Chapter 8

Differential Entropy

Peng-Hua Wang

Graduate Inst. of Comm. Engineering

National Taipei University

Chapter Outline

8.1 Definitions

Definitions

Definition 1 (Differential entropy) *The differential entropy $h(X)$ of a continuous random variable X with pdf $f(X)$ is defined as*

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

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

Example

$$X \sim U(0, a), \quad h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} dx = \log a.$$

Differential Entropy of Gaussian

Example. If $X \sim N(0, \sigma^2)$ with pdf $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, then

$$
\begin{aligned}
h_a(x) &= -\int \phi(x) \log_a \phi(x) dx \\
&= -\int \phi(x) \left(\log_a \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2} \log_a e \right) dx \\
&= \frac{1}{2} \log_a (2\pi\sigma^2) + \frac{\log_a e}{2\sigma^2} E_\phi[X^2] = \frac{1}{2} \log_a (2\pi e \sigma^2) \quad \square
\end{aligned}
$$

Differential Entropy of Gaussian

Remark. If a random variable with pdf $f(x)$ has zero mean and variance σ^2, then

$$
\begin{aligned}
&-\int f(x)\log_a \phi(x)dx \\
=&-\int f(x)\left(\log_a \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2}\log_a e\right)dx \\
=&\frac{1}{2}\log_a(2\pi\sigma^2) + \frac{\log_a e}{2\sigma^2}E_f[X^2] = \frac{1}{2}\log_a(2\pi e\sigma^2)
\end{aligned}
$$

Gaussian has Maximal Differential Entropy

Suppose that a random variable X with pdf $f(x)$ has zero mean and variance σ^2, what is its maximal differential entropy?
Let $\phi(x)$ be the pdf of $N(0, \sigma^2)$.

$$h(X) + \int f(x) \log \phi(x) dx = \int f(x) \log \frac{\phi(x)}{f(x)} dx$$

$$\leq \log \left(\int f(x) \frac{\phi(x)}{f(x)} dx \right) \quad \text{(convexity of logarithm)}$$

$$= \log \int \phi(x) dx = 0$$

That is,

$$h(X) \leq - \int f(x) \log \phi(x) dx = \frac{1}{2} \log(2\pi e \sigma^2)$$

and equality holds if $f(x) = \phi(x)$. $\square$

8.2 AEP for Continuous Random Variables

AEP

Theorem 1 (AEP) *Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with common pdf $f(x)$. Then,*

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

in probability.

Definition 2 (Typical Set) *For $\epsilon > 0$ the typical set $A_\epsilon^{(n)}$ with respect to $f(x)$ is defined as*

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

AEP

Definition 3 (Volume) *The volume* $\mathrm{Vol}(A)$ *of a set* $A \subset \mathcal{R}^n$ *is defined as*

$$\mathrm{Vol}(A) = \int_A dx_1 dx_2 \ldots dx_n$$

Theorem 2 (Properties of typical set) *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.*

8.4 Joint and Conditional Differential Entropy

Definitions

Definition 4 (Differential entropy) *The differential entropy of jointly distributed random variables* $X_1, X_2, \ldots, X_n$ *is defined as*

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

where $f(x^n) = f(x_1, x_2, \ldots, x_n)$ *is the joint pdf.*

Definition 5 (Conditional differential entropy) *The conditional differential entropy of jointly distributed random variables* X, Y *with joint pdf* $f(x, y)$ *is defined as, if it exists,*

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

Multivariate Normal Distribution

Theorem 3 (Entropy of a multivariate normal) *Let* $X_1, X_2, \ldots, X_n$ *have a multivariate normal distribution with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* $\mathbf{K}$. *Then*

$$h(X_1, X_2, \ldots, X_n) = \frac{1}{2} \log(2\pi e)^n |\mathbf{K}|$$

Proof. The joint pdf of a multivariate normal distribution is

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{K}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Multivariate Normal Distribution

Therefore,

$$h(X_1, X_2, \ldots, X_n) = -\int \phi(\mathbf{x}) \log_a \phi(\mathbf{x}) d\mathbf{x}$$

$$= \int \phi(\mathbf{x}) \left[\frac{1}{2} \log_a (2\pi)^n |\mathbf{K}| + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \log_a e \right] d\mathbf{x}$$

$$= \frac{1}{2} \log_a (2\pi)^n |\mathbf{K}| + \frac{1}{2} (\log_a e) \underbrace{E\left[(\mathbf{x} - \boldsymbol{\mu})^t \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]}_{=n}$$

$$= \frac{1}{2} \log_a (2\pi)^n |\mathbf{K}| + \frac{1}{2} n \log_a e$$

$$= \frac{1}{2} \log_a (2\pi e)^n |\mathbf{K}| \quad \square$$

Multivariate Normal Distribution

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^t$ be a random vector. If $\mathbf{K} = E[\mathbf{Y}\mathbf{Y}^t]$, then $E[\mathbf{Y}^t\mathbf{K}^{-1}\mathbf{Y}] = n$.

