4 Information in Physical Systems

What is information? A good answer is that information is what you don't already know. You do not learn much from being told that the sun will rise tomorrow morning; you learn a great deal if you are told that it will not. Information theory quantifies this intuitive notion of surprise. Its primary success is an explanation of how noise and energy limit the amount of information that can be represented in a physical system, which in turn provides insight into how to efficiently manipulate information in the system.

In the last chapter we met some of the many ways that devices can introduce noise into a signal, effectively adding unwanted information to it. This process can be abstracted into the concept of a *communications channel* that accepts an input and then generates an output. A telephone connection is a channel, as is the writing and subsequent reading of bits on a disk drive. In all cases there is assumed to be a set of known input symbols (such as 0 and 1), possibly a device that maps them into other symbols in order to satisfy constraints of the channel, the channel itself which has some probability for modifying the message due to noise or other errors, and possibly a decoder that turns the received symbols into an output set. We will assume that the types of messages and types of channel errors are sufficiently stationary to be able to define probability distributions $p(x)$ to see an input message x, and $p(y|x)$ for the channel to deliver a y if it is given an input x. This also assumes that the channel has no memory so that the probability distribution depends only on the current message. These are important assumptions: the results of this chapter will not apply to non-stationary systems.

4.1 INFORMATION

Let x be a random variable that takes on X possible values indexed by $i = 1, \ldots, X$, and let the probability of seeing the ith value be p_i. For example, x could be the letters of the alphabet, and p_i could be the probability to see letter i. How much information is there on average in a value of x drawn from this distribution? If there is only one possible value for x then we learn very little from successive observations because we already know everything; if all values are equally likely we learn as much as possible from each observation because we start out knowing nothing. An information functional $H(p)$ (a functional is a function of a function) that captures this intuitive notion should have the following reasonable properties:

- $H(p)$ is continuous in p. Small changes in the distribution should lead to small changes in the information.
- $H(p) \geq 0$, and $H(p) = 0$ if and only if just one p_i is non-zero. You always learn something unless you already know everything.
- $H(p) \leq C(X)$, where $C(X)$ is a constant that depends on the number of possible values X, with $H(p) = C(X)$ when all values are equally likely, and $X' > X \Rightarrow C(X') > C(X)$. The more options there are, the less you know about what will happen next.
- If x is drawn from a distribution p and y is independently drawn from a distribution q, then $H(p, q) = H(p) + H(q)$, where $H(p, q)$ is the information associated with seeing a pair (x, y). The information in independent events is the sum of the information in the events individually.

While it might appear that this list is not sufficient to define $H(p)$, it can be shown [Ash, 1990] that these desired properties are uniquely satisfied by the function

$$H(p) = -\sum_{i=1}^{X} p_i \log p_i \quad . \tag{4.1}$$

This is the definition of the *entropy* of a probability distribution, the same definition that was used in the last chapter in statistical mechanics. To make the dependence on x clear, we will usually write this as $H(x)$ instead of $H(p(x))$ or $H(p)$. The choice of the base of the logarithm is arbitrary; if the base is 2 then the entropy is measured in *bits*, and if it is base e then the entropy units are called *nats* for the *natural logarithm*. Note that to change an entropy formula from bits to nats you just change the logarithms from $\log_2$ to $\log_e$, and so unless otherwise noted the base of the logarithms in this chapter is arbitrary.

Now consider a string of N samples $(x_1, \ldots, x_N)$ drawn from p, and let N_i be the number of times that the ith value of x was actually seen. Because of the independence of the observations, the probability to see a particular string is the product of the individual probabilities

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} p(x_n) \quad . \tag{4.2}$$

This product of terms can be regrouped in terms of the possible values of x,

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{X} p_i^{N_i} \quad . \tag{4.3}$$

Taking the log and multiplying both sides by $-1/N$ then lets this be rewritten as

$$\begin{aligned} -\frac{1}{N} \log p(x_1, \ldots, x_N) &= -\frac{1}{N} \log \prod_{i=1}^{X} p_i^{N_i} \\ &= -\sum_{i=1}^{X} \frac{N_i}{N} \log p_i \\ &\approx -\sum_{i=1}^{X} p_i \log p_i \end{aligned}$$

$$= H(x) \quad . \tag{4.4}$$

The third line follows from the Law of Large Numbers (Section 3.2.4): as $N \to \infty$, $N_i/N \to p_i$. Equation (4.4) can be inverted to show that

$$p(x_1, \ldots, x_N) \approx 2^{-NH(x)} \tag{4.5}$$

(taking the entropy to be defined base 2). Something remarkable has happened: the probability of seeing a particular long string is independent of the elements of that string. This is called the *Asymptotic Equipartition Property* (*AEP*). Since the probability of occurrence for a string is a constant, its inverse $1/p = 2^{NH(x)}$ gives the effective number of strings of that length. However, the actual number of strings is larger, equal to

$$X^N = 2^{N \log_2 X} \quad . \tag{4.6}$$

The difference between these two values is what makes data compression possible. It has two very important implications [Blahut, 1988]:

- Since samples drawn from the distribution can on average be described by $H(x)$ bits rather than $\log_2 X$ bits, a coder can exploit the difference to store or transmit the string with $NH(x)$ bits. This is *Shannon's First Coding Theorem*, also called the *Source Coding Theorem* or the *Noiseless Coding Theorem*.
- The compressibility of a typical string is made possible by the vanishing probability to see rare strings, the ones that violate the Law of Large Numbers. In the unlikely event that such a string appears the coding will fail and a longer representation must be used. Because the Law of Large Numbers provides an increasingly tight bound on this occurrence as the number of samples increases, the failure probability can be made arbitrarily small by using a long enough string. This is the *Shannon–McMillan Theorem*.

Because the entropy is a maximum for a flat distribution, an efficient coder will represent information with this distribution. This is why phone modems would "hiss": they make best use of the telephone channel if the information being sent appears to be as random as possible. The value of randomness in improving a system's performance will recur throughout this book, particularly in Chapter 14.

We see that the entropy (base 2) gives the average number of bits that are required to describe a sample drawn from the distribution. Since the entropy is equal to

$$-\sum_{i=1}^{X} p_i \log p_i = \langle -\log p_i \rangle \tag{4.7}$$

it is natural to interpret $-\log p_i$ as the information in seeing event p_i, and the entropy as the expected value of that information.

Entropy can be applied to systems with more degrees of freedom. The joint entropy for two variables with a joint distribution $p(x, y)$ is

$$H(x, y) = -\sum_x \sum_y p(x, y) \log p(x, y) \quad . \tag{4.8}$$

This can be rewritten as

$$\begin{aligned} H(x,y) &= -\sum_x \sum_y p(x,y) \log p(x,y) \\ &= -\sum_x \sum_y p(x,y) \log[p(x|y)p(y)] \\ &= -\sum_x \sum_y p(x,y) \log p(x|y) - \sum_x \sum_y p(x,y) \log p(y) \\ &= -\sum_x \sum_y p(x,y) \log p(x|y) - \sum_y p(y) \log p(y) \\ &= H(x|y) + H(y) \end{aligned} \tag{4.9}$$

by using *Bayes' rule* $p(x,y) = p(x|y)p(y)$. The entropy in a conditional distribution $H(x|y)$ is the expected value of the information $\langle -\log p(x|y) \rangle$. The entropy of both variables equals the entropy of one of them plus the entropy of the other one given the observation of the first.

The *mutual information* between two variables is defined to be the information in them taken separately minus the information in them taken together

$$\begin{aligned} I(x,y) &= H(x) + H(y) - H(x,y) \\ &= H(y) - H(y|x) \\ &= H(x) - H(x|y) \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{4.10}$$

(these different forms are shown to be equal in Problem 4.2). This measures how many bits on average one sample tells you about the other. It vanishes if the variables are independent, and it is equal to the information in one of them if they are completely dependent. The mutual information can be viewed as an information-theoretic analog of the cross-correlation function $\langle x(t)y(t) \rangle$, but the latter is useful only for measuring the overlap among signals from linear systems [Gershenfeld, 1993].

In a sequence of N values $(x_1, x_2, \ldots, x_N)$ the *joint* (or *block entropy*)

$$H_N(x) = -\sum_{x_1} \sum_{x_2} \cdots \sum_{x_N} p(x_1, x_2, \ldots, x_N) \log p(x_1, x_2, \ldots, x_N) \tag{4.11}$$

is the average number of bits needed to describe the string. The limiting rate at which this grows

$$h(x) = \lim_{N \to \infty} \frac{1}{N} H_N(x) = \lim_{N \to \infty} H_{N+1} - H_N \tag{4.12}$$

is called the *source entropy*. It is the rate at which the system generates new information.

So far we've been discussing random variables that can take on a discrete set of values; defining entropy for continuous variables requires some care. If x is a real number, then $p(x)\ dx$ is the probability to see a value between x and $x + dx$. The information in such an observation is given by its logarithm, $-\log[p(x)\ dx] = -\log p(x) - \log dx$. As $dx \to 0$ this will diverge! The divergence is in fact the correct answer, because a single real number can contain an infinite amount of information if it can be specified to any

resolution. The *differential entropy* is the part of the entropy that does not diverge:

$$H(x) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx \quad . \tag{4.13}$$

Unlike the discrete entropy this can be positive or negative. The particular value of the differential entropy is not meaningful, because we have ignored the diverging part due to the infinitesimal limit, but differences between differential entropies are meaningful, because the diverging parts would cancel.

