Ch. 9 [Cover]

Differential Entropy

Def

$$h(X) = -\int_S f(x)\log_2 f(x)\,dx$$

where S is the support set of the random variable.

Example uniform distribution

$$h(x) = -\int_0^a \frac{1}{a}\log_2\frac{1}{a}\,dx = \log_2 a$$

Note that for $a<1$, $\log_2 a < 0$

hence, differential entropy can be negative.

Example 9.1.2 [Normal Distribution]

Let $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$

$$h(\phi) = -\int \phi \ln \phi$$

$$= -\int \phi(x)\left[-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right]$$

$$= \frac{1}{2\sigma^2}E[x^2] + \ln\sqrt{2\pi\sigma^2}\int \phi(x)\,dx$$

$$= \frac{\sigma^2}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2$$

$$= \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma^2$$

$$= \frac{1}{2}\ln 2\pi e\sigma^2 \text{ nats}$$

$$\cancel{= \frac{1}{2}\ln 2\pi e\sigma}$$

changing the base of the logarith

$$h(\phi) = \frac{1}{2}\log_2 2\pi e\sigma^2 \text{ bits}$$

Def: Joint Differential Entropy

$$h(X_1, X_2, \ldots, X_n)$$

$$= -\int f(x_1, x_2, \ldots, x_n)\log_2 f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2 \ldots dx_n$$

Def: Conditional differential Entropy

$$h(X|Y) = -\int f(x,y)\log_2 f(x|y)\,dx\,dy$$

Since $f(x|y) = \frac{f(x,y)}{f(y)}$

$$\Rightarrow h(X|Y) = h(X,Y) - h(Y)$$

Def: Relative Entropy

The relative Entropy between two densities f and g is:

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

Def: Mutual information

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

$$I(X;Y) = h(X) - h(X/Y)$$
$$= h(Y) - h(Y/X)$$

Also

$$I(X;Y) = D\left[f(x,y) \| f(x)f(y)\right]$$

Properties

1. $D(f\|g) \geq 0$
   with equality iff $f = g$
2. $I(X;Y) \geq 0$
   with equality iff X and Y are independent
3. $h(X|Y) \leq h(X)$ with equality if X and Y are independent.
4. $h(X_1, X_2, \dots, X_n) \leq \sum h(X_i)$
   with equality iff $X_1, X_2, \dots, X_n$ are independent.
5. $h(X+c) = h(X)$
   "Translation does not change" the differential Entropy
6. $h(aX) = h(X) + \log_2 |a|$
7. Let A be a Matrix, then
   $h(AX) = h(X) + \log_2 |A|$

~~Example~~ [Entropy of a multivariate normal distribution]

Let $X_1, X_2, \dots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix K. [$\mathcal{N}_n(\mu, K)$]

Then,

$$h(X_1, X_2, \dots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2}\log_2 (2\pi e)^n |K| \text{ bits}$$

Proof:

$$h(f) = -\int f(\vec{x}) \ln f(\vec{x})\, d\vec{x}$$

where $f(\vec{x}) = \dfrac{1}{(\sqrt{2\pi})^n |K|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\mu)^T K^{-1}(\vec{x}-\mu)}$

$\therefore$

$$h(f) = -\int f(\vec{x}) \left[-\frac{1}{2}(\vec{x}-\mu)^T K^{-1}(\vec{x}-\mu) - \ln (\sqrt{2\pi})^n |K|^{1/2}\right] d\vec{x}$$

$$= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(K^{-1})_{ij}(x_j - \mu_j)\right] + \frac{1}{2}\ln (2\pi)^n |K|$$

$$= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(x_j - \mu_j)(K^{-1})_{ij}\right] + \frac{1}{2}\ln (2\pi)^n |K|$$

$$= \frac{1}{2}\sum_{i,j} E[(x_j - \mu_j)(x_i - \mu_i)](K^{-1})_{ij} + \frac{1}{2}\ln (2\pi)^n |K|$$

$$= \frac{1}{2}\sum_j \sum_i K_{ji}(K^{-1})_{ij} + \frac{1}{2}\ln (2\pi)^n |K|$$

$$= \frac{1}{2}\sum_j I_{jj} + \frac{1}{2}\ln (2\pi)^n |K|$$

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

$$= \frac{1}{2}\ln (2\pi e)^n |K| \text{ nats}$$

$$= \frac{1}{2}\log_2 (2\pi e)^n |K| \text{ bits}.$$

Theorem 9.6.5:

Let $X \in \mathbb{R}^n$ be a R.V. with zero mean and covariance $K = E[XX^T]$,

then $h(X) \leq \frac{1}{2}\log_2 (2\pi e)^n |K|$

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

$\Rightarrow$ multivariate normal distribution maximizes the entropy over all distributions with the same covariance.

Proof: Let $g(x)$ be any density satisfying

$$\int g(x)\, x_i x_j \, dx = K_{ij} \rightarrow \text{Covariance}$$

Let ϕ_K be the density of a $\mathcal{N}(0, K)$.

