Elements of Information Theory
Thomas M. Cover, Joy A. Thomas

Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 9

Differential Entropy

We now introduce the concept of *differential entropy*, which is the entropy of a continuous random variable. Differential entropy is also related to the shortest description length, and is similar in many ways to the entropy of a discrete random variable. But there are some important differences, and there is need for some care in using the concept.

9.1 DEFINITIONS

Definition: Let X be a random variable with cumulative distribution function $F(x) = \Pr(X \leq x)$. If $F(x)$ is continuous, the random variable is said to be continuous. Let $f(x) = F'(x)$ when the derivative is defined. If $\int_{-\infty}^{\infty} f(x) = 1$, then $f(x)$ is called the *probability density function* for X. The set where $f(x) > 0$ is called the *support set* of X.

Definition: The *differential entropy* $h(X)$ of a continuous random variable X with a density $f(x)$ is defined as

$$h(X) = -\int_S f(x) \log f(x)\, dx\,, \tag{9.1}$$

where S is the support set of the random variable.

As in the discrete case, the differential entropy depends only on the probability density of the random variable, and hence the differential entropy is sometimes written as $h(f)$ rather than $h(X)$.

Remark: As in every example involving an integral, or even a density, we should include the statement *if it exists*. It is easy to construct examples of random variables for which a density function does not exist or for which the above integral does not exist.

Example 9.1.1 (*Uniform distribution*): Consider a random variable distributed uniformly from 0 to a, so that its density is $1/a$ from 0 to a and 0 elsewhere. Then its differential entropy is

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a}\, dx = \log a\,. \tag{9.2}$$

Note: For $a < 1$, $\log a < 0$, and the differential entropy is negative. Hence, unlike discrete entropy, differential entropy can be negative. However, $2^{h(X)} = 2^{\log a} = a$ is the volume of the support set, which is always non-negative, as we expect.

Example 9.1.2 (*Normal distribution*): Let $X \sim \phi(x) = (1/\sqrt{2\pi\sigma^2}) \times e^{-x^2/2\sigma^2}$. Then calculating the differential entropy in nats, we obtain

$$\begin{aligned} h(\phi) &= -\int \phi \ln \phi && (9.3)\\ &= -\int \phi(x)\left[-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right] && (9.4)\\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 && (9.5)\\ &= \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2 && (9.6)\\ &= \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma^2 && (9.7)\\ &= \frac{1}{2}\ln 2\pi e\sigma^2 \text{ nats}\,. && (9.8) \end{aligned}$$

Changing the base of the logarithm, we have

$$h(\phi) = \frac{1}{2}\log 2\pi e\sigma^2 \text{ bits}\,. \tag{9.9}$$

9.2 THE AEP FOR CONTINUOUS RANDOM VARIABLES

One of the important roles of the entropy for discrete random variables is in the AEP, which states that for a sequence of i.i.d. random variables, $p(X_1, X_2, \ldots, X_n)$ is close to $2^{-nH(X)}$ with high probability. This enables

us to define the typical set and characterize the behavior of typical sequences.

We can do the same for a continuous random variable.

Theorem 9.2.1: *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables drawn i.i.d. according to the density $f(x)$. Then*

$$-\frac{1}{n}\log f(X_1, X_2, \ldots, X_n) \to E[-\log f(X)] = h(X) \quad \textit{in probability}. \tag{9.10}$$

Proof: The proof follows directly from the weak law of large numbers. □

This leads to the following definition of the typical set.

Definition: For $\epsilon > 0$ and any n, we define the *typical set* $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows:

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \ldots, x_n) \in S^n : \left| -\frac{1}{n}\log f(x_1, x_2, \ldots, x_n) - h(X) \right| \le \epsilon \right\}, \tag{9.11}$$

where $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n f(x_i)$.

The properties of the typical set for continuous random variables parallel those for discrete random variables. The analog of the cardinality of the typical set for the discrete case is the volume of the typical set in the continuous case.

