Chapter 9

Differential Entropy

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 \( \mu_i \) and variances \( \sigma_i^2, i = 1, 2 \).

Solution: Differential Entropy.

(a) Exponential distribution.

\[
h(f) = - \int_0^\infty \lambda e^{-\lambda x} \ln \lambda - \lambda x \, dx \quad (9.1)
\]
\[
= - \ln \lambda + 1 \text{ nats.} \quad (9.2)
\]
\[
= \log \frac{e}{\lambda} \text{ bits.} \quad (9.3)
\]

(b) Laplace density.

\[
h(f) = - \int_{-\infty}^{\infty} \frac{1}{2} \lambda e^{-\lambda|x|} \ln \left( \frac{1}{2} + \ln \lambda - \lambda |x| \right) \, dx \quad (9.4)
\]
\[
= - \ln \frac{1}{2} - \ln \lambda + 1 \quad (9.5)
\]
\[
= \ln \frac{2e}{\lambda} \text{ nats.} \quad (9.6)
\]
\[
= \log \frac{2e}{\lambda} \text{ bits.} \quad (9.7)
\]

125
(c) Sum of two normal distributions.

The sum of two normal random variables is also normal, so applying the result
derived the class for the normal distribution, since \( X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \),

\[
h(f) = \frac{1}{2} \log 2\pi e (\sigma^2_1 + \sigma^2_2) \text{ bits.} \tag{9.8}
\]

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[4]:

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

where \( |K| \) denotes the determinant of \( K \).

**Hint:** Let \( Z = X_1 \phi, \) where \( X_1 \sim N(0, K_1) \), \( X_2 \sim N(0, K_2) \) and \( \theta = \text{Bernoulli}(\lambda) \). Then use \( H(Z | \theta) \leq H(Z) \).

**Solution:** Concavity of Determinants. Let \( X_1 \) and \( X_2 \) be normally distributed \( n \)-vectors, \( X_i \sim \phi_K(x), \) \( i = 1, 2 \). Let the random variable \( \theta \) have distribution \( \Pr\{\theta = 1\} = \lambda, \Pr\{\theta = 2\} = 1 - \lambda, \) \( 0 \leq \lambda \leq 1 \). Let \( \theta, X_1, \) and \( X_2 \) be independent and let \( Z = X_\theta \). Then \( Z \) has covariance \( K_Z = \lambda K_1 + (1-\lambda)K_2 \). However, \( Z \) will not be multivariate normal. However, since a normal distribution maximizes the entropy for a given variance, we have

\[
\frac{1}{2} \ln (2\pi e)^n |\lambda K_1 + (1-\lambda)K_2| \geq h(Z) \geq h(Z|\theta) = \frac{1}{2} \ln (2\pi e)^n |K_1| + (1-\lambda) \frac{1}{2} \ln (2\pi e)^n |K_2|.
\]

Thus

\[
|\lambda K_1 + (1-\lambda)K_2| \geq |K_1|^\lambda |K_2|^{1-\lambda}, \tag{9.9}
\]
as desired.

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

\[
\left( \begin{array}{c} X \\ Y \end{array} \right) \sim \mathcal{N}_2 \left( 0, \begin{bmatrix} \sigma^2 & \rho \sigma^2 \\ \rho \sigma^2 & \sigma^2 \end{bmatrix} \right).
\]

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

**Solution:** Mutual information for correlated normals.

\[
\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}_2 \left( 0, \begin{bmatrix} \sigma^2 & \rho \sigma^2 \\ \rho \sigma^2 & \sigma^2 \end{bmatrix} \right) \tag{9.11}
\]

Using the expression for the entropy of a multivariate normal derived in class

\[
h(X,Y) = \frac{1}{2} \log (2\pi e)^2 |K| = \frac{1}{2} \log (2\pi e)^2 \sigma^4 (1-\rho^2).
\]

Since \( X \) and \( Y \) are individually normal with variance \( \sigma^2 \),

\[
h(X) = h(Y) = \frac{1}{2} \log 2\pi e \sigma^2. \tag{9.13}
\]

\[
\]
Hence
\[ I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2} \log(1 - \rho^2). \] (9.14)

(a) \( \rho = 1 \). In this case, \( X = Y \), and knowing \( X \) implies perfect knowledge about \( Y \). Hence the mutual information is infinite, which agrees with the formula.
(b) \( \rho = 0 \). In this case, \( X \) and \( Y \) are independent, and hence \( I(X;Y) = 0 \), which agrees with the formula.
(c) \( \rho = -1 \). In this case, \( X = -Y \), and again the mutual information is infinite as in the case when \( \rho = 1 \).

4. Uniformly distributed noise. Let the input random variable \( X \) to a channel be uniformly distributed over the interval \(-1/2 \leq x \leq +1/2\). Let the output of the channel be \( Y = X + Z \), where the noise random variable is uniformly distributed over the interval \(-a/2 \leq z \leq +a/2\).

(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 \(-1/2 \leq x \leq +1/2\). 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 \(-1/2 \leq x \leq +1/2\).

