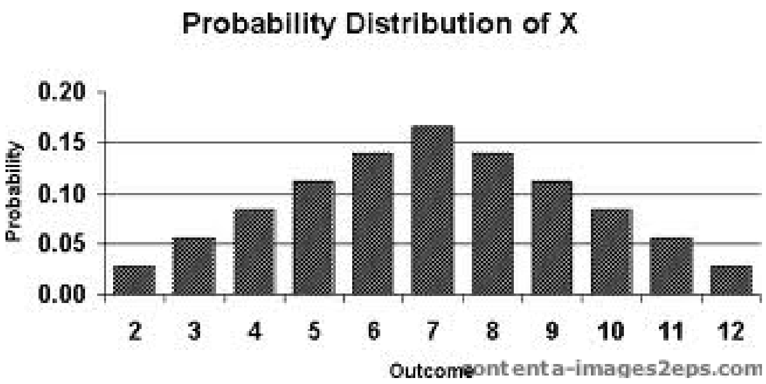
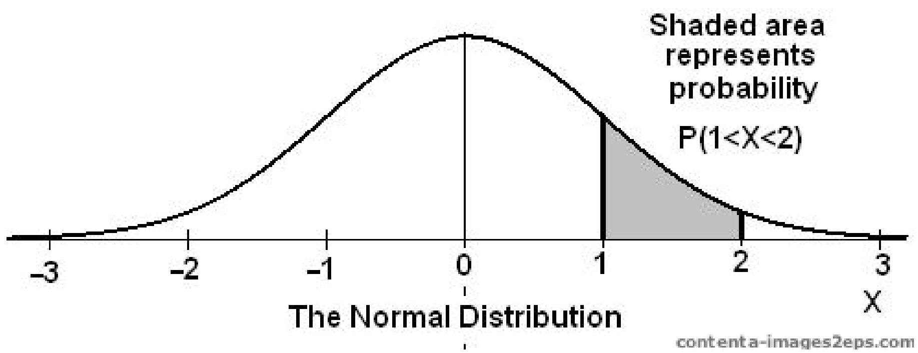
Lecture 17: Differential Entropy

- Differential entropy
- AEP for differential entropy
- Quantization
- Maximum differential entropy
- Estimation counterpart of Fano's inequality

Dr. Yao Xie, ECE587, Information Theory, Duke University

From discrete to continuous world





Differential entropy

- defined for continuous random variable

- differential entropy:

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

S is the support of probability density function (PDF)

- sometimes denote as $h(f)$

Uniform distribution

- $f(x) = 1/a$, $x \in [0, a]$
- differential entropy:

$$h(X) = \int_0^a 1/a \log(a) dx = \log a \text{ bits}$$

- for $a < 1$, $\log a < 0$, differential entropy can be negative! (unlike the discrete world)
- interpretation: volume of support set is $2^{h(X)} = 2^{\log a} = a > 0$

Normal distribution

- $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$, $x \in \mathbb{R}$

- Differential entropy:

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

Calculation:

Some properties

- $h(X + c) = h(X)$
- $h(aX) = h(X) + \log |a|$
- $h(AX) = h(X) + \log |\det(A)|$

AEP for continuous random variable

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

$$p(X_1, \ldots, X_n) \to 2^{-nH(X)}$$

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

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

- Proof from weak law of large number

Size of typical set

- Discrete world: number of typical sequences

$$|A_\epsilon^{(n)}| \approx 2^{nh(X)}$$

- Continuous world: **volume** of typical set

- Volume of set A:

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

- $p(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large

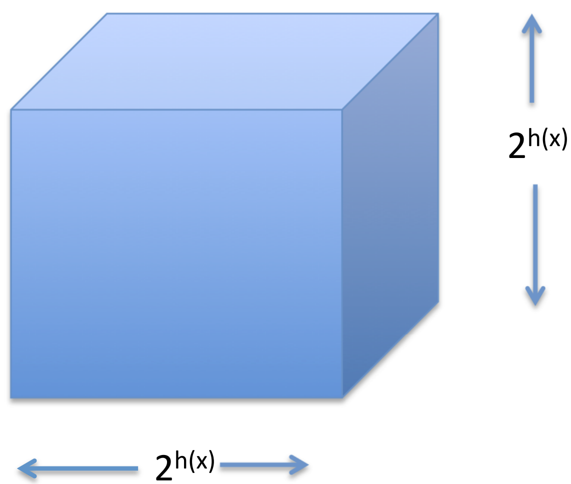
- Vol$(A) \leq 2^{n(h(X)+\epsilon)}$ for n sufficiently large

- Vol$(A) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ for n sufficiently large

- Proofs: very similar to the discrete world

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

- Volume of smallest set that contains most of the probability is $2^{nh(X)}$
- for n-dimensional space, this means that each dim has measure $2^{h(X)}$

