

Lecture 11: Continuous-valued signals and differential entropy

Biology 429

Carl Bergstrom

September 20, 2008

Sources: Parts of today's lecture follow Chapter 8 from Cover and Thomas (2007). Some components of what follows — particularly the statements of theorems and definitions — come directly from that text.

Thus far in the course, we have dealt exclusively with discrete-valued signals, i.e., with communication using a finite alphabet. Many signals — and in particular, many biological signals — may take on continuous values. For example, the volume of a cry, the interspike interval in a spike train, and the length of a peacock's tail are all continuous-valued signals. How do we compute the entropy of a continuous valued signal?

When doing so, we immediately face a problem. Using a noiseless channel, one can send an infinite amount of information using a single continuous valued signal. Suppose I have an n -bit message that I want to send. I can send it using n symbols across a discrete-valued binary channel — or I can send it with a single signal across a continuous-valued channel, as follows. Instead of sending e.g. the message

1011000010111011010000101111011000000111111

using n bits, I can send a single real-valued number, equal to the fraction

0.1011000010111011010000101111011000000111111

through my continuous-valued channel. In this way, as I let get n get large, I can send arbitrarily large amounts of information through a noiseless channel with a single message.






Given that any continuous-valued signal thus can be said to carry infinite information, how can we talk about the entropy of continuous signals?

First, we note that in practice, the encoding scheme above works only in an arbitrarily noiseless continuous valued channel. If noise prevents us from distinguishing values within k of one another, we will in practice be limited to sending $-\log k$ bits per symbol through our continuous channel.

Thus by considering noise, we resolve the paradox of why it seems that we could send infinite information in finite time through a continuous-valued channel. We'll come back to noise later. For now, though, we want to develop the machinery to quantify the entropy of continuous-valued signals, even the in the absence of noise.

Rates of entropy increase

To do so, we will think of continuous-valued signal spaces as being the limit of a sequence of increasingly finely subdivided discrete-valued alphabets. Consider a discrete-valued alphabet with n symbols; this can carry at most $\log n$ bits of information (in the case that all values are equally likely). Thus the entropy of a continuous distribution on $[0, 1]$ with $n = 1/k$ divisions of length k is at most $\log n = -\log k$.

	k	$-\log k$	$H(X)$
	1/2	1	1
	1/4	2	2
	1/8	3	3
	1/16	4	4
	1/32	5	5

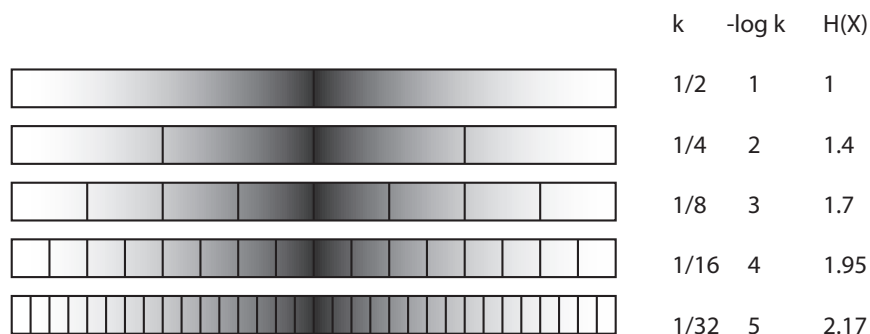
$$dH/d(-\log k) = 1$$

Now halve the size of the bins to $k/2$ each. The entropy can now be at most $-\log k/2 = 1 - \log k$ bits. Halve again to $k/4$, and the entropy can be at most $\log k/4 = 2 - \log k$. Halve again: entropy is now at most $3 - \log k$. And so forth: each time we double the number of divisions, the entropy of what bin we are in can increase by at most 1 bit. In other words, the rate at which entropy increases with the increasingly fine subdivisions of our interval is at most 1 bit per doubling of the number of subdivisions. But this rate is a derivative:

$$\frac{d[\text{max entropy}]}{d[-\log \text{subdivision width}]} = 1$$

This is the maximum rate at which entropy can increase with increasingly small subdivisions. What probability distribution achieves this rate? Well, to achieve this rate, each bin has to be equally likely to occur, and that has to be true for each level of subdivision. So the uniform probability distribution achieves this rate.

What happens if we have some other distribution? I've sketched an example below



$$\lim dH/d(-\log k) = 0.20$$

Now (in this particular example) the first division increases the entropy by 1 bit, but subsequent divisions increase the entropy by less than a bit and in this made-up example, the rate at which entropy increases with increasingly small subdivisions is asymptotically 0.2.

