Compression

Math Club Mini Talk

Raymond Tana

Compression

***Compression**: the technique of representing data through encodings, usually in the pursuit of reducing how large (measured by number of bits) the representation is.*

What Matters?

- **Topology**: sets only matter up to homotopy.
- **Geometry**: sets matter up to isometry.
- **Compression**: data matters only up to encoding.

Compression in 3 Lights

1. Information Theory:
   `Incompressibility = Information Content`
2. Computability:
   `Compressibility is not Computable`
3. Euclidean Geometry:
   `Dimension determined by Incompressibility of Points`

Information Theory

What's the best the compressor can do to minimize the size of this file?

Example 1: text file full of 2^{10} zeros.

Answer: "`print '0' 2^10 times`"

Example 2: text file full of results of 2^{10} fair, 2-sided coin flips.

Answer: verbatim

Information Content

$$H(\text{file}) = \text{shortest length of code for the file}$$

- **All zeros**: low information content

$$H(\text{file}) \approx \log(\text{length}(\text{file}))$$

- **Coin flips**: high information content

$$H(\text{file}) \approx \text{length}(\text{file})$$

Uncomputability of Compression

Optimizing for code-lengths is a task that requires searching for which codes will map to the desired data.

Check code words c shorter than length(z) to see whether they serve as possible representations of z under a fixed, universal, lossless compression algorithm.

Run the decoding algorithm on c and wait to see when that algorithm outputs something. If the output is z, then yes! Otherwise (two cases: either the algorithm fails to halt, or the output is not z), c is not a code/representation for z.

Dimension via Pointwise

For simple sets X (rational intervals, middle-1/3 Cantor set, ...),

$$\dim(X) = \sup_{x \in X} \lim_{n \to \infty} \frac{H(0.x_1 x_2 \cdots x_n)}{n}$$