GI12/4C59: Information Theory

Lectures 4–6

Massimiliano Pontil

Outline

1. Convex functions
2. Entropy
3. Relative entropy
4. Joint entropy
5. Mutual information
6. Conditional entropy and mutual information

About these lectures

Theme of lectures 4–6: We introduce the basic definitions and quantities needed to develop the theory. We provide the intuition behind each notion and begin to speculate on their role in Information Theory.

Math required: Lectures 1–3, familiarity with convex functions (reviewed Today).

Some elements of convex analysis

We recall some basic facts on convex analysis

1. Convex sets
2. Convex functions
3. Constrained minimization and Lagrange multipliers

Convex sets

A set $\mathcal{D} \subset \mathbb{R}^n$ is said convex if the line segment joining every pair of points is in $\mathcal{D}$, that is, for every $x, t \in \mathcal{D}$ we have that

$$\lambda x + (1-\lambda)t \in \mathcal{D}, \quad \lambda \in [0,1]$$

- $\mathbb{R}^n$ is a convex set
- The sets $[a,b]^n, (a,b]^n, [a,b)^n, (a,b)^n$, are convex.
- If S and T are two convex sets then $S \cap T$ is convex but $S \cup T$, in general, is not.
- If S and T are two convex sets then the product set $S \times T = \{z = (s,t) : s \in S, t \in T\}$ is convex.

Convex functions

Let $\mathcal{D} \subseteq \mathbb{R}^n$ be a convex set. A function $g : \mathcal{D} \to \mathbb{R}$ is said convex if for every $x, t \in \mathcal{D}$ and $\lambda \in [0,1]$

$$g(\lambda x + (1-\lambda)t) \leq \lambda g(x) + (1-\lambda) g(t).$$

g is said strictly convex if it is convex and, in the above inequality, the equality holds only for $\lambda = 0$ or 1.

- A convex function always lies below any cord.

A function g is said concave if $-g$ is convex.

Characterization of convex functions

Let $g : \mathbb{R} \to \mathbb{R}$. If its second order derivative g'' exists everywhere and it is everywhere (positive) nonnegative, then g is (strictly) convex.

Example: Let $g(x) = -\log x$, $x \in (0, \infty)$. Then g is strictly convex because

$$g''(x) = \frac{1}{x^2} > 0$$

More generally, let $g : \mathbb{R}^n \to \mathbb{R}$. If the second order partial derivatives of g exist for every $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and the Hessian matrix

$$J_{ij}(x) = \frac{\partial^2 g(x)}{\partial x_i \partial x_j}$$

is (positive) nonnegative definite for every $x \in \mathbb{R}^n$, then g is (strictly) convex.

Sum of convex functions

If $\mathcal{D} \subseteq \mathbb{R}$ are convex sets, the functions $g_i : \mathcal{D} \to \mathbb{R}$ are convex (concave), then the function $g : \mathcal{D}^n \to \mathbb{R}$ defined by

$$g(x_1, \ldots, x_n) = \sum_{i=1}^{n} g_i(x_i), \quad x_i \in \mathcal{D},\ i = 1, \ldots, n$$

is convex (concave) on $\mathcal{D}^n$. Can you proof this? (easy)

Example: Let $h : [0, \infty) \to \mathbb{R}$ be defined as $h(x) = x \log x$. Since h is convex (check!), the function $g : [0, \infty)^n \to \mathbb{R}$ defined by

$$g(x_1, \ldots, x_n) = \sum_{k=1}^{n} x_n \log x_n$$

is also convex.

Composition of convex functions

If $f: \mathbb{R} \to \mathbb{R}$ is a convex function and $g = ax + b$ then $f(g(\cdot))$ is convex.

Proof: let $x_1, x_2 \in \mathbb{R}$ and $\lambda \in [0,1]$. We have

$$\begin{aligned} f(g(\lambda x_1 + (1-\lambda)x_2)) &= f(a(\lambda x_1 + (1-\lambda)x_2) + b) \\ &= f(\lambda(ax_1 + b) + (1-\lambda)(ax_2 + b)) \\ &\leq \lambda f(g(x_1)) + (1-\lambda) f(g(x_2)) \end{aligned}$$

Jensen inequality

If X is a r.v. and $f: \mathbb{R} \to \mathbb{R}$ a convex function then

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

In addition, if f is strictly convex the equality holds if and only if X is a constant.

- We show the proof in the discrete case. This can be easily extended to continuous r.v. (by a continuity step).

