Information Theory: Compression and Predictability

UCLA Directed Reading Program

Dylan Wilbur

June 12, 2024

Introduction to Information Theory

- How can we quantify information?
- What are the theoretical limits of data compression?
- How can we measure and quantify uncertainty?

What Does Entropy Look Like?

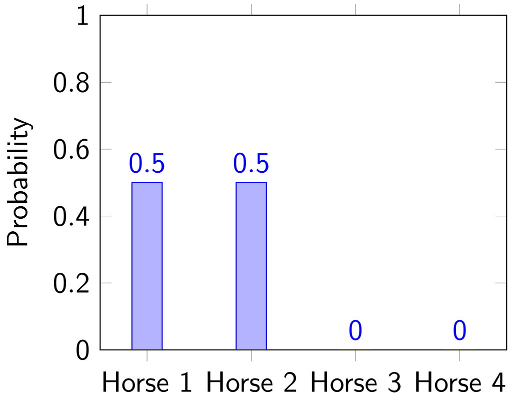


Figure: $H = -\sum p_i \log_2(p_i) = 0.5 + 0.5 + 0 + 0 = 1$

(Note: we define $0 \log 0$ here to be 0)

What Does Entropy Look Like?

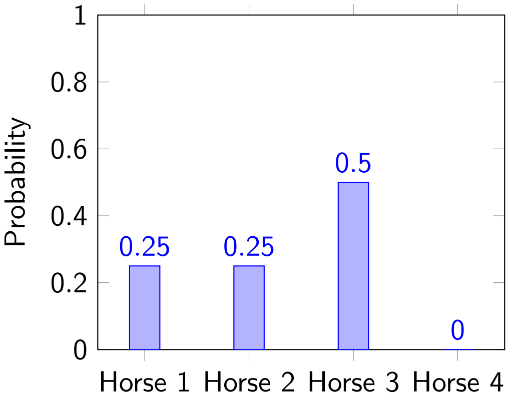


Figure: $H = -\sum p_i \log_2(p_i) = 1.5$

What Does Entropy Look Like?

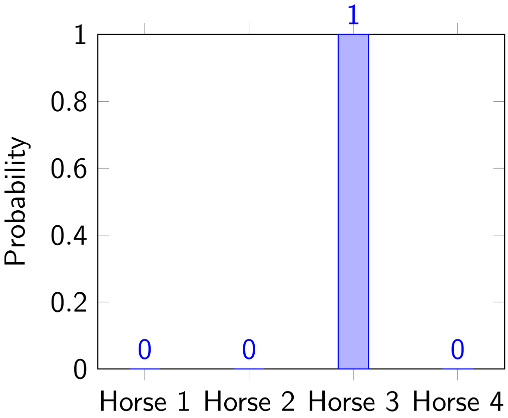


Figure: $H = -\sum p_i \log_2(p_i) = 0$

What Does Entropy Look Like?

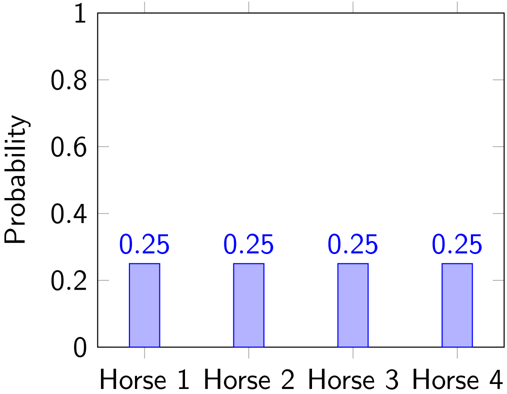


Figure: $H = -\sum p_i \log_2(p_i) = 2$

What Does Entropy Look Like?

