Chapter 5

Differential Entropy and Gaussian Channels

Po-Ning Chen, Professor

Institute of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 30010, R.O.C.

Continuous sources

- Model

$$\{X_t \in \mathcal{X}, t \in I\}$$

  - Discrete sources
    - Both $\mathcal{X}$ and I are discrete.
  - Continuous sources
    - Discrete-time continuous sources
      - $\mathcal{X}$ is continuous; I is discrete.
    - Waveform sources
      - Both $\mathcal{X}$ and I are continuous.

- We have so far examined information measures and their operational characterization for **discrete-time discrete-alphabet** systems. In this chapter, we turn our focus to **discrete-time continuous-alphabet** (real-valued) sources.

Information content of continuous sources

- If the random variable takes on values in a continuum, the minimum number of bits per symbol needed to losslessly describe it must be infinite.
- This is illustrated in the following example and validated in Lemma 5.2.

  Example 5.1

  – Consider a real-valued random variable X that is uniformly distributed on the unit interval, i.e., with pdf given by

  $$f_X(x) = \begin{cases} 1 & \text{if } x \in [0,1); \\ 0 & \text{otherwise.} \end{cases}$$

  – Given a positive integer m, we can discretize X by uniformly quantizing it into m levels by partitioning the support of X into equal-length segments of size $\Delta = \frac{1}{m}$ (Δ is called the quantization step-size) such that:

  $$q_m(X) = \frac{i}{m}, \quad \text{if } \frac{i-1}{m} \leq X < \frac{i}{m},$$

  for $1 \leq i \leq m$.

  – Then the entropy of the quantized random variable $q_m(X)$ is given by

  $$H(q_m(X)) = -\sum_{i=1}^{m} \frac{1}{m} \log_2 \left(\frac{1}{m}\right) = \log_2 m \quad \text{(in bits)}.$$

– Since the entropy $H(q_m(X))$ of the quantized version of X is a lower bound to the entropy of X (as $q_m(X)$ is a function of X) and satisfies in the limit

$$\lim_{m\to\infty} H(q_m(X)) = \lim_{m\to\infty} \log_2 m = \infty,$$

we obtain that the entropy of X is infinite. □

- The above example indicates that to compress a continuous source without incurring any loss or distortion requires an **infinite** number of bits.
- Thus when studying continuous sources, the entropy measure is **limited** in its **effectiveness** and the introduction of a new measure is necessary.
- Such a new measure is obtained upon **close examination** of the **entropy of a uniformly quantized real-valued random-variable** minus the **quantization accuracy** as the accuracy increases without bound.

Lemma 5.2 Consider a real-valued random variable X with support $[a, b)$ and pdf f_X such that

$$\begin{cases} \text{(i) } -f_X \log_2 f_X \text{ is (Riemann-)integrable, } and \\ \text{(ii) } -\int_a^b f_X(x) \log_2 f_X(x) dx \text{ is finite.} \end{cases}$$

Then a uniform quantization of X with an n-bit accuracy (i.e., with a quantization step-size of $\Delta = 2^{-n}$) yields an entropy approximately equal to

$$-\int_a^b f_X(x) \log_2 f_X(x) dx + n \quad \text{bits}$$

for n sufficiently large. In other words,

$$\lim_{n\to\infty} [H(q_n(X)) - n] = -\int_a^b f_X(x) \log_2 f_X(x) dx$$

where $q_n(X)$ is the uniformly quantized version of X with quantization step-size $\Delta = 2^{-n}$.

Proof:

Step 1: Mean-value theorem.

Let $\Delta = 2^{-n}$ be the quantization step-size, and let

$$t_i := \begin{cases} a + i\Delta, & i = 0, 1, \cdots, j-1 \\ b, & i = j \end{cases}$$

where

$$j = \left\lceil \frac{b-a}{\Delta} \right\rceil .$$

From the mean-value theorem, we can choose $x_i \in [t_{i-1}, t_i]$ for $1 \leq i \leq j$ such that

$$p_i := \int_{t_{i-1}}^{t_i} f_X(x)dx = f_X(x_i)(t_i - t_{i-1}) = \Delta \cdot f_X(x_i).$$

Step 2: Definition of $h^{(n)}(X)$.

Let

$$h^{(n)}(X) := -\sum_{i=1}^{j}[f_X(x_i)\log_2 f_X(x_i)]\Delta = -\sum_{i=1}^{j}[f_X(x_i)\log_2 f_X(x_i)]2^{-n}.$$

Since $-f_X(x)\log_2 f_X(x)$ is (Riemann-)integrable,

$$h^{(n)}(X) \to -\int_a^b f_X(x)\log_2 f_X(x)dx \quad \text{as } n \to \infty.$$

Therefore, given any $\varepsilon > 0$, there exists N such that for all $n > N$,

$$\left|-\int_a^b f_X(x)\log_2 f_X(x)dx - h^{(n)}(X)\right| < \varepsilon.$$

Step 3: Computation of $H(q_n(X))$. The entropy of the (uniformly) quantized version of X, $q_n(X)$, is given by

$$\begin{aligned} H(q_n(X)) &= -\sum_{i=1}^{j} p_i \log_2 p_i \\ &= -\sum_{i=1}^{j} (f_X(x_i)\Delta) \log_2 (f_X(x_i)\Delta) \\ &= -\sum_{i=1}^{j} (f_X(x_i)2^{-n}) \log_2 (f_X(x_i)2^{-n}) \end{aligned}$$

where the p_i's are the probabilities of the different values of $q_n(X)$.

Step 4: $H(q_n(X)) - h^{(n)}(X)$. From Steps 2 and 3,

$$\begin{aligned} H(q_n(X)) - h^{(n)}(X) &= -\sum_{i=1}^{j} [f_X(x_i)2^{-n}] \log_2(2^{-n}) \\ &= n \sum_{i=1}^{j} \int_{t_{i-1}}^{t_i} f_X(x)dx \\ &= n \int_a^b f_X(x)dx \\ &= n. \end{aligned}$$

Hence, we have that for $n > N$,

$$\begin{aligned} \left[-\int_a^b f_X(x) \log_2 f_X(x)dx + n \right] - \varepsilon &< H(q_n(X)) = h^{(n)}(X) + n \\ &< \left[-\int_a^b f_X(x) \log_2 f_X(x)dx + n \right] + \varepsilon, \end{aligned}$$

yielding that

$$\lim_{n\to\infty} [H(q_n(X)) - n] = -\int_a^b f_X(x) \log_2 f_X(x)dx. \qquad \square$$

Lemma 5.2 actually holds not limited for support $[a, b)$ but for any support S_X.

Theorem 5.3 [340, Theorem 1](Rényi 1959) For any real-valued random variable with pdf f_X, if

$$-\sum_{i=1}^{j} p_i \log_2 p_i$$

is finite, where the p_i's are the probabilities of the different values of uniformly quantized $q_n(X)$ over support S_X, then

$$\lim_{n\to\infty} [H(q_n(X)) - n] = -\int_{S_X} f_X(x) \log_2 f_X(x) dx$$

provided the integral on the right-hand side exists.

This suggests that $\int_{S_X} f_X(x) \log_2 \frac{1}{f_X(x)} dx$ could be a good information measure for continuous-alphabet sources.

5.1 Differential entropy

Definition 5.4 (Differential entropy) The differential entropy (in bits) of a continuous random variable X with pdf f_X and support S_X is defined as

$$h(X) := -\int_{S_X} f_X(x) \cdot \log_2 f_X(x) dx = E[-\log_2 f_X(X)],$$

when the integral exists.

Example 5.5 A continuous random variable X with support $S_X = [0, 1)$ and pdf $f_X(x) = 2x$ for $x \in S_X$ has differential entropy equal to

$$\begin{aligned} \int_0^1 -2x \cdot \log_2(2x) dx &= \left. \frac{x^2(\log_2 e - 2\log_2(2x))}{2} \right|_0^1 \\ &= \frac{1}{2\ln 2} - \log_2(2) = -0.278652 \text{ bits.} \end{aligned}$$

Next, we have that $q_n(X)$ is given by

$$q_n(X) = \frac{i}{2^n}, \quad \text{if } \frac{i-1}{2^n} \leq X < \frac{i}{2^n},$$

for $1 \leq i \leq 2^n$. Hence,

$$\Pr\left\{q_n(X) = \frac{i}{2^n}\right\} = \frac{(2i-1)}{2^{2n}},$$

which yields

$$\begin{aligned} H(q_n(X)) &= -\sum_{i=1}^{2^n} \frac{2i-1}{2^{2n}} \log_2\left(\frac{2i-1}{2^{2n}}\right) \\ &= \left[-\frac{1}{2^{2n}} \sum_{i=1}^{2^n} (2i-1)\log_2(2i-1) + 2\log_2(2^n)\right]. \end{aligned}$$

n	$H(q_n(X))$	$H(q_n(X)) - n$
1	0.811278 bits	-0.188722 bits
2	1.748999 bits	-0.251000 bits
3	2.729560 bits	-0.270440 bits
4	3.723726 bits	-0.276275 bits
5	4.722023 bits	-0.277977 bits
6	5.721537 bits	-0.278463 bits
7	6.721399 bits	-0.278600 bits
8	7.721361 bits	-0.278638 bits
9	8.721351 bits	-0.278648 bits

Example 5.7 (Differential entropy of a uniformly distributed random variable) Let X be a continuous random variable that is uniformly distributed over the interval (a, b), where $b > a$; i.e., its pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in (a, b); \\ 0 & \text{otherwise.} \end{cases}$$

So its differential entropy is given by

$$h(X) = -\int_a^b \frac{1}{b-a} \log_2 \frac{1}{b-a} = \log_2(b-a) \quad \text{bits.}$$

- Note that if $(b-a) < 1$ in the above example, then $h(X)$ is *negative*.

Example 5.8 (Differential entropy of a Gaussian random variable) Let $X \sim \mathcal{N}(\mu, \sigma^2)$; i.e., X is a Gaussian (or normal) random variable with finite mean μ, variance $\mathrm{Var}(X) = \sigma^2 > 0$ and pdf

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

for $x \in \mathbb{R}$. Then its differential entropy is given by

$$\begin{aligned} h(X) &= \int_{\mathbb{R}} f_X(x) \left[\frac{1}{2} \log_2(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \log_2 e \right] dx \\ &= \frac{1}{2} \log_2(2\pi\sigma^2) + \frac{\log_2 e}{2\sigma^2} E[(X-\mu)^2] \\ &= \frac{1}{2} \log_2(2\pi\sigma^2) + \frac{1}{2} \log_2 e \\ &= \frac{1}{2} \log_2(2\pi e\sigma^2) \quad \text{bits.} \end{aligned} \tag{5.1.1}$$

- Note that for a Gaussian random variable, its differential entropy is only a function of its variance σ^2 (it is functionally independent from its mean μ).
- This is similar to the differential entropy of a uniform random variable, which only depends on difference $(b-a)$ but not the mean $(a+b)/2$.

Definition 5.9 (Joint differential entropy) If $X^n = (X_1, X_2, \cdots, X_n)$ is a continuous random vector of size n (i.e., a vector of n continuous random variables) with joint pdf f_{X^n} and support $S_{X^n} \subseteq \mathbb{R}^n$, then its joint differential entropy is defined as

$$\begin{aligned} h(X^n) &:= -\int_{S_{X^n}} f_{X^n}(x_1, x_2, \cdots, x_n) \log_2 f_{X^n}(x_1, x_2, \cdots, x_n)\, dx_1\, dx_2\, \cdots\, dx_n \\ &= E[-\log_2 f_{X^n}(X^n)] \end{aligned}$$

when the n-dimensional integral exists.

