CHAPTER 8

THE ENTROPY OF THE NORMAL DISTRIBUTION

INTRODUCTION

The "normal distribution" or "Gaussian distribution" or Gaussian probability density function is defined by

$$N(x;\, \mu,\, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}}\, e^{-(x-\mu)^2/2\sigma^2} \,. \tag{8.1}$$

This density function, which is symmetrical about the line $x = \mu$, has the familiar bell shape shown in Figure 8.1. The two parameters, μ and σ^2, each have special significance; μ is the mean and σ^2 the variance of the distribution. All probability density functions must be normalized to unity, and it is shown in most textbooks on advanced calculus that

$$\int_{-\infty}^{\infty} N(x;\, \mu,\, \sigma)\, dx = 1 \,. \tag{8.2}$$

The expectation of x, $E(x)$, is equal to the mean; that is,

$$E(x) = \int_{-\infty}^{\infty} x N(x;\, \mu,\, \sigma)\, dx = \mu \,. \tag{8.3}$$

The expectation of $(x-\mu)^2$, $E(x-\mu)^2$, is equal to the variance; that is,

$$E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 N(x;\, \mu,\, \sigma)\, dx = \sigma^2 \,. \tag{8.4}$$

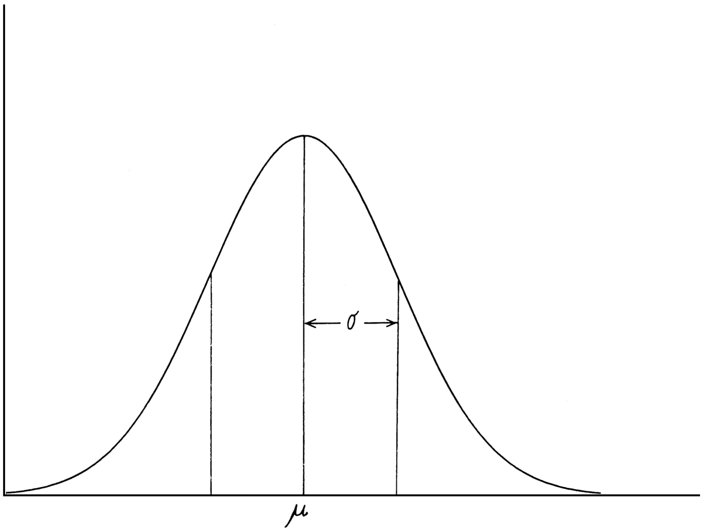

Figure 8.1 The normal distribution with mean, μ, and variance σ^2: $N(x;\, \mu,\, \sigma)$. About 2/3 of the area under the curve lies within one standard deviation, σ, of the mean.

The latter two equations, if unfamiliar, may be found in all textbooks on mathematical statistics, or may be verified directly by the reader.

The differential entropy of the normal distribution can be found without difficulty. From the definition of differential entropy given in Chapter 7, and using Equation (8.1),

$$H = -\int_{-\infty}^{\infty} (2\pi\sigma^2)^{-\frac{1}{2}}\, e^{-(x-\mu)^2/2\sigma^2} \ln\Big[(2\pi\sigma^2)^{-\frac{1}{2}}\, e^{-(x-\mu)^2/2\sigma^2} \Big] dx$$

$$H = \frac{1}{2}\ln(2\pi\sigma^2) \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-\frac{1}{2}}\, e^{-(x-\mu)^2/2\sigma^2}\, dx$$

$$+ \frac{1}{2\sigma^2}\int_{-\infty}^{\infty} (2\pi\sigma^2)^{-\frac{1}{2}}\, (x-\mu)^2\, e^{-(x-\mu)^2/2\sigma^2}\, dx\,.$$

Introducing Equations (8.2) and (8.4),

$$H = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\,.$$

Writing $\frac{1}{2}$ as $(\frac{1}{2}\ln e)$,

$$H = \frac{1}{2}\ln(2\pi e\sigma^2)\,, \tag{8.5}$$

which is the simple result we sought. We note that the differential entropy of the Gaussian probability density function depends only on the variance and not on the mean.

