Functional equations

Tom Leinster*

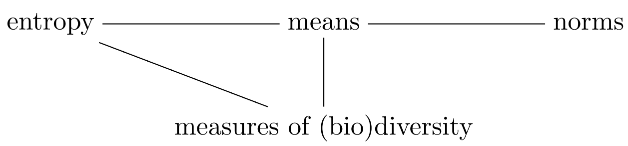
Spring 2017

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 22 March 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 $\mathbf{p} = (p_1, \dots, p_n)$ be a probability distribution on $\{1, \dots, n\}$ (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 $i \in \{1, \dots, n\}$ such that $p_i \neq 0$; equivalently, can sum over all $i \in \{1, \dots, n\}$ 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 $\{1, \dots, n\}$.
- 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 $\mathbf{p} = (0, \dots, 0, 1, 0, \dots, 0)$. Least uniform distribution.
- $H(\mathbf{p}) \leq \log n$ for all $\mathbf{p}$, with equality iff $\mathbf{p} = (1/n, \dots, 1/n)$. 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

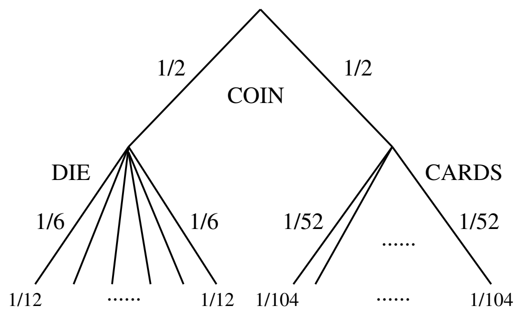
A: 0, B: 10, C: 110, D: 111.

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

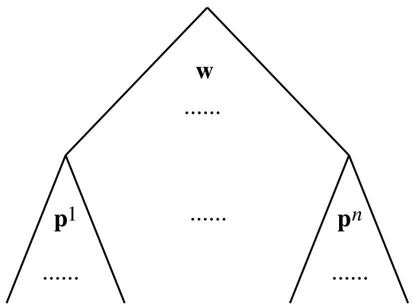
$$\Delta_n = \{\text{probability distributions on } \{1, \dots, n\}\}.$$

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.

Week III (21 Feb)

Recap of last time: Δ_n, H, chain rule.

Information is a slippery concept to reason about. One day it will seem intuitively clear that the distribution $(0.5, 0.5)$ is 'more informative' than $(0.9, 0.1)$, and the next day your intuition will say the opposite. So to make things concrete, it's useful to concentrate on one particular framework: coding.

Slogan:

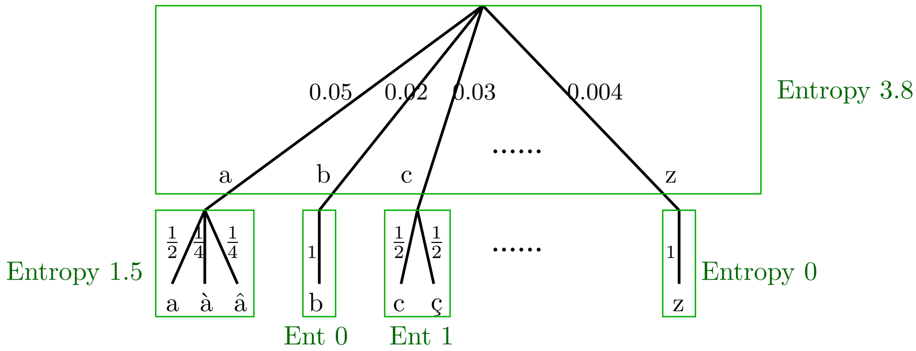
Entropy is average number of bits/symbol in an optimal encoding.

Consider language with alphabet A, B, ..., used with frequencies $p_1, p_2, \ldots, p_n$. Want to encode efficiently in binary.

- Revisit Example 2.2(ii).
- In general: if $p_1, \ldots, p_n$ are all powers of 2, there is an unambiguous encoding where ith letter is encoded as string of length $\log_2(1/p_i)$. So mean bits/symbol $= \sum_i p_i \log_2(1/p_i) = H(\mathbf{p})$.

- Shannon's first theorem: for any $\mathbf{p} \in \Delta_n$, in any unambiguous encoding, mean bits/symbol $\geq H(\mathbf{p})$; moreover, if you're clever, can do it in $< H(\mathbf{p}) + \varepsilon$ for any $\varepsilon > 0$.
- When p_is aren't powers of 2, do this by using *blocks* of symbols rather than individual symbols.

Chain rule: how many bits/symbol on average to encode French, including accents?



We need

$$\underbrace{3.8}_{\text{bits for actual letters}} + \underbrace{\big(0.05 \times 1.5 + 0.02 \times 0 + 0.03 \times 1 + \cdots + 0.004 \times 0\big)}_{\text{bits for accents}}$$

bits/symbol. (Convention: *letters* are a, b, c, ...; *symbols* are a, à, â, b, c, ç, ...) By the chain rule, this is equal to the entropy of the composite distribution

$$\big(0.05 \times \tfrac{1}{2}, 0.05 \times \tfrac{1}{4}, 0.05 \times \tfrac{1}{4}, 0.02 \times 1, \ \ldots, \ 0.004 \times 1\big).$$

Relative entropy

Let $\mathbf{p}, \mathbf{r} \in \Delta_n$. The **entropy of p relative to r** is

$$H(\mathbf{p} \,\|\, \mathbf{r}) = \sum_{i\colon p_i > 0} p_i \log\Big(\frac{p_i}{r_i}\Big).$$

Also called **Kullback–Leibler divergence**, **relative information**, or **information gain**.

First properties:

- $H(\mathbf{p} \,\|\, \mathbf{r}) \geq 0$. Not obvious, as $\log(p_i/r_i)$ is sometimes positive and sometimes negative. For since log is concave,

$$H(\mathbf{p} \,\|\, \mathbf{r}) = -\sum_{i\colon p_i>0} p_i \log\Big(\frac{r_i}{p_i}\Big) \geq -\log\Big(\sum_{i\colon p_i>0} p_i \frac{r_i}{p_i}\Big) \geq -\log 1 = 0.$$

- $H(\mathbf{p} \| \mathbf{r}) = 0$ if and only if $\mathbf{p} = \mathbf{r}$. Evidence so far suggests relative entropy is something like a distance. That's wrong in that it's not a metric, but it's not too terribly wrong. Will come back to this.

- $H(\mathbf{p} \,\|\, \mathbf{r})$ can be arbitrarily large (even for fixed n). E.g.

$$H((1/2, 1/2) \,\|\, (t, 1-t)) \to \infty \text{ as } t \to \infty,$$

  and in fact $H(\mathbf{p} \,\|\, \mathbf{r}) = \infty$ if $p_i > 0 = r_i$ for some i.

- Write

$$\mathbf{u}_n = (1/n, \ldots, 1/n) \in \Delta_n.$$

  Then

$$H(\mathbf{p} \,\|\, \mathbf{u}_n) = \log n - H(\mathbf{p}).$$

  So entropy is pretty much a special case of relative entropy.

- $H(\mathbf{p} \,\|\, \mathbf{r}) \neq H(\mathbf{r} \,\|\, \mathbf{p})$. E.g.

$$\begin{aligned} &H(\mathbf{u}_2 \,\|\, (0,1)) = \infty, \\ &H((0,1) \,\|\, \mathbf{u}_2) = \log 2 - H((0,1)) = \log 2. \end{aligned}$$

  Will come back to this too.

Coding interpretation Convenient fiction: for each 'language' $\mathbf{p}$, there is an encoding for $\mathbf{p}$ using $\log(1/p_i)$ bits for the ith symbol, hence with mean bits/symbol $= H(\mathbf{p})$ exactly. Call this 'machine $\mathbf{p}$'.

