Information Theory
Chap 8 & 9

Differential Entropy & Gaussian Channel

Presenter : JinHyundong

Contents

- Chap 8
  - Recap
  - Continuous Random Variable
  - Differential Entropy
- Chap 9
  - Gaussian Channel
  - Gaussian Channel with Constraint
  - Bandlimited Channel

Chap 8

Differential Entropy

Recap : Case of Discrete RV

- $H(X) = -\sum p(x) \log p(x) = -E_p \log p(x) \;\; for\; x \in X$
- $H(X,Y) = -\sum p(x,y) \log p(x,y)$
- $H(Y|X) = \sum p(x) H(Y|X = x) = -\sum \sum p(x,y) \log p(y|x)$
- $H(X,Y) = H(X) + H(Y|X)$
- $D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)}$
- $I(X;Y) = \sum \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
- $Let\; P_e = \Pr\{\hat{X}(Y) \neq X\}, then\; H(P_e) + P_e \log|\chi| \geq H(X|Y)$

Recap : Case of Discrete RV

- Converge in probability $if\ \forall\ \epsilon > 0, \Pr\{|X_n - X| > \epsilon\} \to 0$
- Converge in mean square $if\ E(X_n - X)^2 \to 0$
- Converge almost surely $if\ \Pr\left\{\lim_{n\to\infty} X_n = X\right\} = 1$
- $Z_1, \dots, Z_n$ are i.i.d. $\sim(\mu, \sigma^2), \overline{Z_n} = \frac{1}{n}\sum Z_i\,, \Pr\{|\overline{Z_n} - \mu| > \epsilon\} \le \frac{\sigma^2}{n\epsilon^2}$
- $If\ X_1, X_2, \dots$ are i.i.d.$\sim p(x), then\ -\frac{1}{n}\log p(X_1, X_2, \dots, X_n) \to H(X)$ in probability.
- $A_\epsilon^{(n)} = \{(x_1, \dots, x_n) \in \chi^n | 2^{-n(H(X)+\epsilon)} \le p(x_1, \dots, x_n) \le 2^{-n(H(X)-\epsilon)}\}$

