Elements of Information Theory

George Voutsadakis[1]

[1]Mathematics and Computer Science
Lake Superior State University

LSSU Math 500

1. Differential Entropy
   - Definitions
   - AEP for Continuous Random Variables
   - Relation of Differential Entropy to Discrete Entropy
   - Joint and Conditional Differential Entropy
   - Relative Entropy and Mutual Information
   - Differential and Relative Entropy, and Mutual Information

Subsection 1

Definitions

Probability Density Functions

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

$$\int_{-\infty}^{\infty} f(x) dx = 1,$$

$f(x)$ is called the **probability density function** for X.
The set where $f(x) > 0$ is called the **support set** of X.

Differential Entropy

Definition

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

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

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.
- Therefore, the differential entropy is sometimes written as $h(f)$ rather than $h(X)$.

Example: Uniform Distribution

- Consider a random variable distributed uniformly from 0 to a.

  So its density is $\frac{1}{a}$ from 0 to a and 0 elsewhere.

  Then its differential entropy is

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

Notes:

- 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.
  This is always nonnegative.

Example: Normal Distribution

- Let $X \sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$.

  Then calculating the differential entropy in nats, we obtain

$$\begin{aligned} h(\phi) &= -\int \phi \ln \phi \\ &= -\int \phi(x)[-\tfrac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2}] \\ &= \tfrac{EX^2}{2\sigma^2} + \tfrac{1}{2} \ln 2\pi\sigma^2 \\ &= \tfrac{1}{2} + \tfrac{1}{2} \ln 2\pi\sigma^2 \\ &= \tfrac{1}{2} \ln e + \tfrac{1}{2} \ln 2\pi\sigma^2 \\ &= \tfrac{1}{2} \ln 2\pi e\sigma^2 \text{ nats.} \end{aligned}$$

  Changing the base of the logarithm, we have

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

Subsection 2

AEP for Continuous Random Variables

I.i.d. Sequence and Continuous Entropy

Theorem

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

$$-\frac{1}{n} \log f(X_1, X_2, \ldots, X_n) \rightarrow E[-\log f(X)] = h(X) \text{ in probability.}$$

- The proof follows directly from the weak law of large numbers.

Typical Sets

Definition

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

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

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.

Volumes

- The analog of the cardinality of the typical set for the discrete case is the volume of the typical set for continuous random variables.

Definition

The volume $\text{Vol}(A)$ of a set $A \subseteq \mathbb{R}^n$ is defined as

$$\text{Vol}(A) = \int_A dx_1 dx_2 \cdots dx_n.$$

AEP for Continuous Random Variables

Theorem

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

1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$, for n sufficiently large.
2. $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$ for all n.
3. $\text{Vol}(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$, for n sufficiently large.

1. By the preceding theorem,

$$-\frac{1}{n}\log f(X^n) = -\frac{1}{n}\sum \log f(X_i) \to h(X) \text{ in probability.}$$

   This establishes Part 1.

AEP for Continuous Random Variables (Part 2)

2. For Part 2, we compute

$$\begin{aligned} 1 &= \int_{S^n} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)} dx_1 dx_2 \cdots dx_n \\ &= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \cdots dx_n \\ &= 2^{-n(h(X)+\epsilon)} \mathrm{Vol}(A_\epsilon^{(n)}). \end{aligned}$$

AEP for Continuous Random Variables (Part 3)

3. If n is sufficiently large so that $\Pr(A_\epsilon^{(n)}) > 1-\epsilon$, then

$$\begin{aligned} 1-\epsilon &\leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \\ &\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)} dx_1 dx_2 \cdots dx_n \\ &= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1 dx_2 \cdots dx_n \\ &= 2^{-n(h(X)-\epsilon)} \mathrm{Vol}(A_\epsilon^{(n)}). \end{aligned}$$

Thus, for n sufficiently large, we have

$$(1-\epsilon) 2^{n(h(X)-\epsilon)} \leq \mathrm{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}.$$

Size of the Typical Set

Theorem

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

- Same as in the discrete case.
- 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 $(2^{nh})^{\frac{1}{n}} = 2^h$.
- This provides an interpretation of the differential entropy:

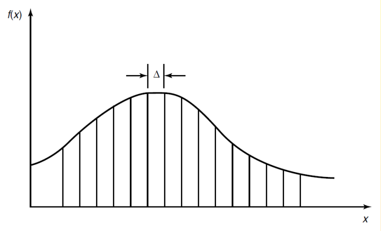
  It is the logarithm of the equivalent side length of the smallest set that contains most of the probability.
- Hence low entropy implies that the random variable is confined to a small effective volume and high entropy indicates that the random variable is widely dispersed.