It has often been demonstrated (for example, Goldman, 1953) that for a given, fixed value of variance, σ^2, the probability density with the greatest value of H is the Gaussian density.

For an n-dimensional Gaussian density defined by

$$\begin{aligned} &N(x_1, x_2, \ldots; \mu_1, \mu_2, ..., \sigma_1, \sigma_2, \ldots) \\ &= \prod_{i=1}^{n} (2\pi e\sigma_i^2)^{-\frac{1}{2}} \exp\Big[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2} \Big]\,, \end{aligned} \tag{8.6}$$

the differential entropy is given by

$$H = (n/2)\ln 2\pi e(\sigma_1^2\sigma_2^2\ldots\sigma_n^2)^{1/n} \tag{8.7}$$

as shown by McEliece (1977). In the limiting case, for $n = 1$, Equation (8.7) reduces to (8.5).

CONVOLUTION OF TWO GAUSSIANS

Suppose that a pure signal is described by $N(x;\, \mu_S,\, \sigma_S)$, and its obfuscating noise by $N(x;\, \mu_N,\, \sigma_N)$. Then, as shown by Equation (7.20), the density function resulting from pure signal in the presence of noise is provided by the convolution

$$p_{SN}(x) = \int_{-\infty}^{\infty} N(x-x';\, \mu_S,\, \sigma_S)\, N(x';\, \mu_N,\, \sigma_N)\, dx'\,. \tag{8.8}$$

In fact, when we carry out the convolution of two Gaussians, the result is a third Gaussian density whose mean is the sum of the means of the two component functions and whose variance is the sum of the variances of the two component functions. That is,

$$p_{SN} = N(x;\, \mu_S + \mu_N,\, (\sigma_S^2 + \sigma_N^2)^{\frac{1}{2}})\,. \tag{8.9}$$

The full demonstration of (8.9) is not usually given in the textbooks because it is rather tedious, but it is provided in the Appendix for completeness. Writing Equation (8.9) explicitly,

$$p_{SN} = [2\pi(\sigma_S^2 + \sigma_N^2)]^{-\frac{1}{2}} \exp - \Bigg[\frac{[x - (\mu_S + \mu_N)]^2}{2(\sigma_S^2 + \sigma_N^2)} \Bigg]\,. \tag{8.10}$$

Using Equations (8.5) and (8.10), we can now write down directly the differential entropy of the two component densities and of the convolution of the two Gaussian components:

$$H_S = \frac{1}{2}\ln(2\pi e\sigma_S^2) \tag{8.11}$$

$$H_N = \frac{1}{2}\ln(2\pi e\sigma_N^2) \tag{8.12}$$

and

$$H_{SN} = \frac{1}{2}\ln[2\pi e(\sigma_S^2 + \sigma_N^2)]\ . \tag{8.13}$$

INFORMATION

Following Equation (7.17), we represent the information of a measurement as a difference (Shannon, 1948):

$$\begin{aligned} \mathscr{I} &= H_{SN} - H_N & (7.18)\\ &= \frac{1}{2}\ln[2\pi e(\sigma_S^2 + \sigma_N^2)] - \frac{1}{2}\ln[2\pi e\sigma_N^2] \end{aligned}$$

$$\mathscr{I} = \frac{1}{2}\ln[1 + \sigma_S^2/\sigma_N^2] \text{ natural units per signal.} \tag{8.14}$$

Remember that (7.18) was put forward as a "reasonable" candidate for the information obtained by making a measurement of a variable that was distributed continuously. Equation (8.14) is just a specific instance of (7.18) where the respective density functions are Gaussian. Equation (8.14) demonstrates various properties that would support its candidacy for an information function. $\mathscr{I}$ increases monotonically with increasing σ_S; the greater the standard deviation of the measured signal, the more uncertain we are about what the signal is, the more information we obtain from the measurement. Moreover, $\mathscr{I}$ increases monotonically with *decreasing* σ_N; the smaller the obfuscating factor, the greater the information obtained from the measurement. And of course, when $\sigma_S^2 = 0$, $\mathscr{I} = \ln 1 = 0$; when the signal (effectively) vanishes, no information is obtained. Or, looked at in another way, when the measurement is certain ($\sigma_S = 0$), no information is obtained.

