Chapter 6

Continuous Sources and Channels

Po-Ning Chen, Professor

Department of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 300, R.O.C.

Continuous Sources and Channels

- 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.

Information Content of Continuous Sources

- A straightforward extension of entropy from discrete sources to continuous sources.

$$\text{Entropy: } H(X) \stackrel{\triangle}{=} \sum_{x \in \mathcal{X}} -P_X(x) \log P_X(x) \text{ nats.}$$

Example 6.1 (extension of entropy to continuous sources) Give a source X with source alphabet $[0, 1)$ and uniform generic distribution.
We can make it discrete by quantizing it into m levels as

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

for $1 \leq i \leq m$.
Then the resultant entropy of the quantized source is

$$H(q_m(X)) = -\sum_{i=1}^{m} \frac{1}{m} \log\left(\frac{1}{m}\right) = \log(m) \text{ nats.}$$

Since the entropy $H(q_m(X))$ of the quantized source is a lower bound to the entropy $H(X)$ of the uniformly distributed continuous source,

$$H(X) \geq \lim_{m \to \infty} H(q_m(X)) = \infty.$$

Information Content of Continuous Sources

- A **quantization** extension of entropy seems **ineffective** for continuous sources, because their entropies are all **infinity**.

  Proof: For any continuous source X, there must exist a non-empty open interval in which the cumulative distribution function $F_X(\cdot)$ is strictly increasing. Now quantize the source into $m+1$ level as follows:

  - Assign one level to the complement of this open interval, and
  - assign m levels to this open interval such that the probability mass on this interval, denoted by a, is equally distributed to these m levels. (This is similar to do equal partitions on the concerned domain of $F_X^{-1}(\cdot)$.)

  Then

$$H(X) \geq H(X^{\Delta}) = -(1-a) \cdot \log(1-a) - a \cdot \log \frac{a}{m},$$

  where X^{Δ} represents the quantized version of X. The lower bound goes to infinity as m tends to infinity. □

Differential Entropy

- Alternative extension of entropy to continuous sources (from its formula)

Definition 6.2 (differential entropy) The differential entropy (in nats) of a continuous source with generic probability density function (pdf) p_X is defined as

$$h(X) \stackrel{\triangle}{=} -\int_{\mathcal{X}} p_X(x) \cdot \log p_X(x) dx.$$

The next example demonstrates the difference (in its quantity) between the entropy and differential entropy.

Example 6.3 A continuous source X with source alphabet $[0, 1)$ and pdf $f(x) = 2x$ has differential entropy equal to

$$\begin{aligned} \int_0^1 -2x \cdot \log(2x) dx &= \left. \frac{x^2(1 - 2\log(2x))}{2} \right|_0^1 \\ &= \frac{1}{2} - \log(2) \approx -0.193 \text{ nats}. \end{aligned}$$

Examples of Differential Entropy

- Note that the differential entropy, unlike the entropy, can be negative in its value.

Example 6.4 (differential entropy of continuous sources with uniform generic distribution) A continuous source X with uniform generic distribution over (a, b) has differential entropy

$$h(X) = \log |b - a| \text{ nats.}$$

Example 6.5 (differential entropy of Gaussian sources) A continuous source X with Gaussian generic distribution of mean μ and variance σ^2 has differential entropy

$$\begin{aligned} h(X) &= \int_{\Re} \phi(x) \left[\frac{1}{2} \log(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \right] dx \\ &= \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} E[(X-\mu)^2] \\ &= \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} \\ &= \frac{1}{2} \log(2\pi\sigma^2 e) \text{ nats,} \end{aligned}$$

where $\phi(x)$ is the pdf of the Gaussian distribution with mean μ and variance σ^2.

Properties of Differential Entropy

- 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 6.6 (AEP for continuous sources) Let $X_1, \ldots, X_n$ be a sequence of sources drawn i.i.d. according to the density $p_X(\cdot)$. Then

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

Proof: The proof is an immediate result of law of large numbers. □

Definition 6.7 (typical set) For $\delta > 0$ and any n given, define the typical set as

$$\mathcal{F}_n(\delta) \stackrel{\triangle}{=} \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n}\log p_X(X_1, \ldots, X_n) - h(X) \right| < \delta \right\}.$$

Definition 6.8 (volume) The *volume* of a set $\mathcal{A}$ is defined as

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

Properties of Differential Entropy

Theorem 6.9 (Shannon-McMillan theorem for continuous sources)

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

Proof: The proof is an extension of Shannon-McMillan theorem for discrete sources, and hence we omit it. □

We can also derive a source coding theorem for continuous sources and obtain that when a continuous source is compressed by quantization, it is beneficial to put most of the quantization effort on its typical set, instead of on the entire source space.

E.g.

- Assigning $(m-1)$-level to elements in $\mathcal{F}_n(\delta)$;
- Assigning 1 level to those elements outside $\mathcal{F}_n(\delta)$.

Properties of Differential Entropy

Remarks:

- If the differential entropy of a continuous source is larger, it is expected that a larger number of quantization levels is required in order to minimize the distortion introduced via quantization.
- So we may conclude that continuous sources with higher differential entropy contain more information in volume.
- *Question:* Which continuous source is the richest in information (i.e., is highest in differential entropy and cost more in quantization)? Answer: Gaussian.

Properties of Differential Entropy

Theorem 6.10 (maximal differential entropy of Gaussian source) The Gaussian source has the largest differential entropy among all continuous sources with identical mean and variance.

Proof: Let $p(\cdot)$ be the pdf of a continuous source X, and let $\phi(\cdot)$ be the pdf of a Gaussian source Y. Assume that these two sources have the same mean μ and variance σ^2. Observe that

$$\begin{aligned}
-\int_{\Re} \phi(y) \log \phi(y) dy &= \int_{\Re} \phi(y) \left[\frac{1}{2} \log(2\pi\sigma^2) + \frac{(y-\mu)^2}{2\sigma^2} \right] dy \\
&= \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} E[(Y-\mu)^2] \\
&= \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} E[(X-\mu)^2] \\
&= -\int_{\Re} p(x) \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \right] dx \\
&= -\int_{\Re} p(x) \log \phi(x) dx.
\end{aligned}$$

Properties of Differential Entropy

Hence,

$$\begin{aligned}
h(Y) - h(X) &= -\int_{\Re} \phi(y) \log \phi(y) dy + \int_{\Re} p(x) \log p(x) dx \\
&= -\int_{\Re} p(x) \log \phi(x) dx + \int_{\Re} p(x) \log p(x) dx \\
&= \int_{\Re} p(x) \log \frac{p(x)}{\phi(x)} dx \\
&\geq \int_{\Re} p(x) \left(1 - \frac{\phi(x)}{p(x)}\right) dx \quad \text{(fundamental inequality)} \\
&= \int_{\Re} (p(x) - \phi(x))\, dx \\
&= 0,
\end{aligned}$$

with equality holds if, and only if, $p(x) = \phi(x)$ for all $x \in \Re$. □

Properties of Differential Entropy

- Properties regarding differential entropy for continuous sources, which are **the same** as in discrete cases.

Lemma 6.11

1. $h(X|Y) \leq h(X)$ with equality holds if, and only if, X and Y are independent.
2. (chain rule for differential entropy)

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

3. $h(X^n) \leq \sum_{i=1}^{n} h(X_i)$ with equality holds if, and only if, $\{X_i\}_{i=1}^{n}$ are independent.

Properties of Differential Entropy

- There are some properties that are conceptually **different** in continuous cases from the original ones in discrete cases.

  Lemma 6.12

  – (In discrete cases) For any one-to-one correspondence mapping f,

  $$H(f(X)) = H(X).$$

  – (In continuous cases) For a mapping $f(x) = ax$ with some non-zero constant a,

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

  Proof: (For continuous cases only) Let $p_X(\cdot)$ and $p_{f(X)}(\cdot)$ be respectively the pdfs of the original source and the mapped source. Then

  $$p_{f(X)}(u) = \frac{1}{|a|} p_X\left(\frac{u}{a}\right).$$

  By taking the new pdf into the formula of differential entropy, we have the desired result. □

Important Notes on Differential Entropy

- The above lemma says that the differential entropy can be increased by a one-to-one correspondence mapping.
- Therefore, when viewing the quantity as a measure of information content of continuous sources, we would yield that information content can be increased by a function mapping, which is somewhat contrary to the intuition.
- Based on this reason, it may not be appropriate to interpret the differential entropy as an index of information content of continuous sources.
- Indeed, it may be better to view this quantity as a measure of **quantization efficiency**.

Important Notes on Differential Entropy

- Some researchers interpret the differential entropy as a convenient intermediate formula of calculating the mutual information and divergence for systems with continuous alphabets, which is in general true.
- Before introducing the formula of relative entropy and mutual information for continuous settings, we point out beforehand that the interpretation, as well as the operational characteristics, of the mutual information and divergence for systems with continuous alphabets is exactly the same as for systems with discrete alphabets.

More Properties on Differential Entropy

Corollary 6.13 For a sequence of continuous sources $X^n = (X_1, \ldots, X_n)$, and a non-singular $n \times n$ matrix A,

$$h(AX^n) = h(X^n) + \log |A|,$$

where $|A|$ represents the determinant of matrix A.

Corollary 6.14 $h(X + c) = h(X)$ and $h(X|Y) = h(X + Y|Y)$.

Operational Meaning of Differential Entropy

Lemma 6.15 Give the pdf $f(x)$ of a continuous source X, and suppose that $-f(x)\log_2 f(x)$ is Riemann-integrable. Then to uniformly quantize the random source within n-bit accuracy, i.e., the quantization width is no greater than 2^{-n}, need approximately $h(X)+n$ bits (as n large enough).

Proof:

Step 1: Mean-value theorem .

Let $\Delta = 2^{-n}$ be the width of two adjacent quantization levels. Let $t_i = i\Delta$ for integer $i \in (-\infty, \infty)$. From mean-value theorem, we can choose $x_i \in [t_{i-1}, t_i]$ such that

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

Operational Meaning of Differential Entropy

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

Let

$$h^{\Delta}(X) \stackrel{\triangle}{=} \sum_{i=-\infty}^{\infty} [f(x_i) \log_2 f(x_i)]\Delta.$$

Since $h(X)$ is Riemann-integrable,

$$h^{\Delta}(X) \to h(X) \quad \text{as } \Delta = 2^{-n} \to 0.$$

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

$$|h(X) - h^{\Delta}(X)| < \varepsilon.$$

Step 3: Computation of $H(X^{\Delta})$.

The entropy of the quantized source X^{Δ} is

$$H(X^{\Delta}) = -\sum_{i=-\infty}^{\infty} p_i \log_2 p_i = -\sum_{i=-\infty}^{\infty} (f(x_i)\Delta) \log_2 (f(x_i)\Delta) \text{ bits.}$$

Operational Meaning of Differential Entropy

Step 4: $H(X^\Delta) - h^\Delta(X)$.

