Chapter 10
Differential Entropy

Department of Information Engineering
The Chinese University of Hong Kong

Real Random Variables

- A real r.v. X with cumulative distribution function (CDF) $F_X(x) = \Pr\{X \leq x\}$ is
  - discrete if $F_X(x)$ increases only at a countable number of values of x;
  - continuous if $F_X(x)$ is continuous, or equivalently, $\Pr\{X = x\} = 0$ for every value of x;
  - mixed if it is neither discrete nor continuous.
- $\mathcal{S}_X$ is the set of all x such that $F_X(x) > F_X(x - \epsilon)$ for all $\epsilon > 0$.
- $$Eg(X) = \int_{\mathcal{S}_X} g(x) dF_X(x),$$

  where the right hand side is a *Lebesgue-Stieltjes integration* which covers all cases (i.e., discrete, continuous, and mixed) for the CDF $F_X(x)$.

Real Random Variables

- A nonnegative function $f_X(x)$ is called a probability density function (pdf) of X if
$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$
for all x.
- If X has a pdf, then X is continuous, but not vice versa.

Jointly Distributed Random Variables

- Let X and Y be two real random variables with joint CDF $F_{XY}(x,y) = \Pr\{X \leq x, Y \leq y\}$.
- Marginal CDF of X: $F_X(x) = F_{XY}(x,\infty)$
- A nonnegative function $f_{XY}(x,y)$ is called a joint pdf of X and Y if

$$F_{XY}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u,v)\, dv du$$

- Conditional pdf of Y given $\{X = x\}$:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

- Conditional CDF of Y given $\{X = x\}$:

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(v|x) dv$$

Variance and Covariance

- Variance of X:

$$\mathrm{var}X = E(X - EX)^2 = EX^2 - (EX)^2$$

- Covariance between X and Y:

$$\mathrm{cov}(X, Y) = E(X - EX)(Y - EY) = E(XY) - (EX)(EY)$$

- Remarks:

  1. $\mathrm{var}(X + Y) = \mathrm{var}X + \mathrm{var}Y + 2\mathrm{cov}(X, Y)$
  2. If $X \perp Y$, then $\mathrm{cov}(X, Y) = 0$, or X and Y are uncorrelated. However, the converse is not true.
  3. If $X_1, X_2, \cdots, X_n$ are mutually independent, then

$$\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{var}X_i$$

Random Vectors

- Let $\mathbf{X} = [X_1\, X_2\, \cdots\, X_n]^\top$.
- Covariance matrix:

$$K_{\mathbf{X}} = E(\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})^\top = [\mathrm{cov}(X_i, X_j)]$$

- Correlation matrix: $\tilde{K}_{\mathbf{X}} = E\mathbf{X}\mathbf{X}^\top = [EX_iX_j]$
- Relations between $K_{\mathbf{X}}$ and $\tilde{K}_{\mathbf{X}}$:

$$\begin{aligned} K_{\mathbf{X}} &= \tilde{K}_{\mathbf{X}} - (E\mathbf{X})(E\mathbf{X})^\top \\ K_{\mathbf{X}} &= \tilde{K}_{\mathbf{X}-E\mathbf{X}} \end{aligned}$$

- These are vector generalizations of

$$\begin{aligned} \mathrm{var} X &= EX^2 - (EX)^2 \\ \mathrm{var} X &= E(X - EX)^2 \end{aligned}$$

Gaussian Distribution

- $\mathcal{N}(\mu, \sigma^2)$ – Gaussian distribution with mean μ and variance σ^2:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

- $\mathcal{N}(\boldsymbol{\mu}, K)$ – multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix K, i.e., the joint pdf of the distribution is given by

$$f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |K|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top K^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad \mathbf{x} \in \Re^n$$

where K is a symmetric positive definite matrix.

10.1 Preliminaries

Definition 10.1 A square matrix K is symmetric if $K^\top = K$.

Definition 10.2 An $n \times n$ matrix K is positive definite if

$$\mathbf{x}^\top K \mathbf{x} > 0$$

for all nonzero column n-vector $\mathbf{x}$, and is positive semidefinite if

$$\mathbf{x}^\top K \mathbf{x} \geq 0$$

for all column n-vector $\mathbf{x}$.

