INFORMATION THEORY & CODING

Differential Entropy

Dr. Rui Wang

Department of Electrical and Electronic Engineering
Southern Univ. of Science and Technology (SUSTech)

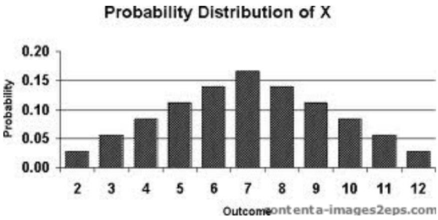
Email: wang.r@sustech.edu.cn

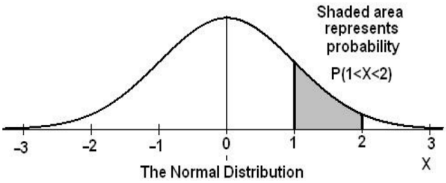
December 11, 2023

Differential Entropy - 1

- Definitions
- AEP for Continuous Random Variables
- Relation of differential entropy to discrete entropy

From discrete to continuous variables





Differential Entropy

Definition

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

Definition

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

$$h(X) = -\int_{\mathcal{S}} f(x) \log f(x) \mathrm{d}x = h(f),$$

where $\mathcal{S}$ is the support set of the random variable.

Example: Uniform distribution

- $f(x) = \frac{1}{a}, x \in [0, a]$
- The differential entropy is:

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \mathrm{d}x = \log a \text{ bits}$$

- for $a < 1$, $h(X) = \log a < 0$, differential entropy can be negative! (unlike discrete entropy)

Example: Uniform distribution

- $f(x) = \frac{1}{a}, x \in [0, a]$
- The differential entropy is:

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \mathrm{d}x = \log a \text{ bits}$$

- for $a < 1$, $h(X) = \log a < 0$, differential entropy can be negative! (unlike discrete entropy)

Example: Normal distribution

- $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-x^2}{2\sigma^2})$, $x \in \mathbb{R}$
- Differential entropy:

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

Calculation:

$$\begin{aligned} h(\phi) &= -\int \phi \log \phi \mathrm{d}x = -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} \log e - \log \sqrt{2\pi\sigma^2} \right] \mathrm{d}x \\ &= \frac{\mathbb{E}(X^2)}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2 = \frac{1}{2} \log e + \frac{1}{2} \log 2\pi\sigma^2 \\ &= \frac{1}{2} \log 2\pi e \sigma^2 \end{aligned}$$

Example: Normal distribution

- $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-x^2}{2\sigma^2})$, $x \in \mathbb{R}$
- Differential entropy:

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

Calculation:

$$\begin{aligned}
h(\phi) &= -\int \phi \log \phi \mathrm{d}x = -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} \log e - \log \sqrt{2\pi\sigma^2} \right] \mathrm{d}x \\
&= \frac{\mathbb{E}(X^2)}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2 = \frac{1}{2} \log e + \frac{1}{2} \log 2\pi\sigma^2 \\
&= \frac{1}{2} \log 2\pi e \sigma^2
\end{aligned}$$

AEP for continuous random variables

- Discrete world: for a sequence of i.i.d. random variables

$$\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X).$$

- Continuous world: for a sequence of i.i.d. random variables

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

Proof follows from the weak law of large numbers.

AEP for continuous random variables

- Discrete world: for a sequence of i.i.d. random variables

$$\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X).$$

- Continuous world: for a sequence of i.i.d. random variables

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

Proof follows from the weak law of large numbers.

Typical set

- Discrete case: number of typical sequences

$$\left|A_\epsilon^{(n)}\right| \approx 2^{nH(X)}$$

- Continuous case: The volume of the typical set

$$\mathsf{Vol}(A) = \int_A \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n, \ A \subset \mathbb{R}^n.$$

Definition

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

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

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

Typical set

Theorem

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

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

Proof. 1.

Similar to the discrete case.
By definition, $-\frac{1}{n}\log f(X^n) = -\frac{1}{n}\sum \log f(X_i) \to h(X)$ in probability. □

Typical set

Theorem

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

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

Proof. 1.

Similar to the discrete case.
By definition, $-\frac{1}{n}\log f(X^n) = -\frac{1}{n}\sum \log f(X_i) \to h(X)$ in probability. □

Typical set

Theorem

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

2. $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$ *for all* n*.*

Poof. 2.

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

□

Typical set

Theorem

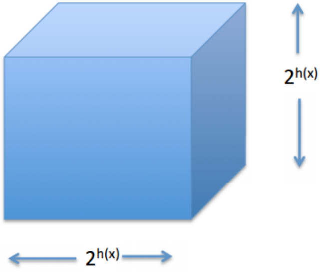
The typical set $A_\epsilon^{(n)}$ *has the following properties:*
3. $\mathsf{Vol}(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ *for* n *sufficiently large.*

Proof. 3.