Now compare the two distributions. In the sense that we discussed at the start of the lecture, a single value from either distribution gives us an infinite precision real number and thus carries infinite information with infinite entropy. But in another sense, the former distribution (the uniform) seems to leave us more uncertain than the latter distribution (the bell-shaped). We want to capture this with some kind of measure, and our idea of looking at this derivative $dH/d(-\log k)$ has the right sort of property. Instead of measuring the total entropy, now that we are working with continuous variables we measure the *rate* at which the total entropy increases as well allow increasingly fine subdivisions. This is the conceptual idea underlying differential entropy and in general the measurement of entropy of continuous variables.

We can now explore the definition of continuous entropy and derive a number of expressions that parallel our expressions for discrete entropy and related quantities. But first, recall two basic definitions for continuous probabilities.

Recall 1 *The cumulative distribution function (CDF), $F(x)$ of a probability distribution \mathcal{F} is the probability that a random draw X from \mathcal{F} is less than*

$F(x)$:

$$F(x) = \text{Prob}(X \leq x)$$

Note that the range of $F(x)$ is $[0, 1]$.

Recall 2 *The probability density function (PDF), $f(x)$ of distribution \mathcal{F} is the derivative of the CDF.*

$$f(x) = F'(x)$$

We will work with well-behaved continuous distributions, such that $F(x)$ is continuous and continuously differentiable, and $\int f(x) = 1$.

Now we define the differential entropy $h(X)$. (Note that we use a lower-case h to indicate that this is the differential entropy rather than the discrete entropy H).

Definition 1 *The differential entropy $h(X)$ of random variable X with probability density function $f(x)$ is given by*

$$h(X) = - \int_S f(x) \log f(x) dx$$

where S is the set of possibility values of the X , i.e., the support set of X .

A few examples:

- Uniform on $[0, 1]$
- Uniform on $[0, a]$
- Normal with mean 0, variance σ^2 .

We also notice that differential entropy is invariant to translations of the random variable x :

Theorem 1 *Translation invariance of differential entropy*

$$h(X + c) = h(X)$$

Proof: the definition of differential entropy involves only the probabilities $f(x)$ and not the values of x itself.

Multiplying a random variable by a constant a changes its differentially entropy by an additive constant $\log |a|$.

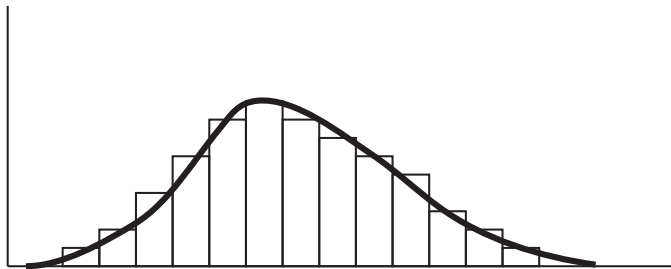
Theorem 2 *Multiplication alters differential entropy by an additive constant.*

$$h(aX) = h(X) + \log |a|$$

The proof is by change of variables, defining a new variable $Y = aX$ and applying the definition of differential entropy.

Relation to discrete entropy

Now we can formally see the connection between this definition and the idea of continuous symbols as the limit of n -division discrete spaces. We take the conventional Riemann integration view of the integral as the limit of a sequence of discrete sums over smaller and smaller subdivisions.



In other words, we break the distribution up into n bins of width Δ , and we consider the new discrete random variable Y^Δ to be the bin into which the value of x falls:

$$Y^\Delta = x_i, \text{ if } i\Delta \leq Y < (i+1)\delta$$

The probability distribution of Y^Δ is then

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

where the latter equality holds by the mean value theorem.

The discrete entropy of Y^Δ is then

$$\begin{aligned}
 H(Y^\Delta) &= -\sum p_i \log p_i \\
 &= -\sum f(x_i)\Delta \log f(x_i)\Delta \\
 &= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \\
 &= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta
 \end{aligned}$$

Since $-\sum \Delta f(x_i) \log f(x_i) \rightarrow -\int f(x) \log f(x)$ by the definition of the Riemann integral, we now have the following theorem:

Theorem 3 *The entropy of an n -bit quantization of a continuous random variable is approximately $h(X) + n$.*

$$H(Y^\Delta) + \log \Delta \rightarrow h(f) \text{ as } \Delta \rightarrow 0$$

This links up our earlier example of an n -bit approximation of a continuous signal with the concept of differential entropy.

Joint differential entropy

Now we are ready to develop a set of additional measures that we can apply to continuous random variables, much as we did in the second lecture of the course for the discrete entropy. If we just think of replacing sums with integrals, none of these are particularly surprising in form. We begin with the joint differential entropy. Recall that in lecture 3 of the course, we defined the joint entropy of a set of discrete random variables as

Recall 3 *The joint entropy $H(X, Y)$ of a pair of discrete random variables with a joint distribution $p(x, y)$ is defined as*

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y).$$

Our definition of joint differential entropy parallels this closely:

Definition 2 The joint differential entropy $h(X, Y)$ of a pair of discrete random variables with a joint density $p(x, y)$ is defined as

$$h(X, Y) = - \int f(x, y) \log f(x, y) dx dy.$$

This generalizes from pairs to collections of n random variables in the obvious way.