Continuous RV & Density Function

Def

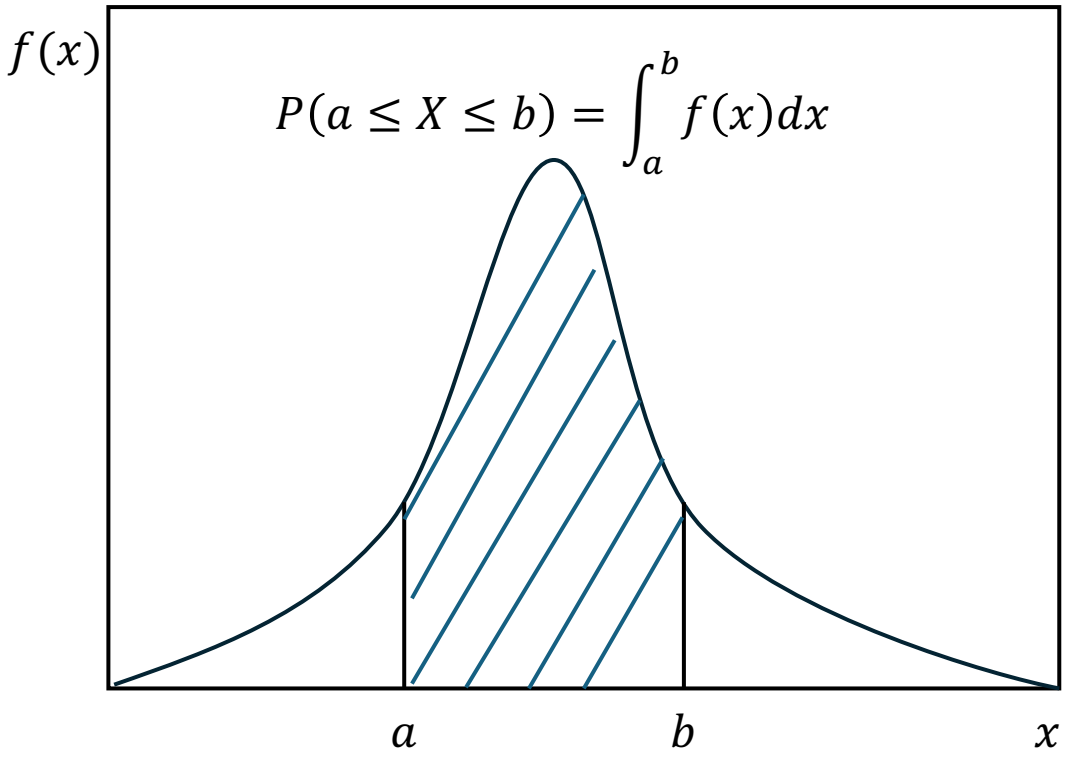
For random variable X

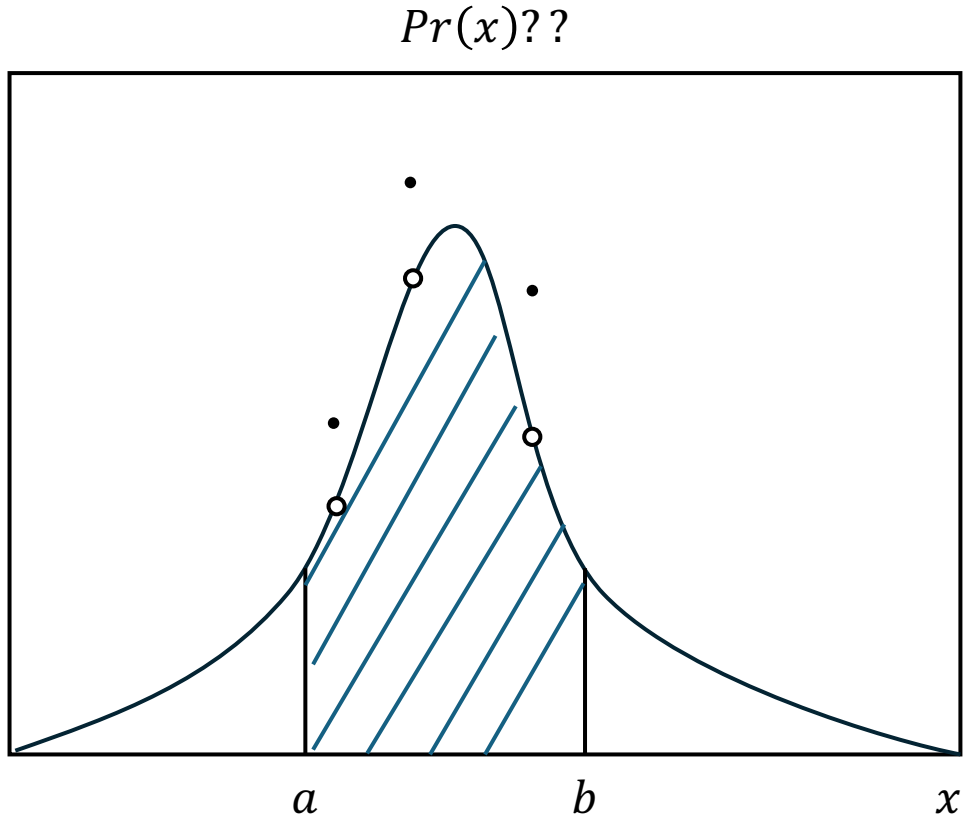
& cumulative distribution function $F(x) = \Pr(X \leq x)$,

If $F(x)$ is continuous, then X is said to be continuous.

If $F(x)$ is absolutely continuous where $F'(x) = \mathrm{f}(x)$ and $\int_{-\infty}^{\infty} \mathrm{f}(x) = 1$, then $\mathrm{f}(x)$ is called the probability density function of X.

Density Function





Expectation of Random Variable

- Discrete

$$E\big(g(X)\big) = \sum p(x)g(x)$$

- Continuous

$$E\big(g(X)\big) = \int_S f(x)g(x)dx$$

Definition of Differential Entropy

Def

Differential Entropy

$$\mathrm{h}(X) = h(f) = -\int_S f(x)\log f(x)\,dx\,, \qquad S : supp(X)$$

Joint Differential Entropy

$$\mathrm{h}(X_1, \dots, X_n) = -\int_S f(x_1, \dots, x_n)\log f(x_1, \dots, x_n)\,dx_1 \dots dx_n$$

Conditional Differential Entropy

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

In general $\mathrm{f(x|y)} = \frac{\mathrm{f(x},y)}{f(y)}$ holds, thus $\mathrm{h(X|Y)} = \mathrm{h(X,Y)} - \mathrm{h(Y)}$.