To understand mutual information for the continuous case we first need *Jensen's Theorem* [Cover & Thomas, 2012]: for a convex function $f(x)$ (one that has a non-negative second derivative, such as $-\log$)

$$\langle f(x) \rangle \geq f(\langle x \rangle) \quad . \tag{4.14}$$

This implies that for two normalized distributions p and q

$$D(p,q) \equiv \int_{-\infty}^{\infty} p \log \frac{p}{q} \tag{4.15}$$

$$= -\int_{-\infty}^{\infty} p \log \frac{q}{p} \tag{4.16}$$

$$\geq -\log \int_{-\infty}^{\infty} p \frac{q}{p}$$

$$= -\log \int_{-\infty}^{\infty} q$$

$$= -\log 1$$

$$= 0 \quad .$$

$D(p,q)$ is non-negative, vanishing if $p = q$. It is called the *Kullback–Leibler distance* between two probability functions, and $D[p(x,y), p(x)p(y)]$ is the continuous analog of the mutual information. The Kullback–Leibler distance arises naturally as a measure of the distance between two distributions, but it is not a true distance function: it is not symmetric in f and g (it can change value if they are interchanged), and it does not satisfy the triangle inequality ($D(p,q) + D(q,r)$ is not necessarily greater than or equal to $D(p,r)$).

4.2 CHANNEL CAPACITY

Claude Shannon is best known for finding a surprisingly simple solution to what had been thought to be a hard problem. The use of telephones grew faster than the available capacity of the phone system and so it became increasingly important to make good use of that capacity, raising an essential question: how many phone calls can be sent through a phone line? This is not easy to answer because a phone line is an analog channel with limited SNR and bandwidth. Clever modulation schemes can let more messages share the same cable; is there any limit to how much of an improvement is possible? Shannon's answer was a simple quantitative yes.

Consider a long string of N symbols $(x_1, x_2, \ldots, x_N)$ drawn independently from $p(x)$

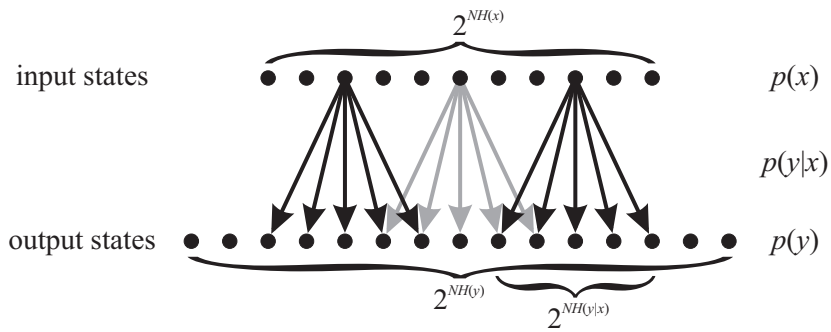

Figure 4.1. Effective number of states input to, added by, and output from a channel.

that are input to a channel specified by $p(y|x)$. On average each sample contains $H(x)$ bits of information, so this input string of N symbols can represent roughly $2^{NH(x)}$ different states. After being sent through the channel an output string $(y_1, y_2, \ldots, y_N)$ can represent $2^{NH(y)}$ states. However, it is possible that because of noise in the channel different input states can produce the same output state and hence garble the message; $2^{NH(y|x)}$ is the average number of different output states that are produced by an input state, the extra information in y given knowledge of x. In order to make sure that each input state typically leads to only one output state it is necessary to reduce the number of allowable output states by the excess information generated by the channel (Figure 4.1)

$$\frac{2^{NH(y)}}{2^{NH(y|x)}} = 2^{N[H(y)-H(y|x)]} = 2^{NI(x,y)} \quad . \tag{4.17}$$

We see that the probability distribution that maximizes the mutual information between the input and the output leads to the maximum number of distinct messages that can reliably be sent through the channel. The *channel capacity* is this maximum bit rate:

$$C = \max_{p(x)} \; I(x, y) \quad . \tag{4.18}$$

Applying the Shannon–McMillan Theorem to the input and output of the channel taken together shows that, if the data rate is below the channel capacity and the block length is long enough, then messages can be decoded with an arbitrarily small error. On the other hand, it is impossible to send data error-free through the channel at a rate greater than the capacity. This is *Shannon's Second Coding Theorem* (also called the *Channel Coding Theorem* or the *Noisy Coding Theorem*). If you're sending information at a rate below the channel capacity you are wasting part of the channel and should seek a better code (Chapter 14 will look at how to do this); if you're sending information near the capacity you are doing as well as possible and there is no point in trying to improve the code; and there is no hope of reliably sending messages much above the capacity.