Example: The function $f(x) = x^2$ is convex so we have:

$$E[X^2] \geq (E[X])^2$$

(recall $var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 \ldots$)

Proof

We need to prove that for every $x_i \in \mathbb{R}$ and $p_i \geq 0$ with $\sum_{i=1}^n p_i$ we have that

$$\sum_{i=1}^n p_i f(x_i) \geq f(\sum_{i=1}^n p_i x_i) \quad (*)$$

and if f is strictly convex the equality holds if and only if all but one p_i are zero.

Proof is by induction: for $n = 2$ (*) is just the definition of convex function. Suppose (*) is true for $n = k - 1$, $k > 3$. Then it is also true for $n = k$ since

$$\begin{aligned}
\sum_{i=1}^k p_i f(x_i) &= (1-p_k)\left(\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} f(x_i)\right) + p_k f(x_k) \\
&\geq (1-p_k) f\left(\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} x_i\right) + p_k f(x_k) \\
&\geq f((1-p_k)\left(\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} x_i\right) + p_k x_k) = f\left(\sum_{i=1}^{k} p_i x_i\right)
\end{aligned}$$

Entropy

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a finite set (alphabet) and X a discrete r.v. with values on $\mathcal{X}$ and probability function $p(x) = P(\{X = x\})$. The entropy of X is defined by

$$H_D(X) := -\sum_{x \in \mathcal{X}} p(x) \log_D p(x)$$

Standard choice: $D = 2$ (here we neglect the subscript 2 in H_2 and $\log_2$) and the entropy is measured in "bits". If $D = e$ the units measure is "nats". Useful conversion formula: $H_a(X) = \log_a(D)\ H_D(X)$.

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

Example: Let $\mathcal{X} = \{0,1\}$ and set $p = P(\{X = 1\})$. Then $H(p) = -p \log p - (1-p)\log(1-p)$. In particular, $H(1/2) = 1$ and $H(1) = H(0) = 0$.

Properties of H

The entropy is a function of the distribution p (it depends on X only through the values $p(x_1), \dots, p(x_n)$). Thus, sometimes we write $H(p)$ instead of $H(X)$.

If p "peaks" at $x^* \in \mathcal{X}$, that is, $p(x) = 1$ if $x = x^*$ and zero otherwise, then $H(p) = 0$. In all other cases $H(p)$ is positive. (Note: we use the convention $0 \log 0 = 0$.

$H(p)$ achieves its maximum when p is the uniform distribution, that is, $p(x) = \frac{1}{n}$ for every $x \in \mathcal{X}$, in which case $H(p) = \log n$ (where $n = |\mathcal{X}|$).

In fact, since the function $f(t) := nt \log t$ is convex (we saw this before), by Jensen's inequality we have that

$$\log \frac{1}{n} = f(\frac{1}{n} \sum_{k=1}^{n} p_k) \leq \frac{1}{n} \sum_{k=1} f(p_k) = -H(p) \quad \Rightarrow H(p) \leq \log n$$

and since f is strictly convex, $H(p) = \log n$ if and only if p is uniform.

Property of H (cont.)

Since the function $-t \log t$ is concave and the sum of concave functions is a concave function, it follows that the entropy is a concave function of the vector point $(p_1, \dots, p_n)$.

In particular if p and q are two probability functions for X then for every $\lambda \in [0, 1]$ we have that:

$$H(\lambda p + (1 - \lambda) q) \leq \lambda H(p) + (1 - \lambda) H(q)$$

Properties of H (summary)

We summarize the properties we have just proved

- $H(X) \in [0, \log n]$ with $H(X) = 0$ if and only if p peaks at some $x \in \mathcal{X}$, and $H(p) = \log n$ if and only if p is the uniform distribution.

- $H(p)$ is a concave function of p.

Interpretation

Interpretation 1: $H(X)$ as a measure of the *uncertainty* of X – the higher the randomness in X the higher the uncertainty of X (or a measure of the *information gained* by measuring X).

Interpretation 2: $H(X)$ as a lower bound on the minimum number of binary questions required to determine the value of X.

Example 1: Let $\mathcal{X} = \{a, b, c, d\}$ and $p(a) = \frac{1}{2}$, $p(b) = \frac{1}{4}$, $p(c) = p(d) = \frac{1}{8}$. Then $H(X) = \frac{7}{4}$. An efficient algorithm to determine X is to ask the following ordered binary questions: Q1 = "Is $X = a$?", Q2 = "Is $X = b$?", Q3 "Is $X = c$?". In this case the expected number of questions asked is $1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{4} = \frac{7}{4}$. We will see that, in general, the *minimum* number of such questions is always between $H(X)$ and $H(X) + 1$.