Definition of KL-distance and Mutual Information

Def

KL-distance (Relative Entropy)

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

Mutual Information (when joint density $f(x,y)$ is given.)

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

In general, $I(X;Y) = h(X) - h(X|Y) = D\big(f(x,y)||f(x)f(y)\big)$

Relation of Differential & Discrete Entropy

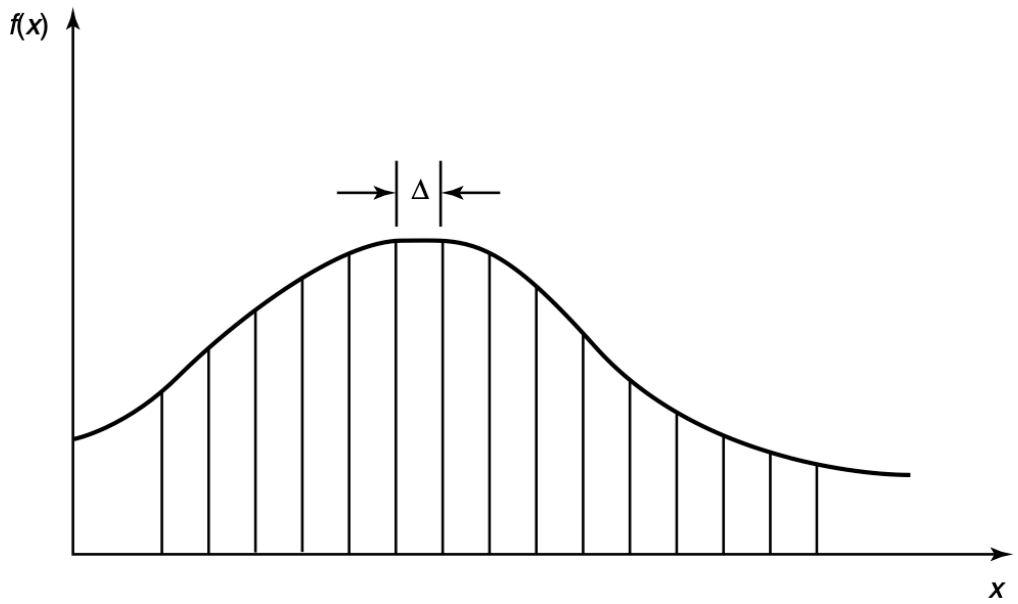


FIGURE 8.1. Quantization of a continuous random variable.

- Divide the range of X into bins of length Δ
- By MVT, $\exists x_i \; such \; that$

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

- Let $X^{\Delta} = x_i, \;\; if \; i\Delta \leq X < (i+1)\Delta$
- Then, $\mathrm{p_i} = \Pr\left(X^{\Delta} = x_i\right) = f(x_i)\Delta$

Relation of Differential & Discrete Entropy

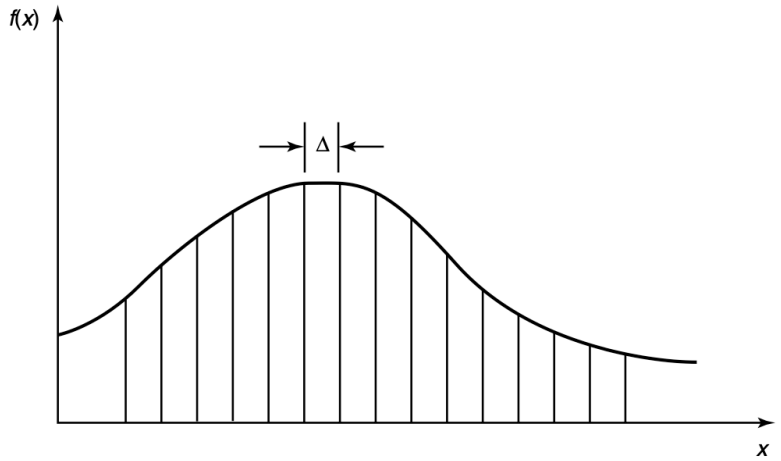


FIGURE 8.1. Quantization of a continuous random variable.

