Functional equations

Tom Leinster*

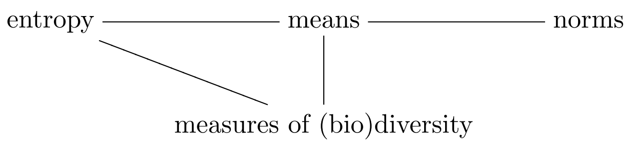
Spring 2017

Preamble

Hello.

Admin: email addresses; sections in outline $\neq$ lectures; pace.

Overall plan: interested in unique characterizations of ...



There are many ways to measure diversity: long controversy.

Ideal: be able to say 'If you want your diversity measure to have properties X, Y and Z, then it must be one of the following measures.'

Similar results have been proved for entropy, means and norms.

This is a tiny part of the field of functional equations!

One ulterior motive: for me to learn something about FEs. I'm not an expert, and towards end, this will get to edge of research (i.e. I'll be making it up as I go along).

Tools:

- native wit
- elementary real analysis
- (new!) some probabilistic methods.

One ref: Aczél and Daróczy, *On Measures of Information and Their Characterizations.* (Comments.) Other refs: will give as we go along.

1 Warm-up

Week I (7 Feb)

Which functions f satisfy $f(x + y) = f(x) + f(y)$? Which functions of two variables can be separated as a product of functions of one variable?

This section is an intro to basic techniques. We may or may not need the actual results we prove.

*School of Mathematics, University of Edinburgh; Tom.Leinster@ed.ac.uk. Last edited on 14 February 2017

The Cauchy functional equation

The Cauchy FE on a function $f\colon \mathbb{R} \to \mathbb{R}$ is

$$\forall x, y \in \mathbb{R}, \quad f(x+y) = f(x) + f(y). \tag{1}$$

There are some obvious solutions. Are they the only ones? Weak result first, to illustrate technique.

Proposition 1.1 *Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. TFAE (the following are equivalent):*

i. f satisfies (1)

ii. there exists $c \in \mathbb{R}$ such that

$$\forall x \in \mathbb{R}, \quad f(x) = cx.$$

If these conditions hold then $c = f(1)$.

Proof (ii)$\Rightarrow$(i) and last part: obvious.

Now assume (i). Differentiate both sides of (1) with respect to x:

$$\forall x, y \in \mathbb{R}, \quad f'(x+y) = f'(x).$$

Take $x = 0$: then $f'(y) = f'(0)$ for all $y \in \mathbb{R}$. So f' is constant, so there exist $c, d \in \mathbb{R}$ such that

$$\forall x \in \mathbb{R}, \quad f(x) = cx + d.$$

Substituting back into (1) gives $d = 0$, proving (ii). □

'Differentiable' is a much stronger condition than necessary!

Theorem 1.2 *As for Proposition 1.1, but with 'continuous' in place of 'differentiable'.*

Proof Let f be a continuous function satisfying (1).

- $f(0+0) = f(0) + f(0)$, so $f(0) = 0$.
- $f(x) + f(-x) = f(x + (-x)) = f(0) = 0$, so $f(-x) = -f(x)$. Cf. group homomorphisms.
- Next, $f(nx) = nf(x)$ for all $x \in \mathbb{R}$ and $n \in \mathbb{Z}$. For $n > 0$, true by induction. For $n = 0$, says $f(0) = 0$. For $n < 0$, have $-n > 0$ and so $f(nx) = -f(-nx) = -(-nf(x)) = nf(x)$.
- In particular, $f(n) = nf(1)$ for all $n \in \mathbb{Z}$.
- For $m, n \in \mathbb{Z}$ with $n \neq 0$, we have

  $$f(n \cdot m/n) = f(m) = mf(1)$$

  but also

  $$f(n \cdot m/n) = nf(m/n),$$

  so $f(m/n) = (m/n)f(1)$. Hence $f(x) = f(1)x$ for all $x \in \mathbb{Q}$.