We have

$$\begin{aligned} H(\mathbf{p} \,\|\, \mathbf{r}) &= \sum p_i \log\Big(\frac{1}{r_i}\Big) - \sum p_i \log\Big(\frac{1}{p_i}\Big) \\ &= (\text{bits/symbol to encode language } \mathbf{p} \text{ using machine } \mathbf{r}) \\ &\quad - (\text{bits/symbol to encode language } \mathbf{p} \text{ using machine } \mathbf{p}) \end{aligned}$$

So relative entropy is the number of extra bits needed if you use the wrong machine. Or: penalty you pay for using the wrong machine. Explains why $H(\mathbf{p} \,\|\, \mathbf{r}) \geq 0$ with equality if $\mathbf{p} = \mathbf{r}$.

If $r_i = 0$ then in machine $\mathbf{r}$, the ith symbol has an infinitely long code word. Or if you like: if $r_i = 2^{-1000}$ then its code word has length 1000. So if also $p_i > 0$ then for language $\mathbf{p}$ encoded using machine $\mathbf{r}$, average bits/symbol $= \infty$. This explains why $H(\mathbf{p} \,\|\, \mathbf{r}) = \infty$.

Taking $\mathbf{r} = \mathbf{u}_n$,

$$H(\mathbf{p} \,\|\, \mathbf{u}_n) = \log n - H(\mathbf{p}) \leq \log n.$$

Explanation: in machine $\mathbf{u}_n$, *every* symbol is encoded with $\log n$ bits, so the average extra bits/symbol caused by using machine $\mathbf{u}_n$ instead of machine $\mathbf{p}$ is $\leq \log n$.

Now a couple of slightly more esoteric comments, pointing in different mathematical directions (both away from functional equations). Tune out if you want...

Measure-theoretic perspective Slogan:

All entropy is relative.

Attempt to generalize definition of entropy from probability measures on finite sets to arbitrary probability measures μ: want to say $H(\mu) = -\int \log(\mu)\, d\mu$, but this makes no sense!

Note that the finite definition $H(p) = -\sum p_i \log p_i$ implicitly refers to counting measure...

However, can generalize *relative* entropy. Given measures μ and ν on measurable space X, define

$$H(\mu \,\|\, \nu) = \int_X \log\Big(\frac{d\mu}{d\nu}\Big)\, d\mu$$

where $\frac{d\mu}{d\nu}$ is Radon–Nikodym derivative. *This* makes sense and is the right definition.

People do talk about the entropy of probability distributions on $\mathbb{R}^n$. For instance, the entropy of a probability density function f on $\mathbb{R}$ is usually defined as $H(f) = -\int_{\mathbb{R}} f(x) \log f(x)\, dx$, and it's an important result that among all density functions on $\mathbb{R}$ with a given mean and variance, the one with the maximal entropy is the normal distribution. (This is related to the central limit theorem.) But here we're implicitly using Lebesgue measure λ on $\mathbb{R}$; so there are two measures in play, λ and $f\lambda$, and $H(f) = H(f\lambda \,\|\, \lambda)$.

Local behaviour of relative entropy Take two close-together distributions $\mathbf{p}, \mathbf{p} + \boldsymbol{\delta} \in \Delta_n$. (So $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ with $\sum \delta_i = 0$.) Taylor expansion gives

$$H(\mathbf{p} + \boldsymbol{\delta} \,\|\, \mathbf{p}) \approx \frac{1}{2} \sum \frac{1}{p_i} \delta_i^2$$

for $\boldsymbol{\delta}$ small. Precisely:

$$H(\mathbf{p} + \boldsymbol{\delta} \,\|\, \mathbf{p}) = \frac{1}{2} \sum \frac{1}{p_i} \delta_i^2 + O(\|\boldsymbol{\delta}\|^3) \text{ as } \boldsymbol{\delta} \to \mathbf{0}.$$

(Here $\|\cdot\|$ is any norm on $\mathbb{R}^n$. It doesn't matter which, as they're all equivalent.) So:

Locally, $H(- \,\|\, -)$ is like a squared distance.

In particular, locally (to second order) it's symmetric.

The square root of relative entropy is *not* a metric on Δ_n: not symmetric and fails triangle inequality. (E.g. put $\mathbf{p} = (0.9, 0.1)$, $\mathbf{q} = (0.2, 0.8)$, $\mathbf{r} = (0.1, 0.9)$. Then $\sqrt{H(\mathbf{p} \,\|\, \mathbf{q})} + \sqrt{H(\mathbf{q} \,\|\, \mathbf{r})} < \sqrt{H(\mathbf{p} \,\|\, \mathbf{r})}$.) But using it as a 'local distance' leads to important things, e.g. Fisher information (statistics), the Jeffreys prior (Bayesian statistics), and the whole subject of information geometry.

Next week: a unique characterization of relative entropy.

Recap: entropy as mean bits/symbol in optimal encoding; definition of relative entropy; relative entropy as extra cost of using wrong machine. Week IV (28 Feb)

Write

$$A_n = \{(\mathbf{p}, \mathbf{r}) \in \Delta_n \times \Delta_n : r_i = 0 \Longrightarrow p_i = 0\}.$$

So $H(\mathbf{p} \,\|\, \mathbf{r}) < \infty \iff (\mathbf{p}, \mathbf{r}) \in A_n$. (Secretly, A stands for 'absolutely continuous'.) Then for each $n \geq 1$, we have the function

$$H(- \,\|\, -) \colon A_n \to \mathbb{R}^+.$$

Properties:

Measurability $H(- \,\|\, -)$ is measurable. If you don't know what that means, ignore: *very* mild condition. Every function that anyone has ever written down a formula for, or ever will, is measurable.

Permutation-invariance E.g. $H((p_1, p_2, p_3) \,\|\, (r_1, r_2, r_3)) = H((p_2, p_3, p_1) \,\|\, (r_2, r_3, r_1))$. It's *that* kind of symmetry, not the kind where you swap $\mathbf{p}$ and $\mathbf{r}$.

Vanishing $H(\mathbf{p} \,\|\, \mathbf{p}) = 0$ for all $\mathbf{p}$. Remember: $H(\mathbf{p} \,\|\, \mathbf{r})$ is *extra* cost of using machine $\mathbf{r}$ (instead of machine $\mathbf{p}$) to encode language $\mathbf{p}$.

Chain rule For all

$$(\mathbf{w}, \widetilde{\mathbf{w}}) \in A_n, \ (\mathbf{p}^1, \widetilde{\mathbf{p}}^1) \in A_{k_1}, \ \ldots, \ (\mathbf{p}^n, \widetilde{\mathbf{p}}^n) \in A_{k_n},$$

we have

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

To understand chain rule for relative entropy:

Example 2.11 Recall the letter/accent tree from last time.

Swiss French and Canadian French are written using the same alphabet and accents, but with slightly different words, hence different frequencies of letters and accents.

Consider Swiss French and Canadian French:

$$\begin{aligned} \mathbf{w} \in \Delta_{26} &: \text{frequencies of letters in Swiss} \\ \widetilde{\mathbf{w}} \in \Delta_{26} &: \text{frequencies of letters in Canadian} \end{aligned}$$

and then

$$\begin{aligned} \mathbf{p}^1 \in \Delta_3 &: \text{frequencies of accents on 'a' in Swiss} \\ \widetilde{\mathbf{p}}^1 \in \Delta_3 &: \text{frequencies of accents on 'a' in Canadian} \\ &\vdots \\ \mathbf{p}^{26} \in \Delta_1 &: \text{frequencies of accents on 'z' in Swiss} \\ \widetilde{\mathbf{p}}^{26} \in \Delta_1 &: \text{frequencies of accents on 'z' in Canadian.} \end{aligned}$$

So

$$\mathbf{w} \circ (\mathbf{p}^1, \ldots, \mathbf{p}^{26}) = \text{frequency distribution of all symbols in Swiss}$$
$$\widetilde{\mathbf{w}} \circ (\widetilde{\mathbf{p}}^1, \ldots, \widetilde{\mathbf{p}}^{26}) = \text{frequency distribution of all symbols in Canadian}$$

where 'symbol' means a letter with (or without) an accent.

Now encode Swiss using Canadian machine. How much extra does it cost (in mean bits/symbol) compared to encoding Swiss using Swiss machine?

$$\text{mean extra cost per symbol} = \text{mean extra cost per letter} + \text{mean extra cost per accent}.$$

And that's the chain rule. Coefficient in sum is w_i, not $\widetilde{w}_i$, because it's Swiss we're encoding.