Definition 5.10 (Conditional differential entropy) Let X and Y be two jointly distributed continuous random variables with joint pdf $f_{X,Y}$ and support $S_{X,Y} \subseteq \mathbb{R}^2$ such that the conditional pdf of Y given X, given by

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

is well defined for all $(x,y) \in S_{X,Y}$, where f_X is the marginal pdf of X. Then the conditional entropy of Y given X is defined as

$$h(Y|X) := -\int_{S_{X,Y}} f_{X,Y}(x,y) \log_2 f_{Y|X}(y|x)\, dx\, dy = E[-\log_2 f_{Y|X}(Y|X)],$$

when the integral exists.

- **Chain rule for differential entropy**:

$$h(X,Y) = h(X) + h(Y|X) = h(Y) + h(X|Y).$$

Definition 5.11 (Divergence or relative entropy) Let X and Y be two continuous random variables with marginal pdfs f_X and f_Y, respectively, such that their supports satisfy $S_X \subseteq S_Y \subseteq \mathbb{R}$. Then the divergence (or relative entropy or Kullback-Leibler distance) between X and Y is written as $D(X\|Y)$ or $D(f_X\|f_Y)$ and defined by

$$D(X\|Y) := \int_{S_X} f_X(x) \log_2 \frac{f_X(x)}{f_Y(x)}\, dx = E\left[\frac{f_X(X)}{f_Y(X)}\right]$$

when the integral exists. The definition carries over similarly in the multivariate case: for $X^n = (X_1, X_2, \cdots, X_n)$ and $Y^n = (Y_1, Y_2, \cdots, Y_n)$ two random vectors with joint pdfs f_{X^n} and f_{Y^n}, respectively, and supports satisfying $S_{X^n} \subseteq S_{Y^n} \subseteq \mathbb{R}^n$, the divergence between X^n and Y^n is defined as

$$D(X^n\|Y^n) := \int_{S_{X^n}} f_{X^n}(x_1, x_2, \cdots, x_n) \log_2 \frac{f_{X^n}(x_1, x_2, \cdots, x_n)}{f_{Y^n}(x_1, x_2, \cdots, x_n)}\, dx_1\, dx_2\, \cdots\, dx_n$$

when the integral exists.

- Note that $D(q_n(X)\|q_n(Y)) \to D(X\|Y)$ for continuous X and Y.

Definition 5.12 (Mutual information) Let X and Y be two jointly distributed continuous random variables with joint pdf $f_{X,Y}$ and support $S_{XY} \subseteq \mathbb{R}^2$. Then the mutual information between X and Y is defined by

$$I(X;Y) := D(f_{X,Y} \| f_X f_Y) = \int_{S_{X,Y}} f_{X,Y}(x,y) \log_2 \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \, dx \, dy,$$

assuming the integral exists, where f_X and f_Y are the marginal pdfs of X and Y, respectively.

- For n and m sufficiently large,

$$\begin{aligned} I(q_n(X); q_m(Y)) &= H(q_n(X)) + H(q_m(Y)) - H(q_n(X), q_m(Y)) \\ &\approx [h(X) + n] + [h(Y) + m] - [h(X,Y) + n + m] \\ &= h(X) + h(Y) - h(X,Y) \\ &= \int_{S_{X,Y}} f_{X,Y}(x,y) \log_2 \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \, dx \, dy. \end{aligned}$$

  Hence,

$$\lim_{n,m \to \infty} I(q_n(X); q_m(Y)) = h(X) + h(Y) - h(X,Y).$$

- This justifies using identical notations for both $I(\cdot;\cdot)$ and $D(\cdot\|\cdot)$ as opposed to the discerning notations of $H(\cdot)$ for entropy and $h(\cdot)$ for differential entropy.

Lemma 5.14 The following properties hold for the information measures of continuous systems.

1. **Non-negativity of divergence:** Let X and Y be two continuous random variables with marginal pdfs f_X and f_Y, respectively, such that their supports satisfy $S_X \subseteq S_Y \subseteq \mathbb{R}$. Then

$$D(f_X \| f_Y) \geq 0$$

with equality iff $f_X(x) = f_Y(x)$ for all $x \in S_X$ except in a set of f_X-measure zero (i.e., $X = Y$ almost surely).

2. **Non-negativity of mutual information:** For any two continuous random variables X and Y,

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

with equality iff X and Y are independent.

3. **Conditioning never increases differential entropy:** For any two continuous random variables X and Y with joint pdf $f_{X,Y}$ and well-defined conditional pdf $f_{X|Y}$,

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

with equality iff X and Y are independent.

4. **Chain rule for differential entropy:** For a continuous random vector $X^n = (X_1, X_2, \cdots, X_n)$,

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

where $h(X_i|X_1, X_2, \ldots, X_{i-1})$:=$h(X_1)$ for $i = 1$.

5. **Chain rule for mutual information:** For continuous random vector $X^n = (X_1, X_2, \cdots, X_n)$ and random variable Y with joint pdf $f_{X^n,Y}$ and well-defined conditional pdfs $f_{X_i,Y|X^{i-1}}$, $f_{X_i|X^{i-1}}$ and $f_{Y|X^{i-1}}$ for $i = 1, \cdots, n$, we have that

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

where $I(X_i; Y|X_{i-1}, \cdots, X_1)$:=$I(X_1; Y)$ for $i = 1$.

6. **Data processing inequality:** For continuous random variables X, Y and Z such that $X \to Y \to Z$, i.e., X and Z are conditional independent given Y,

$$I(X; Y) \geq I(X; Z).$$

7. **Independence bound for differential entropy:** For a continuous random vector $X^n = (X_1, X_2, \cdots, X_n)$,

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

with equality iff all the X_i's are independent from each other.

8. **Invariance of differential entropy under translation:** For continuous random variables X and Y with joint pdf $f_{X,Y}$ and well-defined conditional pdf $f_{X|Y}$,

$$h(X+c) = h(X) \qquad \text{for any constant } c \in \mathbb{R}, \text{ and } h(X+Y|Y) = h(X|Y).$$

The results also generalize in the multivariate case: for two continuous random vectors $X^n = (X_1, X_2, \cdots, X_n)$ and $Y^n = (Y_1, Y_2, \cdots, Y_n)$ with joint pdf f_{X^n,Y^n} and well-defined conditional pdf $f_{X^n|Y^n}$,

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

for any constant n-tuple $c^n = (c_1, c_2, \cdots, c_n) \in \mathbb{R}^n$, and

$$h(X^n + Y^n|Y^n) = h(X^n|Y^n),$$

where the addition of two n-tuples is performed component-wise.

9. **Differential entropy under scaling:** For any continuous random variable X and any non-zero real constant a,

$$h(aX) = h(X) + \log_2 |a|.$$

10. **Joint differential entropy under linear mapping:** Consider the random (column) vector $\underline{X} = (X_1, X_2, \cdots, X_n)^T$ with joint pdf f_{X^n}, where T denotes transposition, and let $\underline{Y} = (Y_1, Y_2, \cdots, Y_n)^T$ be a random (column) vector obtained from the linear transformation $\underline{Y} = \mathbf{A}\underline{X}$, where $\mathbf{A}$ is an invertible (non-singular) $n \times n$ real-valued matrix. Then

$$h(\underline{Y}) = h(Y_1, Y_2, \cdots, Y_n) = h(X_1, X_2, \cdots, X_n) + \log_2 |\det(\mathbf{A})|,$$

where $\det(\mathbf{A})$ is the determinant of the square matrix $\mathbf{A}$.

11. **Joint differential entropy under nonlinear mapping:** Consider the random (column) vector $\underline{X} = (X_1, X_2, \cdots, X_n)^T$ with joint pdf f_{X^n}, and let $\underline{Y} = (Y_1, Y_2, \cdots, Y_n)^T$ be a random (column) vector obtained from the nonlinear transformation

$$\underline{Y} = \underline{g}(\underline{X}) := (g_1(X_1), g_2(X_2), \cdots, g_n(X_n))^T,$$

where each $g_i : \mathbb{R} \to \mathbb{R}$ is a differentiable function, $i = 1, 2, \cdots, n$. Then

$$\begin{aligned} h(\underline{Y}) &= h(Y_1, Y_2, \cdots, Y_n) \\ &= h(X_1, \cdots, X_n) + \int_{\mathbb{R}^n} f_{X^n}(x_1, \cdots, x_n) \log_2 |\det(\mathbf{J})| \; dx_1 \; \cdots \; dx_n, \end{aligned}$$

where $\mathbf{J}$ is the $n \times n$ Jacobian matrix given by

$$\mathbf{J} := \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_n} \end{bmatrix}.$$

Theorem 5.18 (Joint differential entropy of the multivariate Gaussian) If $\underline{X} \sim \mathcal{N}_n(\underline{\mu}, \mathbf{K}_{\underline{X}})$ is a Gaussian random vector with mean vector $\underline{\mu}$ and (positive-definite) covariance matrix $\mathbf{K}_{\underline{X}}$, then its joint differential entropy is given by

$$h(\underline{X}) = h(X_1, X_2, \cdots, X_n) = \frac{1}{2} \log_2 \left[(2\pi e)^n \det(\mathbf{K}_{\underline{X}}) \right]. \tag{5.2.1}$$

In particular, in the univariate case of $n = 1$, (5.2.1) reduces to (5.1.1).

Proof:

- Without loss of generality we assume that $\underline{X}$ has a zero mean vector since its differential entropy is invariant under translation by Property 8 of Lemma 5.14:

$$h(\underline{X}) = h(\underline{X} - \underline{\mu});$$

  so we assume that $\underline{\mu} = \underline{0}$.

- Since the (positive-definite) covariance matrix $\mathbf{K}_{\underline{X}}$ is a real-valued symmetric matrix, then it is orthogonally diagonizable; i.e., there exits a square $(n \times n)$ orthogonal matrix $\mathbf{A}$ (i.e., satisfying $\mathbf{A}^T = \mathbf{A}^{-1}$) such that $\mathbf{A}\mathbf{K}_{\underline{X}}\mathbf{A}^T$ is a diagonal matrix whose entries are given by the eigenvalues of $\mathbf{K}_{\underline{X}}$.
- As a result the linear transformation $\underline{Y} = \mathbf{A}\underline{X} \sim \mathcal{N}_n\left(\underline{0}, \mathbf{A}\mathbf{K}_{\underline{X}}\mathbf{A}^T\right)$ is a Gaussian vector with the diagonal covariance matrix $\mathbf{K}_{\underline{Y}} = \mathbf{A}\mathbf{K}_{\underline{X}}\mathbf{A}^T$ and has therefore independent components.