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

$$= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta)$$

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

$$= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta$$

Relation of Differential & Discrete Entropy

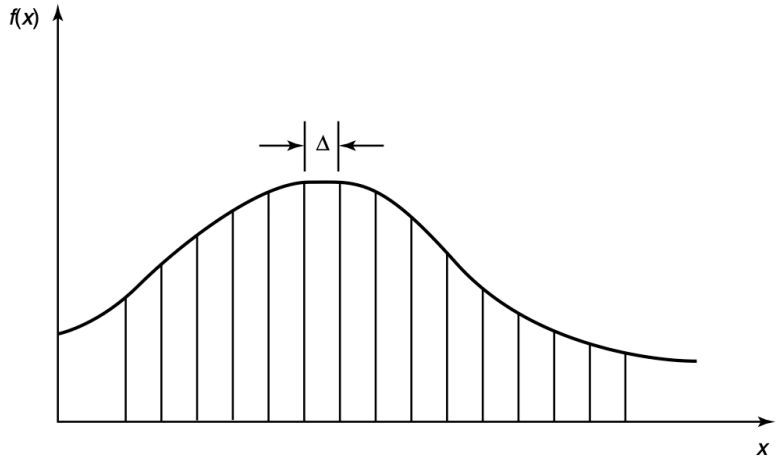


FIGURE 8.1. Quantization of a continuous random variable.

- Thus, if density $f(x)$ of the RV X is Riemann integrable, then

$$H(X^{\Delta}) + \log \Delta \to h(f) = h(X) \quad \text{as } \Delta \to 0$$

- The entropy of an n-bit quantization of a continuous RV X is approximately $h(X) + \mathrm{n}$.

General Definition of Mutual Information

- Mutual information of two continuous random variables is the limit of the mutual information between their quatized versions.

$$I\left(X^{\Delta};Y^{\Delta}\right) = H\left(X^{\Delta}\right) - H\left(X^{\Delta}|Y^{\Delta}\right)$$
$$\approx h(X) - \log\Delta - (h(X|Y) - \log\Delta)$$
$$= I(X;Y)$$

General Definition of Mutual Information

Def

Let χ be the range of a RV X. A partition $\mathcal{P}$ of χ is a finite collection of disjoint sets P_i s.t $\cup_i P_i = \chi$.

The quantization of X by $\mathcal{P}$ is the discrete RV defined by

$$\Pr([X]_{\mathcal{P}} = i) = \Pr(X \in P_i) = \int_{P_i} dF(x)$$

Mutual Information (when joint density $f(x, y)$ is not given.)

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

Some Properties

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

AEP for Continuous Random Variable

Almost same for discrete case :

Let $X_1, \ldots, X_n$ be a sequence of RV with i.i.d. & density $f(x)$.

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

Typical set for Continuous Random Variable

Def

For $\epsilon > 0$ and any n,

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

Since $X_1, \dots, X_n$ are i.i.d. $f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$

The volume of a set $A \subset \mathcal{R}^n$ is defined as

$$Vol(A) = \int_A dx_1 \dots dx_n$$

Some Properties of the Typical set

- $\Pr\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$ for sufficiently large n
- $\mathrm{Vol}\left(A_\epsilon^{(n)}\right) \leq 2^{n(h(X)+\epsilon)}$ for all n
- $\mathrm{Vol}\left(A_\epsilon^{(n)}\right) \geq (1-\epsilon)2^{n(h(X)-\epsilon)}$ for sufficiently large n

Entropy with Normal Distribution

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

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

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

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

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

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

Chap 9

Gaussian Channel

Gaussian Channel

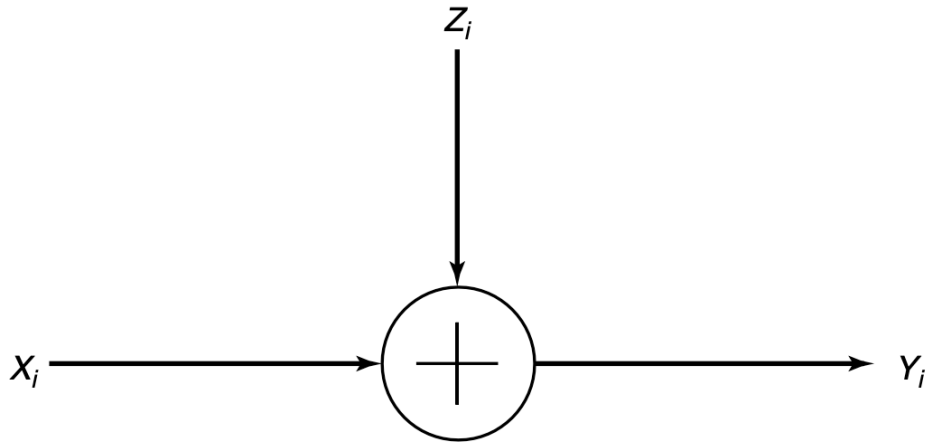


FIGURE 9.1. Gaussian channel.