Conditional differential entropy

Now we move on to *conditional* differential entropy. Our definition again parallels the discrete definition:

Recall 4 The discrete conditional entropy $H(Y|X)$ is defined as

$$H(Y|X) = - \sum_{x \in \mathcal{X}} p(x) H(Y|X = x).$$

Definition 3 The conditional differential entropy $h(Y|X)$ is defined as

$$h(Y|X) = - \int f(x, y) h(Y|X = x) dx dy.$$

We can alternatively write this as $h(Y|X) = - \int f(x, y) \log f(x|y) dx dy$. Since $f(x|y) = f(x, y)/f(y)$, we can recover an expression analogous to what we had for discrete entropy:

$$h(Y|X) = h(X, Y) - h(Y)$$

Relative entropy for continuous random variables

We can also define a relative entropy between two continuous density functions f and g :

Recall 5 The relative entropy $D(p \parallel q)$ between two discrete probability distributions p and q is given by

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Definition 4 The relative entropy $D(p \parallel q)$ between two continuous probability distributions f and g is given by

$$D(f \parallel g) = \int f \log \frac{f}{g}.$$

Mutual information for continuous random variables

And finally, we can define the mutual information between two continuous variables.

Definition 5 The mutual information $I(X; Y)$ measures how much (on average) the realization of random variable Y tells us about the realization of X , i.e., how by how much the entropy of X is reduced if we know the realization of Y .

If X and Y are discrete random variables, the mutual information is given by:

$$I(X; Y) = H(X) - H(X|Y).$$

When X and Y are continuous random variables, the mutual information is given by:

$$I(X; Y) = h(X) - h(X|Y).$$

As before, we can also write this as

$$I(X, Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy.$$

From this formulation, we again see this as symmetric with respect to x and y .

The AEP for continuous random variables

The AEP for discrete random variables states that with arbitrarily high probability, the log probability of a realized sequence of n iid random variables approaches n times the entropy of the random variables as n gets large. Formally,

Recall 6 *The asymptotic equipartition theorem. Let X_1, X_2, X_3, \dots be independent and identically distributed (iid) random variables drawn from some distribution with probability function $p(x)$, and let $H(X)$ be the entropy of this distribution. Then as $n \rightarrow \infty$,*

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \rightarrow H(X)$$

in probability.

The AEP for continuous random variables states the equivalent for the probability density of a sequence of iid continuous random variables.

Theorem 4 *The asymptotic equipartition theorem for continuous random variables Let X_1, X_2, X_3, \dots be independent and identically distributed (iid) random variables drawn from a continuous distribution with probability density $f(x)$, and let $h(X)$ be the entropy of this distribution. Then as $n \rightarrow \infty$,*

$$-\frac{1}{n} \log f(X_1, X_2, \dots, X_n) \rightarrow E[-\log f(X)] = h(X)$$

in probability.

The convergence to expectation is a direct application of the weak law of large numbers; the equality is a consequence of the definition of $h(X)$.

Much of the point of developing the discrete AEP was to let us work with typical sets.

Recall 7 *The (weakly) typical set of sequences A_ϵ^n for a probability distribution $p(x)$ is the set of sequences x_1, x_2, \dots with average probability very close to the entropy of the random variable X :*

$$\left| -\frac{1}{n} \log p(x_1, x_2, \dots) - H(X) \right| \leq \epsilon$$

Again we can talk about typical sets for continuous random variables.

Definition 6 *The (weakly) typical set of sequences A_ϵ^n for a continuous probability distribution $f(x)$ is the set of sequences x_1, x_2, \dots with average probability very close to the entropy of the random variable X :*

$$\left| -\frac{1}{n} \log f(x_1, x_2, \dots) - h(X) \right| \leq \epsilon$$

As with the discrete AEP, almost all realized sequences will be members of the typical set:

$$\Pr[X \in A_\epsilon^n] > 1 - \epsilon.$$

That is, the typical set covers almost all of the probability as n gets large.

When we looked at typical sets for discrete random variables, we counted the size of these sets, i.e., number of members that they contained, and found that this was approximately $2^{nH(X)}$. For continuous random variables (as with any continuous region) the appropriate measure of size of a set is not the number of members (it has an infinite number of members) but the volume of the set. It turns out that the volume of the typical set for a continuous random variable is approximately equal to $2^{nh(x)}$, where the volume of a set S on \mathcal{R}^n is simply integral over the set $\int_S dx_1 dx_2 \dots$. The proof, which we omit here, is provided in Cover and Thomas.

In the case of discrete typical sets, the number of members of the typical set made up only a small fraction of the number of possible typical sequences (unless the distribution was uniform). The same thing holds for continuous typical sets. As n gets large, the *volume* of the typical set is only a small fraction of the volume of the set of all possible sequences.