A few points about channel coding:

- As the transmission rate increases it might be expected that the best-case error rate will also increase; it is surprising that the error rate can remain zero until the capacity is reached (Figure 4.2; Problem 4.3).
- This proves the existence of zero-error codes but it doesn't help find them, and

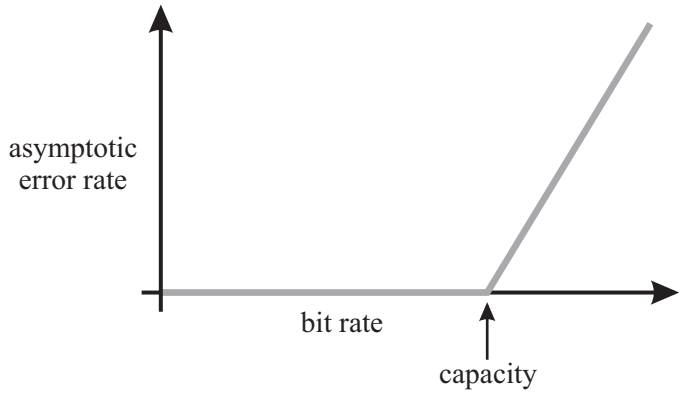

Figure 4.2. Onset of errors at the channel capacity in transmitting a long string.

once they are found they may not be useful. In particular, the coding/decoding effort or latency may become enormous as the rate approaches the capacity. For example, the length of the required code word may become prohibitively long.

- This is not a fundamental limit like the speed of light. The channel capacity holds for long strings of symbols independently drawn from a stationary probability distribution; it does not apply to short strings, non-stationary systems, or correlated variables. These approximations may not be justified, but can nevertheless be useful to make a rough estimate of the properties of a system. High-speed modems, for example, can exceed the theoretical capacity of a phone line (Problem 4.5) by adaptively modeling and coding for the channel errors.
- In many domains, such as broadcasting video, error-free transmission is irrelevant. All that matters is that the errors are not apparent; this is the subject of *lossy compression*. By taking advantage of what is known about human perception much higher bit rates are possible. Although you wouldn't want a money machine to do lossy compression on your bank balance when it communicates with the bank, your ear doesn't respond to a soft sound with a frequency immediately adjacent to that of a louder sound, and your eye cannot recognize the details of families of image textures beyond their statistical properties. These kinds of insights are used in the standards developed by the *Moving Pictures Experts Group* (*MPEG*) for variable lossy compression of video and audio. The MPEG-4 standard goes further to abandon a description based on arbitrary bit patterns and instead decomposes sights and sounds into high-level descriptions of their constituent elements [Koenen, 1999].

4.3 THE GAUSSIAN CHANNEL

In the last chapter we saw that the Central Limit Theorem explains why Gaussian noise is so common. It is therefore natural to consider a channel that adds Gaussian noise: $y_i = x_i + \eta_i$, where η_i is drawn from a Gaussian distribution. This might represent the Johnson noise in the input stage of a telephone amplifier, along with the accumulated effect

of many small types of interference. Gaussian distributions are particularly important in information theory because, for a given mean and variance, they maximize the differential entropy. This makes it easy to calculate the maximum in equation (4.18). To see this, let $\mathcal{N}(x)$ be a Gaussian distribution

$$\mathcal{N}(x) = \frac{1}{\sqrt{2\pi\sigma_{\mathcal{N}}^2}}\, e^{-(x-\mu_{\mathcal{N}})^2/2\sigma_{\mathcal{N}}^2} \quad , \tag{4.19}$$

and let $p(x)$ be an arbitrary distribution with mean μ_p and variance σ_p^2. Then

$$\begin{aligned} &-\int_{-\infty}^{\infty} p(x)\ln\mathcal{N}(x)\,dx \\ &= -\int_{-\infty}^{\infty} p(x)\left[-\ln\sqrt{2\pi\sigma_{\mathcal{N}}^2} - \frac{(x-\mu_{\mathcal{N}})^2}{2\sigma_{\mathcal{N}}^2}\right]\,dx \\ &= \ln\sqrt{2\pi\sigma_{\mathcal{N}}^2} + \frac{\sigma_p^2 + \mu_p^2 - 2\mu_p\mu_{\mathcal{N}} + \mu_{\mathcal{N}}^2}{2\sigma_{\mathcal{N}}^2} \quad . \end{aligned} \tag{4.20}$$

This depends only on the mean and variance of $p(x)$ and so if $q(x)$ has the same mean and variance then

$$-\int_{-\infty}^{\infty} p(x)\ln\mathcal{N}(x)\,dx = -\int_{-\infty}^{\infty} q(x)\ln\mathcal{N}(x)\,dx \quad . \tag{4.21}$$