The most important continuous alphabet channel.

X_i : input

$Z_i \sim \mathcal{N}(0, N)$: i.i.d. gaussian noise

$Y_i = X_i + Z_i$ where i is discrete time

Gaussian Channel without Noise

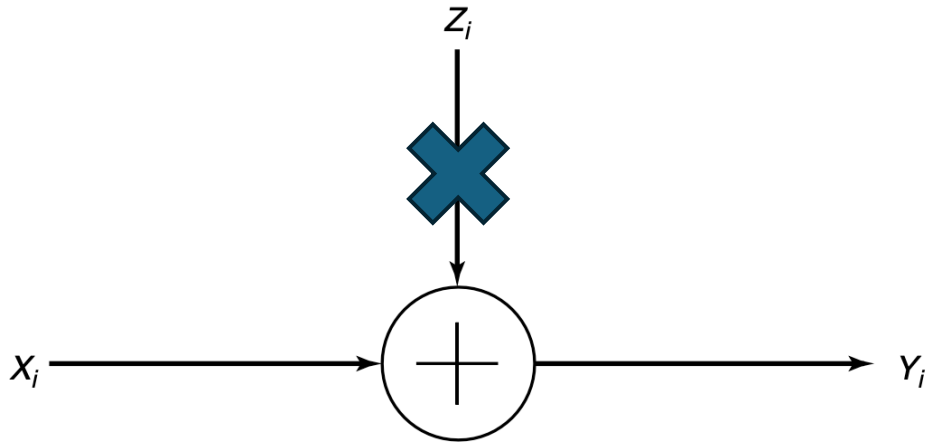


FIGURE 9.1. Gaussian channel.

The capacity of the gaussian channel without noise is infinity since X can take on any real value and transmit it with no error.

Gaussian Channel without constraint

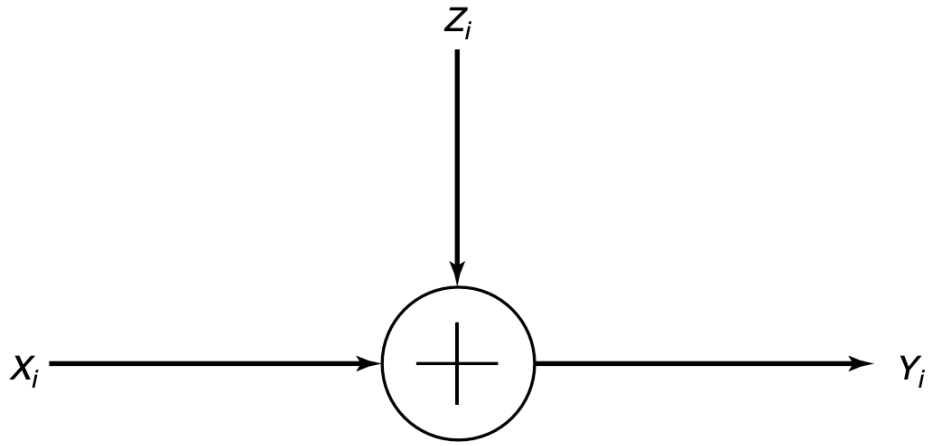


FIGURE 9.1. Gaussian channel.

What happens there exists a noise $Z_i \sim \mathcal{N}(0, N)$?

Gaussian Channel without constraint

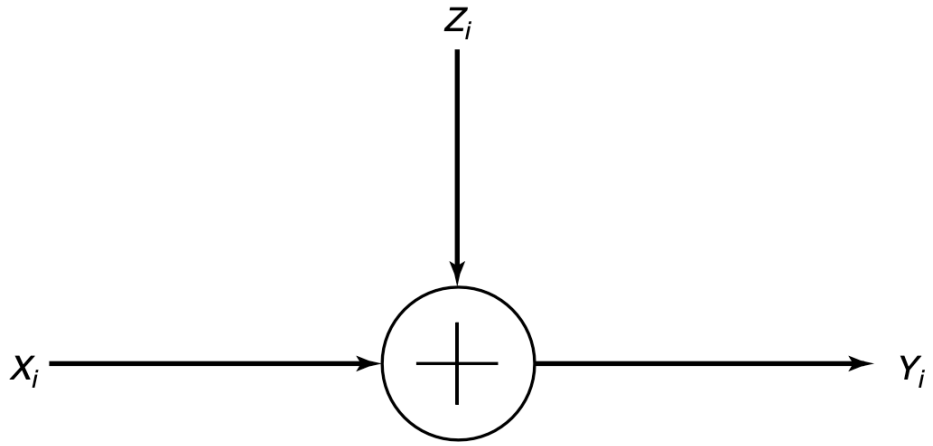


FIGURE 9.1. Gaussian channel.