A brief derivation of Equation (8.14) and its relation to "Shannon's second theorem" is provided by Beck (1976).

In the sensory analysis that follows, it will be helpful to interpret $\mathscr{I}$ as an information, because information is rather a tangible quantity that may conjure a picture in our minds. However, the informational interpretation of Equation (8.14) is not mandatory; no problem of a mathematical nature will be encountered by regarding $\mathscr{I}$ in this equation as simply the difference between differential entropies. In fact, since we shall hold σ_N^2 to be constant, $\mathscr{I}$ may be regarded simply as a differential entropy plus a constant, which definitely conjures no picture. The application of the function $\mathscr{I}$ in the analysis of sensory events will proceed in either case, with informational or entropic interpretation. Information is just a useful "currency" in which we can visualize a sensory neuron as trading. It is rather a concrete matter to state that a certain afferent neuron has relayed *b* bits of information to the brain. However, although less concrete, it is equally valid to say simply that a sensory receptor served by this afferent has reduced its entropy by *b* bits. But we are getting a little ahead of our story.

MORE ON THE INTERPRETATION OF THE INFORMATION FROM CONTINUOUS SOURCES

Using Equation (8.14), we can continue from Chapter 7 the attempt to interpret the information from continuous probability densities into the more intuitive information from discrete probability functions. You will remember that the probability density function for noise was used to limit the number of discrete rectangles into which the probability density for the signal might be divided: the less

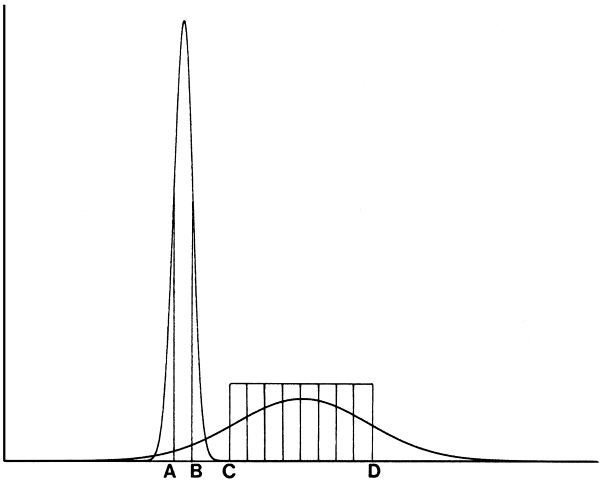

Figure 8.2 Squaring the normal curve (sort of), or "discretizing" the continuum. Two normal distributions are shown, the one on the right-hand side representing the pure signal, and the other representing the noise signal. AB designates the region between $-\sigma_N$ and $+\sigma_N$, and CD designates the region between $-\sigma_S$ and $+\sigma_S$ (see Figure 8.1). It is seen that CD can be divided into 8 rectangles of width AB. We can then regard the 8 rectangles as a histogram defining 8 equally probable, discrete, outcomes to an event. The information obtained from a measurement of the outcome is equal to $\ln(\sigma_S/\sigma_N) = \ln(4/0.5) = \ln 8$, which is approximately equal to the outcome of the original continuous event, $\frac{1}{2}\ln[1+(\frac{1}{2}CD)^2/(\frac{1}{2}AB)^2]$. We can see that as the noise variance, σ^2, becomes smaller, σ becomes smaller, more AB's fit into CD, and the information is greater.

intense the noise, the greater the number of rectangles, and the greater the value of the (discrete) entropy. The process of dividing into narrower and narrower rectangles had to be limited by some natural constraint, so that a unique value of (discrete) entropy could be obtained. The problem was one of "discretizing" the continuum.