From Steps 2 and 3,

$$\begin{aligned} H(X^\Delta) - h^\Delta(X) &= -\sum_{i=-\infty}^{\infty} [f(x_i)\Delta] \log_2 \Delta \\ &= (-\log_2 \Delta) \sum_{i=-\infty}^{\infty} \int_{t_{i-1}}^{t_i} f(x)dx \\ &= (-\log_2 \Delta) \int_{-\infty}^{\infty} f(x)dx = -\log_2 \Delta = n. \end{aligned}$$

Hence,

$$[h(X) + n] - \varepsilon < H(X^\Delta) = h^\Delta(X) + n < [h(X) + n] + \varepsilon,$$

for $n > N$. □

- **Remark 1:** Since $H(X^\Delta)$ is the minimum average number of codeword length for lossless data compression, to uniformly quantize a continuous source upto n-bit accuracy requires approximately $h(X) + n$ bits.
- **Remark 2:** We may conclude that the larger the differential entropy, the average number of bits required to uniformly quantize the source subject to a fixed accuracy is larger.

Operational Meaning of Differential Entropy

- This operational meaning of differential entropy can be used to interpret its properties introduced in the previous subsection. For example, "$h(X + c) = h(X)$" can be interpreted as "A shift in value does not change the quantization efficiency of the original source."

Riemann Integral Versus Lebesgue Integral

Riemann integral:
Let $s(x)$ represent a step function on $[a, b)$, which is defined as that there exists a partition $a = x_0 < x_1 < \cdots < x_n = b$ such that $s(x)$ is constant during (x_i, x_{i+1}) for $0 \leq i < n$.
If a function $f(x)$ is Riemann integrable,

$$\int_a^b f(x) \stackrel{\triangle}{=} \sup_{\left\{s(x) \,:\, s(x) \leq f(x)\right\}} \int_a^b s(x)dx = \inf_{\left\{s(x) \,:\, s(x) \geq f(x)\right\}} \int_a^b s(x)dx.$$

Example of a non-Riemann-integrable function:
$f(x) = 0$ if x is irrational; $f(x) = 1$ if x is rational.
Then

$$\sup_{\left\{s(x) \,:\, s(x) \leq f(x)\right\}} \int_a^b s(x)dx = 0,$$

but

$$\inf_{\left\{s(x) \,:\, s(x) \geq f(x)\right\}} \int_a^b s(x)dx = (b - a).$$

Riemann Integral Versus Lebesgue Integral

Lebesgue integral:
Let $t(x)$ represent a simple function, which is defined as the linear combination of indicator functions for mutually-disjoint partitions.

For example, let $\mathcal{U}_1, \ldots, \mathcal{U}_m$ be the mutually-disjoint partitions of the domain $\mathcal{X}$ and $\cup_{i=1}^m \mathcal{U}_i = \mathcal{X}$. The indicator function of $\mathcal{U}_i$ is $\mathbf{1}(x; \mathcal{U}_i) = 1$ if $x \in \mathcal{U}_i$, and 0, otherwise.

Then $t(x) = \sum_{i=1}^m a_i \mathbf{1}(x; \mathcal{U}_i)$ is a simple function.
If a function $f(x)$ is Lebesgue integrable, then

$$\int_a^b f(x) = \sup_{\left\{t(x) \,:\, t(x) \le f(x)\right\}} \int_a^b t(x)dx = \inf_{\left\{t(x) \,:\, t(x) \ge f(x)\right\}} \int_a^b t(x)dx.$$

The previous example is actually Lebesgue integrable, and its Lebesgue integral is equal to zero.

Example of Quantization Efficiency

Example 6.16 Find the minimum average number of bits required to uniformly quantize the *decay time* (in years) of a radium atom upto 3-digit accuracy, if the half-life of the radium is 80 years. Note that the half-life of a radium atom is the median of its decay time distribution $f(x) = \lambda e^{-\lambda x}$ $(x > 0)$.

Since the median is 80, we obtain:

$$\int_0^{80} \lambda e^{-\lambda x} dx = 0.5,$$

which implies $\lambda = 0.00866$. Also, 3-digit accuracy is approximately equivalent to $\log_2 999 = 9.96 \approx 10$ bit accuracy. Therefore, the number of bits required to quantize the source is approximately

$$h(X) + 10 = \log_2 \frac{e}{\lambda} + 10 = 18.29 \text{ bits}.$$

Relative Entropy and Mutual Information

Definition 6.17 (relative entropy) Define the *relative entropy* between two densities p_X and $p_{\hat{X}}$ by

$$D(X\|\hat{X}) \stackrel{\triangle}{=} \int_{\mathcal{X}} p_X(x) \log \frac{p_X(x)}{p_{\hat{X}}(x)} dx.$$

Definition 6.18 (mutual information) The mutual information with input-output joint density $p_{X,Y}(x,y)$ is defined as

$$I(X;Y) \stackrel{\triangle}{=} \int_{\mathcal{X}\times\mathcal{Y}} p_X(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} dxdy.$$

Relative Entropy and Mutual Information

- Different from the cases for entropy, the properties of relative entropy and mutual information in continuous cases are the same as those in the discrete cases.
- In particular, the mutual information of the quantized version of a continuous channel will converge to the mutual information of the same continuous channel (as the quantization step size goes to zero).
- Hence, some researchers prefer to define the mutual information of a continuous channel directly as the limit of the quantized channel.
- Here, we quote some of the properties on relative entropy and mutual information from discrete settings.

Lemma 6.19

1. $D(X\|\hat{X}) \geq 0$ with equality holds if, and only if, $p_X = p_{\hat{X}}$.
2. $I(X;Y) \geq 0$ with equality holds if, and only if, X and Y are independent.
3. $I(X;Y) = h(Y) - h(Y|X)$.

Rate Distortion Theorem Revisited

- Since the entropy of continuous sources is infinity, to compress a continuous source without distortion is impossible according to Shannon's source coding theorem.
- Thus, one way to characterize the data compression for continuous sources is to encode the original source subject to a constraint on the distortion, which yields the *rate-distortion function* for data compression.
- In concept, the rate-distortion function is the minimum data compression rate (nats required per source letter or nats required per source sample) for which the distortion constraint is satisfied.

Specific Formula of Rate Distortion Theorem

Theorem 6.20 Under the squared error distortion measure, namely

$$\rho(z, \hat{z}) = (z - \hat{z})^2,$$

the rate-distortion function for continuous source Z with zero mean and variance σ^2 satisfies

$$R(D) \leq \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & \text{for } 0 \leq D \leq \sigma^2; \\ 0, & \text{for } D > \sigma^2. \end{cases}$$

Equality holds when Z is Gaussian.

Proof: By Theorem 5.16 (extended to squared-error distortion measure),

$$R(D) = \min_{\left\{p_{\hat{Z}|Z} \,:\, E[(Z-\hat{Z})^2] \leq D\right\}} I(Z; \hat{Z}).$$

So for any $p_{\hat{Z}|Z}$ satisfying the distortion constraint,

$$R(D) \leq I(p_Z, p_{\hat{Z}|Z}).$$

For $0 \leq D \leq \sigma^2$, choose a dummy Gaussian random variable W with zero mean and variance aD, where $a = 1 - D/\sigma^2$, and is independent of Z. Let $\hat{Z} = aZ + W$. Then

$$\begin{aligned} E[(Z - \hat{Z})^2] &= E[(1-a)^2 Z^2] + E[W^2] \\ &= (1-a)^2\sigma^2 + aD = D \end{aligned}$$

which satisfies the distortion constraint. Note that the variance of $\hat{Z}$ is equal to $E[a^2Z^2] + E[W^2] = \sigma^2 - D$. Consequently,

$$\begin{aligned}
R(D) &\le I(Z;\hat{Z}) \\
&= h(\hat{Z}) - h(\hat{Z}|Z) \\
&= h(\hat{Z}) - h(W + aZ|Z) \\
&= h(\hat{Z}) - h(W|Z) \quad \text{(By Corollary 6.14)} \\
&= h(\hat{Z}) - h(W) \quad \text{(By Lemma 6.11)} \\
&= h(\hat{Z}) - \frac{1}{2}\log(2\pi e(aD)) \\
&\le \frac{1}{2}\log(2\pi e(\sigma^2 - D)) - \frac{1}{2}\log(2\pi e(aD)) \\
&= \frac{1}{2}\log\frac{\sigma^2}{D}.
\end{aligned}$$

For $D > \sigma^2$, let $\hat{Z}$ satisfy $\Pr\{\hat{Z} = 0\} = 1$, and be independent of Z. Then $E[(Z - \hat{Z})^2] = E[Z^2] - E[\hat{Z}^2] - 2E[Z]E[\hat{Z}] = \sigma^2 < D$, and $I(Z;\hat{Z}) = 0$. Hence, $R(D) = 0$ for $D > \sigma^2$.

The achievability of the upper bound by Gaussian source will be proved in Theorem 6.22. □

Rate Distortion Function for Binary Source

Theorem 6.21 Fix a memoryless binary source

$$\boldsymbol{Z} = \{Z^n = (Z_1, Z_2, \ldots, Z_n)\}_{n=1}^{\infty}$$

with marginal distribution $P_Z(0) = 1 - P_Z(1) = p$. Assume that the Hamming additive distortion measure is employed. Then the rate-distortion function

$$R(D) = \begin{cases} H_b(p) - H_b(D), & \text{if } 0 \leq D \leq \min\{p, 1-p\}; \\ 0, & \text{if } D > \min\{p, 1-p\}, \end{cases}$$

where $H_b(p) \stackrel{\triangle}{=} -p \cdot \log(p) - (1-p) \cdot \log(1-p)$ is the binary entropy function.

Proof: Assume without loss of generality that $p \leq 1/2$.

