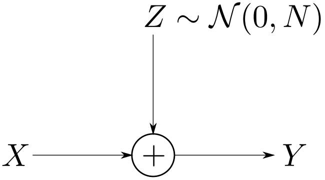
EEE 551 Information Theory (Spring 2022)

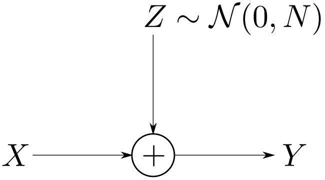
Chapter 8: Differential Entropy

Motivation: Gaussian Channel



$X, Y \in \mathbb{R}$ (continuous-valued)

Motivation: Gaussian Channel



$X, Y \in \mathbb{R}$ (continuous-valued)

Standard entropy cannot be used with continuous variables

$$H(Z) = \infty$$

Brief Review of Continuous Random Variables

- For any random variable $X \in \mathbb{R}$, its cumulative distribution function (CDF) is

$$F_X(x) = \Pr\{X \leq x\}.$$

- $F_X(x)$ is non-decreasing, right-continuous, $F_X(-\infty) = 0$, $F_X(\infty) = 1$
- A random variable is **continuous** if $F_X(x)$ is a continuous function
- The probability density function (PDF) of X is defined by

$$f_X(x) = \frac{dF_X(x)}{dx}$$

- The PDF always satisfies $f_X(x) \geq 0$ and $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = 1$
- For any set $A \subset \mathbb{R}$,

$$\Pr\{X \in A\} = \int_A f_X(x)\,dx$$

- $\displaystyle\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\,f_X(x)\,dx$
- As with PMFs, we often use the notation $f(x) = f_X(x)$

Definition of Differential Entropy

- Let $X \in \mathbb{R}$ be a continuous random variable with PDF $f(x)$
- The **differential entropy** of X is given by

$$
\begin{aligned}
h(X) &= -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx \\
&= \mathbb{E}\big[-\log f(X)\big]
\end{aligned}
$$

- Sometimes written $h(f)$

Example 1: Uniform random variable

Let X be uniform on $[0, a]$

$$f(x) = \begin{cases} \dfrac{1}{a} & x \in [0, a] \\ 0 & \text{otherwise} \end{cases}$$

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

Example 1: Uniform random variable

Let X be uniform on $[0, a]$

$$f(x) = \begin{cases} \dfrac{1}{a} & x \in [0, a] \\ 0 & \text{otherwise} \end{cases}$$

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

Differential entropy is weird!

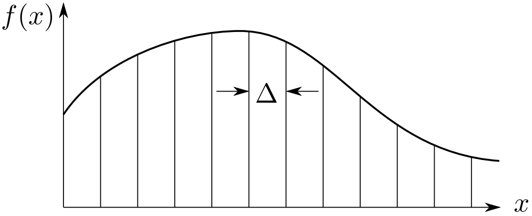
- Can be negative
- For constant c, $h(cX) \neq h(X)$

Example 2: Gaussian random variable

- $X \sim \mathcal{N}(0, \sigma^2)$, i.e. $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$
- $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$

Relationship between differential and discrete entropy

Differential entropy is related to the (discrete) entropy of a quantization of the continuous variable



- Fix $\Delta > 0$
- Define quantized random variable $X^\Delta \in \mathbb{Z}$ as

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

Theorem

$$h(X) = \lim_{\Delta \to 0} H(X^\Delta) + \log \Delta$$

That is, for small Δ, $H(X^\Delta) \approx h(X) - \log \Delta$

Proof:

- For each integer i, choose $a_i \in [i\Delta, (i+1)\Delta)$ so that

$$f(a_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx \qquad \text{(exists by mean value theorem)}$$

- $\Pr\{X^\Delta = i\} = \Pr\{i\Delta \le X < (i+1)\Delta\} = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx = f(a_i)\Delta$

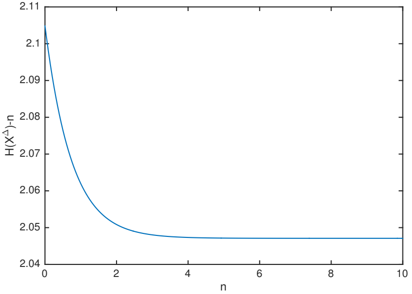
- $$\begin{aligned} H(X^\Delta) &= -\sum_{i=-\infty}^{\infty} f(a_i)\Delta \log\big(f(a_i)\Delta\big) \\ &= -\sum_{i=-\infty}^{\infty} f(a_i)\Delta \log f(a_i) - \sum_{i=-\infty}^{\infty} f(a_i)\Delta \log\Delta \\ &= -\Delta \sum_{i=-\infty}^{\infty} f(a_i)\log f(a_i) - \log\Delta \end{aligned}$$