Subsection 3

Relation of Differential Entropy to Discrete Entropy

Quantization

- Consider a random variable X with density $f(x)$.
- Suppose that 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

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

- We consider the quantized random variable X^Δ, which is defined by

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

Quantization and Entropy

Theorem

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

$$H(X^\Delta) + \log \Delta \to h(f) = h(X), \text{ as } \Delta \to 0.$$

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

- Note that the probability that $X^\Delta = x_i$ is

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

Quantization and Entropy (Cont'd)

- The entropy of the quantized version is

$$\begin{aligned} H(X^\Delta) &= -\sum_{-\infty}^{\infty} p_i \log p_i \\ &= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log (f(x_i)\Delta) \\ &= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \\ &\overset{\sum f(x_i)\Delta = 1}{=} -\sum \Delta f(x_i) \log f(x_i) - \log \Delta. \end{aligned}$$

  If $f(x) \log f(x)$ is Riemann integrable, the first term approaches the integral of $-f(x) \log f(x)$ as $\Delta \to 0$. So in the limit,

$$H(X^\Delta) + \log \Delta \to h(f) = h(X).$$

Examples

1. Let X have uniform distribution on $[0,1]$ and $\Delta = 2^{-n}$.
   Then then $h = 0$ and $H(X^{\Delta}) = n$.
   So n bits suffice to describe X to n bit accuracy.
2. Suppose X is uniformly distributed on $\left[0, \frac{1}{8}\right]$.
   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.
   This agrees with $h(X) = -3$.

Examples (Cont'd)

3. Let $X \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 100$.

   Describing X to n bit accuracy would require on the average

$$n + \frac{1}{2} \log (2\pi e \sigma^2) = n + 5.37 \text{ bits.}$$

- In general, $h(X) + n$ 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 $-\infty$.

  We have $2^{-\infty} = 0$, agreeing with the idea that the volume of the support set of a discrete random variable is zero.

Subsection 4

Joint and Conditional Differential Entropy

Joint Differential Entropy

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

$$h(X_1, X_2, \ldots, X_n) = -\int f(x^n) \log f(x^n) dx^n.$$

Conditional Differential Entropy

Definition

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

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

- Since $f(x|y) = \frac{f(x,y)}{f(y)}$, we can also write

$$\begin{aligned} h(X|Y) &= -\int f(x, y) \log \frac{f(x,y)}{f(y)} dxdy \\ &= -\int f(x, y) \log f(x, y) dxdy + \int f(x, y) \log f(y) dxdy \\ &= -\int f(x, y) \log f(x, y) dxdy + \int f(y) \log f(y) dy \\ &= h(X, Y) - h(Y). \end{aligned}$$

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

Entropy of a Multivariate Normal Distribution

Theorem (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. Then

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

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

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

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

Entropy of a Multivariate Normal Distribution (Cont'd)

- Then

$$
\begin{aligned}
h(f) &= -\int f(\mathbf{x})[-\tfrac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu) - \ln(\sqrt{2\pi})^n |K|^{\frac{1}{2}}]d\mathbf{x} \\
&= \tfrac{1}{2}E[\textstyle\sum_{i,j}(X_i-\mu_i)(K^{-1})_{ij}(X_j-\mu_j)] + \tfrac{1}{2}\ln(2\pi)^n|K| \\
&= \tfrac{1}{2}E[\textstyle\sum_{i,j}(X_i-\mu_i)(X_j-\mu_j)(K^{-1})_{ij}] + \tfrac{1}{2}\ln(2\pi)^n|K| \\
&= \tfrac{1}{2}\textstyle\sum_{i,j}E[(X_j-\mu_j)(X_i-\mu_i)](K^{-1})_{ij} + \tfrac{1}{2}\ln(2\pi)^n|K| \\
&= \tfrac{1}{2}\textstyle\sum_j\sum_i K_{ji}(K^{-1})_{ij} + \tfrac{1}{2}\ln(2\pi)^n|K| \\
&= \tfrac{1}{2}\textstyle\sum_j(KK^{-1})_{jj} + \tfrac{1}{2}\ln(2\pi)^n|K| \\
&= \tfrac{1}{2}\textstyle\sum_j I_{jj} + \tfrac{1}{2}\ln(2\pi)^n|K| \\
&= \tfrac{n}{2} + \tfrac{1}{2}\ln(2\pi)^n|K| \\
&= \tfrac{1}{2}\ln(2\pi e)^n|K| \text{ nats} \\
&= \tfrac{1}{2}\log(2\pi e)^n|K| \text{ bits.}
\end{aligned}
$$