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

□

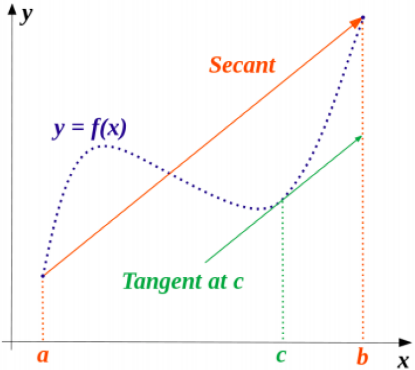
- The volume of the smallest set that contains most of the probability is approximately $2^{nh(X)}$.
- For an n-dim volume, this means that each dim has length $\left(2^{nh(X)}\right)^{\frac{1}{n}} = 2^{h(X)}$.



Mean value theorem (MVT)

If a function f is continuous on the closed interval $[a,b]$, and differentiable on (a,b), then there exists a point $c \in (a,b)$ such that

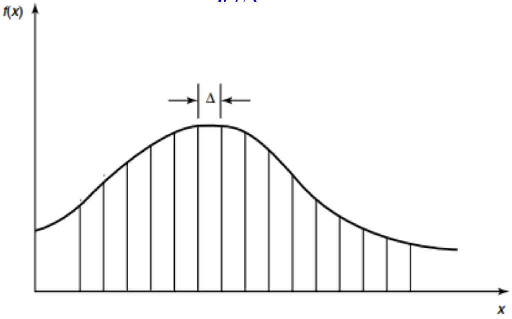
$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



Relation of differential entropy to discrete entropy

- Consider a random variable X with pdf $f(x)$. We divide the range of X into bins of length Δ.
- MVT: there exists a value $x_i \in (i\Delta, (i+1)\Delta)$ within each bin such that

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



Relation of differential entropy to discrete entropy

- Define the quantized random variable as $X^{\Delta} = x_i$ if $i\Delta \leq X \leq (i+1)\Delta$ with pmf

$$p_i = \Pr[X^{\Delta} = x_i] = \int_{i\Delta}^{(i+1)\Delta} f(x)\mathrm{d}x = f(x_i)\Delta.$$

- The entropy of X^{Δ} is

$$H(X^{\Delta}) = -\sum_{-\infty}^{+\infty} p_i \log p_i = -\sum \Delta f(x_i) \log f(x_i) - \log \Delta.$$

- If $f(x)$ is is Riemann integrable, as $\Delta \to 0$,

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

- Define the quantized random variable as $X^\Delta = x_i$ if $i\Delta \le X \le (i+1)\Delta$ with pmf

$$p_i = \Pr[X^\Delta = x_i] = \int_{i\Delta}^{(i+1)\Delta} f(x)\mathrm{d}x = f(x_i)\Delta.$$

- The entropy of X^Δ is

$$H(X^\Delta) = -\sum_{-\infty}^{+\infty} p_i \log p_i = -\sum \Delta f(x_i) \log f(x_i) - \log \Delta.$$

- If $f(x)$ is is Riemann integrable, as $\Delta \to 0$,

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

Relation of differential entropy to discrete entropy

- Define the quantized random variable as $X^\Delta = x_i$ if $i\Delta \leq X \leq (i+1)\Delta$ with pmf

$$p_i = \Pr[X^\Delta = x_i] = \int_{i\Delta}^{(i+1)\Delta} f(x)\mathrm{d}x = f(x_i)\Delta.$$

- The entropy of X^Δ is

$$H(X^\Delta) = -\sum_{-\infty}^{+\infty} p_i \log p_i = -\sum \Delta f(x_i) \log f(x_i) - \log \Delta.$$

- If $f(x)$ is is Riemann integrable, as $\Delta \to 0$,

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

INFORMATION THEORY & CODING

Differential Entropy 2

Dr. Rui Wang

Department of Electrical and Electronic Engineering
Southern Univ. of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

December 27, 2022

Differential Entropy - 2

- Definitions
- AEP for Continuous Random Variables
- Relation of differential entropy to discrete entropy
- Joint and Conditional Differential Entropy
- Relative Entropy and Mutual Information
- Estimation Counterpart of Fano's Inequality

Joint and conditional differential entropy

Definition

The joint differential entropy of $X_1, X_2, ..., X_n$ with pdf $f(x_1, x_2, \ldots, x_n)$ is

$$h(X_1, X_2, \ldots, X_n) = -\int f(x^n) \log f(x^n) \mathrm{d}x^n.$$

Definition

If X, Y have a joint pdf $f(x, y)$, the conditional differential entropy $h(X|Y)$ is

$$h(X|Y) = -\int f(x, y) \log f(x|y) \mathrm{d}x\mathrm{d}y = h(X, Y) - h(Y).$$

Entropy of a multivariate Gaussian

Definition (Multivariate Gaussian Distribution)

If the joint pdf of $X_1, X_2, \dots, X_n$ satisfies

$$f(\mathbf{x}) = f(x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)\right),$$

then $X_1, X_2, \dots, X_n$ are multivariate/joint Gaussian/normal distributed with mean μ and covariance matrix K. Denote as $(X_1, X_2, \dots, X_n) \sim \mathcal{N}_n(\mu, K)$.