Clearly any scalar multiple of relative entropy also has these four properties (measurability, permutation-invariance, vanishing, chain rule).

Theorem 2.12 *Take functions* $\big(I(- \,\|\, -)\colon A_n \to \mathbb{R}^+\big)_{n\geq 1}$. *TFAE:*

i. the functions I satisfy measurability, permutation-invariance, vanishing and the chain rule;

ii. $I(- \,\|\, -) = cH(- \,\|\, -)$ *for some* $c \in \mathbb{R}^+$.

It's hard to believe that Theorem 2.12 is new; it could have been proved in the 1950s. However, I haven't found it in the literature. The proof that follows is inspired by work of Baez and Fritz, who in turn built on and corrected work of Petz. Cf. Baez and Fritz (arXiv:1402.3067) and Petz (cited by B&F).

Take $I(- \,\|\, -)$ satisfying (i). Define $L\colon (0, 1] \to \mathbb{R}^+$ by

$$L(\alpha) = I((1, 0) \,\|\, (\alpha, 1 - \alpha)).$$

Note $((1, 0), (\alpha, 1 - \alpha)) \in A_2$, so RHS is defined. If $I = H$ then $L(\alpha) = -\log \alpha$.

Lemma 2.13 *Let* $(\mathbf{p}, \mathbf{r}) \in A_n$ *with* $p_{k+1} = \cdots = p_n = 0$, *where* $1 \leq k \leq n$. *Then* $r_1 + \cdots + r_k > 0$, *and*

$$I(\mathbf{p} \,\|\, \mathbf{r}) = L(r_1 + \cdots + r_k) + I(\mathbf{p}' \,\|\, \mathbf{r}')$$

where

$$\mathbf{p}' = (p_1, \ldots, p_k), \qquad \mathbf{r}' = \frac{(r_1, \ldots, r_k)}{r_1 + \cdots + r_k}.$$

Proof Case $k = n$ trivial; suppose $k < n$.

Since $\mathbf{p}$ is a probability distribution with $p_i = 0$ for all $i > k$, there is some $i \leq k$ such that $p_i > 0$, and then $r_i > 0$ since $(\mathbf{p}, \mathbf{r}) \in A_n$. Hence $r_1 + \cdots + r_k > 0$. By definition of operadic composition,

$$I(\mathbf{p} \,\|\, \mathbf{r}) = I\Big((1, 0) \circ (\mathbf{p}', \mathbf{r}'') \,\Big\|\, (r_1 + \cdots + r_k,\, r_{k+1} + \cdots + r_n) \circ (\mathbf{r}', \mathbf{r}'')\Big)$$

where $\mathbf{r}''$ is the normalization of $(r_{k+1}, \ldots, r_n)$ if $r_{k+1} + \cdots + r_n > 0$, or is chosen arbitrarily in Δ_{n-k} otherwise. (The set Δ_{n-k} is nonempty since $k < n$.) By chain rule, this is equal to

$$L(r_1 + \cdots + r_k) + 1 \cdot I(\mathbf{p}' \,\|\, \mathbf{r}') + 0 \cdot I(\mathbf{r}'' \,\|\, \mathbf{r}''),$$

and result follows. In order to use the chain rule, we needed to know that various pairs were in A_2, A_k, etc.; that's easily checked. □

Lemma 2.14 *$L(\alpha\beta) = L(\alpha) + L(\beta)$ for all $\alpha, \beta \in (0, 1]$.*

Proof Consider

$$x := I((1, 0, 0) \parallel (\alpha\beta, \alpha(1-\beta), 1-\alpha)).$$

On one hand, Lemma 2.13 with $k = 1$ gives

$$x = L(\alpha\beta) + I((1) \parallel (1)) = L(\alpha\beta)$$

On other, Lemma 2.13 with $k = 2$ gives

$$x = L(\alpha) + I((1, 0) \parallel (\beta, 1-\beta)) = L(\alpha) + L(\beta).$$

Result follows. □

Lemma 2.15 *There is a unique constant $c \in \mathbb{R}^+$ such that $L(\alpha) = -c \log \alpha$ for all $\alpha \in (0, 1]$.*

Proof Follows from Lemma 2.14 and measurability, as in Cor 1.4. That corollary was stated under the hypothesis of continuity, but measurability would have been enough. □

Now we come to a clever part of Baez and Fritz's argument.

Lemma 2.16 *Let $(\mathbf{p}, \mathbf{r}) \in A_n$ and suppose that $p_i > 0$ for all i. Then $I(\mathbf{p} \| \mathbf{r}) = cH(\mathbf{p} \parallel \mathbf{r})$.*

Proof We have $(\mathbf{p}, \mathbf{r}) \in A_n$, so $r_i > 0$ for all i. So can choose $\alpha \in (0, 1]$ such that $r_i - \alpha p_i \geq 0$ for all i.

We will compute the (well-defined) number

$$x := I\Big((p_1, \ldots, p_n, \underbrace{0, \ldots, 0}_{n}) \Big\| (\alpha p_1, \ldots, \alpha p_n, r_1 - \alpha p_1, \ldots, r_n - \alpha p_n)\Big)$$

in two ways. First, by Lemma 2.13 and the vanishing property,

$$x = L(\alpha) + I(\mathbf{p} \parallel \mathbf{p}) = -c \log \alpha.$$

Second, by symmetry and then the chain rule,

$$\begin{aligned}
x &= I((p_1, 0, \ldots, p_n, 0) \parallel (\alpha p_1, r_1 - \alpha p_1, \ldots, p_n, r_n - \alpha p_n)) \\
&= I\Big(\mathbf{p} \circ \big((1, 0), \ldots, (1, 0)\big) \Big\| \mathbf{r} \circ \big((\alpha \tfrac{p_1}{r_1}, 1 - \alpha \tfrac{p_1}{r_1}), \ldots, (\alpha \tfrac{p_n}{r_n}, 1 - \alpha \tfrac{p_n}{r_n})\big)\Big) \\
&= I(\mathbf{p} \parallel \mathbf{r}) + \sum_{i=1}^{n} p_i L\big(\alpha \tfrac{p_i}{r_i}\big) \\
&= I(\mathbf{p} \parallel \mathbf{r}) - c \log \alpha - cH(\mathbf{p} \parallel \mathbf{r}).
\end{aligned}$$

Comparing the two expressions for x gives the result. □

Proof of Theorem 2.12 Let $(\mathbf{p}, \mathbf{r}) \in A_n$. By symmetry, can assume

$$p_1 > 0, \ \ldots, \ p_k > 0, \ p_{k+1} = 0, \ \ldots, \ p_n = 0$$

where $1 \leq k \leq n$. Writing $R = r_1 + \cdots + r_k$,

$$\begin{aligned} I(\mathbf{p} \,\|\, \mathbf{r}) &= L(R) + I((p_1, \ldots, p_k) \,\|\, \tfrac{1}{R}(r_1, \ldots, r_k)) && \text{by Lemma 2.13} \\ &= -c \log R + cH((p_1, \ldots, p_k) \,\|\, \tfrac{1}{R}(r_1, \ldots, r_k)) && \text{by Lemmas 2.15 and 2.16.} \end{aligned}$$

This holds for all $I(- \,\|\, -)$ satisfying the four conditions; in particular, it holds for $cH(- \,\|\, -)$. Hence $I(\mathbf{p} \,\|\, \mathbf{r}) = cH(\mathbf{p} \,\|\, \mathbf{r})$. □

Remark 2.17 Assuming permutation-invariance, the chain rule is equivalent to a very special case:

$$\begin{aligned} &H((tw_1, (1-t)w_1, w_2, \ldots, w_n) \,\|\, (\widetilde{t}\widetilde{w}_1, (1-\widetilde{t})\widetilde{w}_1, \widetilde{w}_2, \ldots, \widetilde{w}_n)) \\ &\qquad = H(\mathbf{w} \,\|\, \widetilde{\mathbf{w}}) + w_1 H((t, 1-t) \,\|\, (\widetilde{t}, 1-\widetilde{t})) \end{aligned}$$

for all $(\mathbf{w}, \widetilde{\mathbf{w}}) \in A_n$ and $((t, 1-t), (\widetilde{t}, 1-\widetilde{t})) \in A_2$. (Proof: induction.)