Definition: The *volume* $\mathrm{Vol}(A)$ of a set $A \in \mathcal{R}^n$ is defined as

$$\mathrm{Vol}(A) = \int_A dx_1\, dx_2 \cdots dx_n. \tag{9.12}$$

Theorem 9.2.2: *The typical set $A_\epsilon^{(n)}$ has the following properties*:

1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ *for n sufficiently large.*
2. $\mathrm{Vol}(A_\epsilon^{(n)}) \le 2^{n(h(X)+\epsilon)}$ *for all n.*
3. $\mathrm{Vol}(A_\epsilon^{(n)}) \ge (1-\epsilon)2^{n(h(X)-\epsilon)}$ *for n sufficiently large.*

Proof: By the AEP, $-\frac{1}{n}\log f(x_1, x_2, \ldots, x_n) = -\frac{1}{n}\sum \log f(x_i) \to h(X)$ in probability, establishing property 1.

Also,

$$1 = \int_{S^n} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n \tag{9.13}$$

$$\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n \tag{9.14}$$

$$\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)}\, dx_1\, dx_2 \ldots dx_n \tag{9.15}$$

$$= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1\, dx_2 \ldots dx_n \tag{9.16}$$

$$= 2^{-n(h(X)+\epsilon)} \operatorname{Vol}(A_\epsilon^{(n)}) . \tag{9.17}$$

Hence we have property 2.

We argue further that the volume of the typical set is at least this large. If n is sufficiently large so that property 1 is satisfied, then

$$1 - \epsilon \leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n \tag{9.18}$$

$$\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)}\, dx_1\, dx_2 \ldots dx_n \tag{9.19}$$

$$= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1\, dx_2 \ldots dx_n \tag{9.20}$$

$$= 2^{-n(h(X)-\epsilon)} \operatorname{Vol}(A_\epsilon^{(n)}) , \tag{9.21}$$

establishing property 3. Thus for n sufficiently large, we have

$$(1 - \epsilon) 2^{n(h(X)-\epsilon)} \leq \operatorname{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)} . \quad \square \tag{9.22}$$

Theorem 9.2.3: *The set $A_\epsilon^{(n)}$ is the smallest volume set with probability $\geq 1 - \epsilon$, to first order in the exponent.*

Proof: Same as in the discrete case. $\square$

This theorem indicates that the volume of the smallest set that contains most of the probability is approximately 2^{nh}. This is an n-dimensional volume, so the corresponding side length is $(2^{nh})^{1/n} = 2^h$. This provides an interpretation of the differential entropy: it is the logarithm of the equivalent side length of the smallest set that contains most of the probability. Hence low entropy implies that the random variable is confined to a small effective volume and high entropy indicates that the random variable is widely dispersed.

Note: Just as the entropy is related to the volume of the typical set, there is a quantity called Fisher information which is related to the

surface area of the typical set. We will say more about this in Section 16.7.

9.3 RELATION OF DIFFERENTIAL ENTROPY TO DISCRETE ENTROPY

Consider a random variable X with density $f(x)$ illustrated in Figure 9.1.

Suppose we divide the range of X into bins of length Δ. Let us assume that the density is continuous within the bins. Then by the mean value theorem, there exists a value x_i within each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx\,. \tag{9.23}$$

Consider the quantized random variable X^Δ, which is defined by

$$X^\Delta = x_i\,, \quad \text{if} \quad i\Delta \le X < (i+1)\Delta \tag{9.24}$$

Then the probability that $X^\Delta = x_i$ is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx \;= f(x_i)\Delta\,. \tag{9.25}$$

