Entropy and the polynomial Freiman–Ruzsa theorem

by

Marcel K. Goh

23 November 2023

Note. This is a gentle introduction to the notion of entropy as it is used in combinatorics, with a view towards understanding the new proof of the polynomial Freiman–Ruzsa conjecture by W. T. Gowers, B. Green, F. Manners, and T. Tao. The preliminary portion of these notes are largely transcribed from lectures given by W. T. Gowers.

1. The Khintchine–Shannon axioms

Let X be a discrete random variable. Its entropy $\mathbf{H}\{X\}$ is a real number (or ∞) that measures the "information content" of X. For example, if X is a constant random variable, then $\mathbf{H}\{X\}$ should be zero (we do not gain any information from knowing the value of X), and if X is uniformly distributed on $\{0,1\}^n$, then $\mathbf{H}\{X\}$ should be proportional to n, since X is determined by n bits of information. It satisfies the following axioms, which are sometimes called the Khinchine–Shannon axioms.

a) (*Invariance.*) If X takes values in A, Y takes values in B, $\phi : A \to B$ is a bijection, and $\mathbf{P}\{Y = \phi(a)\} = \mathbf{P}\{X = a\}$ for all $a \in A$, then $\mathbf{H}\{X\} = \mathbf{H}\{Y\}$.

b) (*Extensibility.*) If X takes values in A and Y takes values in B for a set B such that $A \subseteq B$, and furthermore $\mathbf{P}\{Y = a\} = \mathbf{P}\{X = a\}$ for all $a \in A$, then $\mathbf{H}\{X\} = \mathbf{H}\{Y\}$.

c) (*Continuity.*) The quantity $\mathbf{H}\{X\}$ depends continuously on the probabilities $\mathbf{P}\{X = a\}$.

d) (*Maximisation.*) Over all possible random variables X taking values in a finite set A, the quantity $\mathbf{H}\{X\}$ is maximised for the uniform distribution.

e) (*Additivity.*) For X taking values in A and Y taking values in B, we have the formula

$$\mathbf{H}\{X, Y\} = \mathbf{H}\{X\} + \mathbf{H}\{Y \mid X\},$$

where $\mathbf{H}\{X, Y\} = \mathbf{H}\big\{(X, Y)\big\}$ and

$$\mathbf{H}\{Y \mid X\} = \sum_{x \in A} \mathbf{P}\{X = x\}\, \mathbf{H}\{Y \mid X = x\}.$$

For now, we shall take it on faith that there really exists a function on random variables satisfying these axioms. Later on, when we prove this, we will

find that actually, the axioms only define entropy up to a multiplicative constant, so we shall add the following axiom.

f) (*Normalisation.*) If X is uniformly distributed on $\{0,1\}$, then $\mathbf{H}\{X\} = \log_2(e)$.

It is actually more common (for obvious reasons, especially in computer science) for one to set the entropy of a uniform random variable on $\{0,1\}$ to 1, but we shall follow the convention of Gowers et al., since the eventual goal of these notes is to understand their proof of the polynomial Freiman–Ruzsa conjecture.

Immediate from the definition of the conditional entropy $\mathbf{H}\{Y \mid X\}$ given in the additivity axiom is that if X and Y are independent, then $\mathbf{H}\{Y \mid X\} = \mathbf{H}\{Y\}$. But since the uniform distribution on $\{0,1\}^n$ is the joint distribution $(X_1, \dots, X_n)$ where the X_i are independent and uniformly distributed on $\{0,1\}$, we can additionally use the normalisation axiom to conclude that

$$\mathbf{H}\{X\} = \mathbf{H}\{X_1, \dots, X_n\} = \sum_{i=1}^{n} \mathbf{H}\{X_i\} = \log_2(e) \cdot n,$$

by induction. Induction also yields the chain rule

$$\mathbf{H}\{X_1, \dots, X_n\} = \mathbf{H}\{X_1\} + \mathbf{H}\{X_2 \mid X_1\} + \cdots + \mathbf{H}\{X_n \mid X_1, \dots, X_{n-1}\}.$$

The first proposition establishes the intuitive fact that the entropy of a uniform random variable supported on a set A is at most the entropy of a uniform random variable supported on a superset B of A.

Proposition 1. *Let $A \subseteq B$ with B finite, let X be uniformly distributed on A, and let Y be uniformly distributed on B. Then $\mathbf{H}\{X\} \leq \mathbf{H}\{Y\}$, with equality if and only if $A = B$.*

Proof.

The next proposition is that $\mathbf{H}\{X\}$ is always nonnegative. The following proof is attributed to S. Prendiville.

Proposition 2. *For any random variable X taking values in a finite set A, we have $\mathbf{H}\{X\} \geq 0$.*

Proof. First we consider the case that there exists $n \in \mathbf{N}$ such that for all $x \in A$, there is some c_x such that $\mathbf{P}\{X = x\} = c_x/n$. Let Y be uniform on $[n]$ and let E_x be a partition of $[n]$ such that $|E_x| = c_x$ for all $x \in X$. Then let Z be given by $Z = x$ if $Y \in E_x$. This random variable Z takes values in A and has the exact same distribution as X, so by the invariance axiom, $\mathbf{H}\{X\} = \mathbf{H}\{Z\}$. This means that $Z = f(Y)$ for some function f, meaning that there is a bijection between values y taken by Y and values $\big(y, f(y)\big)$ taken by (Y, Z). So we may apply the invariance axiom again to conclude that $\mathbf{H}\{Y, Z\} = \mathbf{H}\{Y\}$.