Proposition 10.3 A covariance matrix is both symmetric and positive semidefinite.

Diagonalization

- A symmetric matrix K can be diagonalized as

$$K = Q\Lambda Q^\top$$

  where Λ is a diagonal matrix and Q (also $Q^\top$) is an orthogonal matrix, i.e.,

$$Q^{-1} = Q^\top$$

- $|Q| = |Q^\top| = 1$.
- Let $\lambda_i = i$th diagonal element of Λ and $\mathbf{q}_i = i$th column of Q
- $KQ = (Q\Lambda Q^\top)Q = Q\Lambda(Q^\top Q) = Q\Lambda$, or

$$K\mathbf{q}_i = \lambda_i \mathbf{q}_i$$

- That is, $\mathbf{q}_i$ is an eigenvector of K with eigenvalue λ_i.

Proposition 10.4 The eigenvalues of a positive semidefinite matrix are nonnegative.

Proof

1. Consider eigenvector $\mathbf{q} \neq 0$ and corresponding eigenvalue λ of K, i.e.,

$$K\mathbf{q} = \lambda\mathbf{q}$$

2. Since K is positive semidefinite,

$$0 \leq \mathbf{q}^\top K\mathbf{q} = \mathbf{q}^\top(\lambda\mathbf{q}) = \lambda(\mathbf{q}^\top\mathbf{q})$$

3. $\lambda \geq 0$ because $\mathbf{q}^\top\mathbf{q} = \|\mathbf{q}\|^2 > 0$.

Remark Since a covariance matrix is both symmetric and positive semidefinite, it is diagonalizable and its eigenvalues are nonnegative.

Proposition 10.5 Let $\mathbf{Y} = A\mathbf{X}$. Then

$$K_{\mathbf{Y}} = AK_{\mathbf{X}}A^{\top}$$

and

$$\tilde{K}_{\mathbf{Y}} = A\tilde{K}_{\mathbf{X}}A^{\top}.$$

Proposition 10.6 (Decorrelation) Let $\mathbf{Y} = Q^{\top}\mathbf{X}$, where $K_{\mathbf{X}} = Q^{\top}\Lambda Q$. Then $K_{\mathbf{Y}} = \Lambda$, i.e.,

1. the random variables in $\mathbf{Y}$ are uncorrelated

2. $\operatorname{var} Y_i = \lambda_i$ for all i

Corollary 10.7 Any random vector $\mathbf{X}$ can be written as a linear transformation of an uncorrelated vector. Specifically, $\mathbf{X} = Q\mathbf{Y}$, where $K_{\mathbf{X}} = Q^{\top}\Lambda Q$.

Proposition 10.8 Let $\mathbf{X}$ and $\mathbf{Z}$ be independent and $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$. Then

$$K_{\mathbf{Y}} = K_{\mathbf{X}} + K_{\mathbf{Z}}$$

Proposition 10.9 (Preservation of Energy) Let $\mathbf{Y} = Q\mathbf{X}$, where Q is an orthogonal matrix. Then

$$E\sum_{i=1}^{n} Y_i^2 = E\sum_{i=1}^{n} X_i^2$$

10.2 Definition

Definition 10.10 The differential entropy $h(X)$ of a continuous random variable X with pdf $f(x)$ is defined as

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

Remarks

1. Differential entropy is not a measure of the average amount of information contained in a continuous r.v.

2. A continuous random variable generally contains an infinite amount of information.

Example 10.11 Let X be uniformly distributed on $[0, 1)$. Then we can write

$$X = .X_1 X_2 X_3 \cdots ,$$

the dyadic expansion of X, where $X_1, X_2, X_3, \cdots$ is a sequence of fair bits. Then

$$\begin{aligned} H(X) &= H(X_1, X_2, X_3, \cdots) \\ &= \sum_{i=1}^{\infty} H(X_i) \\ &= \sum_{i=1}^{\infty} 1 \\ &= \infty \end{aligned}$$

Relation with Discrete Entropy

- Consider a continuous r.v. X with a continuous pdf $f(x)$.
- Define a discrete r.v. $\hat{X}_\Delta$ by

$$\hat{X}_\Delta = i \quad \text{if} \quad X \in [i\Delta, (i+1)\Delta)$$

- Since $f(x)$ is continuous,

$$p_i = \Pr\{\hat{X}_\Delta = i\} \approx f(x_i)\Delta$$

where $x_i \in [i\Delta, (i+1)\Delta)$.