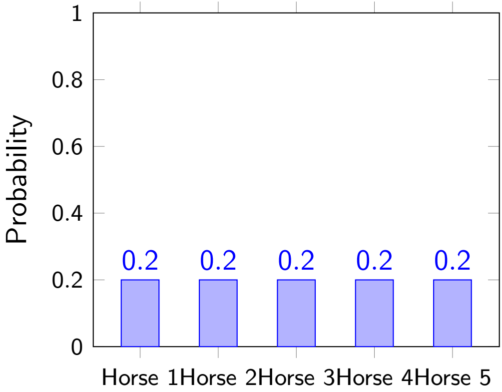


Figure: $H = -\sum p_i \log_2(p_i) = 2.5$

Entropy is Concave

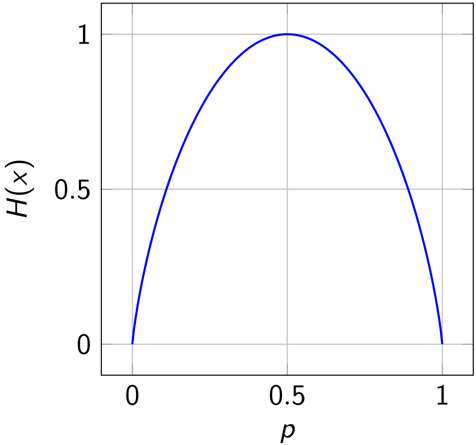


Figure: Plot of $H(p, 1-p) = -\left(p \log_2 p + (1-p)\log_2(1-p)\right)$

How Can We Measure the Difference in Uncertainty Between Two Distributions?

Answer: Relative Entropy

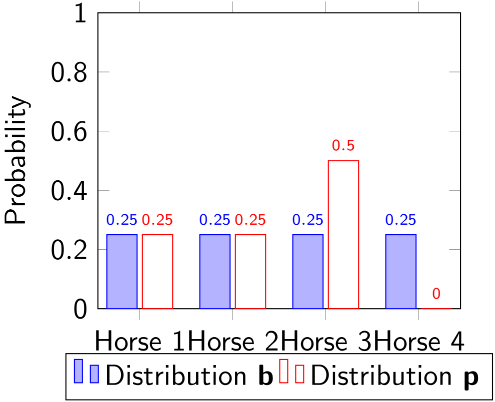


Figure: $D(b \parallel p) = \sum_i b_i \log_2\left(\frac{b_i}{p_i}\right) = 0.5$

Horse Racing and Gambling

Let's generalize this. Consider a sequence of random variables, X_i, representing the result of successive horse races, with distribution $\mathbf{p}$.

- Our betting strategy is defined as a vector $\mathbf{b}$ in which b_i is the amount of money bet on horse i.
- Another vector $\mathbf{o}$ will be the odds vector, so then at the end of the race, our wealth is multiplied by a factor of $b_i o_i$
- Our wealth after n races can be represented by

$$S_n = \prod_{i=1}^{n} S(X_i) \tag{1}$$

where $S(X_i) = b_i o_i$ with probability p_i.

Horse Racing and Gambling

Certainly, we would want to maximize this value S_n. We will take this time to define the doubling rate $W(\mathbf{b}, \mathbf{p})$, or the rate at which S_n grows with n, according to probability distribution $\mathbf{p}$ over the horses

$$S_n = 2^{nW(\mathbf{b},\mathbf{p})} \tag{2}$$

$$W(\mathbf{b}, \mathbf{p}) = E(\log S(X)) = \sum_{k=1}^{m} p_k \log b_k o_k \tag{3}$$

Look familiar? Let's justify this definition.

Derivation

First recall the **Law of Large Numbers** from probability theory:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mathbb{E}[X] \quad \text{as} \quad n \to \infty \tag{4}$$

We can use this to derive:

$$\begin{aligned}
S_n &= \prod_{i=1}^{n} o_i b_i && (5)\\
&= 2^{\log_2(\prod_{i=1}^{n} o_i b_i)} && (6)\\
&= 2^{n(\frac{1}{n}\sum_{i=1}^{n}\log_2(o_i b_i))} && (7)\\
&\to 2^{nE(\log o_i b_i)} && (8)\\
&= 2^{nW(\mathbf{b},\mathbf{p})} && (9)
\end{aligned}$$

Let's Optimize this

Since our wealth S_n grows exponentially with factor $W(\mathbf{b}, \mathbf{p})$, our goal is to maximize this value for given $\mathbf{b}$ and $\mathbf{p}$. How can we use our notion of entropy to achieve this? For a fixed $\mathbf{p}$, we have

$$W(\mathbf{b}) = \sum_i p_i \log b_i o_i \tag{10}$$

$$= \sum_i p_i \log(\frac{b_i}{p_i} p_i o_i) \tag{11}$$

$$= \sum_i p_i \log o_i + \sum_i p_i \log p_i + \sum_i p_i log(\frac{b_i}{p_i}) \tag{12}$$

$$= \sum_i p_i \log o_i - H(\mathbf{p}) - D(\mathbf{p}||\mathbf{b}) \tag{13}$$

Key Takeaways

$$W(\mathbf{b}) = \sum_i p_i \log o_i - H(\mathbf{p}) - D(\mathbf{p}||\mathbf{b}) \qquad (13)$$