We can choose an infinite subset of inputs arbitrarily far apart, so that they are distinguishable at the output with arbitrarily small probability of error.

⇒ Without constraint, it still has infinity capacity!

Gaussian Channel with constraint

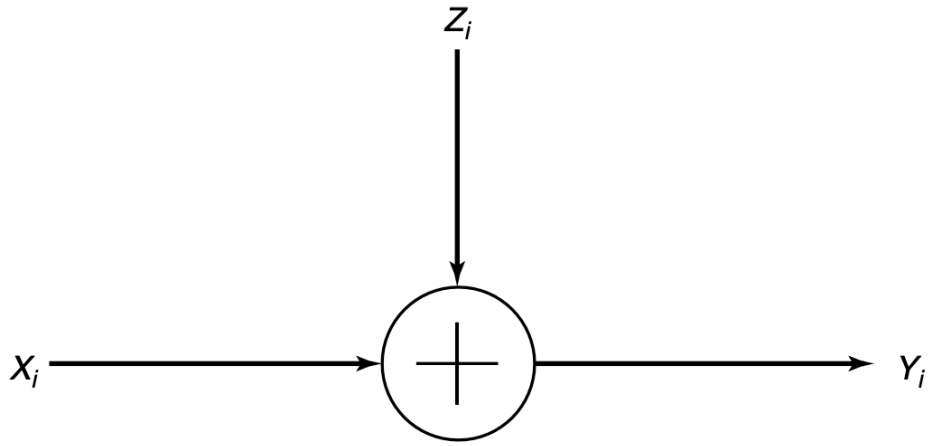


FIGURE 9.1. Gaussian channel.

The most common limitation on input is an average power constraint.

$$\frac{1}{n}\sum x_i^2 \leq P : \text{power constraint}$$

Also, we use quantization to convert the Gaussian channel into a discrete channel which is easier to process.

The Capacity of the Gaussian Channel

Def

The information capacity of the Gaussian channel with power constraint P is given as

$$C = \max_{f(x):EX^2 \leq P} I(X;Y)$$

The Capacity of the Gaussian Channel

$$C = \max_{f(x):EX^2 \le P} I(X;Y)$$

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

Here, $h(Z) = \frac{1}{2}\log 2\pi eN$ as calculated in previous chapter.

The Capacity of the Gaussian Channel

$$EY^2 = E(X+Z)^2 = EX^2 + 2EXEZ + EZ^2 \leq P + N$$

because X & $Z \sim \mathcal{N}(0, N)$ are independent and $EX^2 \leq P$

However, for a fixed variance, the normal distribution maximizes the entropy. (Theorem 8.6.5)

Thus, $h(Y) \leq \frac{1}{2}\log 2\pi e(EY^2)$

The Capacity of the Gaussian Channel

By combining the previous results,

$$
\begin{aligned}
I(X;Y) &= h(Y) - h(Z) \\
&\leq \frac{1}{2}\log 2\pi e(EY^2) - \frac{1}{2}\log 2\pi eN \\
&\leq \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi eN \\
&= \frac{1}{2}\log\left(1+\frac{P}{N}\right)
\end{aligned}
$$

The Capacity of the Gaussian Channel

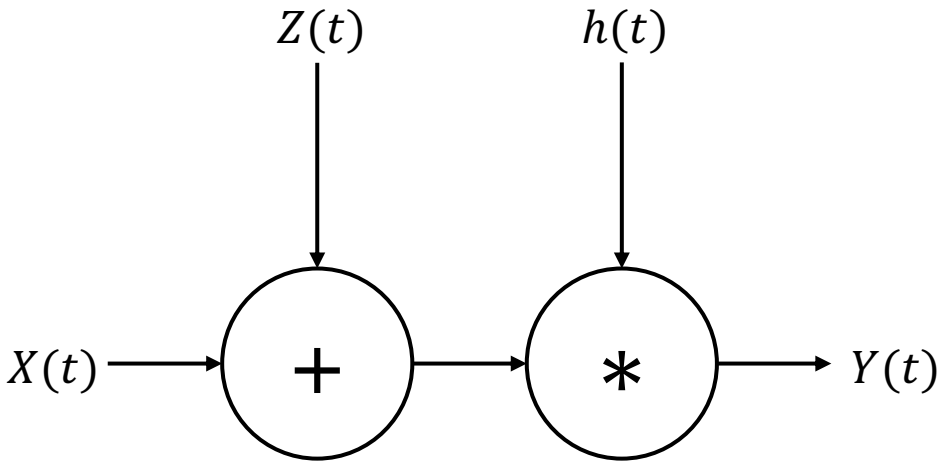
Thus, the information capacity of the Gaussian channel is