Link to data compression

Suppose we wish to represent the elements of $\mathcal{X}$ with variable length codes (for example, binary strings) and let $\ell(x)$ be the length of the code assigned to $x \in \mathcal{X}$.

Later in the course we will see that the entropy plays a key role in this problem. In particular we will show that for every binary code,

$$L = E[\ell(x)] \geq H(X)$$

and any minimizing code for L, that is a code which provides the best compression of X, is always within one bit of the entropy of X,

$$L^* = \min L < H(X) + 1$$

Constrained minimization

We present a different proof that the maximum of H is achieved by the uniform distribution.

Let $g : \mathcal{D} \to \mathbb{R}$ be convex and $g \in C^1$. Suppose we wish to find the minimum of $g : \mathbb{R}^n \to \mathbb{R}$ subject to the constraint that

$$h(x) = 0, \quad h \in C^1.$$

Consider the Lagrangian function $L(x, \mu) = g(x) + \mu h(x)$, where $\mu \in \mathbb{R}$ is called the Lagrange multiplier

Then x_0 is a minimum of g subject to $h(x) = 0$ if and only if

$$\frac{\partial L(x_0, \mu_0)}{\partial x} = \frac{\partial L(x_0, \mu_0)}{\partial \mu} = 0$$

for some $\mu_0 \in \mathbb{R}$.

Maximum entropy problem

Let p be a probability distribution on $\mathcal{X} = \{x_1, \ldots, x_n\}$.

What is the probability distribution which maximizes the entropy?

This problem is equivalent to solve

$$\min\{-H(p) : \sum_{k=1}^{n} p_k = 1, p_k \geq 0\}$$

The Langrangian is

$$L(p, \mu) = \sum_{k=1}^{n} p_k \log p_k + \mu((\sum_{k=1}^{n} p_k) - 1)$$

Maximum entropy problem (cont.)

$L(p, \mu) = \sum_{k=1}^{n} p_k \log p_k + \mu((\sum_{k=1}^{n} p_k) - 1)$. We have: (only in this slides we change notation and measure the entropy in nats).

$$\frac{\partial L(p, \mu)}{\partial p_k} = \log p_k + 1 + \mu = \log e p_k + \mu$$

thus if we set this equation equal to zero we get that $p_k = \frac{2^{-\mu}}{e}$ and using the constraint $\sum_{k=1}^{n} p_k = 1$ we obtain

$$p_k = \frac{1}{n}, \quad k = 1, \ldots, n.$$

Since the entropy is strictly convex, this is the only solution.

Relative Entropy

Let p, q be two probability distributions. The relative entropy of p and q is defined by:

$$D(p \parallel q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Note: D is also called the Kullback Leiber divergence or distance (but it is not a distance since, in general, $D(p \parallel q) \neq D(q \parallel p)$). Note also that D may be infinite, e.g. if $\mathcal{X} = \{0,1\}$, $D((p, 1-p) \parallel (0,1)) = \infty$ for every $p \in (0,1]$.

Remember that we use the convection: $p \log \frac{p}{0} = +\infty$ for every $p > 0$, $0 \log \frac{0}{0} = 0$, $0 \log 0 = 0$ (all these follow from the continuity of the log function).

Interpretation

$D(p \parallel q)$ is a measure of the inefficiency of assuming that the distribution of X is q when the true distribution is p.

Example 1 (cont.): Let $q(a) = \frac{1}{4}$, $q(b) = \frac{1}{2}$, $q(c) = q(d) = \frac{1}{8}$. Then $D(p \parallel q) = \frac{1}{4}$.

If we believe X is distributed according to q, in order to determine X we would ask the binary questions: Q1 = "Is $X = b$?", Q2 = "Is $X = a$", Q3 = "Is $X = c$?" (in this order). Since the true distribution of X is p, the expected number of questions asked is $1 \times \frac{1}{4} + 2 \times \frac{1}{2} + 3 \times \frac{1}{4} = 2 = H(X) + D(p \parallel q)$. We will see that, in general, the *minimum* number of such questions is between $H(X) + D(p \parallel q)$ and $H(X) + D(p \parallel q) + 1$.

Properties of D

We show that

1. $D(p \parallel q) \geq 0$ with equality if and only if $p = q$.