Proof. Denote

$$\mathbf{K} = E[\mathbf{Y}\mathbf{Y}^t] = \begin{pmatrix} | & | & & | \\ \mathbf{k}_1 & \mathbf{k}_2 & \ldots & \mathbf{k}_n \\ | & | & & | \end{pmatrix}$$

and

$$\mathbf{K}^{-1} = \begin{pmatrix} — & \mathbf{a}_1^t & — \\ — & \mathbf{a}_2^t & — \\ & \vdots & \\ — & \mathbf{a}_n^t & — \end{pmatrix}$$

We have $\mathbf{k}_i = E[Y_i\mathbf{Y}]$ and $\mathbf{a}_j^t\mathbf{k}_i = \delta_{ij}$.

Multivariate Normal Distribution

Now,

$$\mathbf{Y}^t\mathbf{K}^{-1}\mathbf{Y} = \mathbf{Y}^t \begin{pmatrix} — & \mathbf{a}_1^t & — \\ — & \mathbf{a}_2^t & — \\ & \vdots & \\ — & \mathbf{a}_n^t & — \end{pmatrix} \mathbf{Y} = (Y_1, Y_2, \ldots, Y_n) \begin{pmatrix} \mathbf{a}_1^t\mathbf{Y} \\ \mathbf{a}_2^t\mathbf{Y} \\ \vdots \\ \mathbf{a}_n^t\mathbf{Y} \end{pmatrix}$$

$$= Y_1\mathbf{a}_1^t\mathbf{Y} + Y_2\mathbf{a}_2^t\mathbf{Y} + \cdots + Y_n\mathbf{a}_n^t\mathbf{Y}$$

and

$$E[\mathbf{Y}^t\mathbf{K}^{-1}\mathbf{Y}] = \mathbf{a}_1^t E[Y_1\mathbf{Y}] + \mathbf{a}_2^t E[Y_2\mathbf{Y}] + \cdots + \mathbf{a}_n^t E[Y_n\mathbf{Y}]$$

$$= \mathbf{a}_1^t\mathbf{k}_1 + \mathbf{a}_2^t\mathbf{k}_2 + \cdots + \mathbf{a}_n^t\mathbf{k}_n = n \quad \square$$

8.5 Relative Entropy and Mutual Information

Definitions

Definition 6 (Relative entropy) *The relative entropy (or KullbackVLeibler distance)* $D(f||g)$ *between two densities* $f(x)$ *and* $g(x)$ *is defined as*

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

Definition 7 (Mutual information) *The mutual information* $I(X;Y)$ *between two random variables with joint density* $f(x,y)$ *is defined as*

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

Example

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

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

Then $h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$ and

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

Therefore,

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

8.6 Properties of Differential Entropy and Related Amounts

Properties

Theorem 4 (Relative entropy)

$$D(f||g) \geq 0$$

with equality iff $f = g$ almost everywhere.

Corollary 1 *1. $I(X;Y) \geq 0$ with equality iff X and Y are independent.*

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

Properties

Theorem 5 (Chain rule for differential entropy)

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

Corollary 2

$$h(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i)$$

Corollary 3 (Hadamard's inequality) *If $\mathbf{K}$ is the covariance matrix of a multivariate normal distribution, then*

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

Properties

Theorem 6 *1.* $h(X + c) = h(X)$

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

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

Gaussian has Maximal Entropy

Theorem 7 *Let the random vector* $\mathbf{X} \in R^n$ *have zero mean and covariance* $\mathbf{K} = E[\mathbf{X}\mathbf{X}^t]$. *Then* $h(\mathbf{X}) \le \frac{1}{2}\log(2\pi e)^n|\mathbf{K}|$. *with equality* $\mathbf{X} \sim N(\mathbf{0}, \mathbf{K})$

Proof. Let $g(\mathbf{x})$ be any density satisfying $\int x_i x_j g(\mathbf{x}) d\mathbf{x} = K_{ij}$. Let $\phi(\mathbf{x})$ be the density of $N(\mathbf{0}, \mathbf{K})$. Then,

$$0 \le D(g||\phi) = \int g \log(g/\phi) = -h(g) - \int g \log \phi$$

$$= -h(g) - \int \phi \log \phi = -h(g) + h(\phi)$$

That is, $h(g) \le h(\phi)$. Equality holds if $g = \phi$. $\square$