- $$\begin{aligned} \lim_{\Delta\to 0} H(X^\Delta) + \log\Delta &= \lim_{\Delta\to 0} -\Delta \sum_{i=-\infty}^{\infty} f(a_i)\log f(a_i) \\ &= -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx \\ &= h(X) \end{aligned}$$

Examples

Let $\Delta = 2^{-n}$. Then X^Δ represents X truncated to n bits after the decimal point

- If $X \sim$ Unif$[0, 1]$, then $H(X^\Delta) \approx h(X) - \log \Delta = n$ bits
- If $X \sim$ Unif$[0, 1/8]$, then $H(X^\Delta) \approx h(X) - \log \Delta = -3 + n$ bits
- If $X \sim \mathcal{N}(0, 1)$, then $H(X^\Delta) \approx h(X) - \log \Delta = 2.047 + n$ bits



Joint and Conditional Differential Entropy

Given $(X_1, X_2, \ldots, X_n) \sim f(x^n)$, the **joint differential entropy** is given by

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

Given $(X, Y) \sim f(x, y)$, the **conditional differential entropy** is given by

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

Joint Differential Entropy Example

Let $X_1, X_2, \ldots, X_n$ have multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}$, i.e. $X^n \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$

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

$$h(X^n) = \frac{1}{2}\log\left[(2\pi e)^n|\mathbf{K}|\right]$$

Relative Entropy and Mutual Information

- Given two densities $f(x)$ and $g(x)$, the **relative entropy** is given by

$$
\begin{aligned}
D(f\|g) &= \int f(x)\log\frac{f(x)}{g(x)}\,dx \\
&= \mathbb{E}_f\left[\log\frac{f(X)}{g(X)}\right]
\end{aligned}
$$

- For $(X,Y)\sim f(x,y)$, the **mutual information** is given by

$$
\begin{aligned}
I(X;Y) &= \int f(x,y)\log\frac{f(x,y)}{f(x)\,f(y)}\,dx\,dy \\
&= D\big(f(x,y)\|f(x)\,f(y)\big) \\
&= h(X)+h(Y)-h(X,Y) \\
&= h(X)-h(X|Y) \\
&= h(Y)-h(Y|X)
\end{aligned}
$$

If X^Δ, Y^Δ are quantized version of X, Y, then

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

Mutual Information Example

Let $(X, Y) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K} = \left[\begin{array}{cc} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{array} \right]$

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

Properties

- $D(f\|g) \geq 0$ with equality iff $f = g$ almost everywhere
- $I(X;Y) \geq 0$ with equality iff X, Y are independent
- $h(X|Y) \leq h(X)$ with equality iff X, Y are independent
- $h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, \ldots, X_{i-1})$
- $h(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i)$ with equality iff $X_1, \ldots, X_n$ are independent
- $h(X + c) = h(X)$
- For any real number a, $h(aX) = h(X) + \log|a|$

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

$$\begin{aligned}
h(Y) &= \mathbb{E}[-\log f_Y(Y)] \\
&= \mathbb{E}\left[-\log \frac{1}{|a|} f_X(Y/a)\right] \\
&= \mathbb{E}\left[-\log f_X(X)\right] + \log|a| \\
&= h(X) + \log|a|
\end{aligned}$$

- For square matrix $\mathbf{A}$, $h(\mathbf{A}\,X^n) = h(X^n) + \log|\det(\mathbf{A})|$

Example:

$$\begin{aligned} I(aX;bY) &= h(aX) + h(bY) - h(aX,bY) \\ &= \big[h(X) + \log|a|\big] + \big[h(Y) + \log|b|\big] - \big[h(X,Y) + \log|a|\cdot|b|\big] \\ &= I(X;Y) \end{aligned}$$

Differential Entropy Maximization

If $X \in \mathbb{R}$ be a random variable with $\mathbb{E}X^2 \leq \sigma^2$, then

$$h(X) \leq \frac{1}{2} \log 2\pi e \sigma^2 \quad \text{with equality iff } X \sim \mathcal{N}(0, \sigma^2)$$

More generally, if $X^n \in \mathbb{R}^n$ is a random vector with covariance matrix $\mathbf{K}$, then

$$h(X^n) \leq \frac{1}{2} \log (2\pi e)^n |\mathbf{K}| \quad \text{with equality iff } X^n \sim \mathcal{N}(0, \mathbf{K})$$

Proof of scalar property

- Let $f(x)$ be the pdf of X
- Let $\phi(x) = \mathcal{N}(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$

$$
\begin{aligned}
0 &\leq D(f \| \phi) \\
&= \int f(x) \log \frac{f(x)}{\phi(x)} \, dx \\
&= -h(f) - \int f(x) \log \phi(x) \, dx \\
&= -h(f) - \int f(x) \left[-\frac{x^2}{2\sigma^2} \log e - \log \sqrt{2\pi\sigma^2} \right] dx \\
&= -h(f) - \mathbb{E}\left[-\frac{X^2}{2\sigma^2} \log e - \log \sqrt{2\pi\sigma^2} \right] \\
&= -h(f) + \frac{\mathbb{E}[X^2]}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2 \\
&\leq -h(f) + \frac{\sigma^2}{2\sigma^2} \log e + \frac{1}{2} \log 2\pi\sigma^2 \\
&= -h(f) + \frac{1}{2} \log 2\pi e\sigma^2
\end{aligned}
$$