2. $D(p \parallel q)$ is a convex function of (p, q), that is if p_1, q_1, p_2, q_2 are probability distributions then for every $\lambda \in [0, 1]$ we have

   $$D(\lambda p_1 + (1-\lambda) p_2 \parallel \lambda q_1 + (1-\lambda) q_2) \leq \lambda D(p_1 \parallel q_1) + (1-\lambda) D(p_2 \parallel q_2)$$

Proof of 1

Recall Jensen inequality: if $f : \mathbb{R} \to \mathbb{R}$ is convex and X is a discrete r.v. then $E[f(X)] \geq f(E[X])$. If f is strictly convex $E[f(X)] = f(E[X])$ if and only if X is a constant.

Let $\mathcal{D} = \{x : p(x) > 0\}$. Since $\log(\cdot)$ is strictly concave, we have that

$$\begin{aligned} -D(p \parallel q) &= -\sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in \mathcal{D}} p(x) \log \frac{q(x)}{p(x)} \leq \log \left(\sum_{x \in \mathcal{D}} p(x) \frac{q(x)}{p(x)} \right) \qquad (\diamond) \\ &= \log \left(\sum_{x \in \mathcal{D}} q(x) \right) \leq \log \left(\sum_{x \in \mathcal{X}} q(x) \right) = \log 1 = 0 \end{aligned}$$

with equality if and only if $p = q$. (because of $(\diamond)$)

The log sum inequality

To proof 2 we use the following inequality: for every non-negative numbers $a_1, \ldots, a_n,\ b_1, \ldots, b_n$,

$$\sum_{k=1}^{n} a_k \log \frac{a_k}{b_k} \geq \left(\sum_{k=1}^{n} a_k \right) \log \frac{\sum_{k=1}^{n} a_k}{\sum_{k=1}^{n} b_k} \qquad (*)$$

with equality if and only if $a_k/b_k = c$ (where c is a constant).

Proof: We set $\alpha_k = b_k / \sum_j b_j$ and $t_k = a_k/b_k$. Since the function $f(t) = t \log t$ is strictly convex, by Jensen inequality we have

$$\sum_{k=1}^{n} \alpha_k f(t_k) \geq f\left(\sum_{k=1}^{n} \alpha_k t_k \right) \qquad \Rightarrow \ (*)$$

which equality if and only if $t_k = c$.

Proof of 2

Recall : $\sum_{k=1}^{n} a_k \log \frac{a_k}{b_k} \geq \left(\sum_{k=1}^{n} a_k \right) \log \frac{\sum_{k=1}^{n} a_k}{\sum_{k=1}^{n} b_k}$ with equality if and only if $a_k/b_k = c$.

We apply (*) to each term (inside the sum) in the relative entropy

$$D\left(\lambda p_1(x) + (1-\lambda) p_2(x)\right) \| \lambda q_1(x) + (1-\lambda) q_2(x)) =$$

$$\sum_x (\underbrace{\lambda p_1(x)}_{a_1} + \underbrace{(1-\lambda) p_2(x)}_{a_2}) \log \frac{\lambda p_1(x) + (1-\lambda) p_2(x)}{\underbrace{\lambda q_1(x)}_{b_1} + \underbrace{(1-\lambda) q_2(x)}_{b_2}}$$

$$\leq \sum_x \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda) p_2(x) \log \frac{(1-\lambda) p_2(x)}{(1-\lambda) q_2(x)}$$

Alternative proof of 1

Inequality (*) can also be used to prove Property 1 above: $D(p \parallel q) \geq 0$ with equality in and only if $p = q$.

Recall : $\displaystyle\sum_{k=1}^{n} a_k \log \frac{a_k}{b_k} \geq \left(\sum_{k=1}^{n} a_k\right) \log \frac{\sum_{k=1}^{n} a_k}{\sum_{k=1}^{n} b_k}$ with equality if and only if $a_k/b_k = c$.

We have

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \geq \left(\sum_{x} p(x)\right) \log \frac{\sum_x p(x)}{\sum_x q(x)} = 1 \log \frac{1}{1} = 0$$

with equality if and only if $\frac{p(x)}{q(x)} = c$, that is if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$ (since, by normalization, $c = 1$).

Two important consequences

If we choose q to be the uniform distribution on $\mathcal{X}$, we have

$$\begin{aligned} 0 &\leq D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log p(x) + p(x) \log n = -H(X) + \log(n) \end{aligned}$$

Thus, the two above properties of D provide an alternate proof of the following facts

- $H(X) \leq \log n$ with equality if and only if p is the uniform distribution.
- $H(p)$ is a concave function of p.