Next week: I'll introduce a family of 'deformations' or 'quantum versions' of entropy and relative entropy. This turns out to be important if we want a balanced perspective on what biodiversity is.

Week V (7 Mar)

Recap: formulas for entropy and cross entropy (in both log(1/something) and −log(something) forms); chain rules for both.

Remark 2.18 I need to make explicit something we've been using implicitly for a while.

Let $\mathbf{p} \in \Delta_n$ and $\mathbf{r} \in \Delta_m$. Then we get a new distribution

$$\begin{aligned} \mathbf{p} \otimes \mathbf{r} &= (p_1 r_1, \ldots, p_1 r_m, \\ &\qquad \vdots \\ &\qquad p_n r_1, \ldots, p_n r_m) \\ &= \mathbf{p} \circ (\underbrace{\mathbf{r}, \ldots, \mathbf{r}}_{n}) \\ &\in \Delta_{nm}. \end{aligned}$$

Probabilistically, this is the joint distribution of independent random variables distributed according to $\mathbf{p}$ and $\mathbf{r}$.

A special case of the chain rule for entropy:

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

and similarly for relative entropy:

$$H(\mathbf{p} \otimes \mathbf{r} \,\|\, \widetilde{\mathbf{p}} \otimes \widetilde{\mathbf{r}}) = H(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) + H(\mathbf{r} \,\|\, \widetilde{\mathbf{r}}).$$

(So H and $H(- \,\|\, -)$ are log-like; like a higher version of Cauchy's functional equation.)

3 Deformed entropies

Shannon entropy is just a single member of a one-parameter family of entropies. In fact, there are *two* different one-parameter families of entropies, both containing Shannon entropy as a member. In some sense, these two families of entropies are equivalent, but they have different flavours.

I'll talk about both families: surprise entropies in this section, then Rényi entropies when we come to measures of diversity later on.

Definition 3.1 Let $q \in [0, \infty)$. The **q-logarithm** $\ln_q : (0, \infty) \to \mathbb{R}$ is defined by

$$\ln_q(x) = \begin{cases} \frac{x^{1-q}-1}{1-q} & \text{if } q \neq 1, \\ \ln(x) & \text{if } q = 1. \end{cases}$$

Notation: log vs. ln.

Then $\ln_q(x)$ is continuous in q (proof: l'Hôpital's rule). Warning:

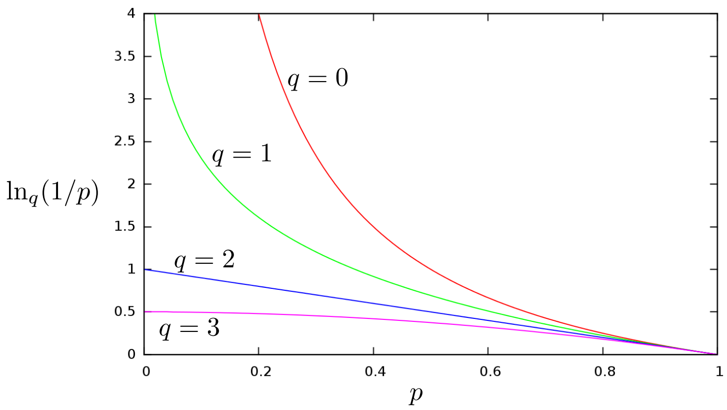
$$\ln_q(xy) \neq \ln_q(x) + \ln_q(y), \qquad \ln_q(1/x) \neq -\ln_q(x).$$

First one inevitable: we already showed that scalar multiples of *actual* logarithm are the only continuous functions that convert multiplication into addition. In fact, there's quite a neat formula for $\ln_q(xy)$ in terms of $\ln_q(x)$, $\ln_q(y)$ and q: exercise!

For a probability $p \in [0, 1]$, can view

$$\ln_q(1/p)$$

as one's 'surprise' at witnessing an event with probability p. (Decreasing in p; takes value 0 at $p = 1$.)



Definition 3.2 Let $q \in [0, \infty)$. The **surprise entropy of order** q of $\mathbf{p} \in \Delta_n$ is

$$S_q(\mathbf{p}) = \sum_{i \colon p_i > 0} p_i \ln_q(1/p_i) = \begin{cases} \frac{1}{1-q}\left(\sum p_i^q - 1\right) & \text{if } q \neq 1, \\ \sum p_i \ln(1/p_i) & \text{if } q = 1. \end{cases}$$

Interpretation: $S_q(\mathbf{p})$ is the *expected surprise* of an event drawn from $\mathbf{p}$.

Usually called 'Tsallis entropies', because Tsallis discovered them in physics in 1988, after Havrda and Charvat discovered and developed them in information theory in 1967, and after Vajda (1968), Daróczy (1970) and Sharma and Mittal (1975) developed them further, and after Patil and Taillie used them as measures of biodiversity in 1982.

Tsallis says the parameter q can be interpreted as the 'degree of non-extensivity'. When we come to diversity measures, I'll give another interpretation.

Special cases:

- $S_0(\mathbf{p}) = |\{i : p_i > 0\}| - 1$
- $S_2(\mathbf{p}) = 1 - \sum_i p_i^2 =$ probability that two random elements of $\{1, \ldots, n\}$ (chosen according to $\mathbf{p}$) are different.

Properties of S_q (for fixed q):

- Permutation-invariance, e.g. $S_q(p_2, p_3, p_1) = S_q(p_1, p_2, p_3)$.
- q-chain rule:

$$S_q(\mathbf{w} \circ (\mathbf{p}^1, \ldots, \mathbf{p}^n)) = S_q(\mathbf{w}) + \sum_{i:\, w_i > 0} w_i^q S_q(\mathbf{p}^i).$$

- q-multiplicativity: as special case of chain rule,

$$S_q(\mathbf{w} \otimes \mathbf{p}) = S_q(\mathbf{w}) + \Big(\sum_i w_i^q\Big) S_q(\mathbf{p}).$$

Theorem 3.3 *Let $q \in [0, \infty) \setminus \{1\}$. Let $\big(I \colon \Delta_n \to \mathbb{R}^+\big)_{n \geq 1}$ be functions. TFAE:*

i. I is permutation-invariant and q-multiplicative in sense above;

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

No regularity condition needed! And don't need full chain rule—just a special case, multiplicativity.

Following proof is extracted from Aczél and Daróczy's book (Theorem 6.3.9), but they make it look way more complicated. The key point is that the multiplicativity property is not symmetric.

Proof (ii)⇒(i) easy. Assume (i). By permutation-invariance,

$$I(\mathbf{p} \otimes \mathbf{r}) = I(\mathbf{r} \otimes \mathbf{p})$$

for all $\mathbf{p} \in \Delta_n$ and $\mathbf{r} \in \Delta_m$. So by q-multiplicativity,

$$I(\mathbf{p}) + \big(\sum p_i^q\big) I(\mathbf{r}) = I(\mathbf{r}) + \big(\sum r_i^q\big) I(\mathbf{p}),$$

hence

$$\big(1 - \sum r_i^q\big) I(\mathbf{p}) = \big(1 - \sum p_i^q\big) I(\mathbf{r}).$$

Now want to get the $\mathbf{p}$s on one side and the $\mathbf{r}$s on the other, to deduce that $I(\mathbf{p})$ is proportional to $1 - \sum p_i^q$. But need to be careful about division by zero. Take $\mathbf{r} = \mathbf{u}_2 = (1/2, 1/2)$: then

$$(1 - 2^{1-q}) I(\mathbf{p}) = \big(1 - \sum p_i^q\big) I(\mathbf{u}_2)$$

for all $\mathbf{p}$. But $q \neq 1$, so $1 - 2^{1-q} \neq 0$. So $I = cS_q$ where $c = I(\mathbf{u}_2) \frac{1-q}{2^{1-q}-1}$. □

Relative entropy generalizes easily too, i.e. extends all along the family from the point $q = 1$. Only ticklish point is $\ln_q(1/x)$ vs. $-\ln_q(x)$.