Subsection 5

Relative Entropy and Mutual Information

Relative Entropy or Kullback-Leibler Distance

Definition

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

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

- 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 $0 \log \frac{0}{0} = 0$).

Mutual Information

Definition

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

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

- From the definition it is clear that

$$\begin{aligned} I(X;Y) &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \\ &= h(X) + h(Y) - h(X,Y). \end{aligned}$$

- Moreover,

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

Mutual Information and Quantization

Claim: The mutual information between two random variables is the limit of the mutual information between their quantized versions.

We have

$$\begin{aligned} I(X^\Delta; Y^\Delta) &= H(X^\Delta) - H(X^\Delta | Y^\Delta) \\ &\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta) \\ &= I(X; Y). \end{aligned}$$

Generalization of Quantization

- We can define mutual information in terms of finite partitions of the range of the random variable.
- Let $\mathcal{X}$ be the range of a random variable X.
- A **partition** $\mathcal{P}$ of $\mathcal{X}$ is a finite collection of disjoint sets P_i, such that

$$\bigcup_i P_i = \mathcal{X}.$$

- The **quantization of** X **by** $\mathcal{P}$, denoted $[X]_{\mathcal{P}}$, is the discrete random variable defined by

$$\Pr([X]_{\mathcal{P}} = i) = \Pr(X \in P_i) = \int_{P_i} dF(x).$$

- For two random variables X and Y with partitions $\mathcal{P}$ and $\mathcal{Q}$, we can calculate the mutual information between the quantized versions of X and Y using the discrete definition.

Generalization of Quantization (Cont'd)

Definition

The **mutual information** between two random variables X and Y is given by

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}),$$

where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

- This definition of mutual information always applies, even to joint distributions with atoms, densities and singular parts.
- By continuing to refine the partitions $\mathcal{P}$ and $\mathcal{Q}$, one finds a monotonically increasing sequence $I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}) \nearrow I$.
- This definition of mutual information is equivalent to:
  - The one given above for random variables that have a density;
  - The one given previously for discrete random variables.

Example (Correlated Gaussian Random Variables)

- Let $(X, Y) \sim \mathcal{N}(0, K)$, where $K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$.

  Then

  $$\begin{aligned} h(X) &= h(Y) = \tfrac{1}{2}\log(2\pi e)\sigma^2; \\ h(X, Y) &= \tfrac{1}{2}\log(2\pi e)^2|K| = \tfrac{1}{2}\log(2\pi e)^2\sigma^4(1-\rho^2). \end{aligned}$$

  Therefore,

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

  - If $\rho = 0$, X and Y are independent.
    Then, the mutual information is 0.
  - If $\rho = \pm 1$, X and Y are perfectly correlated.
    Then the mutual information is infinite.

Subsection 6

Differential and Relative Entropy, and Mutual Information

Nonnegativity of Relative Entropy

Theorem

$D(f\|g) \geq 0$ with equality iff $f = g$ almost everywhere (a.e.).

- Let S be the support set of f. Then

$$\begin{aligned} -D(f\|g) &= \int_S f \log \frac{g}{f} \\ &\leq \log \int_S f \frac{g}{f} \quad \text{(by Jensen's inequality)} \\ &= \log \int_S g \\ &\leq \log 1 = 0. \end{aligned}$$

  We have equality iff we have equality in Jensen's inequality.

  This occurs iff $f = g$ a.e.

Consequences

Corollary

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

- We have $I(X;Y) = D(f(x,y) \| f(x)f(y)) \geq 0$.
  Equality holds iff $f(x,y) = f(x)f(y)$ a.e..
  That is, iff X and Y are independent.