- Thus

$$\begin{aligned}
h(\underline{Y}) &= h(Y_1, Y_2, \cdots, Y_n) \\
&= h(Y_1) + h(Y_2) + \cdots + h(Y_n) \quad \text{(by independence of } \underline{Y}) \\
&= \sum_{i=1}^{n} \frac{1}{2} \log_2 \left[2\pi e \mathrm{Var}(Y_i)\right] \\
&= \frac{n}{2} \log_2(2\pi e) + \frac{1}{2} \log_2 \left[\prod_{i=1}^{n} \mathrm{Var}(Y_i)\right] \\
&= \frac{n}{2} \log_2(2\pi e) + \frac{1}{2} \log_2 \left[\det\left(\mathbf{K}_{\underline{Y}}\right)\right] \\
&= \frac{1}{2} \log_2 \left(2\pi e\right)^n + \frac{1}{2} \log_2 \left[\det\left(\mathbf{K}_{\underline{X}}\right)\right] \qquad (5.2.5) \\
&= \frac{1}{2} \log_2 \left[\left(2\pi e\right)^n \det\left(\mathbf{K}_{\underline{X}}\right)\right],
\end{aligned}$$

where (5.2.5) holds since

$$\begin{aligned}
\det\left(\mathbf{K}_{\underline{Y}}\right) &= \det\left(\mathbf{A}\mathbf{K}_{\underline{X}}\mathbf{A}^T\right) \\
&= \det(\mathbf{A}) \det\left(\mathbf{K}_{\underline{X}}\right) \det(\mathbf{A}^T) \\
&= \det(\mathbf{A})^2 \det\left(\mathbf{K}_{\underline{X}}\right) \\
&= \det\left(\mathbf{K}_{\underline{X}}\right).
\end{aligned}$$

- Now noting that $|\det(\mathbf{A})| = 1$ yield that

  $$h(Y_1, Y_2, \cdots, Y_n) = h(X_1, X_2, \cdots, X_n) + \underbrace{\log_2 |\det(\mathbf{A})|}_{=0} = h(X_1, X_2, \cdots, X_n).$$

  We therefore obtain that

  $$h(X_1, X_2, \cdots, X_n) = \frac{1}{2} \log_2 \left[(2\pi e)^n \det \left(\mathbf{K}_{\underline{X}} \right) \right],$$

  hence completing the proof. □

- **An important fact:** Among all real-valued size-n random vectors (of support $\mathbb{R}^n$) with identical mean vector and covariance matrix, the *Gaussian* random vector has the largest differential entropy. The proof of this fact requires the following inequality.

Corollary 5.19 (Hadamard's inequality) For any real-valued $n \times n$ positive-definite matrix $\mathbf{K} = [K_{i,j}]_{i,j=1,\cdots,n}$,

$$\det(\mathbf{K}) \leq \prod_{i=1}^{n} K_{i,i}$$

with equality iff $\mathbf{K}$ is a diagonal matrix, where $K_{i,i}$ are the diagonal entries of $\mathbf{K}$.

Theorem 5.20 (Maximal differential entropy for real-valued random vectors) Let $\underline{X} = (X_1, X_2, \cdots, X_n)^T$ be a real-valued random vector with support $S_{X^n} \subseteq \mathbb{R}^n$, mean vector $\underline{\mu}$ and covariance matrix $\mathbf{K}_{\underline{X}}$. Then

$$h(X_1, X_2, \cdots, X_n) \leq \frac{1}{2} \log_2 \left[(2\pi e)^n \det(\mathbf{K}_{\underline{\mathbf{X}}}) \right], \tag{5.2.11}$$

with equality iff $\underline{X}$ is Gaussian; i.e., $\underline{X} \sim \mathcal{N}_n \left(\underline{\mu}, \mathbf{K}_{\underline{X}} \right)$.

Proof: We will present the proof in two parts: the scalar or univariate case, and the multivariate case (baed on the univariate case).

(i) *Scalar case* ($n = 1$): Let X be a real-valued random variable with support $S_X \subseteq \mathbb{R}$, mean μ and variance σ^2.

For a Gaussian random variable $Y \sim \mathcal{N}(\mu, \sigma^2)$, we can write

$$\begin{aligned}
0 \;&\leq\; D(X\|Y) \\
&=\; \int_{S_X} f_X(x) \log_2 \frac{f_X(x)}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}\, dx \\
&=\; -h(X) + \int_{S_X} f_X(x) \left[\log_2 \left(\sqrt{2\pi\sigma^2} \right) + \frac{(x-\mu)^2}{2\sigma^2} \log_2 e \right] dx \\
&=\; -h(X) + \frac{1}{2} \log_2(2\pi\sigma^2) + \frac{\log_2 e}{2\sigma^2} \underbrace{\int_{S_X} (x-\mu)^2 f_X(x)\, dx}_{=\sigma^2} \\
&=\; -h(X) + \frac{1}{2} \log_2 \left[2\pi e \sigma^2 \right].
\end{aligned}$$

Thus

$$h(X) \leq \frac{1}{2} \log_2 \left[2\pi e \sigma^2 \right], \tag{5.2.12}$$

with equality iff $X = Y$ (almost surely); i.e., $X \sim \mathcal{N}(\mu, \sigma^2)$.

(ii). *Multivariate case* ($n > 1$): As in the proof of Theorem 5.18, we can use an orthogonal square matrix $\mathbf{A}$ (i.e., satisfying $\mathbf{A}^T = \mathbf{A}^{-1}$ and hence $|\det(\mathbf{A})| = 1$) such that $\mathbf{A}\mathbf{K}_{\underline{X}}\mathbf{A}^T$ is diagonal. Therefore, the random vector generated by the linear map

$$\underline{Z} = \mathbf{A}\underline{X}$$

will have a covariance matrix given by $\mathbf{K}_{\underline{Z}} = \mathbf{A}\mathbf{K}_{\underline{X}}\mathbf{A}^T$ and hence have uncorrelated (but not necessarily independent) components.

Thus

$$
\begin{aligned}
h(\underline{X}) &= h(\underline{Z}) - \underbrace{\log_2 |\det(\mathbf{A})|}_{=0} && (5.2.13)\\
&= h(Z_1, Z_2, \cdots, Z_n) \\
&\leq \sum_{i=1}^{n} h(Z_i) && (5.2.14)\\
&\leq \sum_{i=1}^{n} \frac{1}{2} \log_2 \left[2\pi e \mathrm{Var}(Z_i)\right] && (5.2.15)\\
&= \frac{n}{2} \log_2(2\pi e) + \frac{1}{2} \log_2 \left[\prod_{i=1}^{n} \mathrm{Var}(Z_i)\right] \\
&= \frac{1}{2} \log_2 (2\pi e)^n + \frac{1}{2} \log_2 \left[\det \left(\mathbf{K}_{\underline{Z}}\right)\right] && (5.2.16)\\
&= \frac{1}{2} \log_2 (2\pi e)^n + \frac{1}{2} \log_2 \left[\det \left(\mathbf{K}_{\underline{X}}\right)\right] && (5.2.17)\\
&= \frac{1}{2} \log_2 \left[(2\pi e)^n \det \left(\mathbf{K}_{\underline{X}}\right)\right],
\end{aligned}
$$

where (5.2.15) follows from (5.2.12) (the scalar case above).

Finally, equality is achieved in both (5.2.14) and (5.2.15) iff the random variables Z_1, Z_2, ..., Z_n are Gaussian and independent from each other, or equivalently iff $\underline{X} \sim \mathcal{N}_n\left(\underline{\mu}, \mathbf{K}_{\underline{X}}\right)$. □

Observation 5.21 The following two results can also be shown (the proof is left as an exercise):

1. Among all continuous random variables admitting a pdf with support the interval (a, b), where $b > a$ are real numbers, the uniformly distributed random variable maximizes differential entropy.

2. Among all continuous random variables admitting a pdf with support the interval $[0, \infty)$ and finite mean μ, the exponential distribution with parameter (or rate parameter) $\lambda = 1/\mu$ maximizes differential entropy.

3. Among all continuous random variables admitting a pdf with support $\mathbb{R}$, finite mean μ and finite differential entropy and satisfying $E[|X - \mu|] = \lambda$, where $\lambda > 0$ is a fixed finite parameter, the Laplacian random variable with mean μ, variance $2\lambda^2$ and pdf

$$f_X(x) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}} \text{ for } x \in \mathbb{R}$$

maximizes differential entropy.

5.3 AEP for continuous memoryless sources

- The extension of AEP theorem from discrete cases to continuous cases is not based on "number counting" (which is always infinity for continuous sources), but on "volume measuring."

Theorem 5.23 (AEP for continuous memoryless sources) Let $\{X_i\}_{i=1}^{\infty}$ be a continuous memoryless source (i.e., an infinite sequence of continuous i.i.d. random variables) with pdf $f_X(\cdot)$ and differential entropy $h(X)$. Then

$$-\frac{1}{n}\log f_X(X_1,\ldots,X_n) \to E[-\log_2 f_X(X)] = h(X) \quad \text{in probability.}$$

Proof: The proof is an immediate result of the law of large numbers (e.g., see Theorem 3.3). □

5.3 AEP for continuous memoryless sources

Definition 5.24 (Typical set) For $\delta > 0$ and any n given, define the typical set for the above continuous source as

$$\mathcal{F}_n(\delta) := \left\{ x^n \in \mathbb{R}^n : \left| -\frac{1}{n} \log_2 f_X(X_1, \ldots, X_n) - h(X) \right| < \delta \right\}.$$

Definition 5.25 (Volume) The *volume* of a set $\mathcal{A} \subset \mathbb{R}^n$ is defined as

$$\mathrm{Vol}(\mathcal{A}) := \int_{\mathcal{A}} dx_1 \cdots dx_n.$$

Theorem 5.26 (Consequence of the AEP for continuous memoryless sources) For a continuous memoryless source $\{X_i\}_{i=1}^{\infty}$ with differential entropy $h(X)$, the following hold.

1. For n sufficiently large, $P_{X^n}\{\mathcal{F}_n(\delta)\} > 1 - \delta$.
2. $\mathrm{Vol}(\mathcal{F}_n(\delta)) \leq 2^{n(h(X)+\delta)}$ for all n.
3. $\mathrm{Vol}(\mathcal{F}_n(\delta)) \geq (1-\delta)2^{n(h(X)-\delta)}$ for n sufficiently large.

Proof: The proof is quite analogous to the corresponding theorem for discrete memoryless sources (Theorem 3.4) and is left as an exercise. □

Definition 5.27 (Discrete-time continuous memoryless channels) Consider a discrete-time channel with continuous input and output alphabets given by $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{Y} \subseteq \mathbb{R}$, respectively, and described by a sequence of n-dimensional transition (conditional) pdfs $\{f_{Y^n|X^n}(y^n|x^n)\}_{n=1}^{\infty}$ that govern the reception of $y^n = (y_1, y_2, \cdots, y_n) \in \mathcal{Y}^n$ at the channel output when $x^n = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ is sent as the channel input.

The channel (without feedback) is said to be memoryless with a given (marginal) transition pdf $f_{Y|X}$ if its sequence of transition pdfs $f_{Y^n|X^n}$ satisfies

$$f_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} f_{Y|X}(y_i|x_i) \tag{5.4.1}$$

for every $n = 1, 2, \cdots$, $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$.

- *Average cost constraint* $(t(\cdot), P)$ on any input n-tuple $x^n = (x_1, x_2, \cdots, x_n)$ transmitted over the channel by requiring that

$$\frac{1}{n}\sum_{i=1}^{n} t(x_i) \leq P, \tag{5.4.2}$$

  where $t(\cdot)$ is a given non-negative real-valued function describing the cost for transmitting an input symbol, and P is a given positive number representing the maximal average amount of available resources per input symbol.