Entropy of a pair of r.v.

If X and Y is a pair of discrete r.v. with distribution $p(x, y)$, their *joint entropy* is defined by

$$H(X, Y) := -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) = -E[\log p(X, Y)]$$

The *conditional entropy* of Y given X is defined by

$$H(Y|X) := -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) = -E[\log p(Y|X)]$$

Note: Using the decomposition $p(x, y) = p(x)p(y|x)$ we derive that $H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x)$ where $H(Y|X = x) := H(p(Y|X = x))$.

Chain Rule

$$H(X, Y) := -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

The joint and conditional entropy are related by the formula

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

This result follows by using $\log p(x, y) = \log p(y|x) + \log p(x)$ and taking the expectation.

Likewise we have: $H(X, Y) = H(X|Y) + H(Y)$

Mutual Information

Let X and Y be two r.v. with probability distribution $p(x,y)$ and marginal distributions $p(x)$ and $p(y)$. The mutual information of X and Y is defined by

$$I(X;Y) := D(p(x,y) \parallel p(x)p(y)) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Properties of I

1. *Symmetric*: $I(X;Y) = I(Y;X)$. (trivial)

2. *Nonnegative:* $I(X;Y) \geq 0$ and $I(X;Y) = 0$ if and only if X and Y are independent. (it follows from the property of D)

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

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

5. $I(X;X) = H(X)$ (it follows from 4: $I(X;X) = 2H(X) - H(X,X) = H(X)$)

Proof of property 3

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

We use the decomposition $p(x,y) = p(x|y)p(y)$:

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

$I(X;Y) = H(Y) - H(Y|X)$ is proved as above by interchanging X with Y.

Interpretation of I

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

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

Properties 2 and 3 imply that $H(X) \geq H(X|Y)$ with equality if and only if X and Y are independent. This means that measuring Y reduces (on the average!) the entropy of X.

Example: Let $\mathcal{X} = \mathcal{Y} = \{0,1\}$, $p(0,0) = 0$, $p(0,1) = \frac{3}{4}$, $p(1,0) = p(1,1) = \frac{1}{8}$. Verify that $H(X) = 0.544$, $H(X|Y) = \frac{1}{4}$, $H(X|Y=0) = 0$, $H(X|Y=1) = 1$.

Proof of property 4

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

This follows by combining property 3, $I(X;Y) = H(X) - H(X|Y)$ with the decomposition for the joint entropy, $H(X,Y) = H(Y) + H(X|Y)$.

One more property of I

If we look at the mutual information as a function of $p(x)$ and $p(y|x)$ (the remaining probabilities can be derived from those) we have the following result.

Lemma: $I(X,Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.

Concavity of I in $p(x)$

We have

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_{x \in \mathcal{X}} p(x) H(Y|X=x)$$

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

We know $H(Y)$ is concave in $p(y)$. If we keep $p(y|x)$ fixed then $p(y)$ is linear in $p(x)$ and, so, $H(Y)$ is also concave in $p(x)$.

The second term, $-\sum_{x \in \mathcal{X}} p(x) H(Y|X=x)$ is linear in $p(x)$ so it is concave in $p(x)$.

Convexity of I in $p(y|x)$

Let $p_1(y|x)$ and $p_2(y|x)$ be two conditional distributions and consider their convex combination

$$p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda) p_2(y|x), \quad \lambda \in [0,1]$$

Since $p(x)$ is fixed we have $p_\lambda(x,y) = p(x) p_\lambda(y|x)$ and

$$p_\lambda(y) = \lambda p_1(y) + (1-\lambda) p_2(y)$$

where $p_i(y) = \sum_{x \in \mathcal{X}} p(x) p_i(y|x)$, $i = 1, 2$. Now let $q_\lambda(x,y) = p(x) p_\lambda(y)$ and notice that

$$I(X;Y) = D(p_\lambda \parallel q_\lambda).$$

Since $D(\cdot \parallel \cdot)$ is a convex function then I is a convex function of the conditional distribution.

Link to channel coding

Suppose we wish to send the symbol x, generated with $p(x)$, through a noisy channel with transition probability $p(y|x)$. Unless $p(x|y)$ peaks at some x^*, we won't be able to recover x from y.

However, if we represent x with some "redundant code" it is possibleto recover x from y. The goal is to find an efficient coding strategy which guarantees that this error is small (zero in a limit process).