Now consider the difference in the entropy between a Gaussian distribution $\mathcal{N}$ and another one p with the same mean and variance:

$$\begin{aligned} H(\mathcal{N}) - H(p) &= -\int_{-\infty}^{\infty} \mathcal{N}(x)\ln\mathcal{N}(x)\,dx + \int_{-\infty}^{\infty} p(x)\ln p(x)\,dx \\ &= -\int_{-\infty}^{\infty} p(x)\ln\mathcal{N}(x)\,dx + \int_{-\infty}^{\infty} p(x)\ln p(x)\,dx \\ &= \int_{-\infty}^{\infty} p(x)\ln\frac{p(x)}{\mathcal{N}(x)}\,dx \\ &= D(p,\mathcal{N}) \geq 0 \quad . \end{aligned} \tag{4.22}$$

The differential entropy in any other distribution will be less than that of a Gaussian with the same mean and variance. This differs from the discrete case, where the maximum entropy distribution was a constant, or an exponential if the energy is fixed.

Now return to our Gaussian channel $y = x + \eta$. Typically the input signal will be constrained to have some maximum power $S = \langle x^2 \rangle$. The capacity must be found by maximizing with respect to this constraint:

$$C = \max_{p(x):\langle x^2\rangle \leq S} I(x,y) \quad . \tag{4.23}$$

The mutual information is

$$\begin{aligned} I(x,y) &= H(y) - H(y|x) \\ &= H(y) - H(x+\eta|x) \\ &= H(y) - H(\eta|x) \\ &= H(y) - H(\eta) \quad , \end{aligned} \tag{4.24}$$

where the last line follows because the noise is independent of the signal. The differential entropy of a Gaussian process is straightforward to calculate (Problem 4.4):

$$H(\mathcal{N}) = \frac{1}{2}\log(2\pi e N) \tag{4.25}$$

(where $N = \sigma_{\mathcal{N}}^2$ is the noise power). The mean square channel output is

$$\begin{aligned}\langle y^2\rangle &= \langle (x+\eta)^2\rangle \\ &= \langle x^2\rangle + 2\langle x\rangle \underbrace{\langle \eta\rangle}_{0} + \langle \eta^2\rangle \\ &= S + N \quad .\end{aligned} \tag{4.26}$$

Since the differential entropy of x must be bounded by that of a Gaussian process with the same variance, the mutual information will be a maximum for

$$\begin{aligned}I(x,y) &= H(y) - H(\eta) \\ &\le \frac{1}{2}\log[2\pi e(S+N)] - \frac{1}{2}\log(2\pi e N) \\ &= \frac{1}{2}\log\left(1+\frac{S}{N}\right) \quad .\end{aligned} \tag{4.27}$$

The capacity of a Gaussian channel grows as the logarithm of the ratio of the signal power to the channel noise power.

Real channels necessarily have finite bandwidth. If a signal is sampled with a period of $1/2\Delta f$ then by the *Nyquist Theorem* the bandwidth will be Δf. If the (one-sided, white) noise power spectral density is N_0, the total energy in a time T is $N_0\Delta f T$, and the noise energy per sample is $(N_0\Delta f T)/(2\Delta f T) = N_0/2$. Similarly, if the signal power is S, the signal energy per sample is $S/2\Delta f$. This means that the capacity per sample is

$$\begin{aligned}C &= \frac{1}{2}\log\left(1+\frac{S}{N}\right) \\ &= \frac{1}{2}\log\left(1+\frac{S}{2\Delta f}\frac{2}{N_0}\right) \\ &= \frac{1}{2}\log_2\left(1+\frac{S}{N_0\Delta f}\right) \quad \frac{\text{bits}}{\text{sample}} \quad .\end{aligned} \tag{4.28}$$

If the signal power equals the noise power, then each samples carries 1/2 bit of information.

Since there are $2\Delta f$ samples per second the information rate is

$$\begin{aligned}C &= \Delta f \log\left(1+\frac{S}{N}\right) \\ &= \Delta f \log_2\left(1+\frac{S}{N_0\Delta f}\right) \quad \frac{\text{bits}}{\text{second}} \quad .\end{aligned} \tag{4.29}$$

This is the most important result in this chapter: the capacity of a band-limited Gaussian channel. It increases as the bandwidth and input power increase, and decreases as the noise power increases.

4.4 FISHER INFORMATION

There is a natural connection between the information in a measurement and the accuracy with which the measurement can be made, and so not surprisingly entropy shows up here also. Let $p_\alpha(x)$ be a probability distribution that depends on a parameter α, and let $f(x_1, x_2, ..., x_N)$ be an estimator for the value of α given N measurements of x drawn from $p_\alpha(x)$. The function f is a *biased* estimator if $\langle f(x_1, x_2, \ldots, x_N)\rangle \neq \alpha$, and it is *consistent* if in the limit $N \to \infty$ the probability to see $|f(x_1, x_2, \ldots, x_N) - \alpha| > \epsilon$ goes to 0 for any ϵ. An estimator f_1 *dominates* f_2 if $\langle (f_1(x_1, x_2, \ldots, x_N) - \alpha)^2\rangle \leq \langle (f_2(x_1, x_2, \ldots, x_N) - \alpha)^2\rangle$. This raises the question of what is the minimum variance possible for an unbiased estimator of α? The answer is given by the *Cramér–Rao bound*.