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

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

The definitions for (M, n) code, the rate of error, and achievable are almost same as in chapter 7. Please check the textbook.

The Bandlimited Channels (Continuous Time)



The commonly used channel like a radio network or a telephone line is a bandlimited channel with white noise.

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

Here, $\mathrm{h(t)}$ is the impulse response of a low pass filter and $*$ is the convolution operator.

Representation Theorem

Theorem 9.3.1 by Nyquist and Shannon

$f(t)$ is bandlimited to W (i.e. the spectrum of the function is 0 for all frequencies greater than W).

Then, the $f(t)$ is completely determined by samples of the function spaced $\frac{1}{2W}$ seconds apart.

Proof of the Representation Theorem

Let $F(\omega)$ be the Fourier transform of $f(t)$. Then,

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}d\omega$$
$$= \frac{1}{2\pi}\int_{-2\pi W}^{2\pi W} F(\omega)e^{i\omega t}d\omega$$

If we consider samples spaced $\frac{1}{2W}$ seconds apart,

$$f\left(\frac{n}{2W}\right) = \frac{1}{2\pi}\int_{-2\pi W}^{2\pi W} F(\omega)e^{i\omega\frac{n}{2W}}d\omega$$

Proof of the Representation Theorem

Def (continuous Fourier Series)

Let $\mathrm{f} : [a, b] \to \mathbb{C}$ be an integrable function with $L = b - a$, then the k-th Fourier coefficient of f is defined by

$$a_k = \frac{1}{L}\int_a^b e^{-\frac{2\pi}{L}ikx} f(x)dx$$

The Fourier series of f is given by formally

$$f(x) \sim \sum a_k e^{\frac{2\pi}{L}ikx}$$

Proof of the Representation Theorem

$$f\left(-\frac{n}{2W}\right) = \frac{1}{2\pi}\int_{-2\pi W}^{2\pi W} F(\omega)e^{-i\omega\frac{n}{2W}}d\omega$$

From the above equation, right side is the Fourier coefficient of $\mathrm{F}(\omega)$.

Thus, we can calculate the Fourier coefficients of $\mathrm{F}(\omega)$ from the sampled points.

After that, By using Fourier inversion, we can determine the original function $f(t)$.

Proof of the Representation Theorem

When we only consider the real part, the given function can be explicitly represented in terms of its samples as following:

$$f(t) = \mathrm{g(t)} = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2W}\right) sinc\left(t - \frac{n}{2W}\right)$$

where $sinc(t) = \frac{\sin(2\pi W t)}{2\pi W t}$

The capacity of the Bandlimited Channels

By the Nyquist-Shannon sampling theorem, a bandlimited function has only 2W degrees of freedom per second.

Also, we can say that the most of the power is in bandwidth W and in a finite time interval (0, T).

Then, we can describe any function with 2TW orthonormal bases as almost timelimited & almost bandlimited.

The capacity of the Bandlimited Channels

Let the noise has power spectral density $\frac{N_0}{2}$ (W/hz) and bandwidth W (hz) in time T with power constraint P (W).

Also, We have $2WT$ samples taken $\frac{1}{2W}$ apart.

Then the energy per sample is $\frac{PT}{2WT} = \frac{P}{2W}$, and the noise variance per sample is $\frac{N_0}{2} 2W \frac{T}{2WT} = \frac{N_0}{2}$

The capacity of the Bandlimited Channels

Since the channel capacity is defined as follows,

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

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

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

Since there are 2W samples per second,

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

The capacity of the Bandlimited Channels

A more precise version considers the small fraction of their energy outside the bandwidth W.