We first prove the theorem under $0 \leq D < \min\{p, 1-p\} = p$. Observe that for any binary random variable $\hat{Z}$,

$$H(Z|\hat{Z}) = H(Z \oplus \hat{Z}|\hat{Z}).$$

Also observe that

$$E[\rho(Z, \hat{Z})] \leq D \text{ implies } \Pr\{Z \oplus \hat{Z} = 1\} \leq D.$$

Rate Distortion Function for Binary Source

Then

$$\begin{aligned} I(Z;\hat{Z}) &= H(Z) - H(Z|\hat{Z}) \\ &= H_b(p) - H(Z \oplus \hat{Z}|\hat{Z}) \\ &\geq H_b(p) - H(Z \oplus \hat{Z}) \quad \text{(conditioning never increase entropy)} \\ &\geq H_b(p) - H_b(D), \end{aligned}$$

where the last inequality follows since $H_b(x)$ is increasing for $x \leq 1/2$, and $\Pr\{Z \oplus \hat{Z} = 1\} \leq D$. Since the above derivation is true for any $P_{\hat{Z}|Z}$, we have

$$R(D) \geq H_b(p) - H_b(D).$$

It remains to show that the lower bound is achievable by some $P_{\hat{Z}|Z}$, or equivalently, $H(Z|\hat{Z}) = H_b(D)$ for some $P_{\hat{Z}|Z}$. By defining $P_{Z|\hat{Z}}(0|0) = P_{Z|\hat{Z}}(1|1) = 1 - D$, we immediately obtain $H(Z|\hat{Z}) = H_b(D)$. The desired $P_{\hat{Z}|Z}$ can be obtained by solving

$$\begin{aligned} 1 &= P_{\hat{Z}}(0) + P_{\hat{Z}}(1) \\ &= \frac{P_Z(0)}{P_{Z|\hat{Z}}(0|0)} P_{\hat{Z}|Z}(0|0) + \frac{P_Z(0)}{P_{Z|\hat{Z}}(0|1)} P_{\hat{Z}|Z}(1|0) \\ &= \frac{p}{1-D} P_{\hat{Z}|Z}(0|0) + \frac{p}{D}(1 - P_{\hat{Z}|Z}(0|0)) \end{aligned}$$

and

$$\begin{aligned} 1 &= P_{\hat{Z}}(0) + P_{\hat{Z}}(1) \\ &= \frac{P_Z(1)}{P_{Z|\hat{Z}}(1|0)} P_{\hat{Z}|Z}(0|1) + \frac{P_Z(1)}{P_{Z|\hat{Z}}(1|1)} P_{\hat{Z}|Z}(1|1) \\ &= \frac{1-p}{D}(1 - P_{\hat{Z}|Z}(1|1)) + \frac{1-p}{1-D} P_{\hat{Z}|Z}(1|1), \end{aligned}$$

and yield

$$P_{\hat{Z}|Z}(0|0) = \frac{1-D}{1-2D}\left(1 - \frac{D}{p}\right) \quad \text{and} \quad P_{\hat{Z}|Z}(1|1) = \frac{1-D}{1-2D}\left(1 - \frac{D}{1-p}\right).$$

Now in the case of $p \le D < 1-p$, we can let $P_{\hat{Z}|Z}(1|0) = P_{\hat{Z}|Z}(1|1) = 1$ to obtain $I(Z;\hat{Z}) = 0$ and

$$E[\rho(Z,\hat{Z})] = \sum_{z=0}^{1}\sum_{\hat{z}=0}^{1} P_Z(z) P_{\hat{Z}|Z}(\hat{z}|z)\rho(z,\hat{z}) = p \le D.$$

Similarly, in the case of $D \ge 1-p$, we let $P_{\hat{Z}|Z}(0|0) = P_{\hat{Z}|Z}(0|1) = 1$ to obtain $I(Z;\hat{Z}) = 0$ and

$$E[\rho(Z,\hat{Z})] = \sum_{z=0}^{1}\sum_{\hat{z}=0}^{1} P_Z(z) P_{\hat{Z}|Z}(\hat{z}|z)\rho(z,\hat{z}) = 1-p \le D.$$

Rate Distortion Function for Binary Source

□

- **Remark:** The Hamming additive distortion measure is defined as:

$$\rho_n(z^n, \hat{z}^n) = \sum_{i=1}^{n} z_i \oplus \hat{z}_i,$$

where "$\oplus$" denotes modulo two addition. In such case, $\rho(z^n, \hat{z}^n)$ is exactly the number of bit changes or bit errors after compression.

Rate Distortion Function for Gaussian Sources

Theorem 6.22 Fix a memoryless source

$$\boldsymbol{Z} = \{Z^n = (Z_1, Z_2, \ldots, Z_n)\}_{n=1}^{\infty}$$

with zero-mean Gaussian marginal distribution of variance σ^2. Assume that the squared error distortion measure is employed. Then the rate-distortion function is given by:

$$R(D) = \begin{cases} \dfrac{1}{2}\log\dfrac{\sigma^2}{D}, & \text{if } 0 \leq D \leq \sigma^2; \\ 0, & \text{if } D > \sigma^2. \end{cases}$$

Proof: From Theorem 6.20, it suffices to show that under the Gaussian source, $(1/2)\log(\sigma^2/D)$ is a lower bound to $R(D)$ for $0 \leq D \leq \sigma^2$.

Rate Distortion Function for Gaussian Sources

This can be proved as follows. For Gaussian source Z with $E[(Z-\hat{Z})^2] \le D$,

$$\begin{aligned}
I(Z;\hat{Z}) &= h(Z) - h(Z|\hat{Z}) \\
&= \frac{1}{2}\log(2\pi e\sigma^2) - h(Z-\hat{Z}|\hat{Z}) \quad \text{(Corollary 6.14)} \\
&\ge \frac{1}{2}\log(2\pi e\sigma^2) - h(Z-\hat{Z}) \quad \text{(Lemma 6.11)} \\
&\ge \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log\Big(2\pi e\,\mathrm{Var}[(Z-\hat{Z})]\Big) \quad \text{(Theorem 6.10)} \\
&\ge \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log\Big(2\pi e\,E[(Z-\hat{Z})^2]\Big) \\
&\ge \frac{1}{2}\log(2\pi e\sigma^2) - \frac{1}{2}\log(2\pi eD) \\
&= \frac{1}{2}\log\frac{\sigma^2}{D}.
\end{aligned}$$

□

Channel Coding Theorem for Continuous Sources

- Power constraint on channel input
  - To derive the channel capacity for a memoryless *continuous* channel without any constraint on the inputs is somewhat impractical, especially when the input can be any number on the infinite real line.
  - Such constraint is usually of the form

$$E[t(X)] \leq S$$

or

$$\frac{1}{n}\sum_{i=1}^{n} E[t(X_i)] \leq S \quad \text{for a sequence of random inputs,}$$

where $t(\cdot)$ is a non-negative cost function.

Example 6.23 (average power constraint) $t(x) \stackrel{\triangle}{=} x^2$, i.e., the constraint is that the *average input power* is bounded above by S.

Channel Coding Theorem for Continuous Sources

- As extended from discrete cases, the channel capacity of a discrete-time continuous channel with the input cost constraints is of the form

$$C(S) \stackrel{\triangle}{=} \max_{\{p_X \,:\, E[t(X)] \leq S\}} I(X;Y). \tag{6.3.1}$$

- Claim: $C(S)$ is a concave function of S.

Lemma 6.24 (concavity of capacity-cost function) $C(S)$ is concave, continuous, and strictly increasing in S.

Proof: Let P_{X_1} and P_{X_2} be two distributions that respectively achieve $C(P_1)$ and $C(P_2)$. Denote $P_{X_\lambda} \stackrel{\triangle}{=} \lambda P_{X_1} + (1-\lambda)P_{X_2}$. Then

$$\begin{aligned} C(\lambda P_1 + (1-\lambda)P_2) &= \max_{\{P_X \,:\, E[t(X)] \le \lambda P_1 + (1-\lambda)P_2\}} I(P_X, P_{Y|X}) \\ &\ge I(P_{X_\lambda}, P_{Y|X}) \\ &\ge \lambda I(P_{X_1}, P_{Y|X}) + (1-\lambda) I(P_{X_2}, P_{Y|X}) \\ &= \lambda C(P_1) + (1-\lambda) C(P_2), \end{aligned}$$

where the first inequality holds since

$$\begin{aligned} E_{X_\lambda}[t(X)] &= \int_{\Re} t(x) dP_\lambda(x) \\ &= \lambda \int_{\Re} t(x) dP_{X_1}(x) + (1-\lambda) \int_{\Re} t(x) dP_{X_2}(x) \\ &= \lambda E_{X_1}[t(X)] + (1-\lambda) E_{X_2}[t(X)] \\ &\le \lambda P_1 + (1-\lambda) P_2, \end{aligned}$$

and the second inequality follows from the concavity of mutual information with respect to the first argument. Accordingly, $C(S)$ is concave in S.

Furthermore, it can be easily seen by definition that $C(S)$ is non-decreasing, which, together with its concavity, implies its continuity and strict increasing. □

Channel Coding Theorem for Continuous Sources

- Although the capacity-cost function formula in (6.3.1) is valid for general cost function $t(\cdot)$, we only substantiate it under the *average power constraint* in the next forward channel coding theorem.
- Its validity for a more general case can be similarly proved based on the same concept.

Channel Coding Theorem for Continuous Sources

Theorem 6.25 (forward channel coding theorem for continuous channels under average power constraint) For any $\varepsilon \in (0,1)$, there exist $0 < \gamma < 2\varepsilon$ and a data transmission code sequence $\{\mathscr{C}_n = (n, M_n)\}_{n=1}^{\infty}$ satisfying

$$\frac{1}{n} \log M_n > C(S) - \gamma$$

and for each codeword $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$,

$$\frac{1}{n} \sum_{i=1}^{n} c_i^2 \leq S \tag{6.3.2}$$

such that the probability of decoding error $P_e(\mathscr{C}_n)$ is less than ε for sufficiently large n, where

$$C(S) \stackrel{\triangle}{=} \max_{\{p_X \,:\, E[X^2] \leq S\}} I(X;Y).$$

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

Step 0:

Take a positive γ satisfying

$$\gamma < \min\{2\varepsilon, C(S)\}.$$

Pick $\xi > 0$ small enough such that

$$2[C(S) - C(S - \xi)] < \gamma,$$

where the existence of such ξ is assured by the strict increasing of $C(S)$. Hence, we have $C(S - \xi) - \gamma/2 > C(S) - \gamma > 0$. Choose M_n to satisfy

$$C(S - \xi) - \frac{\gamma}{2} > \frac{1}{n} \log M_n > C(S) - \gamma,$$

for which the choice should exist for all sufficiently large n. Take $\delta = \gamma/8$. Let P_X be the distribution that achieves $C(S - \xi)$; hence, $E[X^2] \leq S - \xi$ and $I(X;Y) = C(S - \xi)$.

Channel Coding Theorem for Continuous Sources

Step 1: Random coding with average power constraint.

Randomly draw $M_n - 1$ codewords according to distribution P_{X^n} with

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

By law of large numbers, each randomly selected codeword

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

satisfies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} c_{mi}^2 \leq S - \varepsilon \quad \text{almost surely}$$

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

Step 2: Coder.

For M_n selected codewords $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{M_n}\}$, replace the codewords that violate the power constraint (i.e., (6.3.2)) by all-zero codeword $\mathbf{0}$. Define the encoder as

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

When receiving an 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

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

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

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

Step 3: Probability of error.

Let λ_m denote the error probability given that codeword m is transmitted. Define

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

Then by following similar argument as (4.3.2) (discrete cases), we get:

$$\begin{aligned} E[\lambda_m] \;\leq\; & P_{X^n}(\mathcal{E}_0) + P_{X^n,Y^n}\left(\mathcal{F}_n^c(\delta)\right) \\ & + \sum_{\substack{m'=1 \\ m' \neq m}}^{M_n} \sum_{\boldsymbol{c}_m \in \mathcal{X}^n} \sum_{y^n \in \mathcal{F}_n(\delta|\boldsymbol{c}_{m'})} P_{X^n,Y^n}(\boldsymbol{c}_m, y^n), \end{aligned}$$

where

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

Note that the additional term $P_{X^n}(\mathcal{E}_0)$ (to discrete cases) 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.

Channel Coding Theorem for Continuous Sources

Finally, by carrying out similar procedure as in the proof of the capacity for discrete channels, we obtain:

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

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

Channel Coding Theorem for Continuous Sources

Theorem 6.26 (converse channel coding theorem for continuous channels) For any data transmission code sequence $\{\mathcal{C}_n = (n, M_n)\}_{n=1}^{\infty}$ satisfying the power constraint, if the ultimate data transmission rate satisfies

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

then its probability of decoding error is bounded away from zero for all n sufficiently large.