- Then

$$
\begin{aligned}
H(\hat{X}_\Delta) &= -\sum_i p_i \log p_i \\
&\approx -\sum_i f(x_i)\Delta \log(f(x_i)\Delta) \\
&= -\sum_i \Delta f(x_i) \log f(x_i) - \sum_i f(x_i)\Delta \log \Delta \\
&= -\sum_i \Delta f(x_i) \log f(x_i) - \log \Delta \\
&\approx h(X) - \log \Delta
\end{aligned}
$$

when Δ is small.

Example 10.12 Let X be uniformly distributed on $[0, a)$. Then

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

Remark $h(X) < 0$ if $a < 1$, so $h(\cdot)$ cannot be a measure of information.

Example 10.13 (Gaussian Distribution) Let $X \sim \mathcal{N}(0, \sigma^2)$. Then

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

Properties of Differential Entropy

Theorem 10.14 (Translation)

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

Theorem 10.15 (Scaling) For $a \neq 0$,

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

Remark on Scaling The differential entropy is

- increased by $\log |a|$ if $|a| > 1$
- decreased by $-\log |a|$ if $|a| < 1$
- unchanged if $a = -1$
- related to the "spread" of the pdf

10.3 Joint Differential Entropy, Conditional (Differential) Entropy, and Mutual Information

Definition 10.17 The joint differential entropy $h(\mathbf{X})$ of a random vector $\mathbf{X}$ with joint pdf $f(\mathbf{x})$ is defined as

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

Corollary If $X_1, X_2, \cdots, X_n$ are mutually independent, then

$$h(\mathbf{X}) = \sum_{i=1}^{n} h(X_i)$$

Theorem 10.18 (Translation) $h(\mathbf{X} + \mathbf{c}) = h(\mathbf{X})$.

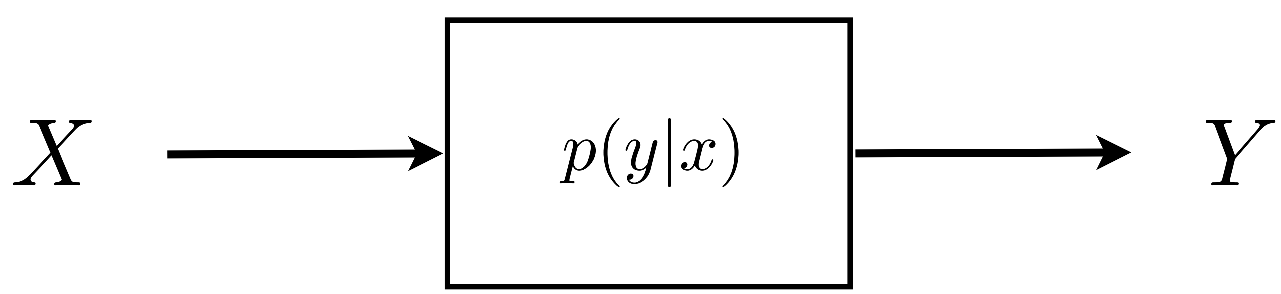
Theorem 10.19 (Scaling) $h(A\mathbf{X}) = h(\mathbf{X}) + \log |\det(A)|$.