- Now f and $x \mapsto f(1)x$ are continuous functions on $\mathbb{R}$ agreeing on $\mathbb{Q}$, hence are equal. □

Remarks 1.3 i. 'Continuous' can be relaxed further still. It was pointed out in class that continuity at 0 is enough. 'Measurable' is also enough (Fréchet, 'Pri la funkcia ekvacio $f(x+y) = f(x) + f(y)$', 1913). Even weaker: 'bounded on some set of positive measure'. But never mind! For this course, I'll be content to assume continuity.

ii. To get a 'weird' solution of Cauchy FE (i.e. not of the form $x \mapsto cx$), need existence of a non-measurable function. So, need some form of choice. So, can't really construct one.

iii. Assuming choice, weird solutions exist. Choose basis B for the vector space $\mathbb{R}$ over $\mathbb{Q}$. Pick $b_0 \neq b_1$ in B and a function $\phi\colon B \to \mathbb{R}$ such that $\phi(b_0) = 0$ and $\phi(b_1) = 1$. Extend to $\mathbb{Q}$-linear map $f\colon \mathbb{R} \to \mathbb{R}$. Then $f(b_0) = 0$ with $b_0 \neq 0$, but $f \not\equiv 0$ since $f(b_1) = 1$. So f cannot be of the form $x \mapsto cx$. But f satisfies the Cauchy functional equation, by linearity.

Variants (got by using the group isomorphism $(\mathbb{R}, +) \cong ((0,\infty), 1)$ defined by exp and log):

Corollary 1.4 *i. Let $f\colon \mathbb{R} \to (0,\infty)$ be a continuous function. TFAE:*

- *$f(x+y) = f(x)f(y)$ for all x, y*
- *there exists $c \in \mathbb{R}$ such that $f(x) = e^{cx}$ for all x.*

ii. Let $f\colon (0,\infty) \to \mathbb{R}$ be a continuous function. TFAE:

- *$f(xy) = f(x) + f(y)$ for all x, y*
- *there exists $c \in \mathbb{R}$ such that $f(x) = c \log x$ for all x.*

iii. Let $f\colon (0,\infty) \to (0,\infty)$ be a continuous function. TFAE:

- *$f(xy) = f(x)f(y)$ for all x, y*
- *there exists $c \in \mathbb{R}$ such that $f(x) = x^c$ for all x.*

Proof For (i), define $g\colon \mathbb{R} \to \mathbb{R}$ by $g(x) = \log f(x)$. Then g is continuous and satisfies Cauchy FE, so $g(x) = cx$ for some constant c, and then $f(x) = e^{cx}$.

(ii) and (iii): similarly, putting $g(x) = f(e^x)$ and $g(x) = \log f(e^x)$. □

Related:

Theorem 1.5 (Erdős?) *Let $f\colon \mathbb{Z}^+ \to (0,\infty)$ be a function satisfying $f(mn) = f(m)f(n)$ for all $m, n \in \mathbb{Z}^+$. (There are loads of solutions: can freely choose $f(p)$ for every prime p. But ...) Suppose that either $f(1) \leq f(2) \leq \cdots$ or*

$$\lim_{n\to\infty} \frac{f(n+1)}{f(n)} = 1.$$

Then there exists $c \in \mathbb{R}$ such that $f(n) = n^c$ for all n.

Proof Omitted. □

Separation of variables

When can a function of two variables be written as a product/sum of two functions of one variable? We'll do sums, but can convert to products as in Corollary 1.4.

Let X and Y be sets and

$$f\colon X \times Y \to \mathbb{R}$$

a function. Or can replace $\mathbb{R}$ by any abelian group. We seek functions

$$g\colon X \to \mathbb{R}, \qquad h\colon Y \to \mathbb{R}$$

such that

$$\forall x \in X, y \in Y, \quad f(x,y) = g(x) + h(y). \tag{2}$$

Basic questions:

A Are there *any* pairs of functions (g, h) satisfying (2)?

B How can we construct all such pairs?

C How many such pairs are there? Clear that if there are any, there are many, by adding/subtracting constants.

I got up to here in the first class, and was going to lecture the rest of this section in the second class, but in the end decided not to. What I actually lectured resumes at the start of Section 2. But for completeness, here's the rest of this section.

Attempt to recover g and h from f. Key insight:

$$f(x,y) - f(x_0,y) = g(x) - g(x_0)$$

$(x, x_0 \in X,\ y \in Y)$. No hs involved!

First lemma: g and h are determined by f, up to additive constant.