For the normal distribution, Equation (8.14) illustrates how the continuous distribution can be rendered, in effect, discrete. Suppose we regard $\sigma_S^2/\sigma_N^2 >> 1$. Then Equation (8.14) becomes effectively,

$$\mathscr{I} = \frac{1}{2}\ln(\sigma_S^2/\sigma_N^2)\,. \tag{8.15}$$

But σ_S and σ_N are the standard deviations of the probability density functions for signal and noise respectively. If we regard σ_S / σ_N, rounded to the nearest integer = n, as the equivalent number of equally probable outcomes to the measurement event, then from the usual equation for discrete entropy,

$$H = \ln(\sigma_S/\sigma_N) = \ln(n)\,. \tag{8.16}$$

From (8.15), of course, $\mathscr{I} = H$. This idea is shown schematically in Figure 8.2 .

THE CENTRAL LIMIT THEOREM

Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ constitute a random sample drawn from an infinite population. We say that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ constitute a *random sample of size n*. Let

$$\bar{\mathbf{x}} = (\mathbf{x}_1 + \mathbf{x}_2 + \ldots + \mathbf{x}_n)/n \tag{8.17}$$

be the sample mean. The Central Limit Theorem states that if random samples of size n are drawn from a large or infinite population with mean μ, and variance σ^2, the sample mean, $\bar{\mathbf{x}}$, is approximately distributed normally with mean μ, and variance σ^2/n. Note that the theorem makes no mention of the nature of the population from which samples are drawn. Even if the population is far from a "normal" or "Gaussian" population, the sample means will still be distributed normally for sample size ≥ 30. If, however, the population is not too different from normal, the distribution of means will be normal for values of n much smaller than 30. The populations we shall be considering in our sensory work are expected to fall into the latter category.

Statistically, no mention need be made about how a sample of size n is obtained. However, in our scientific applications of the Central Limit Theorem it is, indeed, necessary to consider how the sample was obtained. In fact, a measuring device will reach into the large or infinite population and sequentially make n measurements, $x_1, x_2, ..., x_n$. I shall refer to each of these measurements as *one sampling*. That is, *it is necessary to make n samplings* (*or measurements*) *of the population to obtain one sample of size n*. The language is a little unwieldy, but I hope it is clear.

We have seen, now, that if the original population has variance σ^2, the means of samples of size n are normally distributed with variance σ^2/n. Therefore, the differential entropy of the original distribution is given by Equation (8.5) directly, while the differential entropy of the distribution of means of samples of size n is obtained from Equation (8.5) by replacing σ^2 by σ^2/n:

$$H_{\text{mean}} = \frac{1}{2}\ln(2\pi e\sigma^2/n)\,. \tag{8.18}$$

If the precision of the measurement of the means (net result of sampling + computation) is limited by Gaussian noise with variance σ_N^2, the information obtained from such a measurement is given by Equation (8.14) with σ_S^2 replaced by σ_S^2/n:

$$\mathscr{I} = \frac{1}{2}\ln\left[1 + \frac{\sigma_S^2/n}{\sigma_N^2}\right] \text{natural units per measurement.} \tag{8.19}$$

Thus it would appear that the information received by obtaining a measurement of the mean of 10 samplings ($n = 10$) is *less* than the information received by making a measurement based on a single sampling from the population ($n = 1$). However, this is not the interpretation I wish to pursue.

If one looks at the information given by Equation (8.19) as a function of n, the sample size, it is seen that $\mathscr{I}$ is maximum for $n = 1$, and $\mathscr{I} \to 0$ for $n \to \infty$. When $n \to \infty$, the sample variance, $\sigma^2/n \to 0$, implying that one has near-perfect knowledge of the population mean. We have incorporated the idea of Fisher's information (Chapter 5) into Shannon's structure. However, Equation (8.19) was not given by Shannon. I interpret

$$H(n) = \mathscr{I}(n) = \frac{1}{2}\ln\left[1 + \frac{\sigma_S^2/n}{\sigma_N^2}\right] \tag{8.19a}$$

as the information which *can still be gained* about the population mean after n samplings of a population have produced a single sample of size n. That is, $\mathscr{I}(n)$ is an absolute entropy; an uncertainty about the value of the population mean; and a *potential information* that may be received as the process of sampling continues. That is, with increasing n, uncertainty and potential information decrease, while information about the population mean increases. This interpretation of Equation (8.19) will be pursued in the next chapter when we come to model the process of sensation.