Start by defining the *score*:

$$V = \frac{\partial}{\partial\alpha}\log p_\alpha(x) = \frac{1}{p_\alpha(x)}\frac{\partial p_\alpha(x)}{\partial\alpha} \quad . \tag{4.30}$$

The mean value of the score is

$$\begin{aligned}\langle V\rangle &= \int_{-\infty}^{\infty} p_\alpha(x)\frac{1}{p_\alpha(x)}\frac{\partial p_\alpha(x)}{\partial\alpha}\,dx \\ &= \int_{-\infty}^{\infty}\frac{\partial p_\alpha(x)}{\partial\alpha}\,dx \\ &= \frac{\partial}{\partial\alpha}\int_{-\infty}^{\infty} p_\alpha(x)\,dx \\ &= \frac{\partial}{\partial\alpha}1 \\ &= 0 \quad .\end{aligned} \tag{4.31}$$

Therefore the variance of the score is just the mean of its square, $\sigma^2(V) = \langle V^2\rangle$. The variance of the score is called the *Fisher information*:

$$\begin{aligned}J(\alpha) &= \langle V^2\rangle \\ &= \left\langle\left[\frac{\partial\log p_\alpha(x)}{\partial\alpha}\right]^2\right\rangle \\ &= \left\langle\left[\frac{1}{p_\alpha(x)}\frac{\partial p_\alpha(x)}{\partial\alpha}\right]^2\right\rangle \\ &= \int_{-\infty}^{\infty}\frac{1}{p_\alpha(x)}\left[\frac{\partial p_\alpha(x)}{\partial\alpha}\right]^2 dx \quad .\end{aligned} \tag{4.32}$$

The score for a set of independent, identically distributed variables is the sum of the individual scores:

$$\begin{aligned}V(x_1, x_2, \ldots, x_N) &= \frac{\partial}{\partial\alpha}\log p_\alpha(x_1, x_2, \ldots, x_N) \\ &= \frac{\partial}{\partial\alpha}\log\prod_{i=1}^{N} p_\alpha(x_i) \\ &= \sum_{i=1}^{N}\frac{\partial\log p_\alpha(x_i)}{\partial\alpha}\end{aligned}$$

$$= \sum_{i=1}^{N} V(x_i) \tag{4.33}$$

and so the Fisher information for N measurements is

$$\begin{aligned} J_N(\alpha) &= \left\langle \left(\frac{\partial}{\partial \alpha} \log p_\alpha(x_1, x_2, \ldots, x_N) \right)^2 \right\rangle \\ &= \langle V^2(x_1, x_2, \ldots, x_N) \rangle \\ &= \left\langle \left(\sum_{i=1}^{N} V(x_i) \right)^2 \right\rangle \\ &= \sum_{i=1}^{N} \langle V^2(x_i) \rangle \\ &= N J(\alpha) \quad . \end{aligned} \tag{4.34}$$

The sum can be taken out of the expectation because the variables are uncorrelated.

The *Cramér–Rao inequality* states that the mean square error of an unbiased estimator f of α is lower bounded by the reciprocal of the Fisher information:

$$\sigma^2(f) \geq \frac{1}{J(\alpha)} \quad . \tag{4.35}$$

To prove this, start with the *Cauchy–Schwarz inequality*

$$\begin{aligned} \langle (V - \langle V \rangle)(f - \langle f \rangle) \rangle^2 &\leq \langle (V - \langle V \rangle)^2 \rangle \langle (f - \langle f \rangle)^2 \rangle \\ \langle V f - \langle V \rangle f - \langle f \rangle V + \langle V \rangle \langle f \rangle \rangle^2 &\leq \langle V^2 \rangle \langle (f - \langle f \rangle)^2 \rangle \\ \langle V f \rangle^2 &\leq J(\alpha) \sigma^2(f) \quad . \end{aligned} \tag{4.36}$$

The expectation of the left hand side equals one:

$$\begin{aligned} \langle V f \rangle &= \int_{-\infty}^{\infty} p_\alpha(x) \frac{1}{p_\alpha(x)} \frac{\partial p_\alpha(x)}{\partial \alpha} f(x)\, dx \\ &= \int_{-\infty}^{\infty} \frac{\partial p_\alpha(x)}{\partial \alpha} f(x)\, dx \\ &= \frac{\partial}{\partial \alpha} \int_{-\infty}^{\infty} p_\alpha(x) f(x)\, dx \\ &= \frac{\partial}{\partial \alpha} \langle f(x) \rangle \\ &= \frac{\partial \alpha}{\partial \alpha} \\ &= 1 \quad , \end{aligned} \tag{4.37}$$

thus proving the Cramér–Rao inequality. Just like the channel capacity, the Cramér–Rao bound sets a lower limit on what is possible but does not provide any guidance in actually finding the minimum variance estimator. In fact, in practice a biased estimator might be preferable because it could be easier to calculate, or might converge more quickly.