Lemma 1.6 *Let $g\colon X \to \mathbb{R}$ and $h\colon Y \to \mathbb{R}$ be functions. Define $f\colon X \times Y \to \mathbb{R}$ by (2). Let $x_0 \in X$ and $y_0 \in Y$.*

Then there exist $c, d \in \mathbb{R}$ such that $c + d = f(x_0, y_0)$ and

$$g(x) = f(x, y_0) - c \quad \forall x \in X, \tag{3}$$
$$h(y) = f(x_0, y) - d \quad \forall y \in Y. \tag{4}$$

Proof Put $y = y_0$ in (2): then

$$g(x) = f(x, y_0) - c \quad \forall x \in X$$

where $c = h(y_0)$. Similarly

$$h(y) = f(x_0, y) - d \quad \forall y \in Y$$

where $d = g(x_0)$. Now

$$c + d = g(x_0) + h(y_0) = f(x_0, y_0)$$

by (2). □

But given f (and x_0 and y_0), is every pair (g, h) of this form a solution of (2)? Not necessarily (but it's easy to say when)...

Lemma 1.7 *Let $f\colon X \times Y \to \mathbb{R}$ be a function. Let $x_0 \in X$, $y_0 \in Y$, and $c, d \in \mathbb{R}$ with $c + d = f(x_0, y_0)$. Define $g\colon X \to \mathbb{R}$ by (3) and $h\colon Y \to \mathbb{R}$ by (4). If*

$$f(x, y_0) + f(x_0, y) = f(x, y) + f(x_0, y_0) \quad \forall x \in X, y \in Y$$

then

$$f(x, y) = g(x) + h(y) \quad \forall x \in X, y \in Y.$$

Proof For all $x \in X$ and $y \in Y$,

$$g(x) + h(y) = f(x, y_0) + f(x_0, y) - c - d = f(x, y_0) + f(x_0, y) - f(x_0, y_0),$$

etc. □

Can now answer the basic questions.

Existence of decompositions (A):

Proposition 1.8 *Let $f\colon X \times Y \to \mathbb{R}$. TFAE:*

i. there exist $g\colon X \to \mathbb{R}$ and $h\colon Y \to \mathbb{R}$ such that

$$f(x, y) = g(x) + h(y) \quad \forall x \in X, y \in Y$$

ii. $f(x, y') + f(x', y) = f(x, y) + f(x', y')$ for all x, x', y, y'.

Proof (i)⇒(ii): trivial.

(ii)⇒(i): trivial if $X = \emptyset$ or $Y = \emptyset$. Otherwise, choose $x_0 \in X$ and $y_0 \in Y$; then use Lemma 1.7 with $c = 0$ and $d = f(x_0, y_0)$. □

Classification of decompositions (B):

Proposition 1.9 *Let $f\colon X \times Y \to \mathbb{R}$ be a function satisfying the equivalent conditions of Proposition 1.8, and let $x_0 \in X$ and $y_0 \in Y$. Then a pair of functions $(g\colon X \to \mathbb{R}, h\colon Y \to \mathbb{R})$ satisfies (2) if and only if there exist $c, d \in \mathbb{R}$ satisfying $c + d = f(x_0, y_0)$, (3) and (4).*

Proof Follows from Lemmas 1.6 and 1.7. □

Number of decompositions (C) (really: dim of solution-space):

Corollary 1.10 *Let $f\colon X \times Y \to \mathbb{R}$ with X, Y nonempty.* Either *there are no pairs (g, h) satisfying (2),* or *for any pair (g, h) satisfying (2), the set of all such pairs is the 1-dimensional space*

$$\{(g + a, h - a) : a \in \mathbb{R}\}.$$ □

2 Shannon entropy

Week II (14 Feb)

Recap, including Erdős theorem. No separation of variables!

The many meanings of the word *entropy*. Ordinary entropy, relative entropy, conditional entropy, joint entropy, cross entropy; entropy on finite and infinite spaces; quantum versions; entropy in topological dynamics; . . . . Today we stick to the very simplest kind: Shannon entropy of a probability distribution on a finite set.

Let $\mathbf{p} = (p_1, \ldots, p_n)$ be a probability distribution on $\{1, \ldots, n\}$ (i.e. $p_i \geq 0$, $\sum p_i = 1$). The **(Shannon) entropy** of $\mathbf{p}$ is

$$H(\mathbf{p}) = -\sum_{i\colon p_i>0} p_i \log p_i = \sum_{i\colon p_i>0} p_i \log \frac{1}{p_i}.$$

The sum is over all $i \in \{1, \ldots, n\}$ such that $p_i \neq 0$; equivalently, can sum over all $i \in \{1, \ldots, n\}$ but with the convention that $0 \log 0 = 0$.