Definition 3.4 Let $q \in [0, \infty)$. For $\mathbf{p}, \mathbf{r} \in \Delta_n$, the **relative surprise entropy of order** q is

$$S_q(\mathbf{p} \,\|\, \mathbf{r}) = -\sum_{i\colon p_i > 0} p_i \ln_q\Big(\frac{r_i}{p_i}\Big) = \begin{cases} \frac{1}{1-q}\Big(1 - \sum p_i^q r_i^{1-q}\Big) & \text{if } q \neq 1, \\ \sum p_i \log\big(\frac{p_i}{r_i}\big) & \text{if } q = 1. \end{cases}$$

Properties:

- Permutation-invariance (as in case $q = 1$).
- q-chain rule:

$$S_q(\mathbf{w} \circ (\mathbf{p}^1, \ldots, \mathbf{p}^n) \| \widetilde{\mathbf{w}} \circ (\widetilde{\mathbf{p}}^1, \ldots, \widetilde{\mathbf{p}}^n)) = S_q(\mathbf{w} \| \widetilde{\mathbf{w}}) + \sum_{i\colon w_i > 0} w_i^q \widetilde{w}_i^{1-q} S_q(\mathbf{p}^i \| \widetilde{\mathbf{p}}^i).$$

- q-multiplicativity: as special case of chain rule,

$$S_q(\mathbf{w} \otimes \mathbf{p} \,\|\, \widetilde{\mathbf{w}} \otimes \widetilde{\mathbf{p}}) = S_q(\mathbf{w} \,\|\, \widetilde{\mathbf{w}}) + \Big(\sum_{i\colon w_i > 0} w_i^q \widetilde{w}_i^{1-q}\Big) S_q(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}).$$

Again, there's a ludicrously simple characterization theorem that needs no regularity condition. Nor does it need the vanishing condition of Theorem 2.12.

Recall notation:

$$A_n = \{(\mathbf{p}, \mathbf{r}) \in \Delta_n \times \Delta_n : r_i = 0 \implies p_i = 0\}.$$

Then $S_q(\mathbf{p} \,\|\, \mathbf{r}) < \infty \iff (\mathbf{p}, \mathbf{r}) \in A_n$.

Theorem 3.5 *Let $q \in [0, \infty) \setminus \{1\}$. Let $\big(I(- \,\|\, -)\colon A_n \to \mathbb{R}^+\big)_{n \geq 1}$ be functions. TFAE:*

i. $I(- \,\|\, -)$ is permutation-invariant and q-multiplicative in sense above;

ii. $I(- \,\|\, -) = cS_q(- \,\|\, -)$ for some $c \in \mathbb{R}^+$.

Proof (ii)⇒(i) easy. Assume (i). By permutation-invariance,

$$I(\mathbf{p} \otimes \mathbf{r} \,\|\, \widetilde{\mathbf{p}} \otimes \widetilde{\mathbf{r}}) = I(\mathbf{r} \otimes \mathbf{p} \,\|\, \widetilde{\mathbf{r}} \otimes \widetilde{\mathbf{p}})$$

for all $(\mathbf{p}, \widetilde{\mathbf{p}}) \in A_n$ and $(\mathbf{r}, \widetilde{\mathbf{r}}) \in A_m$. So by q-multiplicativity,

$$I(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) + \Big(\sum p_i^q \widetilde{p}_i^{1-q}\Big) I(\mathbf{r} \,\|\, \widetilde{\mathbf{r}}) = I(\mathbf{r} \,\|\, \widetilde{\mathbf{r}}) + \Big(\sum r_i^q \widetilde{r}_i^{1-q}\Big) I(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}),$$

hence

$$\Big(1 - \sum r_i^q \widetilde{r}_i^{1-q}\Big) I(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) = \Big(1 - \sum p_i^q \widetilde{p}_i^{1-q}\Big) I(\mathbf{r} \,\|\, \widetilde{\mathbf{r}}).$$

Take $\mathbf{r} = (1, 0)$ and $\widetilde{\mathbf{r}} = \mathbf{u}_2$: then

$$(1 - 2^{q-1}) I(\mathbf{p} \,\|\, \widetilde{\mathbf{p}}) = I((1, 0) \,\|\, \mathbf{u}_2)\big(1 - \sum p_i^q \widetilde{p}_i^{1-q}\big)$$

for all $(\mathbf{p}, \widetilde{\mathbf{p}}) \in A_n$. But $q \neq 1$, so $1 - 2^{q-1} \neq 0$. So $I(- \,\|\, -) = cS_q(- \,\|\, -)$ where $c = I((1, 0) \,\|\, \mathbf{u}_2) \frac{1-q}{2^{1-q} - 1}$. □

Next week: how to use probability theory to solve functional equations.

4 Probabilistic methods

Week VI (14 Mar)

How to use probability theory to solve functional equations.

References: Aubrun and Nechita, arXiv:1102.2618, S.R.S. Varadhan, Large deviations.

Preview I want to make this broadly accessible, so I need to spend some time explaining the background before I can show how to actually solve functional equations using probability theory. We won't reach the punchline until next time. But to give the *rough* idea...

Functional equations have no stochastic element to them. So how could probability theory possibly help to solve them?

Basic idea: use probability theory to replace *complicated, exact* formulas by *simple, approximate* formulas.

Sometimes, an approximation is all you need.

E.g. (very simple example) multiply out the expression

$$(x+y)^{1000} = (x+y)(x+y)\cdots(x+y).$$

What terms do we obtain?

- Exact but complicated answer: every term is of the form $x^i y^j$ with $0 \leq i \leq 1000$ and $i + j = 1000$, and this term occurs $1000!/i!j!$ times.
- Simple but approximate answer: most terms are of the form $x^i y^j$ where i and j are about 500. (Flip a fair coin 1000 times, and you'll usually get about 500 heads.)

Use probability theory to get descriptions like second. (Different tools of probability theory allow us to be give more or less specific meanings to 'mostly' and 'about', depending on how good an approximation we need.)

Aubrun and Nechita used this method to characterize the p-norms. Can also use their method to characterize means, diversity measures, etc.

We'll need two pieces of background: basic result on large deviations, basic result on convex duality, then how they come together.

Cramér's large deviation theorem

Let X_1, X_2, ... be independent identically distributed (IID) real random variables. Write

$$\overline{X}_r = \tfrac{1}{r}(X_1 + \cdots + X_r)$$

(also a random variable). Fix $x \in \mathbb{R}$, and consider behaviour of $\mathbb{P}(\overline{X}_r \geq x)$ for large r.

- Law of large numbers gives

$$Pr(\overline{X}_r \geq x) \to \begin{cases} 1 & \text{if } x < \mathbb{E}(X), \\ 0 & \text{if } x > \mathbb{E}(X) \end{cases}$$

  where X is distributed identically to $X_1, X_2, \ldots$. But could ask for more fine-grained information: how fast does $\mathbb{P}(\overline{X}_r \geq x)$ converge?

- Central limit theorem: $\overline{X}_r$ is roughly normal for each *individual* large r.
- Large deviation theory is orthogonal to CLT and tells us *rate* of convergence of $\mathbb{P}(\overline{X}_r \geq x)$ as $r \to \infty$.

Rough result: there is a constant $k(x)$ such that for large r,

$$\mathbb{P}(\overline{X}_r \geq x) \approx k(x)^r.$$

If $x < \mathbb{E}(X)$ then $k(x) = 1$. Focus on $x > \mathbb{E}(X)$; then $k(x) < 1$ and $\mathbb{P}(\overline{X}_r \geq x)$ decays exponentially with r.

Theorem 4.1 (Cramér) *The limit*

$$\lim_{r\to\infty} \mathbb{P}(\overline{X}_r \geq x)^{1/r}$$

exists and (we even have a formula!) is equal to

$$\inf_{\lambda \geq 0} \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda x}}.$$

This is a standard result. Nice short proof: Cerf and Petit, arXiv:1002.3496. C&P state it without any kind of hypothesis on finiteness of moments; is that correct?