Theorem 10.20 (Multivariate Gaussian Distribution) Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, K)$. Then

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

The Model of a "Channel" with Discrete Output

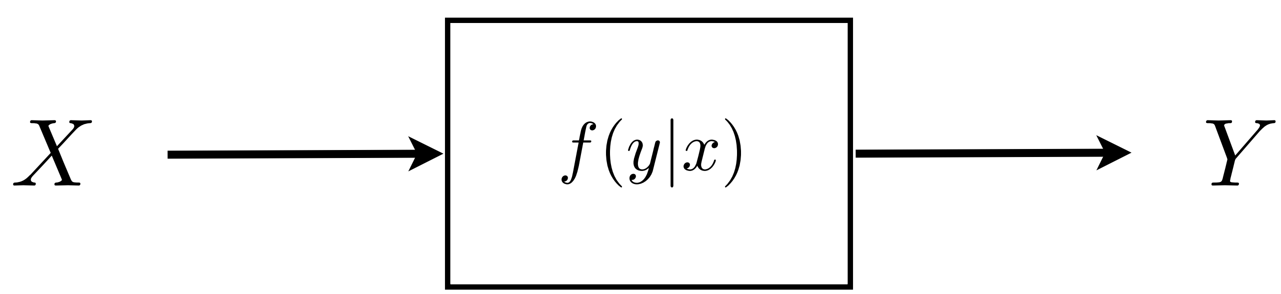
Definition 10.21 The random variable Y is related to the random variable X through a conditional distribution $p(y|x)$ defined for all x means



- X: general
- Y: discrete

The Model of a "Channel" with Continuous Output

Definition 10.22 The random variable Y is related to the random variable X through a conditional pdf $f(y|x)$ defined for all x means



- X: general
- Y: continuous

Conditional Differential Entropy

Definition 10.23 Let X and Y be jointly distributed random variables where Y is continuous and is related to X through a conditional pdf $f(y|x)$ defined for all x. The conditional differential entropy of Y given $\{X = x\}$ is defined as

$$h(Y|X = x) = -\int_{\mathcal{S}_Y(x)} f(y|x) \log f(y|x) dy$$

and the conditional differential entropy of Y given X is defined as

$$h(Y|X) = -\int_{\mathcal{S}_X} h(Y|X = x) dF(x) = -E \log f(Y|X)$$

Proposition 10.24 Let X and Y be jointly distributed random variables where Y is continuous and is related to X through a conditional pdf $f(y|x)$ defined for all x. Then $f(y)$ exists and is given by

$$f(y) = \int f(y|x)dF(x)$$

Proof

1.
$$F_Y(y) = F_{XY}(\infty, y) = \int \int_{-\infty}^{y} f_{Y|X}(v|x)\, dv\, dF(x)$$

2. Since
$$\int \int_{-\infty}^{y} f_{Y|X}(v|x)\, dv\, dF(x) = F_Y(y) \leq 1$$
$f_{Y|X}(v|x)$ is absolutely integrable.

3. By Fubini's theorem, the order of integration in $F_Y(y)$ can be exchanged, and so
$$F_Y(y) = \int_{-\infty}^{y} \left[\int f_{Y|X}(v|x)dF(x) \right] dv$$
proving the proposition.

Proposition 10.24 says that if Y is related to X through a conditional pdf $f(y|x)$, then the pdf of Y exists regardless of the distribution of X. The next proposition is its vector generalization.

Proposition 10.25 Let $\mathbf{X}$ and $\mathbf{Y}$ be jointly distributed random vectors where $\mathbf{Y}$ is continuous and is related to $\mathbf{X}$ through a conditional pdf $f(\mathbf{y}|\mathbf{x})$ defined for all $\mathbf{x}$. Then $f(\mathbf{y})$ exists and is given by

$$f(\mathbf{y}) = \int f(\mathbf{y}|\mathbf{x})dF(\mathbf{x})$$

Mutual Information

Definition 10.26 Let X and Y be jointly distributed random variables where Y is continuous and is related to X through a conditional pdf $f(y|x)$ defined for all x.