Proof: For an (n, M_n) block data transmission code, an encoding function is chosen as:

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

and each index i is equally likely for the average probability of block decoding error criterion. Hence, we can assume that the information message $\{1, 2, \ldots, M_n\}$ is generated from a uniformly distributed random variable, and denote it by W. As a result,

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

Since $W \to X^n \to Y^n$ forms a Markov chain because Y^n only depends on X^n, we obtain by the data processing lemma that $I(W; Y^n) \leq I(X^n; Y^n)$.

Channel Coding Theorem for Continuous Sources

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

$$
\begin{aligned}
I(X^n;Y^n) &\le \max_{\{P_{X^n} \,:\, (1/n)\sum_{i=1}^n E[X_i^2]\le S\}} I(X^n;Y^n) \\
&\le \max_{\{P_{X^n} \,:\, (1/n)\sum_{i=1}^n E[X_i^2]\le S\}} \sum_{j=1}^n I(X_j;Y_j) \\
&= \max_{\{(P_1,P_2,\ldots,P_n) \,:\, (1/n)\sum_{i=1}^n P_i = S\}} \; \max_{\{P_{X^n} \,:\, (\forall\ i)\ E[X_i^2]\le P_i\}} \sum_{j=1}^n I(X_j;Y_j) \\
&\le \max_{\{(P_1,P_2,\ldots,P_n) \,:\, (1/n)\sum_{i=1}^n P_i = S\}} \sum_{j=1}^n \; \max_{\{P_{X^n} \,:\, (\forall\ i)\ E[X_i^2]\le P_i\}} I(X_j;Y_j) \\
&\le \max_{\{(P_1,P_2,\ldots,P_n) \,:\, (1/n)\sum_{i=1}^n P_i = S\}} \sum_{j=1}^n \; \max_{\{P_{X_j} \,:\, E[X_j^2]\le P_j\}} I(X_j;Y_j) \\
&= \max_{\{(P_1,P_2,\ldots,P_n):(1/n)\sum_{i=1}^n P_i = S\}} \sum_{j=1}^n C(P_j) \\
&= \max_{\{(P_1,P_2,\ldots,P_n):(1/n)\sum_{i=1}^n P_i = S\}} n\sum_{j=1}^n \frac{1}{n} C(P_j)
\end{aligned}
$$

$$\begin{aligned} &\leq \max_{\{(P_1,P_2,\ldots,P_n):(1/n)\sum_{i=1}^{n} P_i = S\}} nC\left(\frac{1}{n}\sum_{j=1}^{n} P_j\right) \\ &\quad \text{(by concavity of } C(S)) \\ &= nC(S). \end{aligned}$$

Consequently, by defining $P_e(\mathscr{C}_n)$ as the error of guessing W by observing Y^n via a decoding function

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

which is exactly the average block decoding failure, we get

$$\begin{aligned} \log M_n &= H(W) \\ &= H(W|Y^n) + I(W; Y^n) \\ &\leq H(W|Y^n) + I(X^n; Y^n) \\ &\leq H_b(P_e(\mathscr{C}_n)) + P_e(\mathscr{C}_n) \cdot \log(|\mathcal{W}| - 1) + nC(S), \\ &\quad \text{(by Fano's inequality)} \\ &\leq \log(2) + P_e(\mathscr{C}_n) \cdot \log(M_n - 1) + nC(S), \\ &\quad \text{(by the fact that } (\forall\ t \in [0,1])\ H_b(t) \leq \log(2)) \\ &\leq \log(2) + P_e(\mathscr{C}_n) \cdot \log M_n + nC(S), \end{aligned}$$

which implies that

$$P_e(\mathscr{C}_n) \geq 1 - \frac{C(S)}{(1/n)\log M_n} - \frac{\log(2)}{\log M_n}.$$

So if $\liminf_{n\to\infty}(1/n)\log M_n > C(S)$, then there exists δ with $0 < \delta < 4\varepsilon$ and an integer N such that for $n \geq N$,

$$\frac{1}{n}\log M_n > C(S) + \delta.$$

Hence, for $n \geq N_0 \stackrel{\triangle}{=} \max\{N, 2\log(2)/\delta\}$,

$$P_e(\mathscr{C}_n) \geq 1 - \frac{C(S)}{C(S)+\delta} - \frac{\log(2)}{n(C(S)+\delta)} \geq \frac{\delta}{2(C(S)+\delta)}.$$

□

Remarks:

- This is a weak converse statement.
- Since the capacity is now a function of the cost constraint, it is named *the capacity-cost function*.

Next, we will derive the capacity-cost function of some frequently used channel models.

Memoryless Additive Gaussian Channels

Definition 6.27 (memoryless additive channel) Let

$$X_1, \ldots, X_n \text{ and } Y_1, \ldots, Y_n$$

be the input and output sequences of the channel. Also let $N_1, \ldots, N_n$ be the noise. Then a memoryless additive channel is defined by

$$Y_i = X_i + N_i$$

for each i, where $\{X_i, Y_i, N_i\}_{i=1}^n$ are i.i.d. and X_i is independent of N_i.

Definition 6.28 (memoryless additive Gaussian channel) A memoryless additive channel is called *memoryless additive Gaussian channel*, if the noise is a Gaussian random variable.

Theorem 6.29 (capacity of memoryless additive Gaussian channel under average power constraint) The capacity of a memoryless additive Gaussian channel with noise model $\mathcal{N}(0, \sigma^2)$ and average power constraint is equal to:

$$C(S) = \frac{1}{2} \log \left(1 + \frac{S}{\sigma^2}\right) \text{ nats/channel symbol.}$$

Proof: By definition,

$$\begin{aligned} C(S) &= \max_{\{p_X \,:\, E[X^2] \le S\}} I(X;Y) \\ &= \max_{\{p_X \,:\, E[X^2] \le S\}} (h(Y) - h(Y|X)) \\ &= \max_{\{p_X \,:\, E[X^2] \le S\}} (h(Y) - h(N+X|X)) \\ &= \max_{\{p_X \,:\, E[X^2] \le S\}} (h(Y) - h(N|X)) \\ &= \max_{\{p_X \,:\, E[X^2] \le S\}} (h(Y) - h(N)) \\ &= \left(\max_{\{p_X \,:\, E[X^2] \le S\}} h(Y) \right) - h(N), \end{aligned}$$

where N represents the additive Gaussian noise. We thus need to find an input distribution satisfying $E[X^2] \le S$ that maximizes the differential entropy of Y.

Recall that the differential entropy subject to mean and variance constraint is maximized by Gaussian random variable; also, the differential entropy of a Gaussian random variable with variance σ^2 is $(1/2)\log(2\pi\sigma^2 e)$ nats, and is nothing to do with its mean.

Therefore, by taking X to be a Gaussian random variable having distribution $\mathcal{N}(0, S)$, Y achieves its largest variance $S + \sigma^2$ under the constraint $E[X^2] \leq S$. Consequently,

$$\begin{aligned} C(S) &= \frac{1}{2}\log(2\pi e(S+\sigma^2)) - \frac{1}{2}\log(2\pi e\sigma^2) \\ &= \frac{1}{2}\log\left(1+\frac{S}{\sigma^2}\right) \text{ nats/channel symbol.} \end{aligned}$$

□

Theorem 6.30 (worseness in capacity of Gaussian noise) For all memoryless additive discrete-time continuous channels whose noise has zero-mean and variance σ^2, the capacity subject to average power constraint is lower bounded by

$$\frac{1}{2}\log\left(1+\frac{S}{\sigma^2}\right),$$

which is the capacity for memoryless additive discrete-time Gaussian channel. (This means that the Gaussian noise is the "worst" kind of noise in the sense of channel capacity.)

Proof: Let $p_{Y_g|X_g}$ and $p_{Y|X}$ denote the transition probabilities of the Gaussian channel and some other channel satisfying the cost constraint, respectively. Let N_g and N respectively denote their noises. Then for any Gaussian input p_{X_g},

$$\begin{aligned}
& I(p_{X_g}, p_{Y|X}) - I(p_{X_g}, p_{Y_g|X_g}) \\
= & \int_{\Re}\int_{\Re} p_{X_g}(x)P_N(y-x)\log\frac{p_N(y-x)}{p_Y(y)}dydx \\
& - \int_{\Re}\int_{\Re} p_{X_g}(x)P_{N_g}(y-x)\log\frac{p_{N_g}(y-x)}{p_{Y_g}(y)}dydx \\
= & \int_{\Re}\int_{\Re} p_{X_g}(x)P_N(y-x)\log\frac{p_N(y-x)}{p_Y(y)}dydx \\
& - \int_{\Re}\int_{\Re} p_{X_g}(x)P_N(y-x)\log\frac{p_{N_g}(y-x)}{p_{Y_g}(y)}dydx \\
= & \int_{\Re}\int_{\Re} p_{X_g}(x)P_N(y-x)\log\frac{p_N(y-x)p_{Y_g}(y)}{p_{N_g}(y-x)p_Y(y)}dydx \\
\geq & \int_{\Re}\int_{\Re} p_{X_g}(x)P_N(y-x)\left(1-\frac{p_{N_g}(y-x)p_Y(y)}{p_N(y-x)p_{Y_g}(y)}\right)dydx \\
= & 1-\int_{\Re}\frac{p_Y(y)}{p_{Y_g}(y)}\left(\int_{\Re} p_{X_g}(x)p_{N_g}(y-x)dx\right)dy \\
= & 0,
\end{aligned}$$

with equality holds if, and only if,

$$\frac{p_Y(y)}{p_{Y_g}(y)} = \frac{p_N(y-x)}{p_{N_g}(y-x)}$$

for all x.

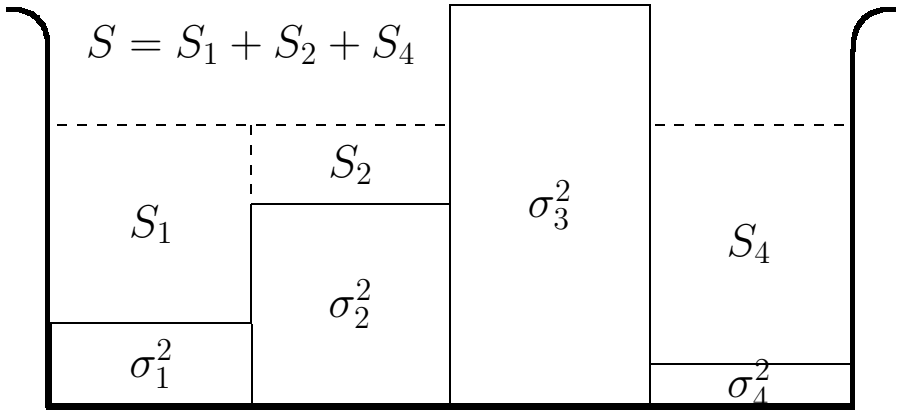
Therefore,

$$\begin{aligned} \max_{\{p_X \,:\, E[X^2] \le S\}} I(p_X, p_{Y_g|X_g}) &= I(p^*_{X_g}, p_{Y_g|X_g}) \\ &\le I(p^*_{X_g}, P_{Y|X}) \\ &\le \max_{\{p_X \,:\, E[X^2] \le S\}} I(p_X, p_{Y|X}). \end{aligned}$$

□

Uncorrelated Parallel Gaussian Channels

- Water-pouring scheme
  - In concept, more *channel input power* should be placed on those channels with smaller *noise power*.
  - The water-pouring scheme not only substantiates the above intuition, but also gives a quantitatively meaning for it.