The entropy of the quantized version is

$$H(X^\Delta) = -\sum_{-\infty}^{\infty} p_i \log p_i \tag{9.26}$$

$$= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \tag{9.27}$$

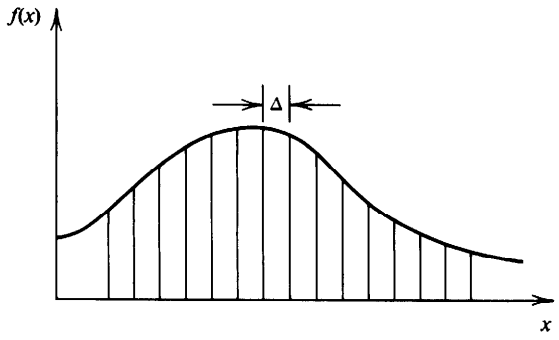

Figure 9.1. Quantization of a continuous random variable.

$$= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \tag{9.28}$$

$$= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta , \tag{9.29}$$

since $\Sigma f(x_i)\Delta = \int f(x) = 1$. If $f(x) \log f(x)$ is Riemann integrable (a condition to ensure the limit is well defined [272]), then the first term approaches the integral of $-f(x) \log f(x)$ by definition of Riemann integrability. This proves the following.

Theorem 9.3.1: *If the density $f(x)$ of the random variable X is Riemann integrable, then*

$$H(X^\Delta) + \log \Delta \rightarrow h(f) = h(X), \quad as\ \Delta \rightarrow 0 . \tag{9.30}$$

Thus the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

Examples:

1. If X has a uniform distribution on $[0, 1]$, and we let $\Delta = 2^{-n}$, then $h = 0$, $H(X^\Delta) = n$ and n bits suffice to describe X to n bit accuracy.
2. If X is uniformly distributed on $[0, \frac{1}{8}]$, then the first 3 bits to the right of the decimal point must be 0. To describe X to n bit accuracy requires only $n - 3$ bits, which agrees with $h(X) = -3$.

In the above two examples, every value of X requires the same number of bits to describe. In general, however $h(X) + n$ is the number of bits *on the average* required to describe X to n bit accuracy.

The differential entropy of a discrete random variable can be considered to be $-\infty$. Note that $2^{-\infty} = 0$, agreeing with the idea that the volume of the support set of a discrete random variable is zero.

9.4 JOINT AND CONDITIONAL DIFFERENTIAL ENTROPY

As in the discrete case, we can extend the definition of differential entropy of a single random variable to several random variables.

Definition: The *differential entropy* of a set $X_1, X_2, \ldots, X_n$ of random variables with density $f(x_1, x_2, \ldots, x_n)$ is defined as

$$h(X_1, X_2, \ldots, X_n)$$

$$= -\int f(x_1, x_2, \ldots, x_n) \log f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n . \tag{9.31}$$

Definition: If X, Y have a joint density function $f(x, y)$, we can define the conditional differential entropy $h(X|Y)$ as

$$h(X|Y) = -\int f(x, y) \log f(x|y)\, dx\, dy\,. \tag{9.32}$$

Since in general $f(x|y) = f(x, y)/f(y)$, we can also write

$$h(X|Y) = h(X, Y) - h(Y)\,. \tag{9.33}$$

But we must be careful if any of the differential entropies are infinite.

The next entropy evaluation is frequently used in the text.

Theorem 9.4.1 (*Entropy of a multivariate normal distribution*): *Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix K. (We use $\mathcal{N}_n(\mu, K)$ or $\mathcal{N}(\mu, K)$ to denote this distribution.) Then*

$$h(X_1, X_2, \ldots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K| \text{ bits}\,, \tag{9.34}$$

where $|K|$ denotes the determinant of K.