1. The mutual information between X and Y is defined as

$$\begin{aligned} I(X;Y) &= \int_{\mathcal{S}_X} \int_{\mathcal{S}_Y(x)} f(y|x) \log \frac{f(y|x)}{f(y)} dy\, dF(x) \\ &= E \log \frac{f(Y|X)}{f(Y)} \end{aligned}$$

2. When both X and Y are continuous and $f(x,y)$ exists,

$$I(X;Y) = E \log \frac{f(Y|X)}{f(Y)} = E \log \frac{f(X,Y)}{f(X)f(Y)}$$

Remarks

- With Proposition 10.26, the mutual information is defined when one r.v. is general and the other is continuous.
- In Ch. 2, the mutual information is defined when both r.v.'s are discrete.
- Thus the mutual information is defined when each of the r.v.'s can be either discrete or continuous.

Conditional Mutual Information

Proposition 10.27 Let X, Y, and T be jointly distributed random variables where Y is continuous and is related to (X, T) through a conditional pdf $f(y|x,t)$ defined for all x and t. The mutual information between X and Y given T is defined as

$$I(X;Y|T) = \int_{\mathcal{S}_T} I(X;Y|T=t)dF(t) = E\log\frac{f(Y|X,T)}{f(Y|T)}$$

where

$$I(X;Y|T=t) = \int_{\mathcal{S}_X(t)}\int_{\mathcal{S}_Y(x,t)} f(y|x,t)\log\frac{f(y|x,t)}{f(y|t)}dy\,dF(x|t)$$

Interpretation of I(X;Y)

- Assume $f(x,y)$ exists and is continuous.
- For all integer i and j, define the intervals

$$A_x^i = [\, i\Delta, (i+1)\Delta) \quad \text{and} \quad A_y^j = [\, j\Delta, (j+1)\Delta)$$