Note that $\log \phi_K(X)$ is a quadratic form

and $\int x_i x_j \phi_K(X) dx = K_{ij}$

then

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

$$\int g \log\left(\frac{g}{\phi_K}\right) \geq 0$$

$$-h(g) - \int g \log \phi_K \geq 0$$

Since $\log \phi_K(X)$ is a quadratic form and

$$\int g(x)\, x_i x_j \, dx = \int \phi_K(x) x_i x_j \, dx = K_{ij}$$

$\Rightarrow$

$$-h(g) - \int g \log \phi_K$$

$$= -h(g) - \int \phi_K \log(\phi_K)$$

$$= -h(g) + h(\phi_K)$$

$$\therefore\ h(g) \leq h(\phi_K)$$

The Gaussian Channel

$$Y_i = X_i + Z_i \qquad Z_i \sim \mathcal{N}(0, N)$$

The noise Z_i is an i.i.d R.V. with Gaussian distribution with variance N.

Special Cases:

① If the noise variance is zero

$\Rightarrow$ Infinite capacity

$\Rightarrow$ No transmission errors

② No constraint on the input.

$\Rightarrow$ choose infinite subset of inputs arbitrarily apart

$\Rightarrow$ Infinite capacity

However, there are always constraints on the inputs in terms of power or energy in addition to limited alphabet size. For any codeword $(x_1, x_2, \cdots, x_n)$ transmitted over the channel, we require $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$

Gaussian Channel Capacity

Def: The information capacity of the Gaussian channel with power constraint P is

~~$C = \max_{p(x)} I(X;Y)$~~

$$\boxed{C = \max_{p(x):\, E[X^2] \le P} I(X;Y)}$$

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y/X) \\ &= h(Y) - h(X+Z|X) \\ &= h(Y) - h(Z|X) \\ &= h(Y) - h(Z) \end{aligned}$$

Since Z is independent of X.

Also, $h(Z) = \frac{1}{2}\log 2\pi e N$

and $E[Y^2] = E[X+Z]^2 = E[X^2] + 2E[X]E[Z] + E[Z^2] = P+N$

$\therefore$ The entropy of Y is bounded by $\frac{1}{2}\log 2\pi e (P+N)$

Applying this to $I(X;Y)$

Ch. 10 {Cover}

$$I(X;Y) = h(Y) - h(Z)$$
$$\leq \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi e N$$
$$= \frac{1}{2}\log\left(1+\frac{P}{N}\right)$$

$\therefore$

$$C = \max_{E(X^2)=P} I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$$

and the maximum is attained when $X \sim \mathcal{N}(0,P)$

Theorem:

The capacity of a Gaussian channel with power constraint P and noise variance N is

$$C = \frac{1}{2}\log\left(1+\frac{P}{N}\right) \text{ bits per transmission}$$

Def: A (M,n) code for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1,2,\ldots,M\}$
2. An encoding function $x:\{1,2,\ldots,M\} \to \mathcal{X}^n$ yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$, satisfying the power constraint P

$$\sum_{i=1}^{n} x_i^2(w) \leq nP$$

where $w = 1,2,\ldots,M$

3. A decoding function $g: \mathcal{Y}^n \to \{1,2,\ldots,M\}$

Definition:

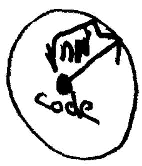
A rate R is said to be achievable for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal prob. of error $\lambda^{(n)}$ tends to be zero.

The capacity of the channel is the supremum of the achievable rates.

Spher Packing Argument

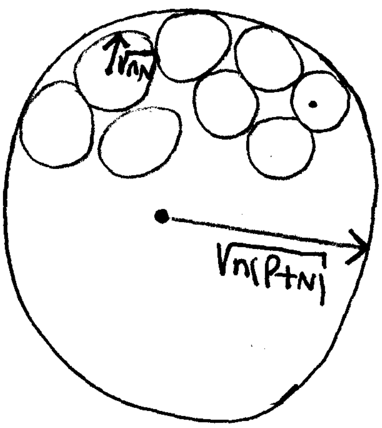
Why we may construct $(2^{nC}, n)$ codes with low Prob. of Error?

- Consider any codeword of length n. the received vector is normally distributed with mean equal to the true codeword and variance equal to the noise variance.
- For a codeword of length n, the n-dimensional noise variance is nN
- Thus, with high Prob., the received vector is contained in a sphere of radius $\sqrt{nN}$ around the true codeword.



- If we receive any vector inside the sphere, we assign it to the given codeword.
- An error will happen if the received vector falls outside the sphere, which has low Prob. of error.

- Similarly, other codewords will be assigned spheres.
- The Total Energy of the Received vectors is $n(P+N)$. So, they lie in a sphere of radius $\sqrt{n(P+N)}$
- The question is: How many non-overlapping spheres you can fit inside the big sphere?