Definition 5.28 The capacity (or capacity-cost function) of a discrete-time continuous memoryless channel with input average cost constraint (t, P) is denoted by $C(P)$ and defined as

$$C(P) := \sup_{F_X : E[t(X)] \leq P} I(X;Y) \qquad \text{(in bits/channel use)} \tag{5.4.3}$$

where the supremum is over all input distributions F_X.

Property A.4 (Properties of the supremum)

3. If $-\infty < \sup \mathcal{A} < \infty$, then $(\forall\ \varepsilon > 0)(\exists\ a_0 \in \mathcal{A})\ a_0 > \sup \mathcal{A} - \varepsilon$.
   (The existence of $a_0 \in (\sup \mathcal{A} - \varepsilon, \sup \mathcal{A}]$ for any $\varepsilon > 0$ under the condition of $|\sup \mathcal{A}| < \infty$ is called the *approximation property for the supremum*.)

Lemma 5.29 (Concavity of capacity) If $C(P)$ as defined in (5.4.3) is finite for any $P > 0$, then it is concave, continuous and strictly increasing in P.

Proof: Fix $P_1 > 0$ and $P_2 > 0$. Then since $C(P)$ is finite for any $P > 0$, then by the 3rd property in Property A.4, there exist two input distributions F_{X_1} and F_{X_2} such that for all $\epsilon > 0$,

$$I(X_i; Y_i) \geq C(P_i) - \epsilon \tag{5.4.4}$$

and

$$E[t(X_i)] \leq P_i \tag{5.4.5}$$

where X_i denotes the input with distribution F_{X_i} and Y_i is the corresponding channel output for $i = 1, 2$.

Now, for $0 \leq \lambda \leq 1$, let X_λ be a random variable with distribution

$$F_{X_\lambda} := \lambda F_{X_1} + (1-\lambda) F_{X_2}.$$

Then by (5.4.5)

$$E_{X_\lambda}[t(X)] = \lambda E_{X_1}[t(X)] + (1-\lambda) E_{X_2}[t(X)] \leq \lambda P_1 + (1-\lambda) P_2. \qquad (5.4.6)$$

Furthermore,

$$\begin{aligned} C(\lambda P_1 + (1-\lambda)P_2) &= \sup_{F_X \,:\, E[t(X)] \leq \lambda P_1 + (1-\lambda)P_2} I(F_X, f_{Y|X}) \\ &\geq I(F_{X_\lambda}, f_{Y|X}) \\ &\geq \lambda I(F_{X_1}, f_{Y|X}) + (1-\lambda) I(F_{X_2}, f_{Y|X}) \\ &= \lambda I(X_1; Y_1) + (1-\lambda) I(X_2 : Y_2) \\ &\geq \lambda C(P_1) - \epsilon + (1-\lambda) C(P_2) - \epsilon, \end{aligned}$$

where the first inequality holds by (5.4.6), the second inequality follows from the concavity of the mutual information with respect to its first argument (cf. Lemma 2.46) and the third inequality follows from (5.4.4).

Letting $\epsilon \to 0$ yields that

$$C(\lambda P_1 + (1-\lambda)P_2) \geq \lambda C(P_1) + (1-\lambda)C(P_2)$$

and hence $C(P)$ is concave in P.

Finally, it can directly be seen by definition that $C(\cdot)$ is non-decreasing, which, together with its concavity, imply that it is continuous and strictly increasing. □

- The most commonly used cost function is the power cost function,

$$t(x) = x^2,$$

resulting in the *average power constraint* P for each transmitted input n-tuple:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P. \tag{5.4.7}$$

- For better understanding, we only focus on the **discrete-time memoryless (additive white) Gaussian (noise) channel** with **average input power constraint** P:

$$Y_i = X_i + Z_i, \quad \text{for } i = 1, 2, \cdots, \tag{5.4.8}$$

where Y_i, X_i and Z_i are the channel output, input and noise at time i, $\{Z_i\}_{i=1}^{\infty}$ i.i.d. Gaussian with mean zero and variance σ^2, and $\{X_i\}_{i=1}^{\infty} \perp \{Z_i\}_{i=1}^{\infty}$.

Derivation of the **Capacity** $C(P)$ for the **discrete-time memoryless (additive white) Gaussian (noise) channel** with **average input power constraint** P.

- For Gaussian distributed Z, we obtain

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X+Z|X) && (5.4.9) \\ &= h(Y) - h(Z|X) && (5.4.10) \\ &= h(Y) - h(Z) && (5.4.11) \\ &= h(Y) - \frac{1}{2}\log_2\left(2\pi e\sigma^2\right), && (5.4.12) \end{aligned}$$

  where (5.4.9) follows from (5.4.8), (5.4.10) holds since differential entropy is invariant under translation (see Property 8 of Lemma 5.14), (5.4.11) follows from the independence of X and Z, and (5.4.12) holds since $Z \sim \mathcal{N}(0,\sigma^2)$ is Gaussian (see (5.1.1)).

- Now since $Y = X + Z$, we have that

$$E[Y^2] = E[X^2] + E[Z^2] + 2E[X]E[Z] = E[X^2] + \sigma^2 + 2E[X](0) \leq P + \sigma^2$$

  since the input in (5.4.3) is constrained to satisfy $E[X^2] \leq P$.

- Thus the variance of Y satisfies

$$\mathrm{Var}(Y) \le E[Y^2] \le P + \sigma^2,$$

and

$$h(Y) \le \frac{1}{2}\log_2\left(2\pi e \mathrm{Var}(Y)\right) \le \frac{1}{2}\log_2\left(2\pi e(P+\sigma^2)\right)$$

where the first inequality follows by Theorem 5.20 since Y is real-valued (with support $\mathbb{R}$).

- Noting that equality holds in the first inequality above iff Y is Gaussian and in the second inequality iff $\mathrm{Var}(Y) = P + \sigma^2$, we obtain that choosing the input X as $X \sim \mathcal{N}(0, P)$ yields $Y \sim \mathcal{N}(0, P+\sigma^2)$ and hence maximizes $I(X;Y)$ over all inputs satisfying $E[X^2] \le P$.

- Thus, the capacity of the discrete-time memoryless Gaussian channel with input average power constraint P and noise variance (or power) σ^2 is given by

$$\begin{aligned} C(P) &= \frac{1}{2}\log_2\left(2\pi e(P+\sigma^2)\right) - \frac{1}{2}\log_2\left(2\pi e\sigma^2\right) \\ &= \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right). \end{aligned} \tag{5.4.13}$$

Definition 5.31 (Fixed-length data transmission code) Given positive integers n and M, and a discrete-time memoryless Gaussian channel with input average power constraint P, a fixed-length data transmission code (or block code) $\mathscr{C}_n = (n, M)$ for this channel with blocklength n and rate $\frac{1}{n}\log_2 M$ message bits per channel symbol (or channel use) consists of:

1. M information messages intended for transmission.
2. An encoding function
$$f : \{1, 2, \ldots, M\} \to \mathbb{R}^n$$
yielding real-valued codewords $\boldsymbol{c}_1 = f(1), \boldsymbol{c}_2 = f(2), \cdots, \boldsymbol{c}_M = f(M)$, where each codeword $\boldsymbol{c}_m = (c_{m1}, \ldots, c_{mn})$ is of length n and satisfies the power constraint P
$$\frac{1}{n}\sum_{i=1}^{n} c_i^2 \leq P,$$
for $m = 1, 2, \cdots, M$.
3. A decoding function $g : \mathbb{R}^n \to \{1, 2, \ldots, M\}$.

Theorem 5.32 (Shannon's coding theorem for the memoryless Gaussian channel) Consider a discrete-time memoryless Gaussian channel with input average power constraint P, channel noise variance σ^2 and capacity $C(P)$ as given by (5.4.13).

- *Forward part (achievability):* For any $\varepsilon \in (0,1)$, there exist $0 < \gamma < 2\varepsilon$ and a sequence of data transmission block code $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$ satisfying

$$\frac{1}{n}\log_2 M_n > C(P) - \gamma$$

with each codeword $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$ in $\mathscr{C}_n$ satisfying

$$\frac{1}{n}\sum_{i=1}^{n} c_i^2 \leq P \tag{5.4.14}$$

such that the probability of error $P_e(\mathscr{C}_n) < \varepsilon$ for sufficiently large n.

- *Converse part:* If for any sequence of data transmission block codes $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$ whose codewords satisfy (5.4.14), we have that

$$\liminf_{n\to\infty} \frac{1}{n}\log_2 M_n > C(P),$$

then the codes' probability of error $P_e(\mathscr{C}_n)$ is bounded away from zero for all n sufficiently large.

Proof of the forward part: The theorem holds trivially when $C(P) = 0$ because we can choose $M_n = 1$ for every n and have $P_e(\mathscr{C}_n) = 0$. Hence, we assume without loss of generality $C(P) > 0$.

Step 0:

- Take a positive γ satisfying $\gamma < \min\{2\varepsilon, C(P)\}$.
- Pick $\xi > 0$ small enough such that $2[C(P) - C(P - \xi)] < \gamma$, where the existence of such ξ is assured by the strictly increasing property of $C(P)$. Hence, we have
$$C(P - \xi) - \frac{\gamma}{2} > C(P) - \gamma > 0.$$
- Choose M_n to satisfy
$$C(P - \xi) - \frac{\gamma}{2} > \frac{1}{n} \log_2 M_n > C(P) - \gamma,$$
for which the choice should exist for all sufficiently large n.
- Take $\delta = \gamma/8$.
- Let F_X be the distribution that achieves $C(P - \xi)$, where $C(P)$ is given by (5.4.13). In this case, F_X is the Gaussian distribution with mean zero and variance $P - \xi$ and admits a pdf f_X. Hence, $E[X^2] \leq P - \xi$ and $I(X;Y) = C(P - \xi)$.

Step 1: Random coding with average power constraint.

Randomly draw M_n codewords according to pdf f_{X^n} with

$$f_{X^n}(x^n) = \prod_{i=1}^{n} f_X(x_i).$$

By law of large numbers, each randomly selected codeword

$$\boldsymbol{c}_m = (c_{m1}, \ldots, c_{mn})$$

satisfies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} c_{mi}^2 = E[X^2] \leq P - \xi$$

for $m = 1, 2, \ldots, M_n$.

Step 2: Code construction.

- For M_n selected codewords $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{M_n}\}$, replace the codewords that violate the power constraint (i.e., (5.4.14)) by an all-zero codeword $\mathbf{0}$.
- Define the encoder as

$$f_n(m) = \boldsymbol{c}_m \quad \text{for} \quad 1 \le m \le M_n.$$

- Given a received output sequence y^n, the decoder $g_n(\cdot)$ is given by

$$g_n(y^n) = \begin{cases} m, & \text{if } (\boldsymbol{c}_m, y^n) \in \mathcal{F}_n(\delta) \\ & \quad \text{and } (\forall\, m' \neq m)\ (\boldsymbol{c}_{m'}, y^n) \notin \mathcal{F}_n(\delta), \\ \text{arbitrary}, & \text{otherwise}, \end{cases}$$

where the set

$$\mathcal{F}_n(\delta) := \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| -\frac{1}{n} \log_2 f_{X^n Y^n}(x^n, y^n) - h(X,Y) \right| < \delta, \right.$$

$$\left| -\frac{1}{n} \log_2 f_{X^n}(x^n) - h(X) \right| < \delta,$$

$$\left. \text{and } \left| -\frac{1}{n} \log_2 f_{Y^n}(y^n) - h(Y) \right| < \delta \right\}.$$

Note that $\mathcal{F}_n(\delta)$ is generated by

$$f_{X^nY^n}(x^n, y^n) = \prod_{i=1}^{n} f_{XY}(x_i, y_i)$$

where $f_{X^nY^n}(x^n, y^n)$ is the joint input-output pdf realized when the memoryless Gaussian channel (with n-fold transition pdf

$$f_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} f_{Y|X}(y_i|x_i))$$

is driven by input X^n with pdf

$$f_{X^n}(x^n) = \prod_{i=1}^{n} f_X(x_i)$$

(where f_X achieves $C(P - \xi)$).