The difference in potential information, and, therefore, the gain in information, when the sample size is increased from n_1 to n_2 is given by $H(n_1) - H(n_2)$. The reader might like to show that for $\sigma_S^2/\, n_2\, \sigma_N^2 >> 1$, the gain in information is equal to $\ln\sqrt{n_2/n_1}$ (cf. Equation (11.10)).

ANALOG CHANNELS

You may remember that in Chapter 5 we left some unfinished business. We discussed the applications of information from discrete systems to communications engineering, but we could not, at that time, examine continuous or analog systems. However, we are now in a position to do so.

Communications systems deal usually with signals such as electrical potentials (voltages) that are transmitted with complex waveforms having, effectively, zero mean value. Equation (8.14) gives information in natural units per sample (of a complex signal). The well-known *sampling theorem*

(Shannon, 1949) states that if a function contains no frequencies higher than W, it is completely determined by giving its ordinates at a series of points spaced $1/(2W)$ seconds apart. The theorem has also been generalized to include the case where the frequency band does not start at zero but at some higher value. W is then a bandwidth. Therefore, if we divide the right-hand side of Equation (8.14) [natural units of information per sample] by $1/(2W)$ [seconds per sample] we obtain

$$C = W\ln(1 + \sigma_S^2/\sigma_N^2) \text{ natural units per second.} \tag{8.20}$$

Shannon has shown (1949), using an argument involving the volumes of spheres in hyperspace, that C is the channel capacity of the channel. If we divide (8.20) by ln 2 we get, of course, bits per second.

The ratio of variances is usually written as P/N, the signal-to-noise ratio, so that

$$C = W\ln(1 + P/N)\,. \tag{8.21}$$

This equation, then, gives the greatest rate at which an analog channel with a given signal-to-noise ratio and Gaussian noise ("white thermal noise") can transmit information. (Remember that the Gaussian distribution has the greatest differential entropy for a given variance.) As an example (from Raisbeck), if an audio circuit for the transmission of speech has a signal-to-noise ratio P/N equal to 36 decibels, and the bandwidth, W, is 4500 Hz, we can immediately calculate the channel capacity, C. Since $P/N = 10^{3.6}$ (note 2, Chapter 3),

$$C = (4500/\ln 2)\ln(1 + 10^{3.6})\,,$$

or about 50,000 bits per second.

APPENDIX: CONVOLUTION OF TWO GAUSSIAN FUNCTIONS

The convolution of the two Gaussian functions $N(x;\mu_S,\sigma_S)$ and $N(x;\mu_N,\sigma_N)$ that is given formally in Equation (8.8) is now carried out explicitly.

$$p_{SN}(x) = \int_{-\infty}^{\infty}\frac{1}{2\pi\,\sigma_S\sigma_N}\exp\left(\frac{-(x-x'-\mu_S)^2}{2\sigma_S^2}\right)\exp\left(\frac{-(x'-\mu_N)^2}{2\sigma_N^2}\right)dx'\,.$$

Changing variable, we set

$$Z = x' - \mu_N\,.$$

$$p_{SN}(x) = \int_{-\infty}^{\infty}\frac{1}{2\pi\,\sigma_S\sigma_N}\exp\left(\frac{-(x-Z-\mu_S-\mu_N)^2}{2\sigma_S^2}\right)\exp\left(\frac{-Z^2}{2\sigma_N^2}\right)dZ\,.$$