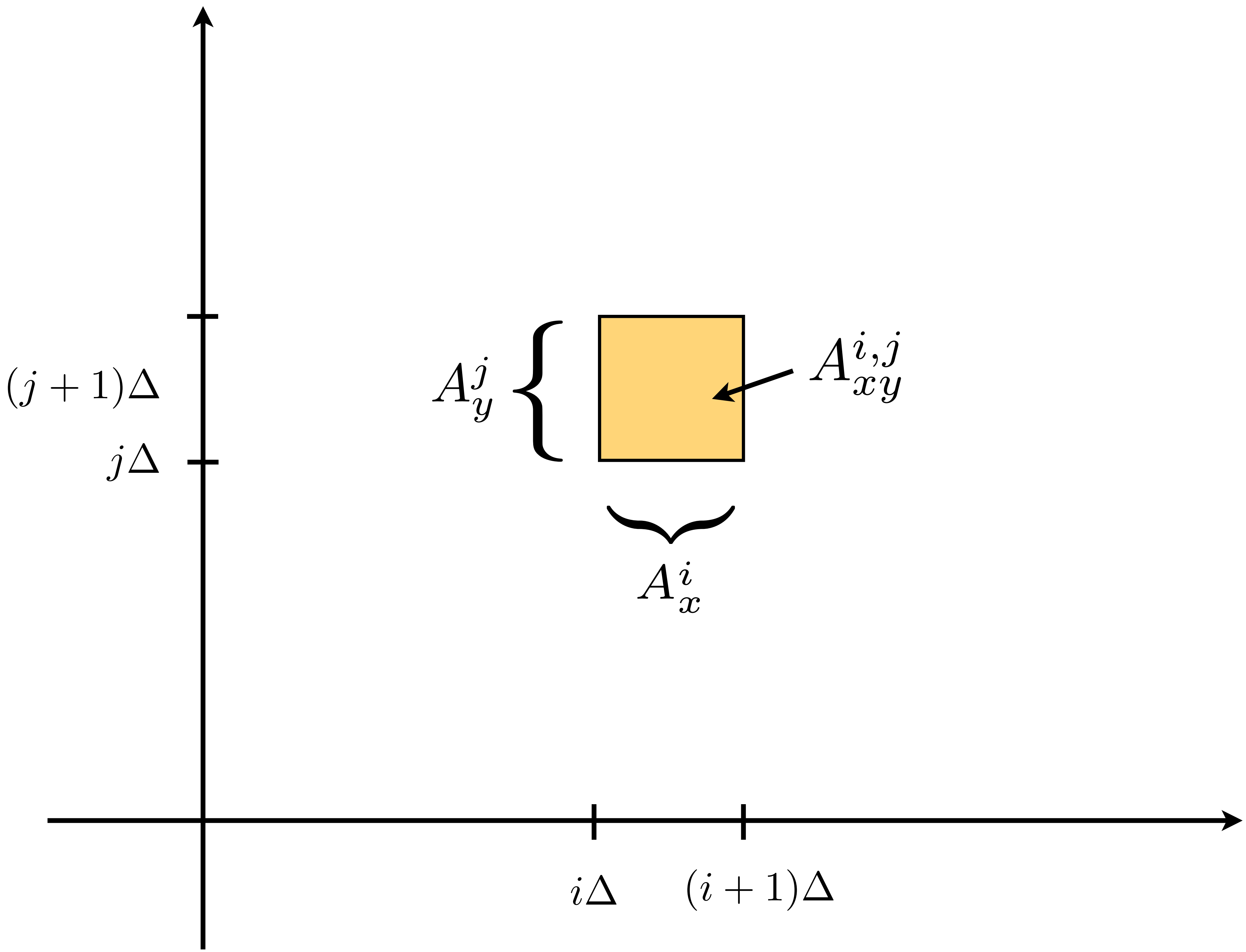
  and the rectangle

$$A_{xy}^{i,j} = A_x^i \times A_y^j$$

- Define discrete r.v.'s

$$\begin{cases} \hat{X}_\Delta = i & \text{if } X \in A_x^i \\ \hat{Y}_\Delta = j & \text{if } Y \in A_y^j \end{cases}$$

- $\hat{X}_\Delta$ and $\hat{Y}_\Delta$ are quantizations of X and Y, resp.
- For all i and j, let $(x_i, y_j) \in A_x^i \times A_y^j$.



$(j+1)\Delta$
$j\Delta$
A_y^j
$A_{xy}^{i,j}$
A_x^i
$i\Delta$
$(i+1)\Delta$

- Then

$$
\begin{aligned}
& I(\hat{X}_\Delta;\hat{Y}_\Delta) \\
= \quad & \sum_i \sum_j \Pr\{(\hat{X}_\Delta,\hat{Y}_\Delta)=(i,j)\} \log \frac{\Pr\{(\hat{X}_\Delta,\hat{Y}_\Delta)=(i,j)\}}{\Pr\{\hat{X}_\Delta=i\}\Pr\{\hat{Y}_\Delta=j\}} \\
\approx \quad & \sum_i \sum_j f(x_i,y_j)\Delta^2 \log \frac{f(x_i,y_j)\Delta^2}{(f(x_i)\Delta)(f(y_j)\Delta)} \\
= \quad & \sum_i \sum_j f(x_i,y_j)\Delta^2 \log \frac{f(x_i,y_j)}{f(x_i)f(y_j)} \\
\approx \quad & \int\int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \\
= \quad & I(X;Y)
\end{aligned}
$$

- Therefore, $I(X;Y)$ can be interpreted as the limit of $I(\hat{X}_\Delta;\hat{Y}_\Delta)$ as $\Delta \to 0$.
- This interpretation continues to be valid for general distribution for X and Y.

Definition 10.28 Let Y be a continuous random variable and X be a discrete random variable, where Y is related to X through a conditional pdf $f(y|x)$. The conditional entropy of X given Y is defined as

$$H(X|Y) = H(X) - I(X;Y)$$

Proposition 10.29 For two random variables X and Y,

1. $h(Y) = h(Y|X) + I(X;Y)$ if Y is continuous
2. $H(Y) = H(Y|X) + I(X;Y)$ if Y is discrete.

Proposition 10.30 (Chain Rule)

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

Theorem 10.31

$$I(X;Y) \geq 0,$$

with equality if and only if X is independent of Y.

Corollary 10.32

$$I(X;Y|T) \geq 0,$$

with equality if and only if X is independent of Y conditioning on T.

Corollary 10.33 (Conditioning Does Not Increase Differential Entropy)

$$h(X|Y) \leq h(X)$$

with equality if and only if X and Y are independent.