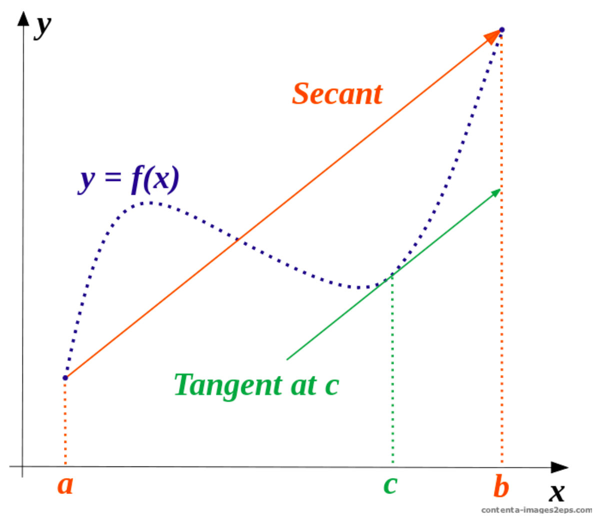


Differential entropy is a the volume of the typical set
Fisher information is a the surface area of the typical set

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

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

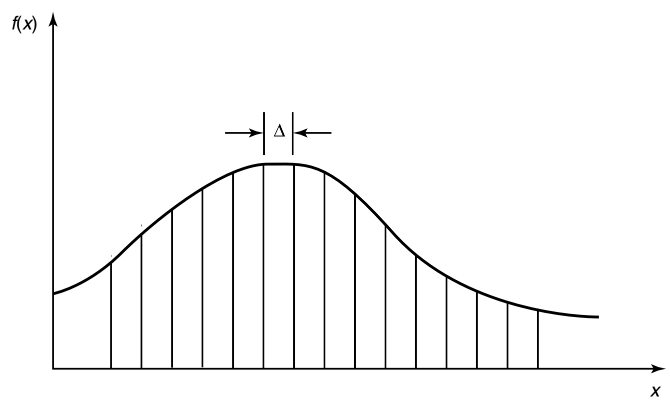


Relation of continuous entropy to discrete entropy

- Discretize a continuous pdf $f(x)$, divide the range of X into bins of length Δ

- MVT: exist a value x_i within each bin such that

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



- Define a random variable $X^\Delta \in \{x_1, \ldots\}$ with probability mass function

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

- Entropy of X^Δ

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

- If $f(x)$ is Riemann integrable, $H(X^\Delta) + \log \Delta \to h(X)$, as $\Delta \to 0$

Implication on quantization

- Let $\Delta = 2^{-n}$, $-\log \Delta = n$
- In general, $h(X) + n$ is the number of bits on average to describe a continuous variable X to n-bit accuracy
- e.g. if X is uniform on $[0, 1/8]$, the first 3 bits must be zero. Hence to describe X to n bit accuracy we need $n - 3$ bits. Agrees with $h(X) = -3$
- if $X \sim \mathcal{N}(0, 100)$, $n + h(X) = n + .5 \log(2\pi e \cdot 100) = n + 5.37$

Joint and conditional differential entropy

- Joint differential entropy

$$h(X_1, \ldots, X_n) = -\int f(x_1, \ldots, x_n) \log f(x_1, \ldots, x_n) dx_1 \cdots dx_n$$

- Conditional differential entropy

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

Relative entropy and mutual information

- Relative entropy
$$D(f||q) = \int f \log \frac{f}{g}$$
Not necessarily finite. Let $0 \log(0/0) = 0$

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

- $I(X;Y) \geq 0$, $D(f||q) \geq 0$

- Same Venn diagram relationship as in the discrete world: chain rule, conditioning reduces entropy, union bound...

Entropy of multivariate Gaussian

Theorem. *Let* $X_1, \ldots, X_n \sim \mathcal{N}(\mu, K)$, μ*: mean vector,* K*: covariance matrix, then*

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

Proof:

Mutual information of multivariate Gaussian

- $(X, Y) \sim \mathcal{N}(0, K)$
$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$$
- $h(X) = h(Y) = \frac{1}{2}\log(2\pi e\sigma^2)$
- $h(X, Y) = \frac{1}{2}\log(2\pi e)^2|K|$
- $h(X; Y) = h(X) + h(Y) - h(X; Y) = -\frac{1}{2}\log(1 - \rho^2)$
- $\rho = \pm 1$, perfectly correlated, mutual information is ∞!

Multivariate Gaussian is maximum entropy distribution

Theorem. *Let $X \in \mathbb{R}^n$ be random vector with zero mean and covariance matrix K. Then*

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

Proof:

Estimation counterpart of Fano's inequality

- Random variable X, estimator $\hat{X}$
- $E(X-\hat{X})^2 \geq \frac{1}{2\pi e}e^{2h(X)}$
- equality iff X is Gaussian and $\hat{X}$ is the mean of X
- corollary: given side information Y and estimator $\hat{X}(Y)$

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

Summary

discrete random variable $\Rightarrow$ continuous random variable
entropy $\Rightarrow$ differential entropy

- Many things similar: mutual information, AEP

- Some things are different in continuous world: $h(X)$ can be negative, maximum entropy distribution is Gaussian.