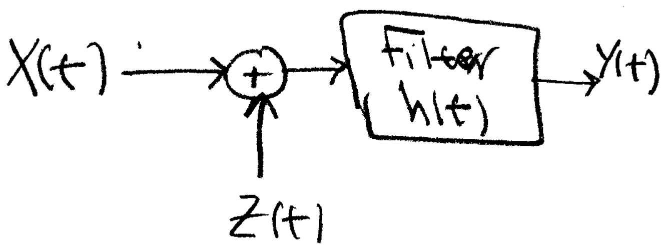
- The Volume of an n-dimensional sphere is $A_n r^n$

$\therefore$ The maximum number of non-intersecting spheres is

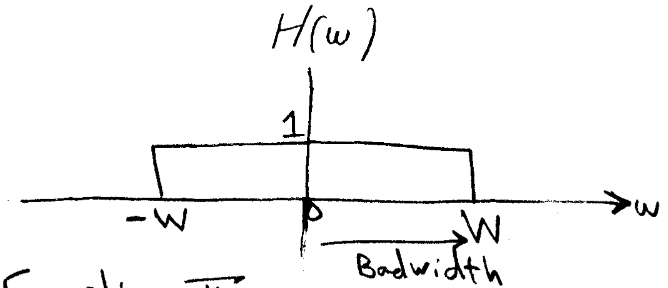
$$\frac{A_n[n(P+N)]^{\frac{n}{2}}}{A_n(nN)^{\frac{n}{2}}} = 2^{\frac{n}{2}\log\left(1+\frac{P}{N}\right)}$$

compare to $2^{nC} \Rightarrow C = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$

10.3 Band-Limited Channels



$$Y(t) = (X(t) + Z(t)) * h(t)$$



Sampling Theorem

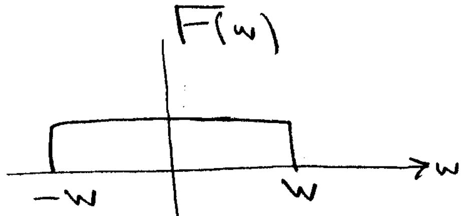
Supposed a function f(t) in band-limited to W, then the function is completely determined by samples of the function spaced $\frac{1}{2W}$ seconds apart.

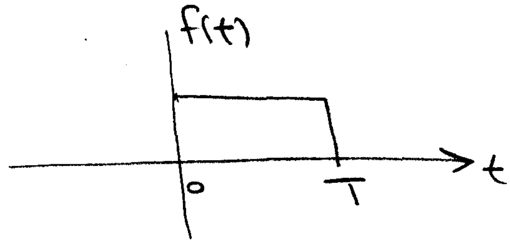
$\Rightarrow$ sampling frequency $f_s = 2W$ Hz.

$$T_s = \frac{1}{2W} \text{ Second}$$

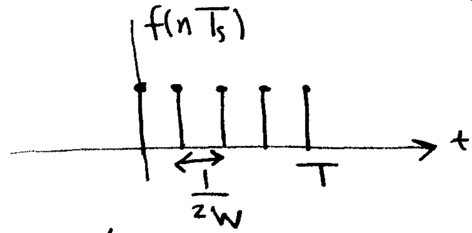
Almost time-limited Almost band-limited functions

— Most of the energy in bandwidth W and most of the energy in a finite time interval $(0, T)$.





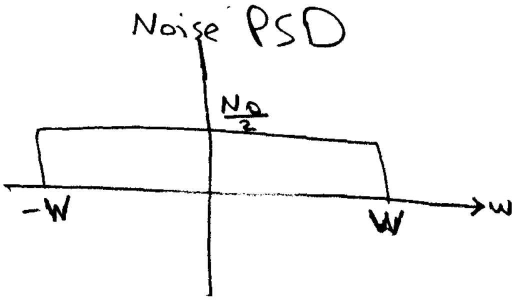
$\Downarrow$ sample at $T_s = \frac{1}{2W}$



$\Rightarrow$ number of sample in f(t) in the period $(0, T)$ is equal to $2WT$

$\Rightarrow$ The sampled function is a vector in a vector space of $2TW$ dimensions

Band-Limited Noise



— If the noise has power spectral density $\frac{N_0}{2}$ and bandwidth W

$\Rightarrow$ The noise power $= \frac{N_0}{2}(2W) = N_0 W$

— The noise samples in one interval T are i.i.d Gaussian R.V with variance equal to $\frac{N_0 W T}{2WT} = \frac{N_0}{2}$

Capacity of Band-Limited Channels

Recal that the capacity of Gaussian channels is

$$C = \frac{1}{2} \log_2\left(1 + \frac{P}{N}\right) \text{ bits per transmission}$$

— Let the channel be used over the time interval $[0, T]$. In this case, the power per sample is $\frac{PT}{2WT} = \frac{P}{2W}$

— The noise variance per sample is $\frac{N_0}{2}$

$\Rightarrow$ The capacity per sample is

$$C = \frac{1}{2} \log\left(1 + \frac{\frac{P}{2W}}{\frac{N_0}{2}}\right)$$

$$C = \frac{1}{2} \log\left(1 + \frac{P}{N_0 W}\right) \text{ bits per sample.}$$

Signal Power

Noise Power

Since there are $2W$ samples each second

$\Rightarrow$ the capacity can be rewritten as

$$C = W \log\left(1 + \frac{P}{N_0 W}\right) \text{ bits per second}$$