Step 3: Conditional probability of error.

- Let $\lambda_m = \lambda_m(\mathscr{C}_n)$ denote the conditional error probability given codeword m is transmitted (with respect to code $\mathscr{C}_n$).
- Define

$$\mathcal{E}_0 := \left\{ x^n \in \mathcal{X}^n : \frac{1}{n} \sum_{i=1}^{n} x_i^2 > P \right\}.$$

- Then

$$\lambda_m(\mathscr{C}_n) \leq \int_{y^n \notin \mathcal{F}_n(\delta|\boldsymbol{c}_m)} f_{Y^n|X^n}(y^n|\boldsymbol{c}_m)\, dy^n + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \int_{y^n \in \mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} f_{Y^n|X^n}(y^n|\boldsymbol{c}_m)\, dy^n,$$

where

$$\mathcal{F}_n(\delta|x^n) := \{ y^n \in \mathcal{Y}^n \ : \ (x^n, y^n) \in \mathcal{F}_n(\delta) \}.$$

- By taking expectation with respect to the m^{th} codeword-selecting distribution $f_{X^n}(\boldsymbol{c}_m)$, we obtain

$$\begin{aligned}
E[\lambda_m] &= \int_{\boldsymbol{c}_m \in \mathcal{X}^n} f_{X^n}(\boldsymbol{c}_m)\lambda_m(\mathcal{C}_n)\, d\boldsymbol{c}_m \\
&= \int_{\boldsymbol{c}_m \in \mathcal{X}^n \cap \mathcal{E}_0} f_{X^n}(\boldsymbol{c}_m)\lambda_m(\mathcal{C}_n)\, d\boldsymbol{c}_m + \int_{\boldsymbol{c}_m \in \mathcal{X}^n \cap \mathcal{E}_0^c} f_{X^n}(\boldsymbol{c}_m)\lambda_m(\mathcal{C}_n)\, d\boldsymbol{c}_m \\
&\le \int_{\boldsymbol{c}_m \in \mathcal{E}_0} f_{X^n}(\boldsymbol{c}_m)\, d\boldsymbol{c}_m + \int_{\boldsymbol{c}_m \in \mathcal{X}^n} f_{X^n}(\boldsymbol{c}_m)\lambda_m(\mathcal{C}_n)\, d\boldsymbol{c}_m \\
&\le P_{X^n}(\mathcal{E}_0) + \int_{\boldsymbol{c}_m \in \mathcal{X}^n} \int_{y^n \notin \mathcal{F}_n(\delta|\boldsymbol{c}_m)} f_{X^n}(\boldsymbol{c}_m) f_{Y^n|X^n}(y^n|\boldsymbol{c}_m)\, dy^n d\boldsymbol{c}_m \\
&\quad + \int_{\boldsymbol{c}_m \in \mathcal{X}^n} \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \int_{y^n \in \mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} f_{X^n}(\boldsymbol{c}_m) f_{Y^n|X^n}(y^n|\boldsymbol{c}_m)\, dy^n d\boldsymbol{c}_m. \\
&= P_{X^n}(\mathcal{E}_0) + P_{X^n,Y^n}\left(\mathcal{F}_n^c(\delta)\right) \\
&\quad + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \int_{\boldsymbol{c}_m \in \mathcal{X}^n} \int_{y^n \in \mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} f_{X^n,Y^n}(\boldsymbol{c}_m, y^n)\, dy^n d\boldsymbol{c}_m. \qquad (5.4.15)
\end{aligned}$$

- Note that the additional term $P_{X^n}(\mathcal{E}_0)$ in (5.4.15) is to cope with the errors due to all-zero codeword replacement, which will be less than δ for all sufficiently large n by the law of large numbers.
- Finally, by carrying out a similar procedure as in the proof of the channel coding theorem for discrete channels (cf. Theorem 4.11), we obtain:

$$\begin{aligned} E[P_e(\boldsymbol{\mathcal{C}}_n)] &\leq P_{X^n}(\mathcal{E}_0) + P_{X^n,Y^n}\left(\mathcal{F}_n^c(\delta)\right) \\ &\quad + M_n \cdot 2^{n(h(X,Y)+\delta)} 2^{-n(h(X)-\delta)} 2^{-n(h(Y)-\delta)} \\ &\leq P_{X^n}(\mathcal{E}_0) + P_{X^n,Y^n}\left(\mathcal{F}_n^c(\delta)\right) + 2^{n(C(P-\xi)-4\delta)} \cdot 2^{-n(I(X;Y)-3\delta)} \\ &= P_{X^n}(\mathcal{E}_0) + P_{X^n,Y^n}\left(\mathcal{F}_n^c(\delta)\right) + 2^{-n\delta}. \end{aligned}$$

Accordingly, we can make the average probability of error, $E[P_e(\boldsymbol{\mathcal{C}}_n)]$, less than $3\delta = 3\gamma/8 < 3\varepsilon/4 < \varepsilon$ for all sufficiently large n. $\square$

Proof of the converse part: Consider an (n, M_n) block data transmission code satisfying the power constraint

$$\frac{1}{n}\sum_{i=1}^{n} c_i^2 \leq P$$

with encoding function

$$f_n : \{1, 2, \ldots, M_n\} \to \mathcal{X}^n$$

and decoding function

$$g_n : \mathcal{Y}^n \to \{1, 2, \ldots, M_n\}.$$

- Since the message W is uniformly distributed over $\{1, 2, \ldots, M_n\}$, we have

$$H(W) = \log_2 M_n.$$

- Since $W \to X^n = f_n(W) \to Y^n$ form a Markov chain (as Y^n only depends on X^n), we obtain by the data processing lemma that

$$I(W; Y^n) \leq I(X^n; Y^n).$$

- We can also bound $I(X^n;Y^n)$ by $C(P)$ as follows:

$$
\begin{aligned}
I(X^n;Y^n) &\le \sup_{F_{X^n}\,:\,(1/n)\sum_{i=1}^n E[X_i^2]\le P} I(X^n;Y^n) \\
&\le \sup_{F_{X^n}\,:\,(1/n)\sum_{i=1}^n E[X_i^2]\le P} \sum_{j=1}^n I(X_j;Y_j) \quad \text{(by Theorem 2.21)} \\
&= \sup_{(P_1,P_2,\ldots,P_n)\,:\,(1/n)\sum_{i=1}^n P_i=P} \; \sup_{F_{X^n}\,:\,(\forall\ i)\ E[X_i^2]\le P_i} \sum_{j=1}^n I(X_j;Y_j) \\
&\le \sup_{(P_1,P_2,\ldots,P_n)\,:\,(1/n)\sum_{i=1}^n P_i=P} \sum_{j=1}^n \; \sup_{F_{X^n}\,:\,(\forall\ i)\ E[X_i^2]\le P_i} I(X_j;Y_j) \\
&= \sup_{(P_1,P_2,\ldots,P_n)\,:\,(1/n)\sum_{i=1}^n P_i=P} \sum_{j=1}^n \; \sup_{F_{X_j}\,:\,E[X_j^2]\le P_j} I(X_j;Y_j) \\
&= \sup_{(P_1,P_2,\ldots,P_n):(1/n)\sum_{i=1}^n P_i=P} \sum_{j=1}^n C(P_j) = \sup_{(P_1,P_2,\ldots,P_n):(1/n)\sum_{i=1}^n P_i=P} n\sum_{j=1}^n \frac{1}{n} C(P_j) \\
&\le \sup_{(P_1,P_2,\ldots,P_n):(1/n)\sum_{i=1}^n P_i=P} nC\left(\frac{1}{n}\sum_{j=1}^n P_j\right) \quad \text{(by concavity of } C(P)) \\
&= nC(P).
\end{aligned}
$$

- Consequently, recalling that $P_e(\mathscr{C}_n)$ is the average error probability incurred by guessing W from observing Y^n via the decoding function $g_n : \mathcal{Y}^n \to \{1, 2, \ldots, M_n\}$, we get

$$\begin{aligned}
\log_2 M_n &= H(W) \\
&= H(W|Y^n) + I(W; Y^n) \\
&\le H(W|Y^n) + I(X^n; Y^n) \\
&\le \underbrace{h_b(P_e(\mathscr{C}_n)) + P_e(\mathscr{C}_n) \cdot \log_2(|\mathcal{W}| - 1)}_{\text{Fano's inequality}} + nC(P) \\
&\le 1 + P_e(\mathscr{C}_n) \cdot \log_2(M_n - 1) + nC(P), \\
&\quad \text{(by the fact that } (\forall\, t \in [0,1])\ h_b(t) \le 1) \\
&< 1 + P_e(\mathscr{C}_n) \cdot \log_2 M_n + nC(P),
\end{aligned}$$

  which implies that

$$P_e(\mathscr{C}_n) > 1 - \frac{C(P)}{(1/n)\log_2 M_n} - \frac{1}{\log_2 M_n} = 1 - \frac{C(P) + 1/n}{(1/n)\log_2 M_n}.$$

5.4 Capacity for discrete memoryless Gaussian chan

- So as identical to the converse proof of Theorem 4.11, we obtain if

$$\liminf_{n\to\infty} \frac{1}{n} \log_2 M_n - C(P) = 1 - \mu > 0,$$

then for any $0 < \epsilon < 1$,

$$P_e(\mathscr{C}_n) \geq (1-\epsilon)\mu \quad \text{for } n \text{ sufficiently large.}$$

□

Theorem 5.33 (Gaussian noise minimizes capacity of additive-noise channels) Every discrete-time continuous memoryless channel with additive noise (admitting a pdf) of mean zero and variance σ^2 and input average power constraint P has its capacity $C(P)$ lower bounded by the capacity of the memoryless Gaussian channel with identical input constraint and noise variance:

$$C(P) \geq \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right).$$

Proof:

- Let $f_{Y|X}$ and $f_{Y_g|X_g}$ denote the transition pdfs of the additive-noise channel and the Gaussian channel, respectively, where both channels satisfy input average power constraint P.
- Let Z and Z_g respectively denote their zero-mean noise variables of identical variance σ^2.