Proof: The probability density function of $X_1, X_2, \ldots, X_n$ is

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}}\, e^{-\frac{1}{2}(x-\mu)^T K^{-1}(x-\mu)}\,. \tag{9.35}$$

Then

$$h(f) = -\int f(\mathbf{x}) \left[-\frac{1}{2} (\mathbf{x} - \mu)^T K^{-1} (\mathbf{x} - \mu) - \ln(\sqrt{2\pi})^n |K|^{1/2} \right] d\mathbf{x} \tag{9.36}$$

$$= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(K^{-1})_{ij} (x_j - \mu_j) \right] + \frac{1}{2} \ln(2\pi)^n |K| \tag{9.37}$$

$$= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(x_j - \mu_j)(K^{-1})_{ij} \right] + \frac{1}{2} \ln(2\pi)^n |K| \tag{9.38}$$

$$= \frac{1}{2} \sum_{i,j} E[(x_j - \mu_j)(x_i - \mu_i)](K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \tag{9.39}$$

$$= \frac{1}{2} \sum_{j} \sum_{i} K_{ji} (K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \tag{9.40}$$

$$= \frac{1}{2} \sum_{j} (KK^{-1})_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \tag{9.41}$$

$$= \frac{1}{2} \sum_{j} I_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \tag{9.42}$$

$$= \frac{n}{2} + \frac{1}{2} \ln(2\pi)^n |K| \tag{9.43}$$

$$= \frac{1}{2} \ln(2\pi e)^n |K| \text{ nats} \tag{9.44}$$

$$= \frac{1}{2} \log(2\pi e)^n |K| \text{ bits} . \quad \square \tag{9.45}$$

9.5 RELATIVE ENTROPY AND MUTUAL INFORMATION

We now extend the definition of two familiar quantities, $D(f\|g)$ and $I(X; Y)$, to probability densities.

Definition: The *relative entropy* (or *Kullback Leibler distance*) $D(f\|g)$ between two densities f and g is defined by

$$D(f\|g) = \int f \log \frac{f}{g} . \tag{9.46}$$

Note that $D(f\|g)$ is finite only if the support set of f is contained in the support set of g. (Motivated by continuity, we set $0 \log \frac{0}{0} = 0$.)

Definition: The *mutual information* $I(X; Y)$ between two random variables with joint density $f(x, y)$ is defined as

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} \, dx \, dy . \tag{9.47}$$

From the definition it is clear that

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) \tag{9.48}$$

and

$$I(X; Y) = D(f(x, y)\|f(x)f(y)) . \tag{9.49}$$

The properties of $D(f\|g)$ and $I(X; Y)$ are the same as in the discrete case. In particular, the mutual information between two random variables is the limit of the mutual information between their quantized versions, since

$$I(X^\Delta; Y^\Delta) = H(X^\Delta) - H(X^\Delta|Y^\Delta) \tag{9.50}$$

$$\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta) \tag{9.51}$$

$$= I(X; Y) . \tag{9.52}$$

Certain authors (e.g., Gallager [120]) prefer to define the mutual information between two continuous random variables directly as the above limit, and not consider differential entropies at all.

9.6 PROPERTIES OF DIFFERENTIAL ENTROPY, RELATIVE ENTROPY AND MUTUAL INFORMATION

Theorem 9.6.1:

$$D(f\|g) \geq 0 . \tag{9.53}$$

with equality iff $f = g$ almost everywhere (a.e.).

Proof: Let S be the support set of f. Then

$$-D(f\|g) = \int_S f \log \frac{g}{f} \tag{9.54}$$

$$\leq \log \int_S f \frac{g}{f} \quad \text{(by Jensen's inequality)} \tag{9.55}$$

$$= \log \int_S g \tag{9.56}$$

$$\leq \log 1 = 0 . \tag{9.57}$$

We have equality iff we have equality in Jensen's inequality, which occurs iff $f = g$ a.e. □

Corollary: $I(X; Y) \geq 0$ *with equality iff X and Y are independent.*

Corollary: $h(X|Y) \leq h(X)$ *with equality iff X and Y are independent.*

Theorem 9.6.2: *Chain rule for differential entropy*:

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}) \tag{9.58}$$

Proof: Follows directly from the definitions. □

Corollary:

$$h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i) , \tag{9.59}$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Proof: Follows directly from Theorem 9.6.2 and the corollary to Theorem 9.6.1. □

Application (*Hadamard's inequality*): If we let $\mathbf{X} \sim \mathcal{N}(0, K)$ be a multivariate normal random variable, then substituting the definitions of entropy in the above inequality gives us

$$|K| \leq \prod_{i=1}^{n} K_{ii}\,, \tag{9.60}$$

which is Hadamard's inequality. A number of determinant inequalities can be derived from information theoretic inequalities in this fashion (Chapter 16).