Remarks For continuous r.v.'s,

1. $h(X), h(X|Y) \geq 0$ DO NOT generally hold;
2. $I(X;Y), I(X;Y|Z) \geq 0$ always hold.

10.4 AEP for Continuous Random Variables

Theorem 10.35 (AEP I for Continuous Random Variables)

$$-\frac{1}{n}\log f(\mathbf{X}) \to h(X)$$

in probability as $n \to \infty$, i.e., for any $\epsilon > 0$, for n sufficiently large,

$$\Pr\left\{\left|-\frac{1}{n}\log f(\mathbf{X}) - h(X)\right| < \epsilon\right\} > 1 - \epsilon.$$

Proof WWLN.

Definition 10.36 The typical set $W^n_{[X]\epsilon}$ with respect to $f(x)$ is the set of sequences $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ such that

$$\left| -\frac{1}{n} \log f(\mathbf{x}) - h(X) \right| < \epsilon$$

or equivalently,

$$h(X) - \epsilon < -\frac{1}{n} \log f(\mathbf{x}) < h(X) + \epsilon$$

where ϵ is an arbitrarily small positive real number. The sequences in $W^n_{[X]\epsilon}$ are called ϵ-typical sequences.

Empirical Differential Entropy:

$$-\frac{1}{n} \log f(\mathbf{x}) = -\frac{1}{n} \sum_{k=1}^{n} \log f(x_k)$$

The empirical differential entropy of a typical sequence is close to the true differential entropy $h(X)$.

Definition 10.37 The volume of a set A in $\Re^n$ is defined as

$$\mathrm{Vol}(A) = \int_A d\mathbf{x}$$

Theorem 10.38 The following hold for any $\epsilon > 0$:

1) If $\mathbf{x} \in W^n_{[X]\epsilon}$, then

$$2^{-n(h(X)+\epsilon)} < f(\mathbf{x}) < 2^{-n(h(X)-\epsilon)}$$

2) For n sufficiently large,

$$\Pr\{\mathbf{X} \in W^n_{[X]\epsilon}\} > 1 - \epsilon$$

3) For n sufficiently large,

$$(1-\epsilon)2^{n(h(X)-\epsilon)} < \mathrm{Vol}(W^n_{[X]\epsilon}) < 2^{n(h(X)+\epsilon)}$$

Remarks

1. The volume of the typical set is approximately equal to $2^{nh(X)}$ when n is large.

2. The fact that $h(X)$ can be negative does not incur any difficulty because $2^{nh(X)}$ is always positive.

3. If the differential entropy is large, then the volume of the typical set is large.

10.5 Informational Divergence

Definition 10.39 Let f and g be two pdf's defined on $\Re^n$ with supports $\mathcal{S}_f$ and $\mathcal{S}_g$, respectively. The informational divergence between f and g is defined as

$$D(f\|g) = \int_{\mathcal{S}_f} f(x) \log \frac{f(x)}{g(x)} dx = E_f \log \frac{f(X)}{g(X)},$$

where E_f denotes expectation with respect to f.

Remark If $D(f\|g) < \infty$, then

$$\mathcal{S}_f \setminus \mathcal{S}_g = \{x : f(x) > 0 \text{ and } g(x) = 0\}$$

has zero Lebesgue measure, i.e., $\mathcal{S}_f$ is essentially a subset of $\mathcal{S}_g$.

Theorem 10.40 (Divergence Inequality) Let f and g be two pdf's defined on $\Re^n$. Then

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

with equality if and only if $f = g$ a.e.

10.6 Maximum Differential Entropy Distributions

The maximization problem:

Maximize $h(f)$ over all pdf f defined on a subset $\mathcal{S}$ of $\Re^n$, subject to

$$\int_{\mathcal{S}_f} r_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = a_i \quad \text{for } 1 \le i \le m \tag{1}$$

where $\mathcal{S}_f \subset \mathcal{S}$ and $r_i(\mathbf{x})$ is defined for all $\mathbf{x} \in \mathcal{S}$.