Remarks 4.2 i. $\mathbb{E}(e^{\lambda X})$ is by definition the **moment generating function** (MGF) $m_X(\lambda)$ of X. For all the standard distributions, the MGF is known and easy to look up, so this inf on the RHS is easy to compute. E.g. dead easy for normal distribution.

ii. The limit is equal to

$$\begin{cases} 1 & \text{if } x \leq \mathbb{E}(X), \\ \inf_{\lambda\in\mathbb{R}} \frac{\mathbb{E}(e^{\lambda X})}{e^{\lambda x}} & \text{if } x \geq \mathbb{E}(X). \end{cases}$$

Not too hard to deduce.

Example 4.3 Let $c_1, \ldots, c_n \in \mathbb{R}$. Let $X, X_1, X_2, \ldots$ take values $c_1, \ldots, c_n$ with probability $1/n$ each. (If some c_is are same, increase probabilities accordingly: e.g. for $7, 7, 8$, our random variables take value 7 with probability $2/3$ and 8 with probability $1/3$.) Then:

- For $x \in \mathbb{R}$,

$$\mathbb{P}(\overline{X}_r \geq x) = \tfrac{1}{n^r}\big|\{(i_1, \ldots, i_r) : c_{i_1} + \cdots + c_{i_r} \geq rx\}\big|.$$

- For $\lambda \in \mathbb{R}$,

$$\mathbb{E}(e^{\lambda x}) = \tfrac{1}{n}\big(e^{c_1\lambda} + \cdots + e^{c_n\lambda}\big)$$

- So by Cramér, for $x \in \mathbb{R}$,

$$\lim_{r\to\infty} \big|\{(i_1, \ldots, i_r) : c_{i_1} + \cdots + c_{i_r} \geq rx\}\big|^{1/r} = \inf_{\lambda\geq 0} \frac{e^{c_1\lambda} + \cdots + e^{c_n\lambda}}{e^{x\lambda}}.$$

So, we've used probability theory to say something nontrivial about a completely deterministic situation.

To get further, need a second tool: convex duality.

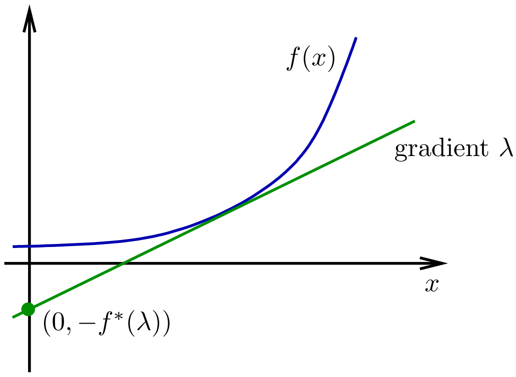
Convex duality

Let $f\colon \mathbb{R} \to [-\infty, \infty]$ be a function. Its **convex conjugate** or **Legendre-Fenchel transform** is the function $f^*\colon \mathbb{R} \to [-\infty, \infty]$ defined by

$$f^*(\lambda) = \sup_{x\in\mathbb{R}}(\lambda x - f(x)).$$

Can show f^* is convex, hence the name.

If f is differentiable, then f^* describes y-intercept of tangent line as a function of its gradient.



In proper generality, f is a function on a real vector space and f^* is a function on its dual space. In even more proper generality, there's a nice paper by Simon Willerton exhibiting this transform as a special case of a general categorical construction (arXiv:1501.03791).

Example 4.4 If $f, g\colon \mathbb{R} \to \mathbb{R}$ are differentiable convex functions with $f' = (g')^{-1}$ then $g = f^*$ and $f = g^*$. (E.g. true if $f(x) = x^p/p$ and $g(x) = x^q/q$ with $1/p + 1/q = 1$—'conjugate exponents'.)

Theorem 4.5 (Fenchel–Moreau) *If $f\colon \mathbb{R} \to \mathbb{R}$ is convex then $f^{**} = f$.*

Since conjugate of anything is convex, can only possibly have $f^{**} = f$ if f is convex. So this is best possible result.

Back to large deviations

Cramér's theorem secretly involves a convex conjugate. Let's see how.

Take IID real random variables $X, X_1, X_2, \ldots$ as before. For $x \geq \mathbb{E}(X)$, Cramér says

$$\lim_{r\to\infty} \mathbb{P}(\overline{X}_r \geq x)^{1/r} = \inf_{\lambda\in\mathbb{R}} \frac{m_X(\lambda)}{e^{\lambda x}}$$

where $m_X(\lambda) = \mathbb{E}(e^{\lambda X})$, i.e. (taking logs)

$$\begin{aligned}\lim_{r\to\infty} \tfrac{1}{r}\log \mathbb{P}(\overline{X}_r \geq x) &= \inf_{\lambda\in\mathbb{R}}(\log m_X(\lambda) - \lambda x) \\ &= -\sup_{\lambda\in\mathbb{R}}(\lambda x - \log m_X(\lambda)) \\ &= -(\log m_X)^*(x)\end{aligned}$$

or equivalently

$$(\log m_X)^*(x) = \lim_{r\to\infty} -\tfrac{1}{r}\log \mathbb{P}(\overline{X}_r \geq x).$$

Now some hand-waving. Ignoring fact that this only holds for $x \in \mathbb{E}(X)$, take conjugate of each side. Then for all $\lambda \geq 0$ (a restriction we need to make the hand-waving work),

$$(\log m_X)^{**}(\lambda) = \sup_{x\in\mathbb{R}}\bigl(\lambda x + \lim_{r\to\infty} \tfrac{1}{r}\log \mathbb{P}(\overline{X}_r \geq x)\bigr).$$

It's a general fact that $\log m_X$ (called the **cumulant generating function**) is convex. Hence $(\log m_X)^{**} = \log m_X$ by Fenchel–Moreau. So (taking exponentials)

$$m_X(\lambda) = \sup_{x\in\mathbb{R}} \lim_{r\to\infty}\Bigl(e^{\lambda x}\,\mathbb{P}(\overline{X}_r \geq x)^{1/r}\Bigr).$$

This really is true. It's a general formula for the moment generating function.

To make the hand-waving respectable: the full formula is

$$(\log m_X)^*(x) = \begin{cases} \lim_{r\to\infty} -\frac{1}{r}\log \mathbb{P}(\overline{X}_r \geq x) & \text{if } x \geq \mathbb{E}(X),\\ \lim_{r\to\infty} -\frac{1}{r}\log \mathbb{P}(\overline{X}_r \leq x) & \text{if } x \leq \mathbb{E}(X).\end{cases}$$

Can prove this by applying Cramér to $-X$ and $-x$. Then can use it to get the formula above for $(\log m_X)^{**}(\lambda)$ when $\lambda \geq 0$.

We'll use a variant:

Theorem 4.6 (Dual Cramér) *For any IID $X, X_1, X_2, \ldots$, for all $\lambda \geq 0$,*

$$m_X(\lambda) = \sup_{x\in\mathbb{R}}\sup_{r\geq 1}\Bigl(e^{\lambda x}\,\mathbb{P}(\overline{X}_r \geq x)^{1/r}\Bigr).$$

Proof See Cerf and Petit, who use it *to prove* Cramér. □

This is the result we'll use (not Cramér's theorem itself). Now let's apply it.

Example 4.7 As before, let $c_1, \ldots, c_n \in \mathbb{R}$ and consider uniform distribution on $c_1, \ldots, c_n$. Dual Cramér gives

$$e^{c_1\lambda} + \cdots + e^{c_n\lambda} = \sup_{x\in\mathbb{R}, r\geq 1}\Bigl(e^{\lambda x}\Bigl|\bigl\{(i_1,\ldots,i_r) : c_{i_1} + \cdots + c_{i_r} \geq rx\bigr\}\Bigr|^{1/r}\Bigr)$$

for all $\lambda \geq 0$.

Remember the background to all this: Aubrun and Nechita used large deviation theory to prove a unique characterization of the p-norms. Since the left-hand side here is a sum of powers (think $\lambda = p$), we can now begin to see the connection.

Next time: we'll exploit this expression for sums of powers. We'll use it to prove a theorem that pins down what's so special about the p-norms, and another theorem saying what's so special about power means.