- Then we have

$$
\begin{aligned}
& I(f_{X_g}, f_{Y|X}) - I(f_{X_g}, f_{Y_g|X_g}) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{X_g}(x) f_Z(y-x) \log_2 \frac{f_Z(y-x)}{f_Y(y)} dy dx \\
&\quad - \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{X_g}(x) f_{Z_g}(y-x) \log_2 \frac{f_{Z_g}(y-x)}{f_{Y_g}(y)} dy dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{X_g}(x) f_Z(y-x) \log_2 \frac{f_Z(y-x)}{f_Y(y)} dy dx \\
&\quad - \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{X_g}(x) f_Z(y-x) \log_2 \frac{f_{Z_g}(y-x)}{f_{Y_g}(y)} dy dx \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{X_g}(x) f_Z(y-x) \log_2 \frac{f_Z(y-x) f_{Y_g}(y)}{f_{Z_g}(y-x) f_Y(y)} dy dx \\
&\geq \int_{\mathcal{X}} \int_{\mathcal{Y}} f_{X_g}(x) f_Z(y-x) (\log_2 e) \left(1 - \frac{f_{Z_g}(y-x) f_Y(y)}{f_Z(y-x) f_{Y_g}(y)}\right) dy dx \\
&= (\log_2 e) \left[1 - \int_{\mathcal{Y}} \frac{f_Y(y)}{f_{Y_g}(y)} \left(\int_{\mathcal{X}} f_{X_g}(x) f_{Z_g}(y-x) dx\right) dy\right] = 0,
\end{aligned}
$$

with equality holding in the inequality iff $f_Y(y)/f_{Y_g}(y) = f_Z(y-x)/f_{Z_g}(y-x)$ for all x.

- Therefore,

$$\begin{aligned}
\frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) &= \sup_{F_X\,:\,E[X^2]\le P} I(F_X, f_{Y_g|X_g}) \\
&= I(f^*_{X_g}, f_{Y_g|X_g}) \\
&\le I(f^*_{X_g}, f_{Y|X}) \\
&\le \sup_{F_X\,:\,E[X^2]\le P} I(F_X, f_{Y|X}) \\
&= C(P).
\end{aligned}$$

□

Capacity of the memoryless fading channel

We further examine the capacity of the memoryless (unity-power) fading channel, which is widely used to model wireless communications channels:

$$Y_i = A_i X_i + Z_i, \qquad \text{for } i = 1, 2, \ldots, \tag{5.4.16}$$

1. *With only decoder side information (DSI):* In this case, as both A and Y are known at the receiver, we can consider (Y, A) as the channel's output and thus aim to maximize

$$I(X; A, Y) = I(X; A) + I(X; Y|A) = I(X; Y|A)$$

where $I(X; A) = 0$ since X and A are independent from each other.

Thus

$$\begin{aligned} C_{DSI}(P) &= \sup_{F_X:\, E[X^2] \le P} I(X; Y|A) \\ &= \sup_{F_X:\, E[X^2] \le P} [h(Y|A) - h(Y|X, A)] \\ &= E_A \left[\frac{1}{2} \log_2 \left(1 + \frac{A^2 P}{\sigma^2} \right) \right] \end{aligned} \tag{5.4.17}$$

where the expectation is taken with respect to the fading distribution. Note that the capacity achieving distribution here is also Gaussian with mean zero and variance P and is independent of the fading coefficient.

Capacity of the memoryless fading channel

In light of the concavity of the logarithm and using Jensen's inequality, we readily obtain that

$$\begin{aligned} C_{DSI}(P) &= E_A\left[\frac{1}{2}\log_2\left(1+\frac{A^2P}{\sigma^2}\right)\right] \\ &\leq \frac{1}{2}\log_2\left(1+\frac{E[A^2]P}{\sigma^2}\right) \\ &= \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) := C_G(P) \end{aligned} \tag{5.4.18}$$

which is the capacity of the AWGN channel with identical SNR, and where the last step follows since $E[A^2]=1$.

Thus we conclude that fading degrades capacity as

$$C_{DSI}(P) \leq C_G(P).$$

2. *With full side information (FSI):* In this case, the transmitter can adaptively adjust its input power according to the value of the fading coefficient. It can be shown using Lagrange multipliers that the capacity in this case is given by

$$\begin{aligned} C_{FSI}(P) &= E_A\left[\sup_{p(\cdot):\, E_A[p(A)]=P} \frac{1}{2}\log_2\left(1+\frac{A^2p(A)}{\sigma^2}\right)\right] \\ &= E_A\left[\frac{1}{2}\log_2\left(1+\frac{A^2p^*(A)}{\sigma^2}\right)\right] \end{aligned} \tag{5.4.19}$$

where

$$p^*(a) = \max\left(0, \frac{1}{\lambda} - \frac{\sigma^2}{a^2}\right)$$

and λ satisfies

$$E_A[p(A)] = P.$$

The optimal power allotment $p^*(A)$ above is a so-called *water-filling* allotment, which we examine in more detail in the next section)in the case of parallel AWGN channels).

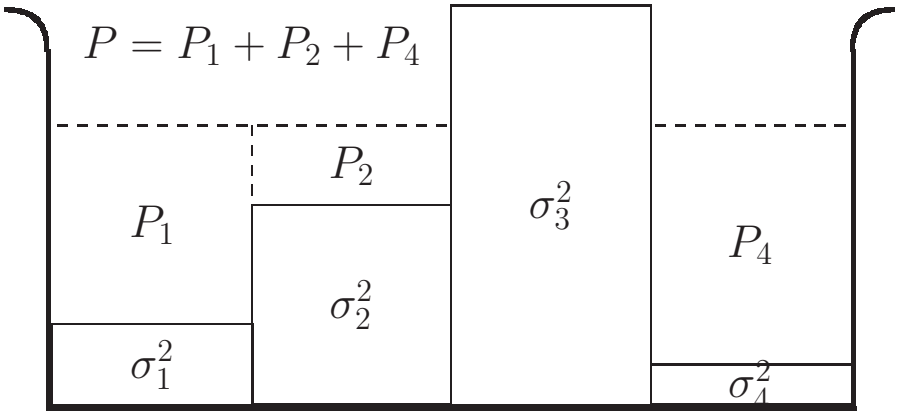
Theorem 5.36 (Capacity of uncorrelated parallel Gaussian channels) The capacity of k uncorrelated parallel Gaussian channels under an overall input power constraint P is given by

$$C(P) = \sum_{i=1}^{k} \frac{1}{2} \log_2 \left(1 + \frac{P_i}{\sigma_i^2}\right),$$

where σ_i^2 is the noise variance of channel i,

$$P_i = \max\{0, \theta - \sigma_i^2\},$$

and θ is chosen to satisfy $\sum_{i=1}^{k} P_i = P$. This capacity is achieved by a tuple of independent Gaussian inputs $(X_1, X_2, \cdots, X_k)$, where $X_i \sim \mathcal{N}(0, P_i)$ is the input to channel i, for $i = 1, 2, \cdots, k$.



Proof:

- By definition,

$$C(P) = \sup_{F_{X^k} \,:\, \sum_{i=1}^{k} E[X_k^2] \le P} I(X^k; Y^k).$$

- Since the noise random variables $Z_1, \ldots, Z_k$ are independent from each other,

$$\begin{aligned}
I(X^k; Y^k) &= h(Y^k) - h(Y^k|X^k) \\
&= h(Y^k) - h(Z^k + X^k|X^k) \\
&= h(Y^k) - h(Z^k|X^k) \\
&= h(Y^k) - h(Z^k) \\
&= h(Y^k) - \sum_{i=1}^{k} h(Z_i) \\
&\le \sum_{i=1}^{k} h(Y_i) - \sum_{i=1}^{k} h(Z_i) \\
&\le \sum_{i=1}^{k} \frac{1}{2} \log_2 \left(1 + \frac{P_i}{\sigma_i^2}\right).
\end{aligned}$$

- Equalities hold above if all the X_i inputs are independent of each other with each input $X_i \sim \mathcal{N}(0, P_i)$ such that $\sum_{i=1}^{k} P_i = P$.

- Thus the problem is reduced to finding the power allotment that maximizes the overall capacity subject to the equality constraint $\sum_{i=1}^{k} P_i = P$ and inequality constraints $P_i \geq 0$, $i = 1, \ldots, k$.

- By using the Lagrange multipliers technique and verifying the KKT condition (see Example B.21 in Appendix B.8), the maximizer $(P_1, \ldots, P_k)$ of

$$\max \left\{ \sum_{i=1}^{k} \frac{1}{2} \log_2 \left(1 + \frac{P_i}{\sigma_i^2} \right) + \sum_{i=1}^{k} \lambda_i P_i - \nu \left(\sum_{i=1}^{k} P_i - P \right) \right\}$$

can be found by taking the derivative of the above equation (with respect to P_i) and setting it to zero, which yields

$$\lambda_i = \begin{cases} -\dfrac{1}{2\ln(2)} \dfrac{1}{P_i + \sigma_i^2} + \nu = 0, & \text{if } P_i > 0; \\ -\dfrac{1}{2\ln(2)} \dfrac{1}{P_i + \sigma_i^2} + \nu \geq 0, & \text{if } P_i = 0. \end{cases}$$

Hence,

$$\begin{cases} P_i = \theta - \sigma_i^2, & \text{if } P_i > 0; \\ P_i \geq \theta - \sigma_i^2, & \text{if } P_i = 0, \end{cases} \quad \text{(equivalently, } P_i = \max\{0, \theta - \sigma_i^2\}),$$

where $\theta := \log_2 e/(2\nu)$ is chosen to satisfy $\sum_{i=1}^{k} P_i = P$. □

Observation 5.37

- According to the water-filling principle, one needs to use capacity-achieving Gaussian inputs and allocate more power to less noisy channels for the optimization of channel capacity. However, Gaussian inputs do not fit digital communication systems in practice.
- One may then wonder what is the optimal power allocation scheme when the channel inputs are practically dictated to be discrete in value, such as inputs used in conjunction with binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK), or 16 quadrature- amplitude modulation (16-QAM) signaling.
- Surprisingly under certain conditions, the answer is different from the water-filling principle.
- The optimal power allocation for parallel AWGN channels with inputs constrained to be discrete is established in 2006 (Lozano, Tulino & Verdú), resulting in a new graphical power allocation interpretation called the mercury/water-filling principle.

- Furthermore, it was found in 2012 (Wang, Chen & Wang) that when the channel's additive noise is no longer Gaussian, the mercury adjustment fails to interpret the optimal power allocation scheme and a new two-phase water-filling principle was observed.

5.6 Capacity of correlated parallel Gaussian channels

Theorem 5.38 (Capacity of correlated parallel Gaussian channels) The capacity of k correlated parallel Gaussian channels with positive-definite noise covariance matrix $\mathbf{K}_Z$ under overall input power constraint P is given by

$$C(P) = \sum_{i=1}^{k} \frac{1}{2} \log_2 \left(1 + \frac{P_i}{\lambda_i}\right),$$

where λ_i is the i-th eigenvalue of $\mathbf{K}_Z$,

$$P_i = \max\{0, \theta - \lambda_i\},$$

and θ is chosen to satisfy $\sum_{i=1}^{k} P_i = P$. This capacity is achieved by a tuple of zero-mean Gaussian inputs $(X_1, X_2, \cdots, X_k)$ with covariance matrix $\mathbf{K}_X$ having the same eigenvectors as $\mathbf{K}_Z$, where the i-th eigenvalue of $\mathbf{K}_X$ is P_i, for $i = 1, 2, \cdots, k$.

Proof:

- In correlated parallel Gaussian channels, the input power constraint becomes

$$\sum_{i=1}^{k} E[X_i^2] = \mathrm{tr}(\mathbf{K}_X) \leq P,$$

  where $\mathrm{tr}(\cdot)$ denotes the trace of the $k \times k$ matrix $\mathbf{K}_X$.