Ways of thinking about entropy:

- Disorder.
- Uniformity. Will see that uniform distribution has greatest entropy among all distributions on $\{1, \ldots, n\}$.
- Expected surprise. Think of $\log(1/p_i)$ as your surprise at learning that an event of probability p_i has occurred. The smaller p_i is, the more surprised you are. Then $H(\mathbf{p})$ is the expected value of the surprise: how surprised you expect to be!
- Information. Similar to expected surprise. Think of $\log(1/p_i)$ as the information that you gain by observing an event of probability p_i. The smaller p_i is, the rarer the event is, so the more remarkable it is. Then $H(\mathbf{p})$ is the average amount of information per event.
- Lack of information (!). Dual viewpoints in information theory. E.g. if $\mathbf{p}$ represents noise, high entropy means more noise. Won't go into this.
- Genericity. In context of thermodynamics, entropy measures how generic a state a system is in. Closely related to 'lack of information'.

First properties:

- $H(\mathbf{p}) \geq 0$ for all $\mathbf{p}$, with equality iff $\mathbf{p} = (0, \ldots, 0, 1, 0, \ldots, 0)$. Least uniform distribution.
- $H(\mathbf{p}) \leq \log n$ for all $\mathbf{p}$, with equality iff $\mathbf{p} = (1/n, \ldots, 1/n)$. Most uniform distribution. Proof that $H(\mathbf{p}) \leq \log n$ uses concavity of log:

$$H(\mathbf{p}) = \sum_{i\colon p_i>0} p_i \log\left(\tfrac{1}{p_i}\right) \leq \log\left(\sum_{i\colon p_i>0} p_i \tfrac{1}{p_i} \right) \leq \log n.$$

- $H(\mathbf{p})$ is continuous in $\mathbf{p}$. (Uses $\lim_{x\to 0+} x \log x = 0$.)

Remark 2.1 Base of logarithm usually taken to be e (for theory) or 2 (for examples and in information theory/digital communication). Changing base of logarithm scales H by constant factor—harmless!

Examples 2.2 Use $\log_2$ here.

i. Uniform distribution on 2^k elements:

$$H(\tfrac{1}{2^k}, \ldots, \tfrac{1}{2^k}) = \log_2(2^k) = k.$$

Interpretation: knowing results of k fair coin tosses gives k bits of information.

ii.

$$\begin{aligned} H(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}) &= \tfrac{1}{2}\log_2 2 + \tfrac{1}{4}\log_2 4 + \tfrac{1}{8}\log_2 8 + \tfrac{1}{8}\log_2 8 \\ &= \tfrac{1}{2}\cdot 1 + \tfrac{1}{4}\cdot 2 + \tfrac{1}{8}\cdot 3 + \tfrac{1}{8}\cdot 3 \\ &= 1\tfrac{3}{4}. \end{aligned}$$

Interpretation: consider a language with alphabet A, B, C, D, with frequencies $1/2, 1/4, 1/8, 1/8$. We want to send messages encoded in binary. Compare Morse code: use short code sequences for common letters. The most efficient unambiguous code encodes a letter of frequency 2^{-k} as a binary string of length k: e.g. here, could use

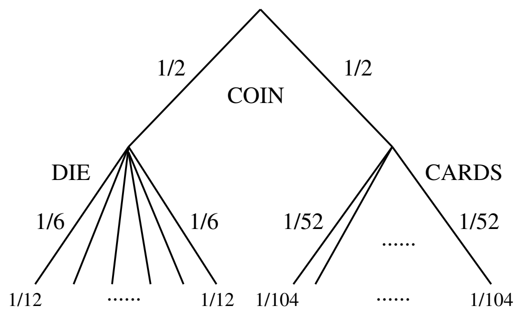
A: 0, B: 10, C: 110, D: 111.