The Cramér–Rao inequality measures how tightly a distribution constrains a parameter. To relate it to the differential entropy $H(x)$ of a distribution $p(x)$, consider what

happens when a random Gaussian variable η is added to x. The new probability distribution is found by convolution:

$$p(\underbrace{x+\eta}_{\equiv y}) = \int_{-\infty}^{\infty} p(x) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x)^2/2\sigma^2}\, dx \quad . \tag{4.38}$$

Differentiating,

$$\frac{\partial p}{\partial \sigma^2} = \int_{-\infty}^{\infty} p(x) \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{(y-x)^2 - \sigma^2}{2\sigma^4} \right] e^{-(y-x)^2/2\sigma^2}\, dx$$

$$\frac{\partial^2 p}{\partial y^2} = \int_{-\infty}^{\infty} p(x) \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{(y-x)^2 - \sigma^2}{\sigma^4} \right] e^{-(y-x)^2/2\sigma^2}\, dx$$

$$\Rightarrow \frac{\partial p}{\partial \sigma^2} = \frac{1}{2} \frac{\partial^2 p}{\partial y^2} \quad . \tag{4.39}$$

This has the form of a diffusion equation: the added noise smooths out the distribution. Now taking the gradient of the differential entropy with respect to the noise variance, we see that

$$\begin{aligned}
\frac{\partial H}{\partial \sigma^2} &= -\frac{\partial}{\partial \sigma^2} \int_{-\infty}^{\infty} p(y) \log p(y)\, dy \\
&= -\frac{\partial}{\partial \sigma^2} \underbrace{\int_{-\infty}^{\infty} p(y)\, dy}_{0} - \int_{-\infty}^{\infty} \frac{\partial p}{\partial \sigma^2} \log p(y)\, dy \\
&= -\frac{1}{2} \int_{-\infty}^{\infty} \frac{\partial^2 p}{\partial y^2} \log p(y)\, dy \qquad \left(\int_{-\infty}^{\infty} u\, dv = uv|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} v\, du \right) \\
&= -\frac{1}{2} \frac{\partial p(y)}{\partial y} \log p(y) \Big|_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{p(y)} \left(\frac{\partial p(y)}{\partial y} \right)^2 dy \\
&= 0 + \frac{1}{2} J(y) \quad .
\end{aligned} \tag{4.40}$$

The first term on the left vanishes because although the logarithm diverges as $p \to 0$ when $y \to \infty$, the slope $\partial p / \partial y$ must be vanishing faster than logarithmically for the probability distribution to be normalized. Taking the limit $\sigma \to 0$,

$$\left. \frac{\partial H}{\partial \sigma^2} \right|_{\sigma^2 = 0} = \frac{1}{2} J(x) \quad . \tag{4.41}$$

The growth rate of the differential entropy with respect to the variance of an added Gaussian variable is equal to the Fisher information of the distribution. This is *de Bruijn's identity*. It can be interpreted as saying that the entropy measures the information in the volume of a distribution, and the Fisher information measures the information associated with its surface (as probed by smoothing it with the noise).

4.5 INFORMATION AND THERMODYNAMICS

We introduced entropy through statistical mechanics in Section 3.4, and in this chapter developed it as a powerful tool for analyzing probability distributions. The connection between thermodynamics and information theory is much deeper, providing a great example of how hard it can be to draw a clear boundary between basic and applied research in the evolution of significant ideas.

The important concept of the maximum efficiency of a heat engine was introduced by Sadi Carnot in 1824, motivated by the practical problem of understanding the limits on the performance of steam engines. This led to the macroscopic definition of entropy $\delta Q = TdS$ by Lord Kelvin (then William Thomson) and Rudolf Clausius around 1850–1860. Clausius named entropy for the Greek word for continuous transformation.