Uncorrelated Parallel Gaussian Channels

Theorem 6.31 (capacity for parallel additive Gaussian channels) The capacity of k parallel additive Gaussian channels under an overall input power constraint S, is

$$C(S) = \sum_{i=1}^{k} \frac{1}{2} \log \left(1 + \frac{S_i}{\sigma_i^2} \right),$$

where σ_i^2 is the noise variance of channel i, $S_i = \max\{0, \theta - \sigma_i^2\}$, and θ is chosen to satisfy $\sum_{i=1}^{k} S_i = S$.

This capacity is achieved by a set of independent Gaussian input with zero mean and variance S_i.

Proof: By definition,

$$C(S) = \max_{\left\{p_{X^k} \,:\, \sum_{i=1}^k E[X_i^2] \le S\right\}} I(X^k; Y^k).$$

Since noise $N_1, \ldots, N_k$ are independent,

$$\begin{aligned}
I(X^k; Y^k) &= h(Y^k) - h(Y^k|X^k) \quad \left(= h(Y^k) - h(N^k + X^k|X^k) = h(Y^k) - h(N^k|X^k) \right) \\
&= h(Y^k) - h(N^k) \\
&= h(Y^k) - \sum_{i=1}^k h(N_i) \\
&\le \sum_{i=1}^k h(Y_i) - \sum_{i=1}^k h(N_i) \\
&= \sum_{i=1}^k I(X_i; Y_i) \\
&\le \sum_{i=1}^k \frac{1}{2} \log\left(1 + \frac{S_i}{\sigma_i^2}\right)
\end{aligned}$$

with equality holds if each input is a Gaussian random variable with zero mean

and variance S_i, and the inputs are independent, where S_i is the individual power constraint applied on channel i with $\sum_{i=1}^{k} S_i = S$.

So the problem is reduced to finding the power allotment that maximizes the capacity subject to the constraint $\sum_{i=1}^{k} S_i = S$. By using the Lagrange multiplier technique, the maximizer of

$$\max \left\{ \sum_{i=1}^{k} \frac{1}{2} \log \left(1 + \frac{S_i}{\sigma_i^2} \right) + \lambda \left(\sum_{i=1}^{k} S_i - S \right) \right\}$$

can be found by taking the derivative (w.r.t. S_i) of the above equation and let it be zero, which yields

$$\begin{cases} \frac{1}{2} \frac{1}{S_i + \sigma_i^2} + \lambda = 0, & \text{if } S_i > 0; \\ \frac{1}{2} \frac{1}{S_i + \sigma_i^2} + \lambda \leq 0, & \text{if } S_i = 0. \end{cases}$$

Hence,

$$\begin{cases} S_i = \theta - \sigma_i^2, & \text{if } S_i > 0; \\ S_i \geq \theta - \sigma_i^2, & \text{if } S_i = 0, \end{cases}$$

where $\theta = -1/(2\lambda)$. □

Rate Distortion for Parallel Gaussian Sources

A theorem on rate-distortion function parallel to that on capacity-cost function can also be established.

Theorem 6.32 (rate distortion for parallel Gaussian sources) Given k mutually independent Gaussian sources with variance $\sigma_1^2, \ldots, \sigma_k^2$. The overall rate-distortion function for additive squared error distortion, namely

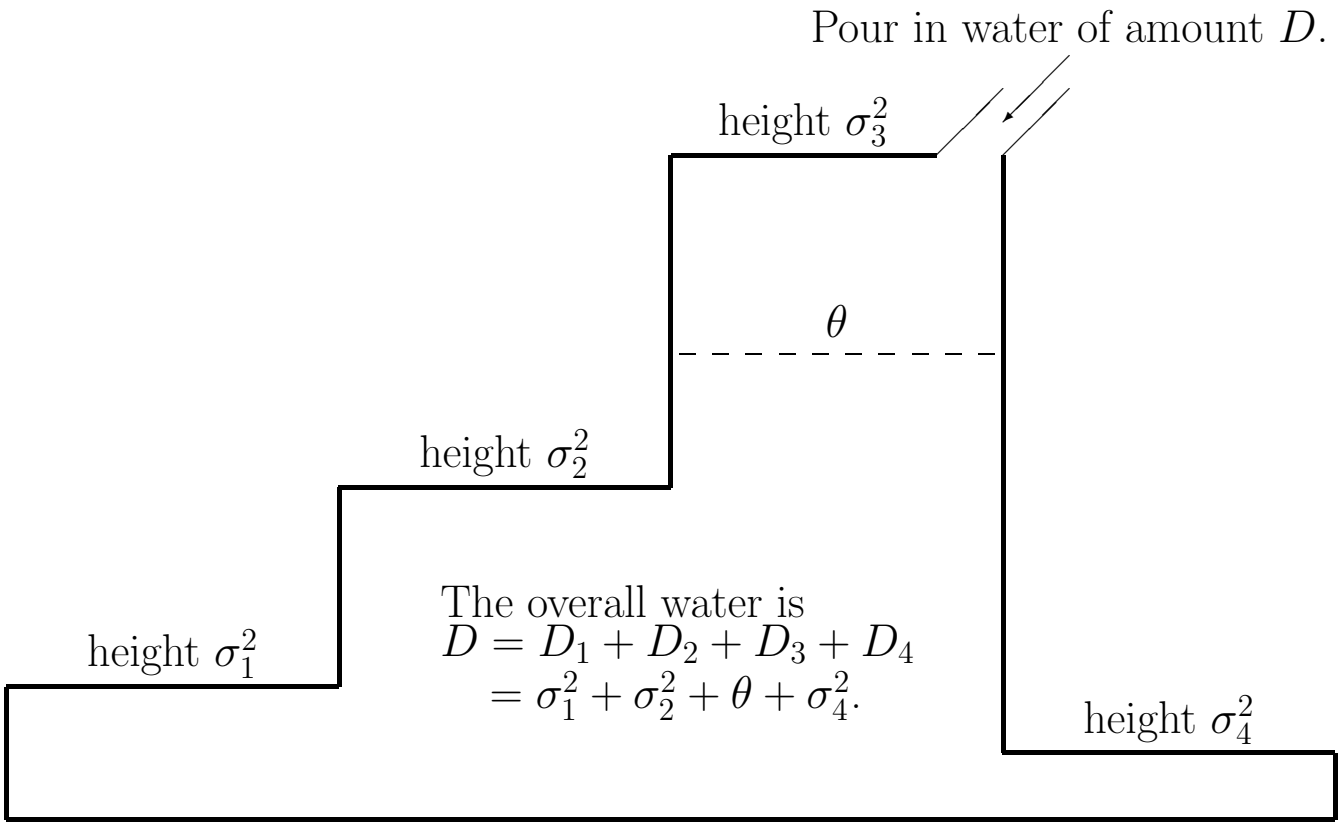
$$\sum_{i=1}^{k} E[(Z_i - \hat{Z}_i)^2] \leq D$$

is given by

$$R(D) = \sum_{i=1}^{k} \frac{1}{2} \log \frac{\sigma_i^2}{D_i},$$

where $D_i = \min\{\theta, \sigma_i^2\}$ and θ is chosen to satisfy $\sum_{i=1}^{k} D_i = D$.

Rate Distortion for Parallel Gaussian Sources



The water-pouring for lossy data compression of parallel Gaussian sources.

Correlated Parallel Additive Gaussian Channels

Theorem 6.33 (capacity for correlated parallel additive Gaussian channels) The capacity of k parallel additive Gaussian channels under an overall input power constraint S, is

$$C(S) = \sum_{i=1}^{k} \frac{1}{2} \log \left(1 + \frac{S_i}{\lambda_i}\right),$$

where λ_i is the i-th eigenvalue of the positive-definite noise covariance matrix $\mathbf{K}_N$,

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

and θ is chosen to satisfy $\sum_{i=1}^{k} S_i = S$.

This capacity is achieved by a set of independent Gaussian input with zero mean and variance S_i.

- A matrix $\mathbf{K}_{k\times k}$ is *positive definite* if for every $x_1, \ldots, x_k$,

$$[x_1, \ldots, x_k]\mathbf{K} \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix} \geq 0,$$

  with equality holds only when $x_i = 0$ for $1 \leq i \leq k$.

- The matrix $\mathbf{\Lambda}$ in the decomposition of a positive-definite matrix $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^t$, is a diagonal matrix with non-zero diagonal components.

Proof: Let $\mathbf{K}_X$ be the covariance matrix of the input, $X_1, \ldots, X_k$. The input power constraint then becomes

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

where $tr(\cdot)$ represent the traverse of the $k \times k$ matrix $\mathbf{K}_X$. Assume that the input is independent of the noise. Then

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

Since $h(N^k)$ is not determined by the input, the capacity-finding problem is reduced to maximize $h(Y^k)$ over all possible inputs satisfying the power constraint.

Now observe that the covariance matrix of Y^k is $\mathbf{K}_Y = \mathbf{K}_X + \mathbf{K}_N$, which implies that the differential entropy of Y^k is upper bounded by

$$h(Y^k) \leq \frac{1}{2} \log((2\pi e)^k |\mathbf{K}_X + \mathbf{K}_N|).$$

It remains to find the $\mathbf{K}_X$ (if it is possible) under which the above upper bound is achieved, and also this achievable upper bound is maximized.

Correlated Parallel Additive Gaussian Channels

Decompose $\mathbf{K}_N$ into its diagonal form as

$$\mathbf{K}_N = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^t,$$

where superscript "t" represents the transpose operation on matrices, $\mathbf{Q}\mathbf{Q}^t = \mathbf{I}_{k\times k}$, and $\mathbf{I}_{k\times k}$ represents the identity matrix of order k. Note that since $\mathbf{K}_N$ is positive definite, $\mathbf{\Lambda}$ is a diagonal matrix with positive diagonal components equal to the eigenvalues of $\mathbf{K}_N$. Then

$$\begin{aligned} |\mathbf{K}_X + \mathbf{K}_N| &= |\mathbf{K}_X + \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^t| \\ &= |\mathbf{Q}| \cdot |\mathbf{Q}^t\mathbf{K}_X\mathbf{Q} + \mathbf{\Lambda}| \cdot |\mathbf{Q}^t| \\ &= |\mathbf{Q}^t\mathbf{K}_X\mathbf{Q} + \mathbf{\Lambda}| \\ &= |\mathbf{A} + \mathbf{\Lambda}|, \end{aligned}$$

where $\mathbf{A} \stackrel{\triangle}{=} \mathbf{Q}^t\mathbf{K}_X\mathbf{Q}$. Since $tr(\mathbf{A}) = tr(\mathbf{K}_X)$, the problem is further transformed to maximize $|\mathbf{A} + \mathbf{\Lambda}|$ subject to $tr(\mathbf{A}) \leq S$.