Then code messages are unambiguous: e.g. 11010011110 can only be CBADB. Since $k = \log_2(1/2^{-k})$, mean number of bits per letter is then $\sum_i p_i \log(1/p_i) = H(\mathbf{p}) = 1\frac{3}{4}$.

iii. That example was special in that all the probabilities were integer powers of 2. But... Can still make sense of this when probabilities aren't powers of 2 (Shannon's first theorem). E.g. frequency distribution $\mathbf{p} = (p_1, \ldots, p_{26})$ of letters in English has $H(\mathbf{p}) \approx 4$, so can encode English in about 4 bits/letter. So, it's as if English had only 16 letters, used equally often.

Will now explain a more subtle property of entropy. Begin with example.

Example 2.3 Flip a coin. If it's heads, roll a die. If it's tails, draw from a pack of cards. So final outcome is either a number between 1 and 6 or a card. There are $6 + 52 = 58$ possible final outcomes, with probabilities as shown (assuming everything unbiased):



How much information do you expect to get from observing the outcome?

- You know result of coin flip, giving $H(1/2, 1/2) = 1$ bit of info.
- With probability 1/2, you know result of die roll: $H(1/6, \dots, 1/6) = \log_2 6$ bits of info.
- With probability 1/2, you know result of card draw: $H(1/52, \dots, 1/52) = \log_2 52$ bits.

In total:

$$1 + \tfrac{1}{2}\log_2 6 + \tfrac{1}{2}\log_2 52$$

bits of info. This suggests

$$H\big(\underbrace{\tfrac{1}{12}, \dots, \tfrac{1}{12}}_{6}, \underbrace{\tfrac{1}{104}, \dots, \tfrac{1}{104}}_{52}\big) = 1 + \tfrac{1}{2}\log_2 6 + \tfrac{1}{2}\log_2 52.$$

Can check true! Now formulate general rule.

The chain rule Write

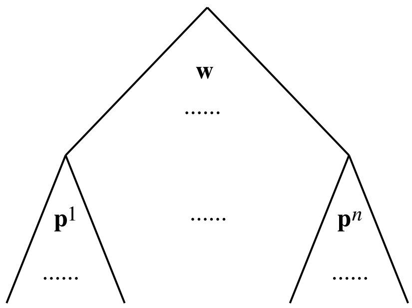
$$\Delta_n = \{\text{probability distributions on } \{1, \dots, n\}\}.$$

Geometrically, this is a simplex of dimension $n-1$. Given

$$\mathbf{w} \in \Delta_n, \quad \mathbf{p}^1 \in \Delta_{k_1}, \dots, \mathbf{p}^n \in \Delta_{k_n},$$

get composite distribution

$$\mathbf{w} \circ (\mathbf{p}^1, \dots, \mathbf{p}^n) = (w_1 p^1_1, \dots, w_1 p^1_{k_1}, \ \dots, \ w_n p^n_1, \dots, w_n p^n_{k_n}) \in \Delta_{k_1 + \cdots + k_n}$$



(For cognoscenti: this defines an operad structure on the simplices.)

Easy calculation[1] proves **chain rule**:

$$H(\mathbf{w} \circ (\mathbf{p}^1, \dots, \mathbf{p}^n)) = H(\mathbf{w}) + \sum_{i=1}^n w_i H(\mathbf{p}^i).$$

Special case: $\mathbf{p}^1 = \cdots = \mathbf{p}^n$. For $\mathbf{w} \in \Delta_n$ and $\mathbf{p} \in \Delta_m$, write

$$\mathbf{w} \otimes \mathbf{p} = \mathbf{w} \circ (\underbrace{\mathbf{p}, \dots, \mathbf{p}}_{n}) = (w_1 p_1, \dots, w_1 p_m, \ \dots, \ w_n p_1, \dots, w_n p_m) \in \Delta_{nm}.$$

[1] This is completely straightforward, but can be made even more transparent by first observing that the function $f(x) = -x \log x$ is a 'nonlinear derivation', i.e. $f(xy) = xf(y) + f(x)y$. In fact, $-x \log x$ is the *only* measurable function F with this property (up to a constant factor), since if we put $g(x) = F(x)/x$ then $g(xy) = g(y) + g(x)$ and so $g(x) \propto \log x$.

This is joint probability distribution if the two things are independent. Then chain rule implies **multiplicativity**:

$$H(\mathbf{w} \otimes \mathbf{p}) = H(\mathbf{w}) + H(\mathbf{p}).$$

Interpretation: information from two independent observations is sum of information from each.