Theorem 9.6.3:

$$h(X + c) = h(X)\,. \tag{9.61}$$

Translation does not change the differential entropy.

Proof: Follows directly from the definition of differential entropy. □

Theorem 9.6.4:

$$h(aX) = h(X) + \log|a|\,. \tag{9.62}$$

Proof: Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, and

$$h(aX) = -\int f_Y(y) \log f_Y(y)\, dy \tag{9.63}$$

$$= -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \tag{9.64}$$

$$= -\int f_X(x) \log f_X(x) + \log|a| \tag{9.65}$$

$$= h(X) + \log|a|\,, \tag{9.66}$$

after a change of variables in the integral. □

Similarly we can prove the following corollary for vector-valued random variables:

Corollary:

$$h(A\mathbf{X}) = h(\mathbf{X}) + \log|A|\,, \tag{9.67}$$

where $|A|$ is the absolute value of the determinant.

We will now show that the multivariate normal distribution maximizes the entropy over all distributions with the same covariance.

Theorem 9.6.5: *Let the random vector $\mathbf{X} \in \mathbf{R}^n$ have zero mean and covariance $K = E\mathbf{X}\mathbf{X}^t$, i.e., $K_{ij} = EX_iX_j$, $1 \le i, j \le n$. Then $h(\mathbf{X}) \le \frac{1}{2}\log(2\pi e)^n|K|$, with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.*

Proof: Let $g(\mathbf{X})$ be any density satisfying $\int g(\mathbf{x})x_ix_j\,d\mathbf{x} = K_{ij}$, for all i, j. Let ϕ_K be the density of a $\mathcal{N}(0, K)$ vector as given in 9.35, where we set $\mu = 0$. Note that $\log \phi_K(\mathbf{x})$ is a quadratic form and $\int x_ix_j\phi_K(\mathbf{x})\,d\mathbf{x} = K_{ij}$. Then

$$0 \le D(g\|\phi_K) \tag{9.68}$$

$$= \int g \log(g/\phi_K) \tag{9.69}$$

$$= -h(g) - \int g \log \phi_K \tag{9.70}$$

$$= -h(g) - \int \phi_K \log \phi_K \tag{9.71}$$

$$= -h(g) + h(\phi_K)\,, \tag{9.72}$$

where the substitution $\int g \log \phi_K = \int \phi_K \log \phi_K$ follows from the fact that g and ϕ_K yield the same moments of the quadratic form $\log \phi_K(\mathbf{x})$. □

9.7 DIFFERENTIAL ENTROPY BOUND ON DISCRETE ENTROPY

Of all distributions with the same variance, the normal maximizes the entropy. So the entropy of the normal gives a good bound on the differential entropy in terms of the variance of the random variable. We will use this bound to give a bound on the discrete entropy of a random variable. It will not be in terms of the variance of the random variable itself, since a discrete random variable can have arbitrarily small variance and still have high discrete entropy. Instead, the bound is in terms of an integer-valued random variable with the same probabilities (and hence the same entropy).