Lemma [Hadamard's inequality] Any positive definite $k \times k$ matrix $\mathbf{K}$ satisfies

$$|\mathbf{K}| \leq \prod_{i=1}^{k} K_{ii},$$

where K_{ii} is the component of matrix $\mathbf{K}$ locating at i^{th} row and i^{th} column. Equality holds if, and only if, the matrix is diagonal.

By observing that $\mathbf{A} + \mathbf{\Lambda}$ is positive definite (because $\mathbf{\Lambda}$ is positive definite), together with the above lemma, we have

$$|\mathbf{A} + \mathbf{\Lambda}| \leq \prod_{i=1}^{k} (A_{ii} + \lambda_i),$$

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 $\mathbf{K}_N$. Thus, the maximum value of $|\mathbf{A} + \mathbf{\Lambda}|$ under $tr(\mathbf{A}) \leq S$ is achieved by a diagonal $\mathbf{A}$ with

$$\sum_{i=1}^{k} A_{ii} = S.$$

Correlated Parallel Additive Gaussian Channels

Finally, we can adopt the Lagrange multiplier technique as used previously to obtain:

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

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

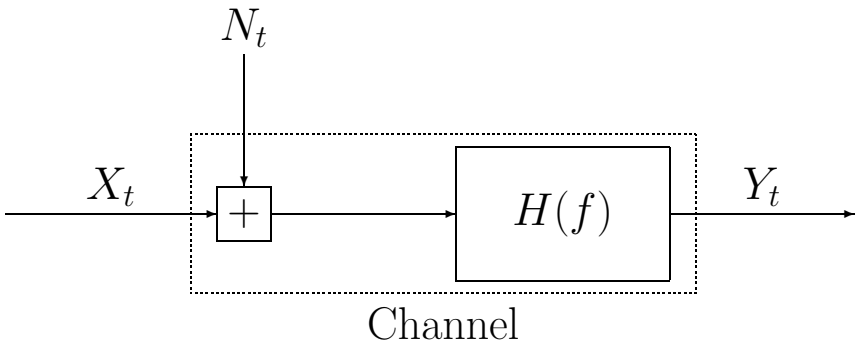
Waveform channels will be discussed next!

Bandlimited Waveform Channels with AWGN

- A common model for communication over a radio network or a telephone line is a band-limited channel with white noise, which is a continuous-time channel, modelled as

$$Y_t = (X_t + N_t) * h(t),$$

where "$*$" represents the convolution operation, X_t is the waveform source, Y_t is the waveform output, N_t is the white noise, and $h(t)$ is the band-limited filter.



Bandlimited Waveform Channels with AWGN

Coding over waveform channels

- For a fixed interval $[0, T)$, select M different functions (waveform codebook)

$$c_1(t), c_2(t), \ldots, c_M(t),$$

  for each informational message.

- Based on the received function $y(t)$ for $t \in [0, T)$ at the channel output, a decision on the input message is made.

Sampling theorem

- Sampling a band-limited signal at a sampling rate $1/(2W)$ is sufficient to reconstruct the signal from the samples, if W is the bandwidth of the signal.

Conventional Remark: Based on the sampling theorem, one can sample the filtered waveform signal $c_m(t)$ and the filtered waveform noise N_t, and reconstruct them distortionlessly by the sampling frequency $2W$ (from the receiver Y_t point of view).

Bandlimited Waveform Channels with AWGN

My Remarks: Assume for the moment that the channel is noiseless.

- The input waveform codeword $c_j(t)$ is *time-limited* in $[0, T)$; so it cannot be *band-limited*.
- However, the receiver can only observe a band-limited version $\tilde{c}_j(t)$ of $c_j(t)$ due to the ideal band-limited filter $h(t)$. Notably, $\tilde{c}_j(t)$ is **no longer** *time-limited*.
- By sampling theorem, the *band-limited* but *time-unlimited* $\tilde{c}_j(t)$ can be distortionlessly reconstructed by its (possibly infinitely many) samples at sampling rate $1/(2W)$.
- *Implicit System Constraint*: Yet, the receiver can only use those samples within time $[0, T)$ to guess what the transmitter originally sent out.
  - Notably, these $2WT$ samples may not reconstruct $\tilde{c}_j(t)$ without distortion.
- As a result, the waveform codewords $\{c_j(t)\}_{j=1}^{M}$ are chosen such that their residual signals $\{\tilde{\tilde{c}}_j(t)\}_{j=1}^{M}$, after experiencing the *ideal lowpass filter* $h(t)$ and the implicit $2WT$*-sample constraint*, are more "resistent" to noise.

What we learn from these remarks.

- The time-limited waveform $c_j(t)$ would pass through an ideal lowpass filter and be sampled, and only $2WT$ samples survives at the receiver end without noise.
- $\tilde{\tilde{X}}_t$ (not X_t) is the true residual signal that survives at the receiver end without noise.
- The power constraint in the capacity-cost function is actually applied on $\tilde{\tilde{X}}_t$ (the true signal that can be reconstructed by the $2WT$ samples seen at the receiver end), rather than the transmitted signal X_t.
  - Indeed, the signal-to-noise ratio concerned in most communication problems is the ratio of the signal power **survived at the receiver end** against the noise power experienced by this received signal.
  - Do not misunderstand that the signal power in this ratio is the transmitted power at the transmitter end.

Bandlimited Waveform Channels with AWGN

The noise process

- How about the band-limited AWGN $\tilde{N}_t = N_t * h(t)$?
- $\tilde{N}_t$ is no longer white?
- However, with the right sampling rate, the samples are still **uncorrelated**.
- Consequently, the noises experienced by the $2WT$ signal samples are still independent Gaussian distributed (cf. The next slide).

$$Y_t = (X_t + N_t) * h(t) = \tilde{X}_t + \tilde{N}_t.$$

$$Y_{k/(2W)} = \tilde{X}_{k/(2W)} + \tilde{N}_{k/(2W)} \text{ for } 0 \le k < 2WT.$$

$$
\begin{aligned}
E[\tilde{N}_{i/(2W)}\tilde{N}_{k/(2W)}] &= E\left[\left(\int_{\Re} h(\tau)N_{i/(2W)-\tau}d\tau\right)\left(\int_{\Re} h(\tau')N_{k/(2W)-\tau'}d\tau'\right)\right] \\
&= \int_{\Re}\int_{\Re} h(\tau)h(\tau')E\left[N_{i/(2W)-\tau}N_{k/(2W)-\tau'}\right]d\tau' d\tau \\
&= \int_{\Re}\int_{\Re} h(\tau)h(\tau')\frac{N_0}{2}\delta\left(\frac{i}{2W}-\frac{k}{2W}-\tau+\tau'\right)d\tau' d\tau \\
&= \frac{N_0}{2}\int_{\Re} h(\tau)h(\tau-(i-k)/(2W))d\tau \\
\left(\int_{-\infty}^{\infty}|H(f)|^2 df = 1\right) \quad &= \frac{N_0}{2}\int_{\Re}\left(\int_{-W}^{W}\frac{1}{\sqrt{2W}}e^{j2\pi f\tau}df\right)\left(\int_{-W}^{W}\frac{1}{\sqrt{2W}}e^{j2\pi f'(\tau-(i-k)/(2W))}df'\right)d\tau \\
&= \frac{N_0}{4W}\int_{-W}^{W}\int_{-W}^{W}\left(\int_{\Re} e^{j2\pi(f+f')\tau}d\tau\right)e^{-j2\pi f'(i-k)/(2W)}df' df \\
&= \frac{N_0}{4W}\int_{-W}^{W}\int_{-W}^{W}\delta(f+f')e^{-j2\pi f'(i-k)/(2W)}df' df \\
&= \frac{N_0}{4W}\int_{-W}^{W} e^{j2\pi f(i-k)/(2W)}df \quad \left(=\frac{N_0}{2}\frac{\sin(2\pi W(i-k)/(2W))}{\pi(i-k)/(2W)}\right) \\
&= \frac{N_0}{2}\frac{\sin(\pi(i-k))}{\pi(i-k)} \quad \text{(With the right sample rate, } W \text{ is cancelled out.)} \\
&= \begin{cases} N_0/2, & \text{if } i=k; \\ 0, & \text{if } i\neq k. \end{cases}
\end{aligned}
$$

Hence, the **capacity-cost** function of this channel subject to input waveform width T (and implicit system constraint) is equal to:

$$\begin{aligned} C_T(S) &= \max_{\left\{p_{X^{2WT}} \,:\, \sum_{i=0}^{2WT-1} E[\tilde{X}^2_{i/(2W)}] \le S\right\}} I(\tilde{X}^{2WT}; Y^{2WT}) \\ &= \sum_{i=0}^{2WT-1} \frac{1}{2} \log\left(1 + \frac{S_i}{\sigma_i^2}\right) \\ &= \sum_{i=0}^{2WT-1} \frac{1}{2} \log\left(1 + \frac{S/(2WT)}{(N_0/2)}\right) \\ &= \sum_{i=0}^{2WT-1} \frac{1}{2} \log\left(1 + \frac{S}{WTN_0}\right) \\ &= WT \cdot \log\left(1 + \frac{S}{WTN_0}\right), \end{aligned}$$

where the input samples $\tilde{X}^{2WT}$ that achieves the capacity is also i.i.d. Gaussian distributed.

- It can be proved similarly to "**white** $N_t \Rightarrow$ **i.i.d.** $\tilde{N}^{2WT}$" that a white Gaussian process X_t can render an i.i.d. Gaussian distributed filtered samples X^{2WT}, if a right sampling rate is employed.

Bandlimited Waveform Channels with AWGN

Remark on notations

- $C_T(S)$ denotes the capacity-cost function subject to input waveform width T. This notation is specifically used in waveform channels.
- $C_T(S)$ should not be confused with the notation of $C(S)$, which is used to represent the capacity-cost function of a *discrete-time* channels, where the channel input is transmitted only at each sampled time instance, and hence no duration T is involved.
- One, however, can measure $C(S)$ by the unit of *bits per sample period* to relate the quantity to the usual unit of data transmission speed, such as *bits per second*.
- To obtain the maximum reliable data transmission speed over waveform channel in units of *bits per second*, we require to calculate:

$$\frac{C_T(S)}{T}.$$

Example 6.34 (telephone line channel) Suppose telephone signals are bandlimited to 4 kHz. Given signal-to-noise (SNR) ratio of 20dB (namely, $S/(WN_0) = 20dB$) and $T = 1$ millisecond, the capacity of bandlimited Gaussian waveform channel is equal to:

$$\begin{aligned} C_T(S) &= WT \log\left(1 + \frac{S}{WTN_0}\right) \\ &= 4000 \times \left(1 \times 10^{-3}\right) \times \log\left(1 + \frac{100}{1 \times 10^{-3}}\right). \end{aligned}$$

Therefore, the maximum reliable transmission speed is:

$$\begin{aligned} \frac{C_T(S)}{T} &= 46051.7 \text{ nats per second} \\ &= 66438.6 \text{ bits per second.} \end{aligned}$$

Bandlimited Waveform Channels with AWGN

Remarks:

- It needs to pay special attention that the capacity-cost formula used above is calculated based on a prohibitively simplified channel model, i.e., the noise is additive white Gaussian.
- The capacity formula from such a simplified model only provides a *lower bound* to the true channel capacity, since AWGN is the *worst* noise in the sense of capacity.
- So the true channel capacity is possibly higher than the quantity obtained in the above example!