Vector property follows similarly

General Definition of Mutual Information

Consider the mixed (neither discrete nor continuous) random variable

$$X = \begin{cases} 0 & \text{w.p. } 1/2 \\ \mathcal{N}(0,1) & \text{w.p. } 1/2 \end{cases}$$

General Definition of Mutual Information

Consider the mixed (neither discrete nor continuous) random variable

$$X = \begin{cases} 0 & \text{w.p. } 1/2 \\ \mathcal{N}(0,1) & \text{w.p. } 1/2 \end{cases}$$

X has neither a finite discrete entropy (∞) nor a differential entropy ($-\infty$)

- Let $\mathcal{X}$ be the range of random variable X (discrete, continuous, or mixed)
- A partition $\mathcal{P} = (P_1, \ldots, P_K)$ of $\mathcal{X}$ is a finite collection of disjoint sets such that $\bigcup_i P_i = \mathcal{X}$
- The quantization of X by $\mathcal{P}$ is the discrete random variable $[X]_\mathcal{P}$ where

$$\Pr\{[X]_\mathcal{P} = i\} = \Pr\{X \in P_i\}$$

- Similarly define $[Y]_\mathcal{Q}$ as a quantization of Y by partition $\mathcal{Q}$ of $\mathcal{Y}$
- The **mutual information** is given by

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

AEP for Continuous Random Variables

If $X_1, X_2, \ldots, X_n$ are i.i.d. with pdf $f(x)$, then

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

Proof:

$$-\frac{1}{n} \log f(X_1, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log f(X_i) \to \mathbb{E}[-\log f(X)] = h(X).$$

Typical Set

Given pdf $f(x)$ with support set S, typical set is given by

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

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

Properties of the Continuous Typical Set

- $\Pr\{X^n \in A_\epsilon^{(n)}\} \to 1$ as $n \to \infty$
- $\mathrm{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$
- $\mathrm{Vol}(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ for sufficiently large n

where for any set $A \subset \mathbb{R}^n$, $\mathrm{Vol}(A) = \int_A dx_1\, dx_2 \cdots dx_n$

Proofs: Property 1 follows from AEP

Property 2:

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

Property 3 follows similarly

Joint AEP for Continuous Variables

Given joint pdf $f(x, y)$, define the jointly typical set

$$A_\epsilon^{(n)} = \Big\{(x^n, y^n) : \left|-\tfrac{1}{n}\log f(x^n) - h(X)\right| \le \epsilon, \\ \left|-\tfrac{1}{n}\log f(y^n) - h(Y)\right| \le \epsilon, \\ \left|-\tfrac{1}{n}\log f(x^n, y^n) - h(X, Y)\right| \le \epsilon\Big\}$$

Properties

- If $(X^n, Y^n) \overset{\text{iid}}{\sim} f(x, y)$, then $\Pr\left\{(X^n, Y^n) \in A_\epsilon^{(n)}\right\} \to 1$ as $n \to \infty$
- $\mathrm{Vol}(A_\epsilon^{(n)}) \le 2^{n(h(X,Y)+\epsilon)}$,
  $\mathrm{Vol}(A_\epsilon^{(n)}) \ge (1-\epsilon)2^{n(h(X,Y)-\epsilon)}$ for sufficiently large n
- If $(\tilde{X}^n, \tilde{Y}^n) \overset{\text{iid}}{\sim} f(x)\, f(y)$, then

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \le 2^{-n(I(X;Y)-3\epsilon)}$$

$$\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} \ge (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \text{ for sufficiently large } n$$

Proof of Property 3

Let $(\tilde{X}^n, \tilde{Y}^n) \overset{\text{iid}}{\sim} f(x)\, f(y)$

$$\begin{aligned}
\Pr\left\{(\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right\} &= \int_{A_\epsilon^{(n)}} f(x^n)\, f(y^n)\, dx^n\, dy^n \\
&\le \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)}\, 2^{-n(h(Y)-\epsilon)}\, dx^n\, dy^n \\
&= \mathrm{Vol}(A_\epsilon^{(n)})\, 2^{-n(h(X)-\epsilon)}\, 2^{-n(h(Y)-\epsilon)} \\
&\le 2^{n(h(X,Y)+\epsilon)}\, 2^{-n(h(X)-\epsilon)}\, 2^{-n(h(Y)-\epsilon)} \\
&= 2^{-n(I(X;Y)-3\epsilon)}
\end{aligned}$$

Lower bound follows similarly