Let X be a discrete random variable on the set $\mathcal{X} = \{a_1, a_2, \dots\}$ with

$$\Pr(X = a_i) = p_i . \tag{9.73}$$

Theorem 9.7.1:

$$H(p_1, p_2, \dots) \leq \frac{1}{2} \log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i\right)^2 + \frac{1}{12}\right). \tag{9.74}$$

Moreover, for every permutation σ,

$$H(p_1, p_2, \dots) \leq \frac{1}{2} \log(2\pi e)\left(\sum_{i=1}^{\infty} p_{\sigma(i)} i^2 - \left(\sum_{i=1}^{\infty} i p_{\sigma(i)}\right)^2 + \frac{1}{12}\right). \tag{9.75}$$

Proof: Define two new random variables. The first, X_0, is an integer-valued discrete random variable with the distribution

$$\Pr(X_0 = i) = p_i . \tag{9.76}$$

Let U be a random variable uniformly distributed on the range $[0, 1]$, independent of X_0. Define the continuous random variable $\tilde{X}$ by

$$\tilde{X} = X_0 + U . \tag{9.77}$$

The distribution of the r.v. $\tilde{X}$ is shown in Figure 9.2.

It is clear that $H(X) = H(X_0)$, since discrete entropy depends only on the probabilities and not on the values of the outcomes. Now

$$H(X_0) = -\sum_{i=1}^{\infty} p_i \log p_i \tag{9.78}$$

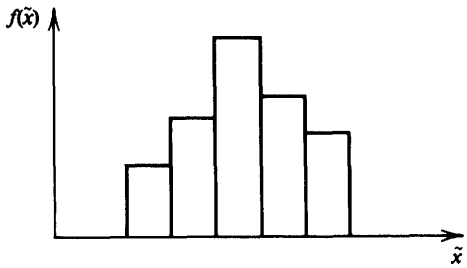

Figure 9.2. Distribution of $\tilde{X}$.

$$= -\sum_{i=1}^{\infty} \left(\int_i^{i+1} f_{\tilde{X}}(x)\,dx \right) \log \left(\int_i^{i+1} f_{\tilde{X}}(x)\,dx \right) \tag{9.79}$$

$$= -\sum_{i=1}^{\infty} \int_i^{i+1} f_{\tilde{X}}(x) \log f_{\tilde{X}}(x)\,dx \tag{9.80}$$

$$= -\int_1^{\infty} f_{\tilde{X}}(x) \log f_{\tilde{X}}(x)\,dx \tag{9.81}$$

$$= h(\tilde{X}), \tag{9.82}$$

since $f_{\tilde{X}}(x) = p_i$ for $i \le x < i+1$.

Hence we have the following chain of inequalities:

$$H(X) = H(X_0) \tag{9.83}$$

$$= h(\tilde{X}) \tag{9.84}$$

$$\le \frac{1}{2} \log(2\pi e) \operatorname{Var}(\tilde{X}) \tag{9.85}$$

$$= \frac{1}{2} \log(2\pi e)(\operatorname{Var}(X_0) + \operatorname{Var}(U)) \tag{9.86}$$

$$= \frac{1}{2} \log(2\pi e) \left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i \right)^2 + \frac{1}{12} \right). \quad \square \tag{9.87}$$

Since entropy is invariant with respect to permutation of $p_1, p_2, \ldots$, we can also obtain a bound by a permutation of the p_i's. We conjecture that a good bound on the variance will be achieved when the high probabilities are close together, i.e, by the assignment $\ldots, p_5, p_3, p_1, p_2, p_4, \ldots$ for $p_1 \ge p_2 \ge \cdots$.

How good is this bound? Let X be a Bernoulli random variable with parameter $\frac{1}{2}$, which implies that $H(X) = 1$. The corresponding random variable X_0 has variance $\frac{1}{4}$, so the bound is

$$H(X) \le \frac{1}{2} \log(2\pi e) \left(\frac{1}{4} + \frac{1}{12} \right) = 1.255 \text{ bits}. \tag{9.88}$$