Week VII (21 Mar)

Last time: Cramér's theorem and its convex dual. Crucial result: Example 4.7.

Today: we'll use this to give a unique characterization of the p-norms.

I like theorems arising from 'mathematical anthropology'. You observe some group of mathematicians and notice that they seem to place great importance on some particular object: for instance, algebraic topologists are always talking about simplicial sets, representation theorists place great emphasis on characters, certain kinds of analyst make common use of the Fourier transform, and other analysts often talk about the p-norms.

Then you can ask: why do they attach such importance to *that* object, not something slightly different? Is it the *only* object that enjoys the properties it enjoys? If not, why do they use the object they use and not some other object enjoying those properties? (Are they missing a trick?) And if it *is* the only object with those properties, we should be able to prove it!

We'll do this now for the p-norms, which are very standard in analysis.

Following theorem and proof are from Aubrun and Nechita, arXiv:1102.2618. I believe they were the pioneers of this probabilistic method for solving functional equations. See also *Tricki*, 'The tensor power trick'.

Notation: for a set I (which will always be finite for us), write

$$\mathbb{R}^I = \{\text{functions } I \to \mathbb{R}\} = \{\text{families } (x_i)_{i\in I} \text{ of reals}\}.$$

E.g. $\mathbb{R}^{\{1,\ldots,n\}} = \mathbb{R}^n$. In some sense we might as well *only* use $\mathbb{R}^n$, because every $\mathbb{R}^I$ is isomorphic to $\mathbb{R}^n$. But the notation will be smoother if we allow arbitrary finite sets.

Definition 4.8 Let I be a set. A **norm** on $\mathbb{R}^I$ is a function $\mathbb{R}^I \to \mathbb{R}^+$, written $\mathbf{x} \mapsto \|\mathbf{x}\|$, satisfying:

i. $\|\mathbf{x}\| = 0 \iff \mathbf{x} = 0$;

ii. $\|c\mathbf{x}\| = |c|\,\|\mathbf{x}\|$ for all $c \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^I$;

iii. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

There are many norms, but the following family comes up more often than any other.

Example 4.9 Let I be any finite set and $p \in [1,\infty]$. The ***p*-norm** $\|\cdot\|_p$ on $\mathbb{R}^I$ is defined for $\mathbf{x} = (x_i)_{i\in I} \in \mathbb{R}^I$ by

$$\|\mathbf{x}\|_p = \begin{cases} \left(\sum_{i\in I} |x_i|^p\right)^{1/p} & \text{if } p < \infty,\\ \max_{i\in I} |x_i| & \text{if } p = \infty.\end{cases}$$

Then $\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p$.

Aubrun and Nechita prove and use following result, which they don't quite make explicit:

Proposition 4.10 *For $p \in [1,\infty)$ and $\mathbf{x} \in (0,\infty)^n$,*

$$\|\mathbf{x}\|_p = \sup_{u>0,\, r\geq 1} \left(u \cdot \left|\{(i_1,\ldots,i_r) : x_{i_1}\cdots x_{i_r} \geq u^r\}\right|^{1/rp} \right).$$

Proof In Example 4.7, put $c_i = \log x_i$ and $\lambda = p$. Also substitute $u = e^x$, noting that the letter x now means something else. Then

$$x_1^p + \cdots + x_n^p = \sup_{u>0, r\geq 1} \Big(u^p \cdot \big|\{(i_1, \ldots, i_r) : x_{i_1} \cdots x_{i_r} \geq u^r\}\big|^{1/r} \Big).$$

Then take pth root of each side. □

Fix $p \in [1, \infty]$. For each finite set I, have the p-norm on $\mathbb{R}^I$. These enjoy some special properties. Special properties of the p-norms:

- We have
$$\|(x_3, x_2, x_1)\|_p = \|(x_1, x_2, x_3)\|_p \qquad (5)$$
(etc.) and
$$\|(x_1, x_2, x_3, 0)\|_p = \|(x_1, x_2, x_3)\|_p \qquad (6)$$
(etc). Generally, any injection $f\colon I \to J$ induces a map $f_*\colon \mathbb{R}^I \to \mathbb{R}^J$, defined by relabelling according to f and padding out with zeros, i.e.
$$(f_*\mathbf{x})_j = \begin{cases} x_i & \text{if } j = f(i) \text{ for some } i \in I, \\ 0 & \text{otherwise} \end{cases}$$
($\mathbf{x} \in \mathbb{R}^I$, $j \in J$). Then
$$\|f_*\mathbf{x}\|_p = \|\mathbf{x}\|_p \qquad (7)$$
for all injections $f\colon I \to J$ and $\mathbf{x} \in \mathbb{R}^I$. For permutations f of $\{1, \ldots, n\}$, (7) gives equations such as (5); for inclusion $\{1, \ldots, n\} \hookrightarrow \{1, \ldots, n, n+1\}$, (7) gives equations such as (6).

- For $A, B, x, y, z \in \mathbb{R}$, we have
$$\|(Ax, Ay, Az, Bx, By, Bz)\|_p = \|(A, B)\|_p \, \|(x, y, z)\|_p$$
(etc). Generally, for $\mathbf{x} = (x_i)_{i\in I} \in \mathbb{R}^I$ and $\mathbf{y} \in \mathbb{R}^J$, define
$$\mathbf{x} \otimes \mathbf{y} = (x_i y_j)_{i\in I, j\in J} \in \mathbb{R}^{I\times J}.$$
(If you identify $\mathbb{R}^I \otimes \mathbb{R}^J$ with $\mathbb{R}^{I\times J}$, as you can for finite sets, then $\mathbf{x} \otimes \mathbf{y}$ means what you think.) Then
$$\|\mathbf{x} \otimes \mathbf{y}\|_p = \|\mathbf{x}\|_p \, \|\mathbf{y}\|_p \qquad (8)$$
for all finite sets I and J, $\mathbf{x} \in \mathbb{R}^I$, and $\mathbf{y} \in \mathbb{R}^J$.

That's all!

Definition 4.11 A **system of norms** consists of a norm $\|\cdot\|$ on $\mathbb{R}^I$ for each finite set I, satisfying (7). This just guarantees that the norms on $\mathbb{R}^I$, for different sets I, hang together nicely. It is **multiplicative** if (8) holds. That's the crucial property of the p-norms.

Example 4.12 For each $p \in [1, \infty]$, the p-norm $\|\cdot\|_p$ is a multiplicative system of norms.

Theorem 4.13 (Aubrun and Nechita) *Every multiplicative system of norms is equal to $\|\cdot\|_p$ for some $p \in [1, \infty]$.*

Rest of today: prove this.

Let $\|\cdot\|$ be a multiplicative system of norms.

Step 1: elementary results

Lemma 4.14 *Let I be a finite set and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^I$.*

i. If $y_i = \pm x_i$ for each $i \in I$ then $\|\mathbf{x}\| = \|\mathbf{y}\|$.

ii. If $y_i = |x_i|$ for each $i \in I$ then $\|\mathbf{x}\| = \|\mathbf{y}\|$.

iii. If $0 \leq x_i \leq y_i$ for each $i \in I$ then $\|\mathbf{x}\| \leq \|\mathbf{y}\|$.