Where are the functional equations?

For each $n \geq 1$, have function $H \colon \Delta_n \to \mathbb{R}^+ = [0, \infty)$. Faddeev[2] showed:

Theorem 2.4 (Faddeev, 1956) *Take functions* $\left(I \colon \Delta_n \to \mathbb{R}^+\right)_{n \geq 1}$. *TFAE:*

i. the functions I *are continuous and satisfy the chain rule;*

ii. $I = cH$ *for some* $c \in \mathbb{R}^+$.

That is: up to a constant factor, Shannon entropy is uniquely characterized by continuity and chain rule.

Should we be disappointed to get *scalar multiples* of H, not H itself? No: recall that different scalar multiples correspond to different choices of the base for log.

Rest of this section: proof of Faddeev's theorem.

Certainly (ii)$\Rightarrow$(i). Now take I satisfying (i).

Write $\mathbf{u}_n = (1/n, \ldots, 1/n) \in \Delta_n$. Strategy: think about the sequence $\left(I(\mathbf{u}_n)\right)_{n \geq 1}$. It should be $(c \log n)_{n \geq 1}$ for some constant c.

Lemma 2.5 *i.* $I(\mathbf{u}_{mn}) = I(\mathbf{u}_m) + I(\mathbf{u}_n)$ *for all* $m, n \geq 1$.

ii. $I(\mathbf{u}_1) = 0$.

Proof For (i), $\mathbf{u}_{mn} = \mathbf{u}_m \otimes \mathbf{u}_n$, so

$$I(\mathbf{u}_{mn}) = I(\mathbf{u}_m \otimes \mathbf{u}_n) = I(\mathbf{u}_m) + I(\mathbf{u}_n)$$

(by multiplicativity). For (ii), take $m = n = 1$ in (i). □

Theorem 1.5 (Erdős) *would* now tell us that $I(\mathbf{u}_n) = c \log n$ for some constant c (putting $f(n) = \exp(I(\mathbf{u}_n))$). But to conclude that, we need one of the two alternative hypotheses of Theorem 1.5 to be satisfied. We prove the second one, on limits. This takes some effort.

Lemma 2.6 $I(1, 0) = 0$.

Proof We compute $I(1, 0, 0)$ in two ways. First,

$$I(1,0,0) = I((1,0) \circ ((1,0), \mathbf{u}_1)) = I(1,0) + 1 \cdot I(1,0) + 0 \cdot I(\mathbf{u}_1) = 2I(1,0).$$

Second,

$$I(1,0,0) = I((1,0) \circ (\mathbf{u}_1, \mathbf{u}_2)) = I(1,0) + 1 \cdot I(\mathbf{u}_1) + 0 \cdot I(\mathbf{u}_2) = I(1,0)$$

since $I(\mathbf{u}_1) = 0$. Hence $I(1,0) = 0$. □

[2] Dmitry Faddeev, father of the physicist Ludvig Faddeev.

To use Erdős, need $I(\mathbf{u}_{n+1}) - I(\mathbf{u}_n) \to 0$ as $n \to \infty$. Can *nearly* prove that:

Lemma 2.7 *$I(\mathbf{u}_{n+1}) - \frac{n}{n+1} I(\mathbf{u}_n) \to 0$ as $n \to \infty$.*

Proof We have

$$\mathbf{u}_{n+1} = \left(\tfrac{n}{n+1}, \tfrac{1}{n}\right) \circ (\mathbf{u}_n, \mathbf{u}_1),$$

so by the chain rule and $I(\mathbf{u}_1) = 0$,

$$I(\mathbf{u}_{n+1}) = I\left(\tfrac{n}{n+1}, \tfrac{1}{n}\right) + \tfrac{n}{n+1} I(\mathbf{u}_n).$$

So

$$I(\mathbf{u}_{n+1}) - \tfrac{n}{n+1} I(\mathbf{u}_n) = I\left(\tfrac{n}{n+1}, \tfrac{1}{n+1}\right) \to I(1,0) = 0$$

as $n \to \infty$, by continuity and Lemma 2.6. □

To improve this to $I(\mathbf{u}_{n+1}) - I(\mathbf{u}_n) \to 0$, use a general result that has nothing to do with entropy:

Lemma 2.8 *Let $(a_n)_{n \geq 1}$ be a sequence in $\mathbb{R}$ such that $a_{n+1} - \frac{n}{n+1} a_n \to 0$ as $n \to \infty$. Then $a_{n+1} - a_n \to 0$ as $n \to \infty$.*

Proof Omitted; uses Cesàro convergence.[3]

Although I omitted this proof in class, I'll include it here. I'll follow the argument in Feinstein, The Foundations of Information Theory, *around p.7.*

It is enough to prove that $a_n/(n+1) \to 0$ as $n \to \infty$. Write $b_1 = a_1$ and $b_n = a_n - \frac{n-1}{n} a_{n-1}$ for $n \geq 2$. Then $n a_n = n b_n + (n-1) a_{n-1}$ for all $n \geq 2$, so

$$n a_n = n b_n + (n-1) b_{n-1} + \cdots + 1 b_1$$

for all $n \geq 1$. Dividing through by $n(n+1)$ gives

$$\frac{a_n}{n+1} = \tfrac{1}{2} \cdot \mathrm{mean}(b_1, b_2, b_2, b_3, b_3, b_3, \ldots, \underbrace{b_n, \ldots, b_n}_{n}).$$

Since $b_n \to 0$ as $n \to \infty$, the sequence

$$b_1, b_2, b_2, b_3, b_3, b_3, \ldots, \underbrace{b_n, \ldots, b_n}_{n}, \ldots$$

also converges to 0. Now a general result of Cesàro states that if a sequence (x_r) converges to ℓ then the sequence $(\bar{x}_r)$ also converges to ℓ, where $\bar{x}_r = (x_1 + \cdots + x_r)/r$. Applying this to the sequence above implies that

$$\mathrm{mean}(b_1, b_2, b_2, b_3, b_3, b_3, \ldots, \underbrace{b_n, \ldots, b_n}_{n}) \to 0 \text{ as } n \to \infty.$$

Hence $a_n/(n+1) \to 0$ as $n \to \infty$, as required. □

We can now deduce what $I(\mathbf{u}_n)$ is:

Lemma 2.9 *There exists $c \in \mathbb{R}^+$ such that $I(\mathbf{u}_n) = c \log n$ for all $n \geq 1$.*

[3] Xīlíng Zhāng pointed out that this is also a consequence of Stolz's lemma—or as Wikipedia calls it, the Stolz–Cesàro theorem.

Proof We have $I(\mathbf{u}_{mn}) = I(\mathbf{u}_m) + I(\mathbf{u}_n)$, and by last two lemmas,

$$I(\mathbf{u}_{n+1}) - I(\mathbf{u}_n) \to 0 \text{ as } n \to \infty.$$

So can apply Erdős's theorem (1.5) with $f(n) = \exp(I(\mathbf{u}_n))$ to get $f(n) = n^c$ for some constant $c \in \mathbb{R}$. So $I(\mathbf{u}_n) = c \log n$, and $c \geq 0$ since I maps into $\mathbb{R}^+$.□

We now know that $I = cH$ on the *uniform* distributions $\mathbf{u}_n$. It might seem like we still have a mountain to climb to get to $I = cH$ for *all* distributions. But in fact, it's easy.

Lemma 2.10 $I(\mathbf{p}) = cH(\mathbf{p})$ *whenever* $p_1, \ldots, p_n$ *are rational.*

Proof Write

$$\mathbf{p} = \left(\frac{k_1}{k}, \ldots, \frac{k_n}{k} \right)$$

where $k_1, \ldots, k_n \in \mathbb{Z}$ and $k = k_1 + \cdots + k_n$. Then

$$\mathbf{p} \circ (\mathbf{u}_{k_1}, \ldots, \mathbf{u}_{k_n}) = \mathbf{u}_k.$$

Since I satisfies the chain rule and $I(\mathbf{u}_r) = cH(\mathbf{u}_r)$ for all r,

$$I(\mathbf{p}) + \sum_{i=1}^{n} p_i \cdot cH(\mathbf{u}_{k_i}) = cH(\mathbf{u}_k).$$

But since cH also satisfies the chain rule,

$$cH(\mathbf{p}) + \sum_{i=1}^{n} p_i \cdot cH(\mathbf{u}_{k_i}) = cH(\mathbf{u}_k),$$

giving the result. □

Theorem 2.4 follows by continuity.

Next time: relative entropy.