- Note that for fixed $\mathbf{p}$, our doubling rate only relies on $\mathbf{b}$ in our relative entropy. Therefore, we can achieve a maximal doubling rate with $D(\mathbf{p}||\mathbf{b}) = 0$, or $\mathbf{b} = \mathbf{p}$ (Kelly Betting)
  - Our choice of betting strategy does not rely on the odds!
- Under optimal betting strategy, our doubling rate is equal to $\sum_i p_i \log o_i - H(\mathbf{p})$. Therefore, our growth rate is inversely proportional to the entropy.
  - Why does this make sense? Intuitively, we can make more money on more predictable races.

How Does This Relate to Data Compression?

Suppose we have a sequence of data we want to compress. A string of n characters can be thought of as a sequence of horse races. Perhaps this is:

- A genome sequence, ATGTCCA. . . . We can represent this as a sequence of horse races with a sample space of 4, like our earlier examples.
- The English language: a horse race with 27 outcomes(including a space)
- A bit string, 0111010. . . . WLOG, horse races with *n* horses can be represented in this way.

A Good Gambler is a Good Data Compressor!

Let's consider a horse race with $\mathbf{p} =< \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8} >$. We calculate the entropy of this distribution as

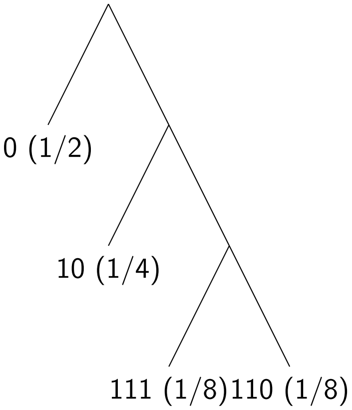
$$H = -\sum p_i \log_2(p_i) = \frac{7}{4} \tag{14}$$

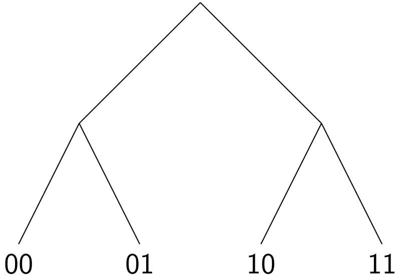
Betting on this horse race, our optimal growth rate would be

$$W = \sum_i p_i \log o_i - 7/4 \tag{15}$$

depending on the odds $\mathbf{o}$ given by a bookie. Similar to how the entropy of the distribution limits our growth rate, the entropy of our distribution also limits how much we can compress a sequence of results! Let's see this in action.

Huffman Coding our Sequence





- On the left, we have an average string length of $\frac{7}{4}$, and on the right we have an average length of 2.
- This is the average number of binary questions needed to answer "Is our random variable $X = x$?"
- The best we can encode our results is $H(\mathbf{p})$

Comparing results

- Sequence: 2, 1, 1, 1, 1, 2, 1, 1, 1, 2, 4, 3, 3, 2, 2, 4, 1, 1, 4, 1, 4, 2, 1, 1, 1, 2, 4, 1, 1, 2, 3, 1, 1, 2, 1, 2, 1, 1, 2, 3, 1, 1, 1, 2, 3, 2, 2, 1, 2, 1
- Scheme 1:
  0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 0
  1 0 1 1 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 1 1 1 0 0 0 0 0 1 1 1 0 0
  0 0 1 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 0 1 0
  0 0
- Scheme 2:
  1 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 1 1 1 1 1 1 0 1 0 1 1 0
  0 0 1 1 0 0 1 1 0 0 0 0 1 0 1 1 1 0 0 0 1 0 1 1 1 0 0 0 1 1 0 1
  1 1 1 0 0 0 1 0 0 1 0 1 1 1 0 0 0 0 1 0 1 0 1 1 1 0 0 0 1 0 1 1
  1 1 0 0 0 0 1 0 1 0 0

Information Theory Takeaways

- Predictable sequences are compressible.
- High predictability implies low entropy.
- Techniques and ideas shown here will scale

Acknowledgements

- My mentor, Robert Miranda
- Elements of Information Theory, Cover and Thomas
- A Mathematical Theory of Communication, Claude Shannon

Next Steps

- Future study: application to stock markets, machine learning
- Questions? Email me at dylan.m.w@icloud.com
- dylanwilbur.github.io