The asymmetry of relative entropy

(notes by Michael Baer)

This handout illustrates that, in addition to being asymmetric, relative entropy can actually be infinite going in one direction, but finite going the other.

Relative entropy, or Kullback-Leibler divergence, is defined as

$$D(p||q) = \sum_x p(x) \lg \frac{p(x)}{q(x)}$$

(where lg is log base 2) and is often thought of as a metric or distance measurement. Its notation, $D(p||q)$, is very metric-like, and some texts even refer to it as "K-L distance." Because of this, most information theoretic texts caution that $D(p||q)$ need not be equal to $D(q||p)$. For example, the textbook by Cover and Thomas calculates an instance in which $D(p||q)$ is approximately 0.2075 bits, while $D(q||p)$ is approximately 0.1887 bits.

This is a good method of driving this point home, but some students might be left with the impression that, although different, perhaps these "distances" are never too far apart. Perhaps they are like driving times; although it might take more time to enter Manhattan from New Jersey than to go the opposite direction during morning commute hours, those times will still be somewhat proportional to the overall "distance" between the two entities. This handout intends to show that, in fact, one can have an instance in which $D(p||q)$ is finite but $D(q||p)$ is infinite. The analogous situation would be if one could drive from Manhattan to New Jersey in a few minutes, but could never return! This handout will further explore precisely what this means in a lossless source coding context.

One might question how $D(q||p)$ could be infinite, since all terms in the necessary summation are, by definition, finite. However, if we use probability mass functions with infinite support, then the sum of finite terms can be infinite. Let us select perhaps the two simplest and most common infinite-support probability mass functions,

$$p(x) = 2^{-x} \qquad \text{and} \qquad q(x) = \frac{6}{\pi^2 x^2}$$

for $x \in \{1, 2, 3, \ldots\}$, and find the two divergences involving them. First:

$$\begin{aligned} D(p||q) &= \sum_{x=1}^{\infty} p(x) \lg \frac{p(x)}{q(x)} = \sum_{x=1}^{\infty} 2^{-x} \left(-x + \lg \frac{\pi^2 x^2}{6} \right) \\ &= -\sum_{x=1}^{\infty} x 2^{-x} + \sum_{x=1}^{\infty} 2^{-x} \lg \frac{\pi^2 x^2}{6} \\ &= -2 + \lg \frac{\pi^2}{6} + 2 \sum_{x=1}^{\infty} 2^{-x} \lg x \\ &\approx 0.1833. \end{aligned}$$

This approximation is easy to arrive at due to the fast convergence of the final summation. Thus the two probability mass functions seem rather close. Contrast this with:

$$\begin{aligned} D(q||p) &= \sum_{x=1}^{\infty} q(x) \lg \frac{q(x)}{p(x)} = \sum_{x=1}^{\infty} \frac{6}{\pi^2 x^2} \left(\lg \frac{6}{\pi^2 x^2} + x \right) \\ &= \frac{6}{\pi^2} \lg \frac{6}{\pi^2} \sum_{x=1}^{\infty} x^{-2} - \frac{12}{\pi^2} \sum_{x=1}^{\infty} x^{-2} \lg x + \frac{6}{\pi^2} \sum_{x=1}^{\infty} x^{-1} \\ &= \infty \end{aligned}$$

which can be arrived at by noting that only the harmonic sum term (the third additive term) diverges.

What does this astonishing asymmetry imply? In layman's terms, this means that, while the power law $q(x)$ is a good "guess" for the geometric distribution $p(x)$, the geometric distribution, as "guess" for the power law distribution, is not just poor, but totally unsuitable for any information theoretic application.

For a specific application, consider lossless (zero-error) source coding. In this application, $D(q||p)$ expresses the additional bits necessary for coding assuming probability distribution p when the correct distribution is q. For example, the unary code $\{0, 10, 100, \ldots\}$ corresponds to $p(x)$ (since the codeword lengths are equal to $-\lg p(x)$), so if one uses the unary code for coding a variable with distribution $q(x)$, the average codeword length has $D(q||p)$ more bits than it would were the correct source code used. For these p and q, that means infinitely more bits on average. By contrast, using an optimal code for $q(x)$ to code a random variable with probability mass function $p(x)$ results in only about 0.1833 more bits per symbol. (Note that this is not precise because $q(x)$ is not dyadic and expected codeword length is thus not equal to entropy even using this optimal code. This, however, is a minor quibble.)

In fact, $D(r||q)$ is relatively well-behaved for any r. This is connected to the idea of a *universal code*. (Such static codes should not be confused with the concept of universal source coding, which is adaptive in nature.) A universal code is one for which there exists a fixed, finite, positive c such that, given any monotonic p with finite $H(p)$, the code either uses no more than c expected bits for random variables with distribution p (if $H(p) < 1$) or performs no more than c times worse than the optimal code for that random variable. In other words, it is "good enough" for *any* random variable with a monotonic probability mass function. If the code has length $l(x)$ and $s = 2^{-l(x)}$, this means that, for any p with $H(p) < \infty$, either $H(p) < 1$ with $D(p||s) \leq c - H(p)$, or $D(p||s) \leq (c-1)H(p)$. The most widely used universal code, the Elias γ code, has a distribution of codewords quite close to $\lg q(x)$. For details, see P. Elias, "Universal Codeword Sets and Representations of the Integers," IEEE Trans. Inf. Theory, vol. IT-21, No. 2, 194–203 (March 1975).