SUMMARY OF CHAPTER 9

$$h(X) = h(f) = -\int_S f(x) \log f(x)\,dx. \tag{9.89}$$

$$f(X^n) \doteq 2^{-nh(X)}, \text{ a.e.} \tag{9.90}$$

$$\operatorname{Vol}(A_\epsilon^{(n)}) \doteq 2^{nh(X)}. \tag{9.91}$$

$$H([X]_{2^{-n}}) \approx h(X) + n . \tag{9.92}$$

$$h(\mathcal{N}(0, \sigma^2)) = \frac{1}{2} \log 2\pi e\sigma^2 . \tag{9.93}$$

$$h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K| . \tag{9.94}$$

$$D(f\|g) = \int f \log \frac{f}{g} \geq 0 . \tag{9.95}$$

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}) . \tag{9.96}$$

$$h(X|Y) \leq h(X) . \tag{9.97}$$

$$h(aX) = h(X) + \log|a| . \tag{9.98}$$

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} \geq 0 . \tag{9.99}$$

$$\max_{E\mathbf{X}\mathbf{X}^t = K} h(\mathbf{X}) = \frac{1}{2} \log(2\pi e)^n |K| . \tag{9.100}$$

$2^{H(X)}$ is the effective alphabet size for a discrete random variable.
$2^{h(X)}$ is the effective support set size for a continuous random variable.
2^C is the effective alphabet size of a channel of capacity C.

PROBLEMS FOR CHAPTER 9

1. *Differential entropy.* Evaluate the differential entropy $h(X) = -\int f \ln f$ for the following:
   (a) The exponential density, $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.
   (b) The Laplace density, $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$.
   (c) The sum of X_1 and X_2, where X_1 and X_2 are independent normal random variables with means μ_i and variances σ_i^2, $i = 1, 2$.

2. *Concavity of determinants.* Let K_1 and K_2 be two symmetric nonnegative definite $n \times n$ matrices. Prove the result of Ky Fan [103]:

$$|\lambda K_1 + \bar{\lambda} K_2| \geq |K_1|^{\lambda} |K_2|^{\bar{\lambda}}, \quad \text{for } 0 \leq \lambda \leq 1, \ \bar{\lambda} = 1 - \lambda ,$$

   where $|K|$ denotes the determinant of K.
   Hint: Let $\mathbf{Z} = \mathbf{X}_\theta$, where $\mathbf{X}_1 \sim N(0, K_1)$, $\mathbf{X}_2 \sim N(0, K_2)$ and $\theta =$ Bernoulli(λ). Then use $H(\mathbf{Z}|\theta) \leq H(\mathbf{Z})$.

3. *Mutual information for correlated normals.* Find the mutual information $I(X; Y)$, where

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(0, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right).$$

Evaluate $I(X; Y)$ for $\rho = 1$, $\rho = 0$, and $\rho = -1$, and comment.

4. *Uniformly distributed noise.* Let the input random variable X for a channel be uniformly distributed over the interval $-1/2 \le x \le +1/2$. Let the output of the channel be $Y = X + Z$, where the noise random variable is uniformly distributed over the interval $-a/2 \le z \le +a/2$.
   (a) Find $I(X; Y)$ as a function of a.
   (b) For $a = 1$ find the capacity of the channel when the input X is peak-limited; that is, the range of X is limited to $-1/2 \le x \le +1/2$. What probability distribution on X maximizes the mutual information $I(X; Y)$?
   (c) (Optional) Find the capacity of the channel for all values of a, again assuming that the range of X is limited to $-1/2 \le x \le +1/2$.

5. *Quantized random variables.* Roughly how many bits are required on the average to describe to 3 digit accuracy the decay time (in years) of a radium atom if the half-life of radium is 80 years? Note that half-life is the median of the distribution.

6. *Scaling.* Let $h(\mathbf{X}) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}$. Show $h(A\mathbf{X}) = \log|\det(A)| + h(\mathbf{X})$.

HISTORICAL NOTES

Differential entropy and discrete entropy were introduced in Shannon's original paper [238]. The general rigorous definition of relative entropy and mutual information for arbitrary random variables was developed by Kolmogorov [156] and Pinsker [212], who defined mutual information as $\sup_{P,Q} I([X]_P; [Y]_Q)$, where the supremum is over all finite partitions P and Q. The differential entropy bound on discrete entropy was developed independently by J. Massey (unpublished) and by F. Willems (unpublished).