Theorem (Entropy of a multivariate normal distribution)

Let $X_1, X_2, \dots, X_n$ have multivariate normal distribution with mean μ and covariance matrix K. Then

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

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

Definition

The relative entropy $D(f||g)$ between two pdfs f and g is

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

Note: $D(f||g)$ is finite only if the support set of f is contained in the support set of g.

Definition

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

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

Relative entropy and mutual information

Definition

The relative entropy $D(f||g)$ between two pdfs f and g is

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

Note: $D(f||g)$ is finite only if the support set of f is contained in the support set of g.

Definition

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

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

By definition, it is clear that

$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y).$$

and

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

Mutual information between correlated Gaussian r.v.s

- Let $(X, Y) \sim \mathcal{N}(0, K)$, where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

- $h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$
- $h(X, Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}(\log 2\pi e)^2\sigma^4(1-\rho^2)$
- $I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2}\log(1-\rho^2)$

if $\rho = 0$, X and Y are independent, the mutual information is 0.

if $\rho \pm 1$, X and Y are perfectly correlated, the mutual information is infinite.

Mutual information between correlated Gaussian r.v.s

- Let $(X, Y) \sim \mathcal{N}(0, K)$, where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

- $h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$
- $h(X, Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}(\log 2\pi e)^2\sigma^4(1-\rho^2)$
- $I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2}\log(1-\rho^2)$

if $\rho = 0$, X and Y are independent, the mutual information is 0.

if $\rho \pm 1$, X and Y are perfectly correlated, the mutual information is infinite.

Theorem

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

Proof.

Let $\mathcal{S}$ be the support set of f. Then

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

Theorem

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

Proof.

Let $\mathcal{S}$ be the support set of f. Then

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

□

Properties of differential entropy

- $I(X;Y) \geq 0$ with equality iff X and Y are independent.
- $h(X|Y) \leq h(X)$ with equality iff X and Y are independent.

Theorem (Chain rule for differential entropy)

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

- $h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i)$, with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Properties of differential entropy

Theorem (Translation does not change the differential entropy)

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

Theorem

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

Proof.

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

$$h(aX) = -\int f_Y(y) \log f_Y(y) \mathrm{d}y = -\int \frac{1}{|a|} f_X(\frac{y}{a}) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) \mathrm{d}y$$

$$= -\int f_X(x) \log f_X(x) \mathrm{d}x + \log|a| = h(X) + \log|a|$$

□

Theorem (Translation does not change the differential entropy)

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

Theorem

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

Proof.

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

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

$\square$

Properties of differential entropy

Theorem (Translation does not change the differential entropy)

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

Theorem

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

Corollary.

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

□

Multivariate Gaussian maximizes the entropy

Theorem

Let the random vector $\mathbf{X} \in \mathbb{R}^n$ have zero mean and covariance $K = \mathbb{E}\mathbf{X}\mathbf{X}^t$ (i.e., $K_{ij} = \mathbb{E}X_iX_j$, $1 \le i, j \le n$). Then

$$h(\mathbf{X}) \le \frac{1}{2}\log(2\pi e)^n|K|$$

with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.

Random variable X, estimator $\hat{X}$. The expected prediction error $\mathbf{E}(X - \hat{X})^2$.

Theorem (Estimation error and differential entropy)

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

$$\mathbb{E}(X - \hat{X})^2 \geq \frac{1}{2\pi e} \exp\Big(2h(X)\Big),$$

with equality iff X is Gaussian and $\hat{X}$ is the mean of X.

Theorem (Estimation error and differential entropy)

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

$$\mathbb{E}(X - \hat{X})^2 \geq \frac{1}{2\pi e} \exp\Big(2h(X)\Big),$$

with equality iff X is Gaussian and $\hat{X}$ is the mean of X.

Proof.

We have

$$\begin{aligned}
\mathbb{E}(X - \hat{X})^2 &\geq \min_{\hat{X}} \mathbb{E}(X - \hat{X})^2 \\
&= \mathbb{E}(X - \mathbb{E}(X))^2 \quad \text{mean is the best estimator} \\
&= \mathsf{Var}(X) \\
&\geq \frac{1}{2\pi e} \exp\Big(2h(X)\Big). \quad \text{The Gaussian has maximum entropy}
\end{aligned}$$

□

- Discrete r.v. $\Rightarrow$ continuous r.v.
- entropy $\Rightarrow$ differential entropy.
- Many things similar: mutual information, relative entropy, AEP, chain rule, ...

  Some things different: $h(X)$ can be negative, maximum entropy distribution is Gaussian