Setting

$$X = x - \mu_S - \mu_N\,, \tag{A8.1}$$

$$\begin{aligned}
p_{SN}(x) &= \frac{1}{2\pi\,\sigma_S\sigma_N}\int_{-\infty}^{\infty} e^{-(X-Z)^2/2\sigma_S^2}\,e^{-Z^2/2\sigma_N^2}\,dZ\\
&= \frac{1}{2\pi\,\sigma_S\sigma_N}e^{-X^2/2\sigma_S^2}\int_{-\infty}^{\infty}\exp\left(\frac{2XZ}{2\sigma_S^2}-\frac{Z^2}{2\sigma_S^2}-\frac{Z^2}{2\sigma_N^2}\right)dZ\\
&= \frac{1}{2\pi\,\sigma_S\sigma_N}e^{-X^2/2\sigma_S^2}\int_{-\infty}^{\infty}\exp-\left\{2\left[\frac{-X}{2\sigma_S^2}\right]Z+\left[\frac{1}{2\sigma_S^2}+\frac{1}{2\sigma_N^2}\right]Z^2\right\}dZ\\
&= \frac{1}{2\pi\,\sigma_S\sigma_N}\,e^{bX}\int_{-\infty}^{\infty}e^{-(aZ^2+2bZ)}\,dZ\,,
\end{aligned}$$

where

$$a = \frac{1}{2}\left(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_N^2}\right), \qquad \text{and} \qquad b = -\frac{X}{2\sigma_S^2} . \tag{A8.2}$$

$$p_{SN}(x) = \frac{1}{2\pi\,\sigma_S\sigma_N}\, e^{bX} \int_{-\infty}^{\infty} e^{b^2/a}\, e^{-a(Z+b/a)^2}\, dZ .$$

Changing variable by setting $u = Z + b/a$, $du = dZ$,

$$p_{SN}(x) = \frac{1}{2\pi\,\sigma_S\sigma_N}\, e^{bX} e^{b^2/a} \int_{-\infty}^{\infty} e^{-au^2}\, du .$$

Since $\int_{-\infty}^{\infty} e^{-au^2} du = \sqrt{\pi/a}$, $a > 0$ (see any discussion of the error function),

$$p_{SN}(x) = \frac{\sqrt{\pi/a}}{2\pi\,\sigma_S\sigma_N}\, e^{bX + b^2/a} . \tag{A8.3}$$

b^2/a can be evaluated from Equation (A8.2):

$$b^2/a = \frac{X^2}{4\sigma_S^4} \bigg/ \frac{1}{2}\left(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_N^2}\right) = \left(\frac{\sigma_N^2}{\sigma_S^2 + \sigma_N^2}\right)\frac{X^2}{2\sigma_S^2} .$$

Completing the algebra,

$$bX + b^2/a = -X^2 / 2(\sigma_S^2 + \sigma_N^2) . \tag{A8.4}$$

From the definition of a in (A8.2),

$$\frac{\sqrt{\pi/a}}{2\pi\sigma_S\sigma_N} = \frac{1}{\sqrt{2\pi(\sigma_S^2 + \sigma_N^2)}} . \tag{A8.5}$$

Substituting Equations (A8.4) and (A8.5) into (A8.3), and returning the value for X from (A8.1), we obtain the required result,

$$p_{SN} = [2\pi(\sigma_S^2 + \sigma_N^2)]^{-\frac{1}{2}} \exp - \left[\frac{[x - (\mu_S + \mu_N)]^2}{2(\sigma_S^2 + \sigma_N^2)}\right] . \tag{A8.10}$$

REFERENCES

Beck, A.H.W. 1976. *Statistical Mechanics, Fluctuations and Noise*. Edward Arnold, London.

Goldman, S. 1953. *Information Theory*. Prentice-Hall, Englewood Cliffs, N.J.

McEliece, R.J. 1977. *The Theory of Information and Coding: A Mathematical Framework for Communication.* Addison-Wesley, Reading, Mass.

Raisbeck, G. 1963. *Information Theory: An Introduction for Scientists and Engineers*. M.I.T. Press, Cambridge.

Shannon, C.E. 1948. *A mathematical theory of communication*. Bell System Technical Journal **27**, 623-656.

Shannon, C.E. 1949. *Communication in the presence of noise*. Proceedings of the IRE, **37**, 10-21.