5.6 Capacity of correlated parallel Gaussian channels

- Since in each channel, the input and noise variables are independent from each other, we have

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

- Since $h(Z^k)$ is not determined by the input, determining the system's capacity reduces to maximizing $h(Y^k)$ over all possible inputs $(X_1, \ldots, X_k)$ satisfying the power constraint.

- Now observe that the covariance matrix of Y^k is equal to

$$\mathbf{K}_Y = \mathbf{K}_X + \mathbf{K}_Z,$$

which implies by Theorem 5.20 that the differential entropy of Y^k is upper bounded by

$$h(Y^k) \leq \frac{1}{2}\log_2\left[(2\pi e)^k \det(\mathbf{K}_X + \mathbf{K}_Z)\right],$$

with equality iff Y^k Gaussian. It remains to find out whether we can find inputs $(X_1, \ldots, X_k)$ satisfying the power constraint which achieve the above upper bound and maximize it.

- As in the proof of Theorem 5.18, we can orthogonally diagonalize $\mathbf{K}_Z$ as

$$\mathbf{K}_Z = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T,$$

where $\mathbf{A}\mathbf{A}^T = \mathbf{I}_k$ (and thus $\det(\mathbf{A})^2 = 1$), $\mathbf{I}_k$ is the $k \times k$ identity matrix, and $\mathbf{\Lambda}$ is a diagonal matrix with positive diagonal components consisting of the eigenvalues of $\mathbf{K}_Z$ (as $\mathbf{K}_Z$ is positive definite). Then

$$\begin{aligned}
\det(\mathbf{K}_X + \mathbf{K}_Z) &= \det(\mathbf{K}_X + \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T) \\
&= \det(\mathbf{A}\mathbf{A}^T\mathbf{K}_X\mathbf{A}\mathbf{A}^T + \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T) \\
&= \det(\mathbf{A}) \cdot \det(\mathbf{A}^T\mathbf{K}_X\mathbf{A} + \mathbf{\Lambda}) \cdot \det(\mathbf{A}^T) \\
&= \det(\mathbf{A}^T\mathbf{K}_X\mathbf{A} + \mathbf{\Lambda}) \\
&= \det(\mathbf{B} + \mathbf{\Lambda}),
\end{aligned}$$

where $\mathbf{B}{:=}\mathbf{A}^T\mathbf{K}_X\mathbf{A}$.

- Since for any two matrices $\mathbf{C}$ and $\mathbf{D}$,

$$\mathrm{tr}(\mathbf{C}\mathbf{D}) = \mathrm{tr}(\mathbf{D}\mathbf{C}),$$

we have that

$$\mathrm{tr}(\mathbf{B}) = \mathrm{tr}(\mathbf{A}^T\mathbf{K}_X\mathbf{A}) = \mathrm{tr}(\mathbf{A}\mathbf{A}^T\mathbf{K}_X) = \mathrm{tr}(\mathbf{I}_k\mathbf{K}_X) = \mathrm{tr}(\mathbf{K}_X).$$

- Thus the capacity problem is further transformed

$$\text{to maximizing } \det(\mathbf{B}+\mathbf{\Lambda}) \text{ subject to } \mathrm{tr}(\mathbf{B}) \le P.$$

- By observing that $\mathbf{B}+\mathbf{\Lambda}$ is positive definite (because $\mathbf{\Lambda}$ is positive definite) and using Hadamard's inequality given in Corollary 5.19, we have

$$\det(\mathbf{B}+\mathbf{\Lambda}) \le \prod_{i=1}^{k}(B_{ii}+\lambda_i),$$

  where λ_i is the component of matrix $\mathbf{\Lambda}$ locating at i^{th} row and i^{th} column, which is exactly the i-th eigenvalue of $\mathbf{K}_Z$.

- Thus, the maximum value of $\det(\mathbf{B}+\mathbf{\Lambda})$ under $\mathrm{tr}(\mathbf{B}) \le P$ is realized by a diagonal matrix $\mathbf{B}$ (to achieve equality in Hadamard's inequality) with

$$\sum_{i=1}^{k} B_{ii} = P \text{ and each } B_{ii} \ge 0.$$

Thus,

$$\begin{aligned} C(P) &= \frac{1}{2}\log_2\left[(2\pi e)^k \det(\mathbf{B}+\mathbf{\Lambda})\right] - \frac{1}{2}\log_2\left[(2\pi e)^k \det(\mathbf{\Lambda})\right] \\ &= \sum_{i=1}^{k} \frac{1}{2}\log_2\left(1+\frac{B_{ii}}{\lambda_i}\right). \end{aligned}$$

- Finally, as in the proof of Theorem 5.36, we obtain a water-filling allotment for the optimal diagonal elements of $\mathbf{B}$:

$$B_{ii} = \max\{0, \theta - \lambda_i\},$$

where θ is chosen to satisfy $\sum_{i=1}^{k} B_{ii} = P$. □

5.6 Capacity of correlated parallel Gaussian channels

Observation 5.39 (Capacity of memoryless MIMO channels)

$$\underline{Y}_i = \mathbf{H}_i \underline{X}_i + \underline{Z}_i, \qquad \text{for } i = 1, 2, \ldots, \tag{5.6.1}$$

where $\underline{X}_i$ is the $M \times 1$ transmitted vector, $\underline{Y}_i$ is the $N \times 1$ received vector, and $\underline{Z}_i$ is the $N \times 1$ AWGN vector.

In general, $\underline{Y}_i$, $\mathbf{H}_i$, $\underline{X}_i$ and $\underline{Z}_i$ are complex-valued.

When $\underline{Z} = (Z_1, Z_2, \ldots, Z_N)^T$ is Gaussian with zero-mean and covariance matrix $\mathbf{K}_{\underline{Z}} = \sigma^2 \mathbf{I}_N$, we have

$$f_{\underline{Z}}(\underline{z}) = \begin{cases} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^N Z_i^2\right), & \text{if } \underline{Z} \text{ real-valued} \\ \left(\frac{1}{\pi\sigma^2}\right)^N \exp\left(-\frac{1}{\sigma^2}\sum_{j=1}^N |Z_j|^2\right), & \text{if } \underline{Z} \text{ complex-valued.} \end{cases}$$

Thus, the joint differential entropy for a complex-valued Gaussian $\underline{Z}$ is equal to

$$h(\underline{Z}) = h(Z_1, Z_2, \ldots, Z_N) = \frac{1}{2} \log_2 \left[(2\pi e)^N \det(\mathbf{K}_{\underline{Z}})\right],$$

where the multiplicative factors $1/2$ and 2 in the differential entropy formula in Theorem 5.18 are removed. Accordingly, the multiplicative factor $1/2$ in the capacity formula for real-valued AWGN channels is no longer necessary when a complex-valued AWGN channel is considered.

5.6 Capacity of correlated parallel Gaussian channels

- The noise covariance matrix $\mathbf{K}_{\underline{Z}}$ is often assumed to be given by the identity matrix $\mathbf{I}_N$ (by multiplying the received vector $\underline{Y}_i$ with the whitening matrix $\mathbf{W}$ of $\underline{Z}_i$).

- $$C_{FSI}(P) = E_{\mathbf{H}}\left[\max_{\mathbf{K}_{\underline{X}}:\mathrm{tr}(\mathbf{K}_{\underline{X}})\leq P} \log_2\left(\det(\mathbf{H}\,\mathbf{K}_{\underline{X}}\,\mathbf{H}^{\dagger}+\mathbf{I}_N)\right)\right] \tag{5.6.2}$$

  $$C_{DSI}(P) = \max_{\mathbf{K}_{\underline{X}}:\mathrm{tr}(\mathbf{K}_{\underline{X}})\leq P} E_{\mathbf{H}}\left[\log_2\left(\det(\mathbf{H}\,\mathbf{K}_{\underline{X}}\,\mathbf{H}^{\dagger}+\mathbf{I}_N)\right)\right] \tag{5.6.3}$$

  where "$\dagger$" is the Hermitian (conjugate) transposition operation.

- A key finding is that in virtue of their spatial diversity, such channels can provide significant capacity gains vis-a-vis the traditional single-antenna (with $M = N = 1$) channel.

  – Under Rayleigh $\mathbf{H}$, $C_{DSI}(P)$ scales *linearly* in $\min\{M, N\}$ at high channel SNR values.

5.7 Non-Gaussian discrete-time memoryless channels

- If a discrete-time channel has an additive but non-Gaussian memoryless noise and an input power constraint, then it is often hard to calculate its capacity.
- Hence, we introduce an upper bound and a lower bound on the capacity of such a channel (we assume that the noise admits a pdf).

Definition 5.41 (Entropy power) For a continuous random variable Z with (well-defined) differential entropy $h(Z)$ (measured in bits), its *entropy power* is denoted by Z_e and defined as

$$Z_e := \frac{1}{2\pi e} 2^{2 \cdot h(Z)}.$$

Lemma 5.42 For a discrete-time continuous-alphabet memoryless additive-noise channel with input power constraint P and noise variance σ^2, its capacity satisfies

$$\frac{1}{2} \log_2 \frac{P + \sigma^2}{Z_e} \geq C(P) \geq \frac{1}{2} \log_2 \frac{P + \sigma^2}{\sigma^2}. \tag{5.7.1}$$

Proof: The lower bound in (5.7.1) is already proved in Theorem 5.33. The upper bound follows from

$$I(X;Y) = h(Y) - h(Z) \leq \frac{1}{2} \log_2[2\pi e(P + \sigma^2)] - \frac{1}{2} \log_2[2\pi e Z_e].$$

□

5.7 Non-Gaussian discrete-time memoryless channels

- Whenever two independent Gaussian random variables, Z_1 and Z_2, are added, the power (variance) of the sum is equal to the sum of the powers (variances) of Z_1 and Z_2. This relationship can then be written as

$$2^{2h(Z_1+Z_2)} = 2^{2h(Z_1)} + 2^{2h(Z_2)},$$

  or equivalently

$$\mathrm{Var}(Z_1 + Z_2) = \mathrm{Var}(Z_1) + \mathrm{Var}(Z_2).$$

- **(Entropy-power inequality)** However, when two independent random variables are non-Gaussian, the relationship becomes

$$2^{2h(Z_1+Z_2)} \geq 2^{2h(Z_1)} + 2^{2h(Z_2)}, \tag{5.7.2}$$

  or equivalently

$$Z_e(Z_1 + Z_2) \geq Z_e(Z_1) + Z_e(Z_2). \tag{5.7.3}$$

- It reveals that the sum of two independent random variables may introduce more **(differential) entropy power** than the sum of each individual **entropy power**, except in the Gaussian case.

Observation 5.43 (Capacity bounds in terms of Gaussian capacity and non-Gaussianness)

- It can be readily verified that

$$\frac{1}{2}\log_2\frac{P+\sigma^2}{\mathsf{Z}_e} = \frac{1}{2}\log_2\frac{P+\sigma^2}{\sigma^2} + D(Z\|Z_G)$$

  where $D(Z\|Z_G)$ is the divergence between Z and a Gaussian random variable Z_G of mean zero and variance σ^2.