We will see that the "maximum rate" C at which we can transmit the the coded data x through the channel with arbitrary small probability of error is given by

$$C = max_{p(x)} I(X;Y)$$

Noisy typewriter channel

Let $\mathcal{X} = \mathcal{Y} = \{1, ..., 26\}$ and $p(y|x) = \frac{1}{2}$ if $y = x$ or $y = x + 1$ mod 26, and zero otherwise.

We have $C = \log 13$. In fact

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = \sum_{x \in \mathcal{X}} p(x) 1 = 1$$

and, thus,

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} \{H(Y) - 1\} = \log 26 - 1 = \log 13 \; bits.$$

The maximum is achieved when $p(x)$ is the uniform distribution.

Which code achieves the channel capacity?

Consider a code of unit length: $x(1) = 1$, $x(2) = 3$, $x(3) = 5$, etc. This code has zero probability of error because each codeword is either transmitted as such or as the next symbol in $\mathcal{X}$. This code achieves capacity since its transmission rate is $\log 13$ $bits$

Entropy of more than two r.v.

It is straightforward to extend these concepts to an $n-$tuple of r.v. $X_1, \ldots, X_N$. In particular we have the following chain rule:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \sum_{i=3}^{n} H(X_i|X_{i-1}, \ldots, X_1)$$

which follows by using the chain rule for probability:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1)$$

or, equivalently, $\log p(x_1, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i|x_{i-1}, \ldots, x_1)$

Note: Above, $p(x_i|x_{i-1}, \ldots, x_1)$ is meant to be $p(x_1)$ when $i = 1$ and $p(x_2|x_1)$ when $i = 2$.

Conditional entropy of two joint r.v.

We have

$$\begin{aligned} H(X,Y|Z) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} p(x,y,z) \log p(x,y|z) \\ &= \sum_{z} p(z) \sum_{x,y} p(x,y|z) \log p(x,y|z) \\ &= \sum_{z} p(z) H(X,Y|Z=z) \end{aligned}$$

A direct computation (as in the above case of two joint r.v.) gives

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

compare to $H(X,Y) = H(X) + H(Y|X)$

Conditional mutual information

If X, Y, Z are r.v., the conditional mutual information of X and Y given Z is defined by

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = E\left[\log \frac{p(X,Y,Z)}{p(X|Z)p(Y|Z)}\right]$$

Using the chain rule for the entropy we see that the mutual information satisfies the chain rule:

$$I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\ldots,X_1)$$

Conditional relative entropy

It is defined by the formula

$$\begin{aligned} D(p_{Y|X} \parallel q_{Y|X}) &= \sum_{x\in\mathcal{X}} p(x) D(p(\cdot|x) \parallel q(\cdot|x)) \\ &= \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log \frac{p(y|x)}{p(q|x)} \end{aligned}$$

Note We also denote (abusing notation) $D(p_{Y|X} \parallel q_{Y|X})$ by $D(p(y|x) \parallel q(y|x))$.

Chain rule for relative entropy

$$D(p(x,y) \parallel q(x,y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)).$$

Ordered Markov chain

We say that the r.v. X, Y, Z for a Markov chain *in that order* (we write $X \to Y \to Z$) if $p(z|y,x) = p(z|y)$, that is, Z is conditionally independent of X given Y. Thus, we have

$$p(x,y,z) = p(z|y,x)p(y|x)p(x) = p(x)p(y|x)p(z|y)$$

This condition is equivalent to ask that X and Z are conditionally independent give Y. In fact

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

In addition, we have that $X \to Y \to Z$ implies $Z \to Y \to X$ (check!).

- Particular case: if $Z = f(Y)$ then $X \to Y \to Z$.

Data processing inequality

If X, Y, Z form a Markov chain, no preprocessing of Y (deterministic or random) can increase the information that Y contains about X That is, if $X \to Y \to Z$, then

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

Proof: Using the chaining rule, we have

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

and, also,

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

where $I(Z;X|Y) = 0$ because from hypothesis Z and X are conditionally independent given Y. Thus $I(Z;X)+I(Y;X|Z) = I(Y;X)$ and since $I(X;Y|Z)$ is nonnegative we conclude that $I(Y;X) \geq I(Z;X)$, or, equivalently $I(X;Y) \geq I(X,Z)$.

Note: Similarly, we have that $I(Y;Z) \geq I(X;Z)$.

Bibliography

See Chapter 2 of
T.M. Cover and J.A. Thomas, *The elements of information theory*, Wiley, 1991.