Theorem 10.41 Let

$$f^*(\mathbf{x}) = e^{-\lambda_0 - \sum_{i=1}^m \lambda_i r_i(\mathbf{x})}$$

for all $\mathbf{x} \in \mathcal{S}$, where $\lambda_0, \lambda_1, \cdots, \lambda_m$ are chosen such that the constraints in (1) are satisfied. Then f^* maximizes $h(f)$ over all pdf f defined on $\mathcal{S}$, subject to the constraints in (1).

Corollary 10.42 Let f^* be a pdf defined on $\mathcal{S}$ with

$$f^*(\mathbf{x}) = e^{-\lambda_0 - \sum_{i=1}^m \lambda_i r_i(\mathbf{x})}$$

for all $\mathbf{x} \in \mathcal{S}$. Then f^* maximizes $h(f)$ over all pdf f defined on $\mathcal{S}$, subject to the constraints

$$\int_{\mathcal{S}_f} r_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \int_{\mathcal{S}} r_i(\mathbf{x}) f^*(\mathbf{x}) d\mathbf{x} \quad \text{for } 1 \le i \le m$$

Theorem 10.43 Let X be a continuous random variable with $EX^2 = \kappa$. Then

$$h(X) \le \frac{1}{2}\log(2\pi e\kappa),$$

with equality if and only if $X \sim \mathcal{N}(0, \kappa)$.

Proof

1. Maximize $h(f)$ subject to the constraint

$$\int x^2 f(x)dx = EX^2 = \kappa.$$

2. Then by Theorem 10.41, $f^*(x) = ae^{-bx^2}$, which is the Gaussian distribution with zero mean.

3. In order to satisfy the second moment constraint, the only choices are

$$a = \frac{1}{\sqrt{2\pi\kappa}} \quad \text{and} \quad b = \frac{1}{2\kappa}$$

An Application of Corollary 10.42

Consider the pdf of $\mathcal{N}(0, \sigma^2)$:

$$f^*(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

1. Write

$$f(x) = e^{-\lambda_0} e^{-\lambda_1 x^2}$$

2. Then f^* maximizes $h(f)$ over all f subject to

$$\int x^2 f(x) dx = \int x^2 f^*(x) dx = EX^2 = \sigma^2$$

Theorem 10.44 Let X be a continuous random variable with mean μ and variance σ^2. Then

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

with equality if and only if $X \sim \mathcal{N}(\mu, \sigma^2)$.

Proof

1. Let $X' = X - \mu$.
2. Then $EX' = 0$ and $E(X')^2 = E(X-\mu)^2 = \text{var}X = \sigma^2$.
3. By Theorems 10.14 and 10.43,

$$h(X) = h(X') \leq \frac{1}{2} \log(2\pi e \sigma^2)$$

4. Equality holds if and only if $X' \sim \mathcal{N}(0, \sigma^2)$, or $X \sim \mathcal{N}(\mu, \sigma^2)$.

Remark Theorem 10.43 says that with the constraint $EX^2 = \kappa$, the differential entropy is maximized by the distribution $\mathcal{N}(0, \kappa)$. If we impose the additional constraint that $EX = 0$, then $\mathrm{var}X = EX^2 = \kappa$. By Theorem 10.44, the differential entropy is still maximized by $\mathcal{N}(0, \kappa)$.

Differential Entropy and Spread

1. From Theorem 10.44, we have

$$h(X) \leq \frac{1}{2}\log(2\pi e\sigma^2) = \log\sigma + \frac{1}{2}\log(2\pi e)$$

2. $h(X)$ is at most equal to the logarithm of the standard deviation plus a constant.

3. $h(X) \to \infty$ as $\sigma \to 0$.

Theorem 10.45 Let $\mathbf{X}$ be a vector of n continuous random variables with correlation matrix $\tilde{K}$. Then

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

with equality if and only if $\mathbf{X} \sim \mathcal{N}(0, \tilde{K})$.

Theorem 10.46 Let $\mathbf{X}$ be a vector of n continuous random variables with mean $\boldsymbol{\mu}$ and covariance matrix K. Then

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

with equality if and only if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, K)$.