- Note that $D(Z\|Z_G)$ is called the *non-Gaussianness* of Z (e.g, see Tulino & Verdú 2006) and is a measure of the "non-Gaussianity" of the noise Z.

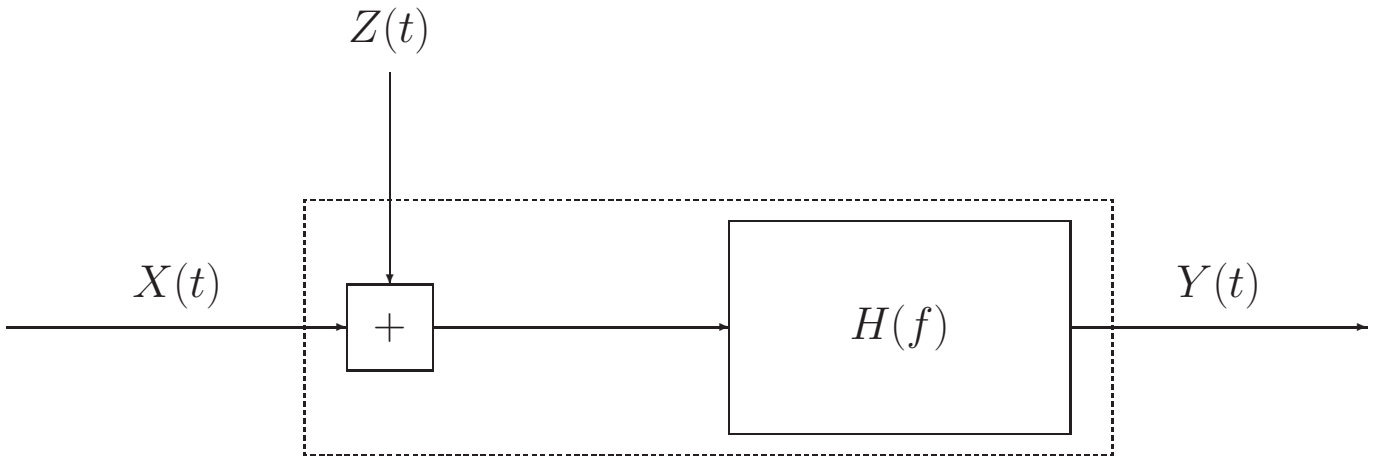
- Thus

$$\frac{1}{2}\log_2\frac{P+\sigma^2}{Z_e} \geq C(P) \geq \frac{1}{2}\log_2\frac{P+\sigma^2}{\sigma^2}$$

  can be written as

$$C_G(P) + D(Z\|Z_G) \geq C(P) \geq C_G(P). \tag{5.7.4}$$

5.8 Capacity of band-limited white Gaussian channel



- The output waveform is given by

$$Y(t) = (X(t) + Z(t)) * h(t), \qquad t \geq 0,$$

where "$*$" represents the convolution operation.

5.8 Capacity of band-limited white Gaussian channel

- (**Time-unlimited**) $X(t)$ is the channel input waveform with average power constraint

$$\lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}E[X^2(t)]dt \le P \tag{5.8.1}$$

and **bandwidth** W cycles per second or Hertz (Hz); i.e., its spectrum or Fourier transform

$$X(f):=\mathcal{F}[X(t)] = \int_{-\infty}^{+\infty}X(t)e^{-j2\pi ft}dt = 0$$

for all frequencies $|f| > W$, where $j = \sqrt{-1}$ is the imaginary unit number.

- $Z(t)$ is the noise waveform of a zero-mean stationary white Gaussian process with power spectral density $N_0/2$; i.e.,

$$\mathrm{PSD}_Z(f) = \mathcal{F}[K_Z(t)] = \int_{-\infty}^{+\infty}K_Z(t)e^{-j2\pi ft}dt = \frac{N_0}{2} \qquad \forall f$$

where $K_Z(\tau):=E[Z(s)Z(s+\tau)]$, $s,\tau\in\mathbb{R}$.

- $h(t)$ is the impulse response of an ideal bandpass filter with cutoff frequencies at $\pm W$ Hz:

$$H(f) = \mathcal{F}[(h(t)] = \begin{cases}1 & \text{if } -W \le f \le W,\\ 0 & \text{otherwise.}\end{cases}$$

5.8 Capacity of band-limited white Gaussian channel

- Note that we can write the channel output as

$$Y(t) = X(t) + \tilde{Z}(t)$$

where $\tilde{Z}(t)$:=$Z(t) * h(t)$ is the filtered noise waveform.

- To determine the capacity (in bits per second) of this continuous-time band-limited white Gaussian channel with parameters, P, W and N_0, we convert it to an **"equivalent" discrete-time channel** with power constraint P by using the well-known **Sampling theorem** (due to Nyquist, Kotelnikov and Shannon), which states that sampling a band-limited signal with bandwidth W at a rate of $1/(2W)$ is sufficient to reconstruct the signal from its samples.

- Since $X(t)$, $\tilde{Z}(t)$ and $Y(t)$ are all **band-limited** to $[-W, W]$, we can thus represent these signals by their samples taken $\frac{1}{2W}$ seconds apart and model the channel by a discrete-time channel described by:

$$Y_n = X_n + \tilde{Z}_n, \quad n = 1, 2, \cdots,$$

where X_n:=$X(\frac{n}{2W})$ are the input samples and $\tilde{Z}_n$ and Y_n are the random samples of the noise $\tilde{Z}(t)$ and output $Y(t)$ signals, respectively.

5.8 Capacity of band-limited white Gaussian channel

- Since $\tilde{Z}(t)$ is a filtered version of $Z(t)$, which is a zero-mean stationary Gaussian process, we obtain that $\tilde{Z}(t)$ is also zero-mean, stationary and Gaussian.
  - This directly implies that the samples $\tilde{Z}_n$, $n = 1, 2, \cdots$, are zero-mean Gaussian identically distributed random variables.
  - Note that

$$\mathrm{PSD}_{\tilde{Z}}(f) = \mathrm{PSD}_Z(f)|H(f)|^2 = \begin{cases} \frac{N_0}{2} & \text{if } -W \le f \le W, \\ 0 & \text{otherwise.} \end{cases}$$

    Hence, covariance function of the filtered noise process is

$$K_{\tilde{Z}}(\tau) = E[\tilde{Z}(s)\tilde{Z}(s+\tau)] = \mathcal{F}^{-1}[\mathrm{PSD}_{\tilde{Z}}(f)] = N_0 W \mathrm{sinc}(2W\tau) \quad \tau \in \mathbb{R}. \tag{5.8.2}$$

  - (5.8.2) implies

$$E[\tilde{Z}_n \tilde{Z}_{n'}] = K_{\tilde{Z}}(\tfrac{n-n'}{2W}) = \begin{cases} N_0 W, & n = n' \\ 0, & n \neq n' \end{cases}$$

5.8 Capacity of band-limited white Gaussian channel

- We conclude that the capacity of the band-limited white Gaussian channel in bits per channel use is given using (5.4.13) by

$$\frac{1}{2}\log_2\left(1+\frac{P}{N_0W}\right) \qquad \text{bits/channel use.}$$

- Given that we are using the channel (with inputs X_n) every $\frac{1}{2W}$ seconds, we obtain that the capacity in bits/second of the band-limited white Gaussian channel is given by

$$C(P)=\frac{\frac{1}{2}\log_2\left(1+\frac{P}{N_0W}\right)}{\frac{1}{2W}}=W\log_2\left(1+\frac{P}{N_0W}\right) \qquad \text{bits/second,} \quad (5.8.3)$$

  where $\frac{P}{N_0W}$ is typically referred to as the signal-to-noise ratio (SNR).

- (5.8.3) is achieved by zero-mean i.i.d. Gaussian $\{X_n\}_{n=-\infty}^{\infty}$ with $E[X_n^2]=P$, which can be obtained by sampling a zero-mean, stationary and Gaussian $X(t)$ with

$$\text{PSD}_X(f)=\begin{cases}\frac{P}{2W} & \text{if } -W\le f\le W,\\ 0 & \text{otherwise.}\end{cases}$$

Examining this $X(t)$ confirms that it satisfies (5.8.1):

$$\frac{1}{T}\int_{-T/2}^{T/2} E[X^2(t)]dt = E[X^2(t)] = K_X(0) = P \cdot \mathrm{sinc}(2W \cdot 0) = P.$$

Example 5.44 (Telephone line channel) Suppose telephone signals are band-limited to 4 KHz. Given an SNR of 40 decibels (dB) – i.e., $10\log_{10}\frac{P}{N_0 W} = 40$ dB – then from (5.8.3), we calculate that the capacity of the telephone line channel (when modeled via the band-limited white Gaussian channel) is given by

$$4000\log_2(1+10000) = 53151.4 \text{ bits/second } = 51.906 \text{ Kbits/second.}$$

By increasing the bandwidth to 1.2 MHz, the capacity becomes

$$1200000\log_2(1+10000) = 15945428 \text{ bits/second } = 15.207 \text{ Mbits/second.}$$

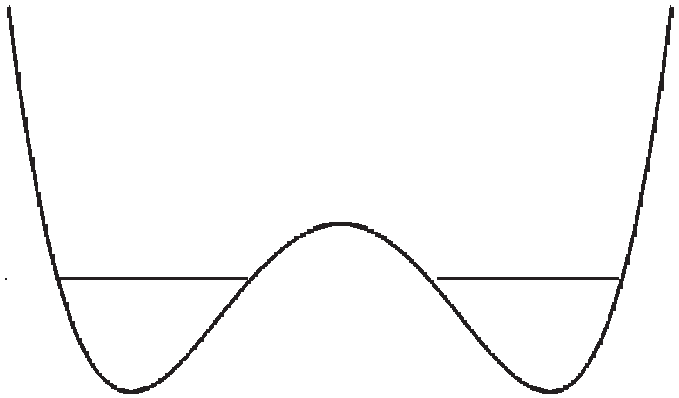
Observation 5.46 (Band-limited colored Gaussian channel) If the above band-limited channel has a stationary *colored* (non-white) additive Gaussian noise, then it can be shown (e.g., see [19]) that the capacity of this channel becomes

$$C(P) = \frac{1}{2}\int_{-W}^{W} \max\left[0, \log_2 \frac{\theta}{\mathrm{PSD}_Z(f)}\right] df,$$

where θ is the solution of

$$P = \int_{-W}^{W} \max\left[0, \theta - \mathrm{PSD}_Z(f)\right] df.$$

5.8 Capacity of band-limited white Gaussian channel



(a) The spectrum of $\mathrm{PSD}_Z(f)$ where the horizontal line represents θ, the level at which water rises to.

(b) The input spectrum that achieves capacity.

Water-pouring for the band-limited colored Gaussian channel.

Key Notes

- Differential entropy and its operational meaning in quantization efficiency
- Maximal differential entropy of Gaussian source, among all sources with the same mean and variance
- The mismatch in properties of entropy and differential entropy
- Relative entropy and mutual information of continuous systems
- Capacity-cost function and its proof
- Calculation of the capacity-cost function for specific channels
  - Memoryless additive Gaussian channels
  - Uncorrelated and correlated parallel Gaussian channels
  - Water-filling scheme (graphical interpretation)
  - Gaussian band-limited waveform channels
- Interpretation of entropy-power (provide an upper bound on capacity of non-Gaussian channels)
  - Operational characteristics of entropy-power inequality