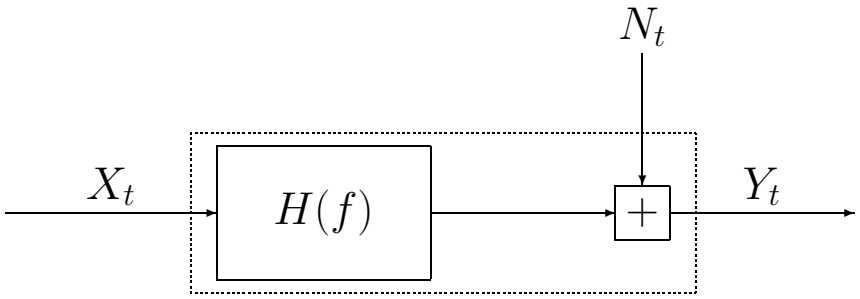
Band-Unlimited Waveform Channels with AWGN

Take $W \to \infty$ in the above formula, we obtain the channel capacity for Gaussian waveform channel of infinite bandwidth is

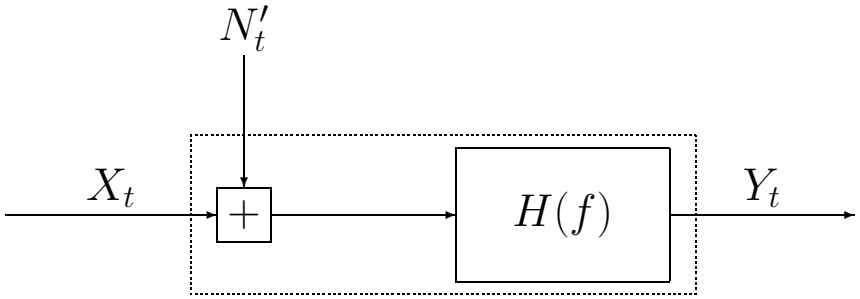
$$C_T(S) \to \frac{S}{N_0} \text{ (nats per } T \text{ unit time)}.$$

- The capacity grows *linearly* with the input power.

Filtered Waveform Stationary Gaussian Channels



The filtered waveform stationary Gaussian channel can be transformed to:



where the power spectral density $\mathrm{PSD}_{N'}(f)$ of N'_t satisfies

$$\mathrm{PSD}_{N'}(f) \stackrel{\triangle}{=} \frac{\mathrm{PSD}_N(f)}{|H(f)|^2},$$

and $\mathrm{PSD}_N(f)$ is the power spectral density of N_t.

Notes

- We cannot use **sampling theorem** in this case (as we did previously) because
  - N'_t is no longer assumed *white*; so the noise samples are not i.i.d.
  - $h(t)$ is not necessarily *band-limited*.

Lemma 6.35 Any real-valued function $v(t)$ defined over $[0, T)$ can be decomposed into

$$v(t) = \sum_{i=1}^{\infty} v_i \Psi_i(t),$$

where the real-valued functions $\{\Psi_i(t)\}$ are any orthonormal set (which can span the space with respect to $v(t)$) of functions on $[0, T)$, namely

$$\int_0^T \Psi_i(t)\Psi_j(t)dt = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{if } i \neq j, \end{cases}$$

and

$$v_i = \int_0^T \Psi_i(t)v(t)dt.$$

Filtered Waveform Stationary Gaussian Channels

Examples

- Orthonormal set is not unique.
  - Sampling theorem employs *sinc* function as the orthonormal set to span a bandlimited signal.
  - Fourier expansion is another example of orthonormal expansion.
- The above two orthonormal sets do not suit here, because their resultant *coefficients* are not necessarily i.i.d.
- We therefore have to introduce the **Karhunen-Loeve expansion**.

Lemma 6.36 (Karhunen-Loeve expansion) Given a stationary random process V_t, and its autocorrelation function

$$\phi_V(t) = E[V_\tau V_{\tau+t}].$$

Let $\{\Psi_i(t)\}_{i=1}^{\infty}$ and $\{\lambda_i\}_{i=1}^{\infty}$ be the eigenfunctions and eigenvalues of $\phi_V(t)$, namely

$$\int_0^T \phi_V(t-s)\Psi_i(s)ds = \lambda_i \Psi_i(t), \quad 0 \le t \le T.$$

Then the expansion coefficients $\{\Lambda_i\}_{i=1}^{\infty}$ of V_t with respect to orthonormal functions $\{\Psi_i(t)\}_{i=1}^{\infty}$ are uncorrelated. In addition, if V_t is Gaussian, then $\{\Lambda_i\}_{i=1}^{\infty}$ are independent Gaussian random variables.

Proof:

$$\begin{aligned}
E[\Lambda_i \Lambda_j] &= E\left[\int_0^T \Psi_i(t) V_t dt \times \int_0^T \Psi_j(s) V_s ds\right] \\
&= \int_0^T \int_0^T \Psi_i(t)\Psi_j(s) E[V_t V_s] dt ds \\
&= \int_0^T \int_0^T \Psi_i(t)\Psi_j(s)\phi_V(t-s) dt ds \\
&= \int_0^T \Psi_i(t) \left(\int_0^T \Psi_j(s)\phi_V(t-s) ds\right) dt \\
&= \int_0^T \Psi_i(t)(\lambda_j \Psi_j(t)) dt \\
&= \begin{cases} \lambda_i, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases}
\end{aligned}$$

□

Filtered Waveform Stationary Gaussian Channels

Theorem 6.37 Give a filtered Gaussian waveform channel with noise spectral density $\mathrm{PSD}_N(f)$ and filter $H(f)$.

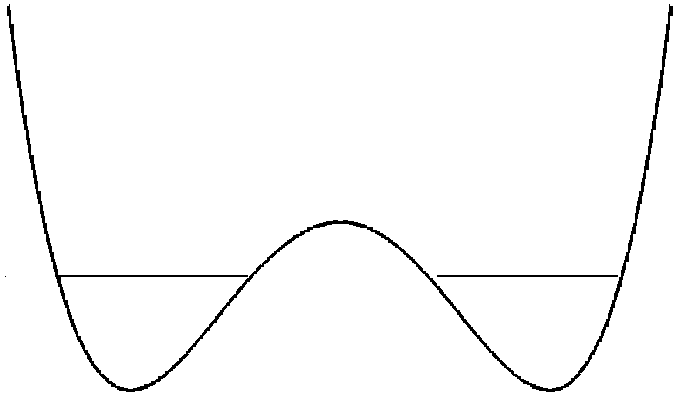
$$C(S) = \lim_{T\to\infty} \frac{C_T(S)}{T} = \frac{1}{2}\int_{-\infty}^{\infty} \max\left[0, \log \frac{\theta}{\mathrm{PSD}_N(f)/|H(f)|^2}\right] df,$$

where θ is the solution of

$$S = \int_{-\infty}^{\infty} \max\left[0, \theta - \frac{\mathrm{PSD}_N(f)}{|H(f)|^2}\right] df.$$

- **Remark:** This also follows the *water-pouring* scheme.

Filtered Waveform Stationary Gaussian Channels



(a) The water pouring scheme. The curve is $\mathrm{PSD}_N(f)/H^2(f)$

(b) The input spectral density that achieves capacity.

Proof: Let the Karhunen-Loeve expansion of the autocorrelation function of N_t' be denoted by $\{\Psi_i(t)\}_{i=1}^{\infty}$

To abuse the notations without ambiguity, we let N_i' and X_i be the Karhunen-Loeve coefficients of N_t' and X_t with respect to $\Psi_i(t)$.

Since $\{N_i'\}_{i=1}^{\infty}$ are independent Gaussian distributed, we obtain that the channel capacity subject to input waveform width T is

$$\begin{aligned} C_T(S) &= \sum_{i=1}^{\infty} \frac{1}{2} \log \left(1 + \frac{\max(0, \theta - \lambda_i)}{\lambda_i} \right) \\ &= \sum_{i=1}^{\infty} \frac{1}{2} \log \left[\max \left(1, \frac{\theta}{\lambda_i} \right) \right] \\ &= \sum_{i=1}^{\infty} \frac{1}{2} \max \left[0, \log \frac{\theta}{\lambda_i} \right], \end{aligned}$$

where θ is the solution of

$$S = \sum_{i=1}^{\infty} \max \left[0, \theta - \lambda_i \right]$$

and $E[(N_i')^2] = \lambda_i$ is the i-th eigenvalue of the autocorrelation function of N_t' (corresponding to eigenfunction $\Psi_i(t)$).

The proof is then completed by applying the Toeplitz distribution theorem. □

[Toeplitz distribution theorem] Consider a zero-mean stationary random process V_t with power spectral density $\mathrm{PSD}_V(f)$ and

$$\int_{-\infty}^{\infty} \mathrm{PSD}_V(f)df < \infty.$$

Denote by $\lambda_1(T), \lambda_2(T), \lambda_3(T), \ldots$ the eigenvalues of the Karhunen-Loeve expansion corresponding to the autocorrelation function of V_t over the time interval of width T. Then for any real-valued continuous function $a(\cdot)$ satisfying

$$a(t) \leq K \cdot t \qquad \text{for } 0 \leq t \leq \max_{f \in \Re} \mathrm{PSD}_V(f)$$

for any finite constant K,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{\infty} a(\lambda_i(T)) = \int_{-\infty}^{\infty} a\big(\mathrm{PSD}_V(f)\big)df.$$

Information Transmission Theorem

Theorem 6.38 (joint source-channel coding theorem) Fix a distortion measure. A DMS can be reproduced at the output of a channel with distortion less than D (by taking sufficiently large blocklength), if

$$\frac{R(D) \text{ nats/source letter}}{T_s \quad \text{seconds/source letter}} < \frac{C(S) \text{ nats/channel usage}}{T_c \quad \text{seconds/channel usage}},$$

where T_s and T_c represent the durations per source letter and per channel input, respectively.

• Note that the units of $R(D)$ and $C(S)$ should be the same, i.e., they should be measured both in nats (by taking natural logarithm), or both in bits (by taking base-2 logarithm).

Theorem 6.39 (joint source-channel coding converse) All data transmission codes will have average distortion larger than D for sufficiently large blocklength, if

$$\frac{R(D)}{T_s} > \frac{C(S)}{T_c}.$$

Example 6.40 (additive white Gaussian noise (AWGN) channel with binary channel input)

- The source is discrete-time binary memoryless with uniform marginal.
- The discrete-time continuous channel has binary input alphabet and real-line output alphabet with Gaussian transition probability.
- Denote by P_b the probability of bit error (i.e, Hamming distortion measure is adopted.)