Solution: Uniformly distributed noise. The probability density function for \( Y = X + Z \) is the convolution of the densities of \( X \) and \( Z \). Since both \( X \) and \( Z \) have rectangular densities, the density of \( Y \) is a trapezoid. For \( a < 1 \) the density for \( Y \) is
\[
pr(y) = \begin{cases} 
(1/2a)(y + (1 + a)/2) & -(1 + a)/2 \leq y \leq -(1 - a)/2 \\
1 & -(1 - a)/2 \leq y \leq +(1 - a)/2 \\
(1/2a)(y - (1 + a)/2) & +(1 - a)/2 \leq y \leq +(1 + a)/2
\end{cases}
\]
and for \( a > 1 \) the density for \( Y \) is
\[
pr(y) = \begin{cases} 
y + (a + 1)/2 & -(a + 1)/2 \leq y \leq -(a - 1)/2 \\
1/a & -(a - 1)/2 \leq y \leq +(a - 1)/2 \\
y - (a + 1)/2 & +(a - 1)/2 \leq y \leq +(a + 1)/2
\end{cases}
\]
(When \( a = 1 \), the density of \( Y \) is triangular over the interval \([-1, +1]\).)

(a) We use the identity \( I(X;Y) = h(Y) - h(Y|X) \). It is easy to compute \( h(Y) \) directly, but it is even easier to use the grouping property of entropy. First suppose that \( a < 1 \). With probability \( 1 - a \), the output \( Y \) is conditionally uniformly distributed in the interval \([- (1 - a)/2, + (1 - a)/2] \); whereas with probability \( a \), \( Y \) has a split triangular density where the base of the triangle has width \( a \). As shown in examples in class,
\[
h(Y) = H(a) + (1 - a) \ln(1 - a) + a(\frac{1}{2} + \ln a)
= -a \ln a - (1 - a) \ln(1 - a) + (1 - a) \ln(1 - a) + \frac{a}{2} + a \ln a = \frac{a}{2} \text{ nats.}
\]
If \( a > 1 \) the trapezoidal density of \( Y \) can be scaled by a factor \( a \), which yields \( h(Y) = \ln a + 1/2a \). Given any value of \( x \), the output \( Y \) is conditionally uniformly distributed over an interval of length \( a \), so the conditional differential entropy in nats is \( h(Y|X) = h(Z) = \ln a \) for all \( a > 0 \). Therefore the mutual information in nats is

\[
I(X;Y) = \begin{cases} 
  a/2 - \ln a & \text{if } a \leq 1 \\
  1/2a & \text{if } a \geq 0 
\end{cases}
\]

As expected, \( I(X;Y) \to \infty \) as \( a \to 0 \) and \( I(X;Y) \to 0 \) as \( a \to \infty \).

(b) As usual with additive noise, we can express \( I(X;Y) \) in terms of \( h(Y) \) and \( h(Z) \):

\[
I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z).
\]

Since both \( X \) and \( Z \) are limited to the interval \([-1/2, +1/2]\), their sum \( Y \) is limited to the interval \([-1, +1]\). The differential entropy of \( Y \) is at most that of a random variable uniformly distributed on that interval; that is, \( h(Y) \leq 1 \). This maximum entropy can be achieved if the input \( X \) takes on its extreme values \( x = \pm 1 \) each with probability \( 1/2 \). In this case, \( I(X;Y) = h(Y) - h(Z) = 1 - 0 = 1 \). Decoding for this channel is quite simple:

\[
\hat{X} = \begin{cases} 
  -1/2 & \text{if } y < 0 \\
  +1/2 & \text{if } y \geq 0
\end{cases}
\]

This coding scheme transmits one bit per channel use with zero error probability. (Only a received value \( y = 0 \) is ambiguous, and this occurs with probability 0.)

(c) When \( a \) is of the form \( 1/m \) for \( m = 2, 3, \ldots \), we can achieve the maximum possible value \( I(X;Y) = \log m \) when \( X \) is uniformly distributed over the discrete points \( \{-1, -1+2/(m-1), \ldots, +1-2/(m-1), +1\} \). In this case \( Y \) has a uniform probability density on the interval \([-1 - 1/(m-1), +1 + 1/(m-1)]\). Other values of \( a \) are left as an exercise.

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.

**Solution:** **Quantized random variables.** The differential entropy of an exponentially distributed random variable with mean \( 1/\lambda \) is \( \log \frac{e}{\lambda} \) bits. If the median is 80 years, then

\[
\int_0^{80} \lambda e^{-\lambda x} \, dx = \frac{1}{2}
\]

or

\[
\lambda = \frac{\ln 2}{80} = 0.00866
\]

and the differential entropy is \( \log e/\lambda \). To represent the random variable to 3 digits \( \approx 10 \) bits accuracy would need \( \log e/\lambda + 10 \) bits = 18.3 bits.
6. *Scaling.* Let \( h(X) = -\int f(x) \log f(x) \, dx \). Show \( h(AX) = \log |\det(A)| + h(X) \).

**Solution:** *Scaling.* Let \( Y = AX \). Then the density of \( Y \) is

\[
g(y) = \frac{1}{|A|} f(A^{-1}y).
\]

Hence

\[
h(AX) = - \int g(y) \ln g(y) \, dy
\]

\[
= - \int \frac{1}{|A|} f(A^{-1}y) \left[ \ln f(A^{-1}y) - \log |A| \right] \, dy
\]

\[
= - \int \frac{1}{|A|} f(x) [\ln f(x) - \log |A|] \, dx
\]

\[
= h(X) + \log |A|.
\]