q) \neq D(q\|p)$  in general. For example, one can be  $+\infty$  and the other can be 0.

**1.3.3 THEOREM (Jensen's Inequality).** *If  $f(u)$  is a convex function defined for  $u$  in a convex subset  $S$  of  $\mathbb{R}^n$  for some  $n \geq 1$  and  $U$  is a random variable taking values in the domain  $S$  of  $f$ , then*

$$\mathbb{E}[f(U)] \geq f(\mathbb{E}[U])$$

The proof comes from noting that a convex function at the centroid of a convex set should be lower than the polytope connecting the points  $\{(u, f(u)) : u \in S\}$ .

**1.3.4 DEFINITION.** A convex function  $f$  defined on a convex set  $C \subset \mathbb{R}^n$  is called **strictly convex** if  $(1 - \eta)f(u_0) + \eta f(u_1) > f((1 - \eta)u_0 + \eta u_1)$  unless either  $u_0 = u_1$  or  $\eta = 0$  or  $\eta = 1$ .

A test for strict convexity is that the second derivative be strictly positive.

**1.3.5 THEOREM.** *It is always the case that  $D(p\|q) \geq 0$ . We have equality if and only if  $p = q$ .*

*Proof.* Write

$$D(p\|q) = \sum_x q(x) \left( \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} \right)$$

and consider  $U$  taking values in  $[0, \infty)$  taking the value  $\frac{p(x)}{q(x)}$  with probability  $q(x)$ . Use Jensen's Inequality for the strictly convex function  $f(u) = u \log u$ . This gives

$$\begin{aligned} D(p\|q) &= \mathbb{E}[f(U)] \\ &\geq f(\mathbb{E}[U]) \\ &= \left( \sum_x q(x) \frac{p(x)}{q(x)} \right) \log \left( \sum_x q(x) \frac{p(x)}{q(x)} \right) \\ &= \left( \sum_x q(x) \frac{p(x)}{q(x)} \right) \log \left( \sum_x p(x) \right) \\ &= 0 \end{aligned}$$

Since  $f$  is strictly convex, we have inequality if and only if  $q$  is a one-point distribution (in which case  $D(p\|q) = \infty$  unless  $p$  is also a one-point distribution on the same point) or if all  $\frac{p(x)}{q(x)}$  are the same, in which case they must all equal 1.  $\square$

**1.3.6 THEOREM.** *It is always true that  $I(X;Y) \geq 0$ . We have equality if and only if  $X$  and  $Y$  are independent.*

*Proof.*  $I(X;Y) = D(p(x,y)\|p(x)p(y))$ , and  $D(p(x,y)\|p(x)p(y)) = 0$  if and only if  $p(x,y)$  and  $p(x)p(y)$  are the same distribution, i.e. if  $X$  and  $Y$  are independent.  $\square$

## 2 ENTROPY RATE

**2.0.7 DEFINITION.** A **stochastic process** is simply a sequence of random variables.

**2.0.8 DEFINITION.** A **discrete-time stationary stochastic process** is a sequence of random variables  $(X_1, X_2, X_3, \dots)$  with the property that for every  $m \geq 1$ ,  $p(x_1, \dots, x_m) = p(x_{n+1}, x_{n+2}, \dots, x_{n+m})$  for all  $n \geq 0$ .



---

2.0.9 THEOREM (Entropy rate). *For any discrete-time stationary stochastic process,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, \dots, X_n) = \lim_{n \rightarrow \infty} H(X_{n+1} \mid X_1, \dots, X_n)$$

*This quantity is called the **entropy rate** of the process.*