Corollary

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

- We have $h(X) - h(X|Y) = I(X;Y) \geq 0$.
  Equality holds iff X and Y are independent.

Chain Rule for Differential Entropy

Theorem (Chain Rule for Differential Entropy)

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

- Follows directly from the definitions.

Corollary

$$h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i),$$

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

- Follows directly from the preceding theorem and the preceding corollary.

Application: Hadamard's Inequality

- Let $\boldsymbol{X} \sim \mathcal{N}(0, K)$ be a multivariate normal random variable.
- Calculating the entropy in the above inequality gives us

$$|K| \leq \prod_{i=1}^{n} K_{ii}.$$

- This is **Hadamard's inequality**.
- A number of determinant inequalities can be derived in this fashion from information-theoretic inequalities.

Translation Invariance

Theorem

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

Translation does not change the differential entropy.

- Follows directly from the definition of differential entropy.

Scaling

Theorem

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

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

$$\begin{aligned} h(aX) &= -\int f_Y(y)\log f_Y(y)dy \\ &= -\int \frac{1}{|a|} f_X(\tfrac{y}{a}) \log\left(\frac{1}{|a|} f_X(\tfrac{y}{a})\right) dy \\ &= -\int f_X(x)\log f_X(x)dx + \log|a| \\ &= h(X) + \log|a|. \end{aligned}$$

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

Corollary

$h(A\boldsymbol{X}) = h(\boldsymbol{X}) + \log|\det(A)|$.

Maximization Property of Normal Distribution

- The multivariate normal distribution maximizes the entropy over all distributions with the same covariance.

Theorem

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

- Let $g(\boldsymbol{x})$ be any density satisfying

$$\int g(\boldsymbol{x})x_ix_jd\boldsymbol{x} = K_{ij}, \quad \text{for all } i, j.$$

Let ϕ_K be the density of a $\mathcal{N}(0, K)$ vector, with

$$f(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^n|K|^{\frac{1}{2}}}e^{-\frac{1}{2}\boldsymbol{x}^TK^{-1}\boldsymbol{x}}.$$

Note that $\log\phi_K(\boldsymbol{x})$ is a quadratic form and $\int x_ix_j\phi_K(\boldsymbol{x})d\boldsymbol{x} = K_{ij}$.

Maximization Property of Normal Distribution (Cont'd)

- Now we have

$$\begin{array}{rcl} 0 & \leq & D(g\|\phi_K) \\ & = & \int g \log \frac{g}{\phi_K} \\ & = & -h(g) - \int g \log \phi_K \\ & = & -h(g) - \int \phi_K \log \phi_K \\ & = & -h(g) + h(\phi_K). \end{array}$$

The equality $\int g \log \phi_K = \int \phi_K \log \phi_K$ holds since g and ϕ_K yield the same moments of the quadratic form $\log \phi_K(\boldsymbol{x})$.

Estimation Error and Differential Entropy

- Let X be a random variable with differential entropy $h(X)$.
- Let $\widehat{X}$ be an estimate of X.
- Let $E(X-\widehat{X})^2$ be the expected prediction error.
- Let $h(X)$ be in nats.

Theorem (Estimation Error and Differential Entropy)

For any random variable X and estimator $\widehat{X}$,

$$E(X-\widehat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)},$$

with equality if and only if X is Gaussian and $\widehat{X}$ is the mean of X.

Estimation Error and Differential Entropy (Proof)

- Let $\widehat{X}$ be any estimator of X. Then

$$\begin{aligned} E(X-\widehat{X})^2 &\geq \min_{\widehat{X}} E(X-\widehat{X})^2 \\ &= E(X-E(X))^2 \quad \text{(mean is best estimator)} \\ &= \operatorname{var}(X) \\ &\geq \tfrac{1}{2\pi e}e^{2h(X)}. \quad (h(X) \leq \tfrac{1}{2}\ln 2\pi e \operatorname{var}(X)) \end{aligned}$$

We have equality only if $\widehat{X}$ is the mean of X and X is Gaussian.

Corollary

Given side information Y and estimator $\widehat{X}(Y)$, it follows that

$$E(X-\widehat{X}(Y))^2 \geq \frac{1}{2\pi e}e^{2h(X|Y)}.$$