Statistical mechanics then grew out of the search for a microscopic explanation for macroscopic thermodynamics. This started with Maxwell's kinetic model of a gas, and the crucial connection $S = k \log \Omega$ was made by Boltzmann in 1877. Boltzmann, through his *H-Theorem*, provided a microscopic explanation for the macroscopic observation that a system moves to the available state with the maximum entropy. One of the (many) paradoxes in statistical mechanics was introduced by Maxwell in 1867: a microscopic creature (later called a *Maxzwell Demon*) could open and close a door between two containers, separating fast from slow gas molecules without doing any work on them. This appears to violate the Second Law of Thermodynamics, because the hot and cold gases could be used to run a heat engine, making a perpetual motion machine. Leo Szilard studied this problem in 1929, reducing it to a single molecule that can be on either side of a partition, arguably the first introduction of the notion of a bit of information [Szilard, 1929]. While Szilard did not explain the paradox of the Demon, Shannon was inspired by this analysis to use entropy as a measure of information to build information theory, which later helped create the very important and practical modern theory of coding [Slepian, 1974].

The real resolution of Maxwell's Demon did not come until 1961, when Rolf Landauer showed that the irreversibility in the Demon arises when it forgets what it has done; any computer that erases information necessarily dissipates energy [Landauer, 1961]. A stored bit can be in one of two states; if you initially have no idea what was stored then the minimum entropy associated with this bit is

$$S = k \log \Omega = k \log 2 \quad . \tag{4.42}$$

A real bit may represent much more entropy than this because many electrons (for example) are used to store it, but this is the minimum possible. Erasing the bit reduces the number of possible microscopic states down to one. It compresses the phase space of the computer, and so the dissipation associated with this erasure is

$$\delta Q = T\ dS = kT \log 2 - kT \log 1 = kT \log 2 \quad . \tag{4.43}$$

No matter how a computer is built, erasing a bit costs a minimum energy on the order of $kT \log 2$.

This result contains a strong assumption that the bit is near enough to thermal equilibrium for statistical mechanics to apply and for temperature to be a meaningful concept. Also, kT at room temperature is on the order of 0.02 eV, far below the energy of bits

stored in common computers. Nevertheless, Landauer's result is very important: whatever sets the energy scale of a stored bit (this might be the size of thermal fluctuations, or quantization in a small system), there is an energy penalty for erasing information. This has significant immediate implications for the design of low-power computers and algorithms [Gershenfeld, 1996].

Charles Bennett later showed in 1973 that it is possible to compute with *reversible computers* that never erase information and so can use arbitrarily little energy, depending on how long you are willing to wait for a sufficiently correct answer [Bennett, 1973]. We will return to this possibility in Section 11.5 when we look at the limits on computer performance.

4.6 SELECTED REFERENCES

[Cover & Thomas, 2012] Cover, Thomas M, & Thomas, Joy A. (2012). *Elements of information theory*. John Wiley & Sons.

A clear modern treatment of information theory.

[Balian, 1991] Balian, Roger. (1991). *From Microphysics to Macrophysics : Methods and Applications of Statistical Physics*. New York: Springer–Verlag. Translated by D. ter Haar and J.F. Gregg, 2 volumes.

Statistical mechanics beautifully introduced from an information-theoretical point of view.

[Slepian, 1974] Slepian, David (ed.). (1974). *Key Papers in the Development of Information Theory*. New York: IEEE Press

The seminal papers in the development of information theory.

[Brush, 1976] Brush, Stephen G. (1976). *The Kind of Motion We Call Heat: A History of the Kinetic Theory of Gases in the 19th Century*. New York: North-Holland. 2 volumes.

The history of statistical mechanics.

[Leff & Rex, 1990] Leff, Harvey S. & Rex, Andrew F. (eds.). (1990). *Maxwell's Demon: Entropy, Information, Computing*. Princeton: Princeton University Press.

Key papers relating entropy and computing.

4.7 PROBLEMS

(4.1) Verify that the entropy function satisfies the required properties of continuity, non-negativity, monotonicity, and independence.

(4.2) Prove the relationships in Equation (4.10).

(4.3) Consider a binary channel that has a small probability ϵ of making a bit error.

- (*a*) What is the probability of an error if a bit is sent independently three times and the value determined by majority voting?
- (*b*) How about if that is done three times, and majority voting is done on the majority voting?

(*c*) If majority voting on majoriy voting on ... on majority voting is done N times, how many bits are needed, and what is the probability of an error? How does this probability depend on ϵ?

(4.4) Calculate the differential entropy of a Gaussian process.

(4.5) A standard telephone line is specified to have a bandwidth of 3300 Hz and an SNR of 20 dB.

(*a*) What is the capacity?

(*b*) What SNR would be necessary for the capacity to be 1 Gbit/s?

(4.6) Let $(x_1, x_2, \ldots, x_n)$ be drawn from a Gaussian distribution with variance σ^2 and unknown mean value x_0. Show that $f(x_1, \ldots x_n) = n^{-1} \sum_{i=1}^{n} x_i$ is an estimator for x_0 that is unbiased and achieves the Cramér–Rao lower bound.