Now by the additivity axiom, $\mathbf{H}\{Y, Z\} = \mathbf{H}\{Y \mid Z\} + \mathbf{H}\{Z\}$, so we have $\mathbf{H}\{Y\} = \mathbf{H}\{Y \mid Z\} + \mathbf{H}\{Z\}$.

If X is a random variable such that $X = f(X)$ for some random variable Y, then we say that X is *determined by* Y or Y *determines* X. The following proposition formalises the idea that we get more information from Y than we get from X.

Proposition 3. *Let X and Y be random variables such that Y determines X. Then*

$$\mathbf{H}\{X\} \leq \mathbf{H}\{Y\}.$$

Proof. The

For random variables X and Y, the *mutual information* $\mathbf{I}\{X : Y\}$ is defined by the equivalent formulas

$$\begin{aligned}\mathbf{I}\{X : Y\} &= \mathbf{H}\{X\} + \mathbf{H}\{Y\} - \mathbf{H}\{X, Y\} \\ &= \mathbf{H}\{X\} - \mathbf{H}\{X \mid Y\} \\ &= \mathbf{H}\{Y\} - \mathbf{H}\{Y \mid X\}.\end{aligned}$$

It measures, roughly speaking, how much information one can get from one variable by looking at the other one. From the formula it is clear that $\mathbf{I}\{X : Y\} = 0$ if and only if X and Y are independent, and we also have the inequality $\mathbf{H}\{X \mid Y\} \leq \mathbf{H}\{X\}$. Given a triple (X, Y, Z) of random variables, we can apply this with $(X \mid Z = z)$ to obtain the inequality

$$\mathbf{H}\{X \mid Y, Z\} \leq \mathbf{H}\{X \mid Z\},$$

which is called *submodularity*. An equivalent statement is

$$\mathbf{H}\{X, Y, Z\} + \mathbf{H}\{Z\} \leq \mathbf{H}\{X, Z\} + \mathbf{H}\{Y, Z\}.$$

If Z takes values in a set C, the *conditional mutual information* is defined by

$$\begin{aligned}\mathbf{I}\{X : Y \mid Z\} &= \sum_{z \in C} p_Z(z)\, \mathbf{I}\big\{(X \mid Z = z) : (Y \mid Z = z)\big\} \\ &= \mathbf{H}\{X \mid Z\} + \mathbf{H}\{Y \mid Z\} - \mathbf{H}\{X, Y \mid Z\} \\ &= \mathbf{H}\{X, Z\} - 2\,\mathbf{H}\{Z\} + \mathbf{H}\{Y, Z\} - \mathbf{H}\{X, Y, Z\} + \mathbf{H}\{Z\} \\ &= \mathbf{H}\{X, Z\} + \mathbf{H}\{Y, Z\} - \mathbf{H}\{X, Y, Z\} - \mathbf{H}\{Z\},\end{aligned}$$

so we see that submodularity is equivalent to the statement $\mathbf{I}\{X : Y \mid Z\} \geq 0$. Analogously to the unconditional case, we have equality if and only if X and Y are independent when conditioned on Z.

The formula for entropy. Now we finally give the formula for the entropy of a random variable. For a discrete random variable X, we let $p_X : S \to [0, 1]$ denote the probability mass function of X, given by $p_X(x) = \mathbf{P}\{X = x\}$ for all $x \in S$. Unless subscripted, log always denotes the natural logarithm.

Proposition 4. *Let X be a random variable taking values in a finite set S. Then $\mathbf{H}\{X\}$ satisfies axioms (a) through (e) as well as the additional normalisation axiom (f) if and only if*

$$\mathbf{H}\{X\} = \mathbf{E}\big\{1/\log p_X(X)\big\} = \sum_{x \in S} p_X(x) \log \frac{1}{p_X(s)},$$

where we adopt the convention that $0 \log(1/0) = 0$.

Proof. [TODO: Prove this.] ∎

Equipped now with a formula for the entropy $\mathbf{H}\{X\}$ of a random variable X, we can prove further useful facts. We saw in the previous section that the entropy of a uniform random variable whose range is $\{0,1\}^n$ is $\log_2(e) \cdot n = \log(2^n)$, and we know, by the maximisation axiom, that this is the largest the entropy of any X taking values in $\{0,1\}^n$ can be. The following proposition generalises this notion.

Proposition 5. *For any random variable X taking values in a finite set S, we have*

$$\mathbf{H}\{X\} \leq \log |S|.$$

Proof. By Jensen's inequality, we have

$$\begin{aligned}
\mathbf{H}\{X\} &= \mathbf{E}\big\{\log(1/p_X(X)\big\} \\
&\leq \log \mathbf{E}\big\{1/p_X(X)\big\} \\
&= \log \sum_{x \in S} \mathbf{P}\{X = x\} \frac{1}{\mathbf{P}\{X = x\}} \\
&= \log |S|. \quad ∎
\end{aligned}$$

References