To consider it, let $W \to \infty$. Then, we obtain

$$C = \frac{P}{N_0} \log e \quad bits\ per\ second$$

Following Topics are skipped…

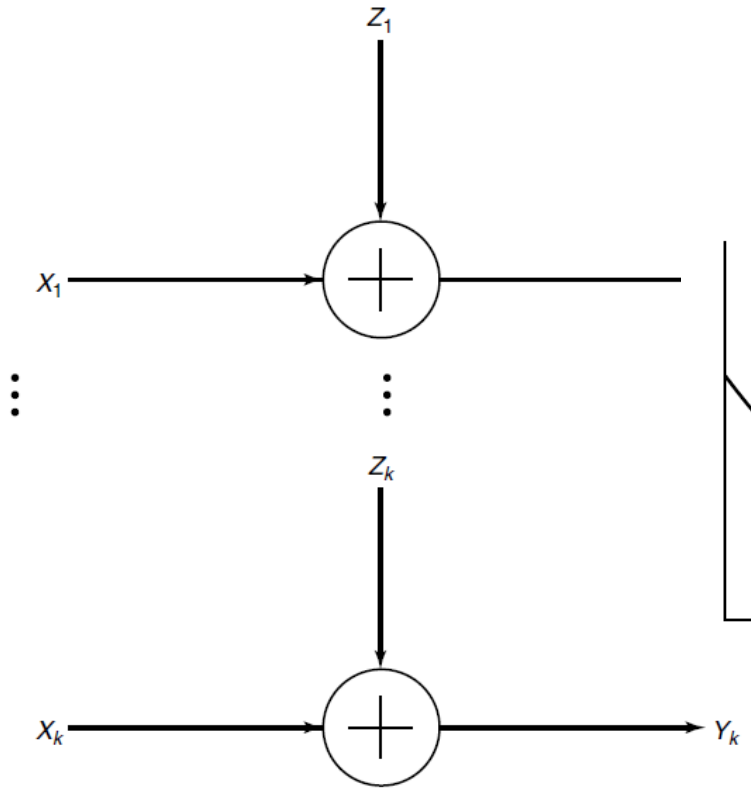


FIGURE 9.3. Parallel Gaussian channels.

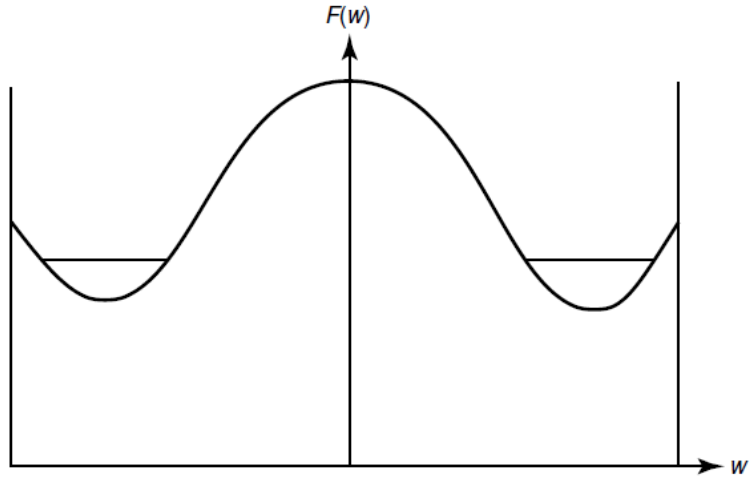


FIGURE 9.5. Water-filling in the spectral domain.

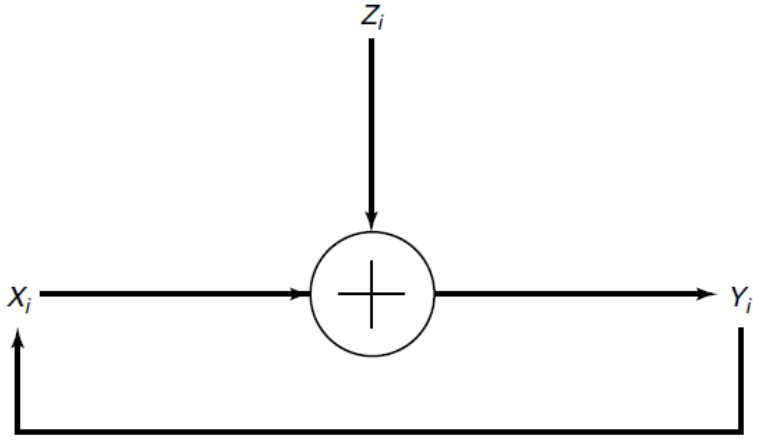


FIGURE 9.6. Gaussian channel with feedback.

Thank You!