Information Transmission Theorem

- The rate-distortion function for binary input and Hamming additive distortion measure is

$$R(D) = \begin{cases} \log(2) - H_b(D), & \text{if } 0 \le D \le \frac{1}{2}; \\ 0, & \text{if } D > \frac{1}{2}. \end{cases}$$

- Due to Butman and McEliece, the channel capacity-cost function for binary-input AWGN channel is

$$\begin{aligned} C(S) &= \frac{S}{\sigma^2} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log \left[\cosh \left(\frac{S}{\sigma^2} + y \sqrt{\frac{S}{\sigma^2}} \right) \right] dy \\ &= \frac{E_b T_c/T_s}{N_0/2} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log \left[\cosh \left(\frac{E_b T_c/T_s}{N_0/2} + y \sqrt{\frac{E_b T_c/T_s}{N_0/2}} \right) \right] dy \\ &= 2R\gamma_b - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log[\cosh(2R\gamma_b + y\sqrt{2R\gamma_b})] dy, \end{aligned}$$

  where $R = T_c/T_s$ is the code rate for data transmission and is measured in the unit of *source letter/channel usage* (or *information bit/channel bit*), and γ_b (often denoted by E_b/N_0) is the signal-to-noise ratio per information bit.

Information Transmission Theorem

- Then from the joint source-channel coding theorem, good codes exist when

$$R(D) < \frac{T_s}{T_c} C(S),$$

or equivalently

$$\log(2) - H_b(P_b) < 2\gamma_b - \frac{1}{R\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log[\cosh(2R\gamma_b + y\sqrt{2R\gamma_b})] dy.$$

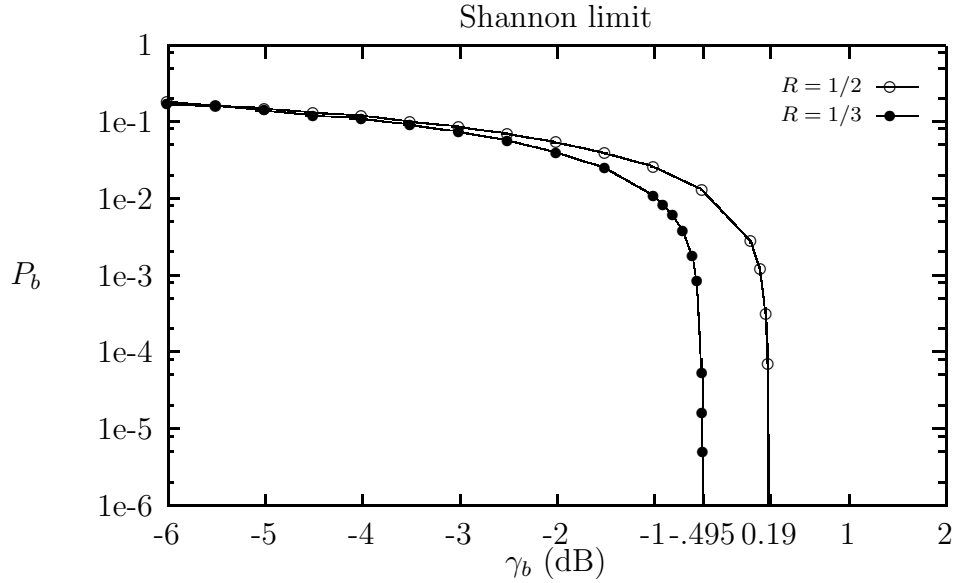
Information Transmission Theorem

- By re-formulating the above inequality as

$$H_b(P_b) > \log(2) - 2\gamma_b + \frac{1}{R\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log[\cosh(2R\gamma_b + y\sqrt{2R\gamma_b})]dy,$$

  a lower bound on the bit error probability as a function of γ_b is established.

- This is the Shannon limit for any code to achieve binary-input Gaussian channel.



The Shannon limits for $(2, 1)$ and $(3, 1)$ codes under binary-input AWGN channel.

Information Transmission Theorem

- The result in the above example becomes important due to the invention of the Turbo coding, for which a near-Shannon-limit performance is first obtained.
- That implies that a near-optimal channel code has been constructed, since in principle, no codes can perform better than the Shannon-limit.

Example 6.41 (AWGN channel with real number input)

- The source is discrete-time binary memoryless with uniform marginal.
- The discrete-time continuous channel has real-line input alphabet and real-line output alphabet with Gaussian transition probability.
- Denote by P_b the probability of bit error (i.e, Hamming distortion measure is adopted.)

Information Transmission Theorem

- The rate-distortion function for binary input and Hamming additive distortion measure is

$$R(D) = \begin{cases} \log(2) - H_b(D), & \text{if } 0 \le D \le \frac{1}{2}; \\ 0, & \text{if } D > \frac{1}{2}. \end{cases}$$

- The channel capacity-cost function is

$$\begin{aligned} C(S) &= \frac{1}{2}\log\left(1 + \frac{S}{\sigma^2}\right) \\ &= \frac{1}{2}\log\left(1 + \frac{E_b T_c/T_s}{N_0/2}\right) \\ &= \frac{1}{2}\log\left(1 + 2R\gamma_b\right) \text{ nats/channel symbol,} \end{aligned}$$

where $R = T_c/T_s$ is the code rate for data transmission and is measured in the unit of *information bit/channel usage*, and $\gamma_b = E_b/N_0$ is the signal-to-noise ratio per information bit.

Information Transmission Theorem

- Then from the joint source-channel coding theorem, good codes exist when

$$R(D) < \frac{T_s}{T_c} C(S),$$

or equivalently

$$\log(2) - H_b(P_b) < \frac{1}{R}\left[\frac{1}{2}\log\left(1 + 2R\gamma_b\right)\right].$$

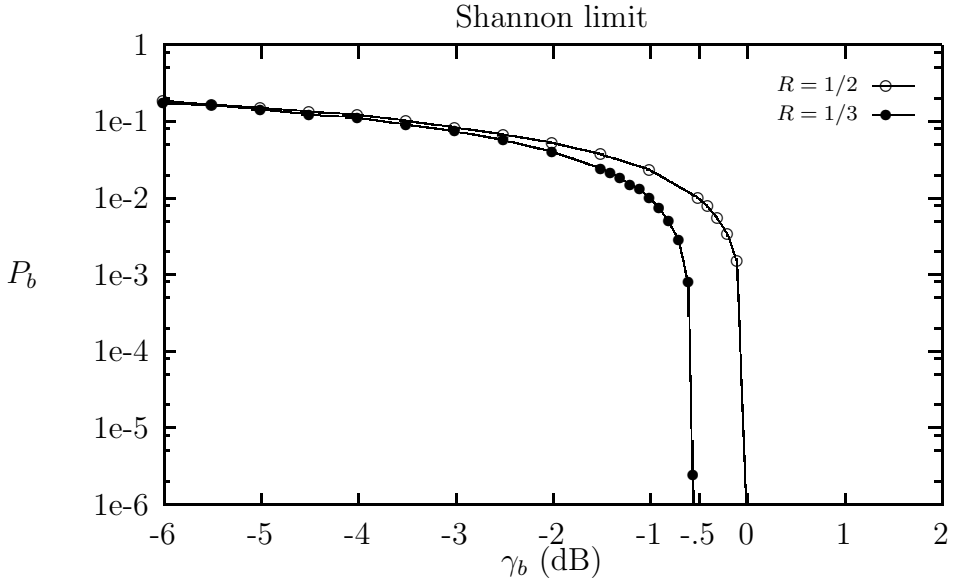
Information Transmission Theorem

- By re-formulating the above inequality as

$$H_b(P_b) > \log(2) - \frac{1}{2R}\log\left(1 + 2R\gamma_b\right),$$

  a lower bound on the bit error probability as a function of γ_b is established.

- This is the Shannon limit for any code to achieve for real-input Gaussian channel.



The Shannon limits for $(2, 1)$ and $(3, 1)$ codes under continuous-input AWGN channels.

Capacity Bounds for Non-Gaussian Channels

- If a channel has additive but non-Gaussian noise and an input power constraint, then it is often hard to calculate the channel capacity, not to mention to derive a close-form capacity formula.
- We then only introduce an upper bound and a lower bound on the capacity for non-Gaussian channels.

Capacity Bounds for Non-Gaussian Channels

Definition 6.42 (entropy power) The entropy power of a random variable N is defined as

$$N_e \stackrel{\triangle}{=} \frac{1}{2\pi e} e^{2 \cdot h(N)}.$$

Lemma 6.43 For a discrete-time continuous-alphabet additive-noise channel, the channel capacity-cost function satisfies

$$\frac{1}{2} \log \frac{S+\sigma^2}{N_e} \geq C(S) \geq \frac{1}{2} \log \frac{S+\sigma^2}{\sigma^2},$$

where S is the bound in input power constraint and σ^2 is the noise power.

Proof: The lower bound is already proved in Theorem 6.30. The upper bound follows from

$$\begin{aligned} I(X;Y) &= h(Y) - h(N) \\ &\leq \frac{1}{2} \log[2\pi e(S+\sigma^2)] - \frac{1}{2} \log[2\pi e N_e]. \end{aligned}$$

□

Capacity Bounds for Non-Gaussian Channels

- The entropy power of a noise N can be viewed as the average noise power of a corresponding Gaussian random variable who has the same differential entropy as N.
- For a Gaussian noise N, its entropy power is equal to

$$N_e = \frac{1}{2\pi e} e^{2h(X)} = \mathrm{Var}(N),$$

  from which the name comes.
- The larger the entropy power, the smaller the upper bound of the capacity-cost function.

Capacity Bounds for Non-Gaussian Channels

- Whenever two independent Gaussian noises, N_1 and N_2, are added, the power (variance) in the sum is equal to the sum of the power (variance) of the two noises, i.e.,

$$e^{2h(N_1+N_2)} = e^{2h(N_1)} + e^{2h(N_2)},$$

or equivalently

$$\mathrm{Var}(N_1 + N_2) = \mathrm{Var}(N_1) + \mathrm{Var}(N_2).$$

- However, when two independent noises are non-Gaussian, the relationship becomes

$$e^{2h(N_1+N_2)} \geq e^{2h(N_1)} + e^{2h(N_2)},$$

or equivalently

$$N_e(N_1 + N_2) \geq N_e(N_1) + N_e(N_2).$$

This is called the *entropy-power inequality*.

  - The entropy-power inequality indicates that the sum of two independent noises may introduce more noise power than the sum of each individual power, except for Gaussian noises.

Key Notes

- Models of discrete-time continuous sources and channels
- Models of waveform sources and channels.
- 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
- Rate-distortion function for continuous sources
  - Its calculation for Gaussian sources under squared error distortion
  - Its calculation for binary sources under Hamming additive distortion measure
- 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

Key Notes

  - Water pouring scheme (graphical interpretation)
  - Band-limited waveform channels with white Gaussian noise
  - Filtered waveform stationary Gaussian channels

- Information-transmission theorem (joint source-channel coding theorem)
  - Shannon limit (BER versus $\mathrm{SNR_b}$)
- Interpretation of entropy-power (provide an upper bound on capacity of non-Gaussian channels)
  - Operational characteristics of entropy-power inequality