Proof *Omitted in class, but here it is.*

i. $\mathbf{x} \otimes (1, -1)$ is a permutation of $\mathbf{y} \otimes (1, -1)$, so

$$\|\mathbf{x} \otimes (1, -1)\| = \|\mathbf{y} \otimes (1, -1)\|,$$

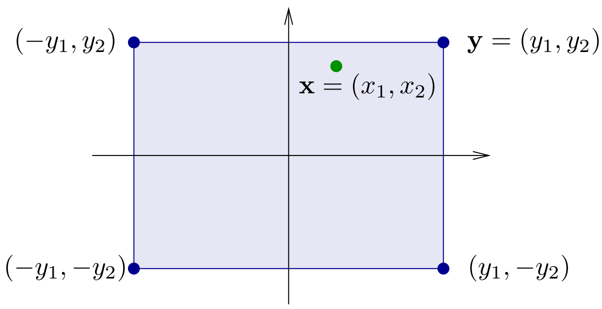
or equivalently

$$\|\mathbf{x}\| \, \|(1, -1)\| = \|\mathbf{y}\| \, \|(1, -1)\|,$$

and so $\|\mathbf{x}\| = \|\mathbf{y}\|$.

ii. Immediate from last part.

iii. For each $\mathbf{a} \in \{1, -1\}^I$, have vector $(a_i y_i)_{i \in I} \in \mathbb{R}^I$. Let S be the set of such vectors. So if I has n elements then S has 2^n elements. Then the convex hull of S is $\prod_{i \in I} [-y_i, y_i]$, which contains $\mathbf{x}$:



But every vector in S has norm $\|\mathbf{y}\|$ (by first part), so $\|\mathbf{x}\| \leq \|\mathbf{y}\|$ by triangle inequality. □

Step 2: finding p Write

$$\mathbf{1}_n = (1, \ldots, 1) \in \mathbb{R}^n.$$

(Thought: $\|\mathbf{1}_n\|_p = n^{1/p}$.) Then

$$\|\mathbf{1}_{mn}\| = \|\mathbf{1}_m \otimes \mathbf{1}_n\| = \|\mathbf{1}_m\| \, \|\mathbf{1}_n\|$$

by multiplicativity, and

$$\|\mathbf{1}_n\| = \|(\underbrace{1, \ldots, 1}_{n}, 0)\| \leq \|\mathbf{1}_{n+1}\|$$

by Lemma 4.14(iii), so by Theorem 1.5, there exists $c \geq 0$ such that $\|\mathbf{1}_n\| = n^c$ for all $n \geq 0$. By triangle inequality, $\|\mathbf{1}_{m+n}\| \leq \|\mathbf{1}_m\| + \|\mathbf{1}_n\|$, which implies $c \in [0, 1]$. Put $p = 1/c \in [1, \infty]$. Then $\|\mathbf{1}_n\| = n^{1/p} = \|\mathbf{1}_n\|_p$.

Step 3: the case $p = \infty$ If $p = \infty$, easy to show $\|\cdot\| = \|\cdot\|_\infty$. *(Omitted in class, but here it is.)* Let $p = \infty$, that is, $c = 0$. By Lemma 4.14, enough to prove $\|\mathbf{x}\| = \|\mathbf{x}\|_\infty$ for all $\mathbf{x} \in \mathbb{R}^n$ such that $x_i \geq 0$ for all i. Choose j such that $x_j = \|\mathbf{x}\|_\infty$. Using Lemma 4.14(iii),

$$\|\mathbf{x}\| \leq \|(x_j, \ldots, x_j)\| = x_j\|\mathbf{1}_n\| = x_j = \|\mathbf{x}\|_\infty.$$

On the other hand, using Lemma 4.14(iii) again,

$$\|\mathbf{x}\| \geq \|(0, \ldots, 0, x_j, 0, \ldots, 0)\| = \|(x_j)\| = x_j\|\mathbf{1}_1\| = x_j = \|\mathbf{x}\|_\infty.$$

So $x_j = \|\mathbf{x}\|_\infty$, as required. Now assume $p < \infty$.

Step 4: exploiting Cramér By Lemma 4.14(ii), enough to prove $\|\mathbf{x}\| = \|\mathbf{x}\|_p$ for each $\mathbf{x} \in \mathbb{R}^n$ such that $x_i > 0$ for all i. (Assume this from now on.) For $\mathbf{w} \in \mathbb{R}^J$ and $t \in \mathbb{R}$, write

$$N(\mathbf{w}, t) = \big|\{j \in J : w_j \geq t\}\big|.$$

For $r \geq 1$, write

$$\mathbf{x}^{\otimes r} = \mathbf{x} \otimes \cdots \otimes \mathbf{x} \in \mathbb{R}^{n^r}.$$

Then Proposition 4.10 states that

$$\|\mathbf{x}\|_p = \sup_{u>0,\, r\geq 1} u \cdot N(\mathbf{x}^{\otimes r}, u^r)^{1/rp}$$

or equivalently

$$\|\mathbf{x}\|_p = \sup_{u>0, r\geq 1} \big\|\underbrace{(u^r, \ldots, u^r)}_{N(\mathbf{x}^{\otimes r}, u^r)}\big\|^{1/r}. \tag{9}$$

Will use this to show $\|\mathbf{x}\| \geq \|\mathbf{x}\|_p$, then $\|\mathbf{x}\| \leq \|\mathbf{x}\|_p$.

Step 5: the lower bound First show $\|\mathbf{x}\| \geq \|\mathbf{x}\|_p$. By (9), enough to show that for each $u > 0$ and $r \geq 1$,

$$\|\mathbf{x}\| \geq \big\|\underbrace{(u^r, \ldots, u^r)}_{N(\mathbf{x}^{\otimes r}, u^r)}\big\|^{1/r}.$$

By multiplicativity, this is equivalent to

$$\|\mathbf{x}^{\otimes r}\| \geq \big\|\underbrace{(u^r, \ldots, u^r)}_{N(\mathbf{x}^{\otimes r}, u^r)}\big\|.$$

But this is clear, since by Lemma 4.14(iii),

$$\|\mathbf{x}^{\otimes r}\| \geq \big\|\underbrace{(\underbrace{u^r, \ldots, u^r}_{N(\mathbf{x}^{\otimes r}, u^r)}, 0, \ldots, 0)}_{n^r}\big\| = \big\|\underbrace{(u^r, \ldots, u^r)}_{N(\mathbf{x}^{\otimes r}, u^r)}\big\|.$$

Step 6: the upper bound Now prove $\|\mathbf{x}\| \leq \theta\|\mathbf{x}\|_p$ for each $\theta > 1$. Since $\min_i x_i > 0$, can choose $u_0, \ldots, u_d$ such that

$$\min_i x_i = u_0 < u_1 < \cdots < u_d = \max_i x_i$$

and $u_k/u_{k-1} < \theta$ for all $k \in \{1, \ldots, d\}$.

For each $r \geq 1$, define $\mathbf{y}_r \in \mathbb{R}^{n^r}$ to be $\mathbf{x}^{\otimes r}$ with each coordinate rounded up to the next one in the set $\{u_1^r, \ldots, u_d^r\}$. Then $\mathbf{x}^{\otimes r} \leq \mathbf{y}_r$ coordinatewise, giving

$$\begin{aligned}\|\mathbf{x}^{\otimes r}\| &\leq \|\mathbf{y}_r\| \\ &= \|(u_1^r, \ldots, u_1^r, \ldots, u_d^r, \ldots, u_d^r)\|\end{aligned}$$

where the number of terms u_k^r is $\leq N(\mathbf{x}^{\otimes r}, u_{k-1}^r)$. So

$$\begin{aligned}\|\mathbf{x}^{\otimes r}\| &\leq \sum_{k=1}^{d} \Big\|\Big(\underbrace{u_k^r, \ldots, u_k^r}_{\leq N(\mathbf{x}^{\otimes r}, u_{k-1}^r) \text{ terms}} \Big)\Big\| \\ &\leq d \max_{1 \leq k \leq d} \Big\|\Big(\underbrace{u_k^r, \ldots, u_k^r}_{\leq N(\mathbf{x}^{\otimes r}, u_{k-1}^r)} \Big)\Big\| \\ &\leq d\theta^r \max_{1 \leq k \leq d} \Big\|\underbrace{\Big(u_{k-1}^r, \ldots, u_{k-1}^r\Big)}_{\leq N(\mathbf{x}^{\otimes r}, u_{k-1}^r)}\Big\| \\ &\leq d\theta^r \|\mathbf{x}\|_p^r\end{aligned}$$

by (9). Hence $\|\mathbf{x}\| = \|\mathbf{x}^{\otimes r}\|^{1/r} \leq d^{1/r}\theta\|\mathbf{x}\|_p$. Letting $r \to \infty$, get $\|\mathbf{x}\| \leq \theta\|\mathbf{x}\|_p$. True for all $\theta > 1$, so $\|\mathbf{x}\| \leq \|\mathbf{x}\|_p$, completing the proof of Theorem 4.13.

Next week: how to measure biological diversity, and how diversity measures are related to norms, means and entropy.